Van Folksonomieën naar Ontologieën

Céline Van Damme

Assistent Vakgroep MOSI, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel
celine.van.damme@vub.ac.be

Abstract

De laatste jaren kent het world wide web een nieuwe categorisatietechniek namelijk folksonomieën. Internetgebruikers gaan hun web bronnen categoriseren met hun eigen keywords of tags. Het samennemen of aggregeren van deze tags resulteert in een vlakke taxonomie of folksonomie. Onderzoekers zijn zich meer en meer bewust van de wetenschappelijke waarde die achter dit fenomeen schuil gaat: folksonomieën kunnen worden verrijkt met bestaande web en lexicale bronnen en anderzijds vormen folksonomieën een vruchtbare bodem voor de creatie van nieuwe ontologieën.

1. Folksonomieën nader toegelicht.

Sedert enkele jaren vindt een nieuwe vorm van categorisatie meer en meer opgang op het web: tagging en de resulterende folksonomie. Internetgebruikers kunnen alle informatie of web bronnen identificeerbaar met een unieke URL categoriseren met hun eigen sleutelwoorden of tags.

De cognitieve overhead voor de actoren of gebruikers is zeer laag. Ze zijn aan geen enkele regel gebonden inzake het geven van keywords. Daarenboven is er een directe return verbonden aan hun inspanning namelijk het terugvinden van hun eigen content. Categorisaties uitgedrukt met eigen sleutelwoorden of tags maken het voor de gebruiker veel eenvoudiger de gewenste informatie terug te vinden.

Het samennemen van alle tags vormt een vlakke bottom-up gecreëerde taxonomie of folksonomie. Het was Thomas vander wal die dit begrip introduceerde door de woorden folk en taxonomie samen te smelten[1]. Een taxonomie bestaat uit een door een groep van experten gekozen “controlled vocabulary” waartussen hiërarchische relaties gedefinieerd zijn en op basis daarvan indexeren ze de documenten. Een klassiek voorbeeld van een taxonomie zijn de Yahoo Directories die web pagina’s classificeren volgens voorgedefinieerde hiërarchische categorieëen.

De ontwikkeling van een taxonomie staat in schril contrast met een folksonomie waar de ontwikkelaars de gebruikers zijn en er een volledige flexibiliteit bestaat bij het kiezen van de sleutelwoorden. Daarenboven is

een folksonomie zeer dynamisch. Het weerspiegelt ten alle tijden de terminologie die leeft bij zijn gebuikers/ontwikkelaars. Daarenboven worden de bronnen onmiddellijk geïndexeerd door de persoon die de web resource consumeert of creëert. Terwijl in het geval van een taxonomie het een hele poos duurt vooraleer een nieuwe term erin wordt opgenomen[2].

Een folksonomie wordt gevisualiseerd door middel van een tag cloud. Deze tag cloud is een soort wolk waarbij de tekstgrootte van de tags recht evenredig is met de frequentie van de tags. Hoe vaker een tag wordt gebruikt, hoe groter de tag in de cloud. Deze tag cloud wordt meestal gebruikt als navigatieinstrument [3]. Door te klikken op een van de tags wordt een overzicht gegenereerd van alle content die gecategorizeerd is door deze tag. Daarenboven zijn de meeste websites met een folksonomie mechanisme (zoals Del.icio.us, Flickr en Technorati) uitgerust met een tag search engine die de gebruikers toelaten op meerdere tags tergelijkertijd te zoeken.

Het feit dat tags en objecten of content publiek beschikbaar zijn laat een nieuwe vorm van navigatie toe namelijk sociale navigatie. Internetgebruikers met dezelfde interesses kunnen elkaar gemakkelijk terugvinden op basis van de tags en objecten. Tags en objecten zijn immers een reflectie van de interesses en kennis van een actor. Vandaar dat een gemeenschappelijke tag of object een signaal kan zijn van overlappende interessevelden[4]. Een folksonomie genereert dus niet alleen return voor het individu in kwestie, maar de hele community ondervindt er voordelen van.

2. Een folksonomie en zijn gebreken.

Er zijn echter een aantal nadelen verbonden aan deze categorisatietechniek. Wanneer we onderstaande tag cloud van nabij bekijken zijn een aantal van deze gebreken zichtbaar. Zo bevat de tag cloud: een probleem van getal (vb. flower en flowers), synoniemen en acroniemen (vb. nyc en newyorkcity), homoniemen (vb. canon, kan een fototoestel of een camera zijn), idiosyncratisch of egocentrisch taggen (vb. “me”)[5]. Tot slot behoren lexicale of tikfouten eveneens tot een van de nadelen, maar deze zijn niet direct zichtbaar in onderstaande tag cloud. In het geval van een kritische massa worden lapsussen snel opgevangen doordat ze te laag in frequentie voorkomen en bijgevolg niet worden afgebeeld in de tag cloud.

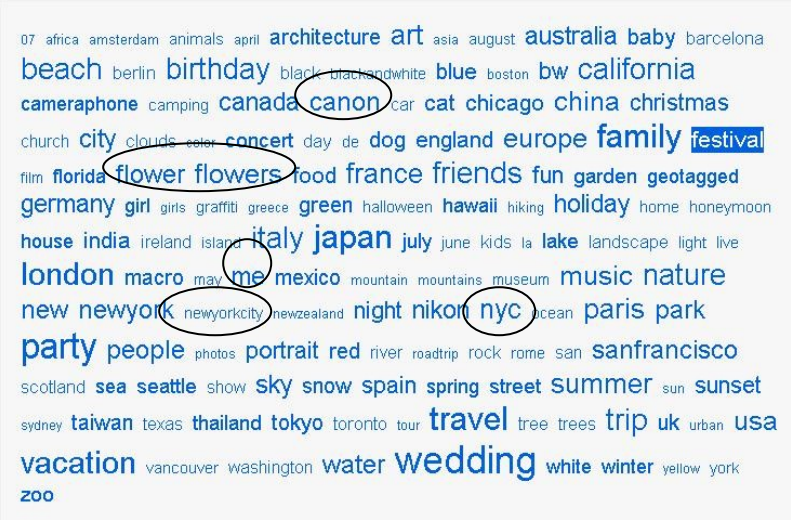

Fig. 2: Tag Cloud Flickr website (op 17 Oktober 2007)

2.1. Mogelijke Oplossingen.

Er zijn twee mogelijkheden om deze tekortkomingen op te vangen: meer input vragen aan de gebruiker of tekortkomingen automatisch detecteren en oplossen.

2.1.1. Meer en of betere input aan gebruiker vragen.

Er kan op verschillende manieren meer of betere input worden verkregen van de gebruiker. Zo kan er feedback worden gegeven om de gebruiker aan te zetten tot consistent woordgebruik bij het taggen. Door middel van autocompleting[6], dit is een techniek die bij elke toetsaanslag woorden suggereert op basis van de reeds ingegeven lettercombinaties, krijgt hij een overzicht van reeds gebruikte tags. Stel de gebruiker wenst de tag *product* te gebruiken, maar volgens de autocompleting heeft tot op heden niemand tag *product* gebruikt, maar wel *produkt*. Dit kan voor de gebruiker een indicatie zijn om de tag *produkt* in plaats van *product* te kiezen. Daarnaast kan er de mogelijkheid worden gegeven om tags manueel te gaan clusteren. De gebruiker kan dan zelf aangeven welke tags bijelkaar horen. Dit is bijvoorbeeld een functionaliteit die voorzien is bij de social bookmark manager Del.icio.us [2]. In [7] dient de gebruiker semantiek aan de tag toe te voegen door een concept aan te duiden nadat hij een tag heeft ingegeven. Deze concepten worden opgehaald uit een database waar een thesaurus zoals Wordnet gestockeerd is.

Hoewel het vragen van bijkomende informatie waardevol kan zijn, kan dit volgens ons een nadelig effect teweegbrengen. Meer informatie vragen, vergt immers een grotere inspanning met mogelijk een lagere responsgraad als gevolg. Het grote voordeel van taggen zit ons immers is in de eenvoud en lage cognitieve overhead.

2.1.1. Automatisch.

Een andere optie bestaat erin om geen extra informatie te vragen aan de gebruiker, maar om dit automatisch te detecteren. In volgende paragrafen beschrijven we een aantal interessante technieken die kunnen gebruikt worden tijdens dit proces.

Stemming algoritmes die een woord herleiden naar zijn stam vb. *boeken* wordt *boek* en *spelen* wordt *speel*[8], kunnen bijvoorbeeld worden ingezet om similaire woorden te vinden. Deze algoritmes zijn echter afhankelijk van de taal. Vandaar dat er telkens dient achterhaald te worden tot welke taal de tag behoort opdat de betekenis van de tags niet verloren zou gaan. Zo zou de Engelse tag *cheese omgevormd* worden tot *chees* bij een Nederlandstalig stemmingsalgoritme. Het toepassen van Nederlandstalige stemming algoritmes op Nederlandse tags is daarenboven echter niet zo evident. De Nederlandse taal is immers doorspekt met anglicismen die niet altijd de Nederlandse schrijfregels volgen.

Een andere optie is het berekenen van de afstand tussen de woorden. Deze similariteit wordt ook wel de Levenshtein similariteit genoemd die nagaat hoeveel letters dienen aangepast te worden (invoegen, verwijderen of veranderen) om het ene woord te transformeren in een ander woord. Vervolgens moet deze som gedeeld worden door het aantal letters in het te transfereren woord. Deze verhouding genereert een waarde die tussen 0 en 1 ligt. Hoe dichter bij 0 hoe meer verschillend de woorden zijn en hoe dichter bij 1 hoe similairder [9]. Door middel van *trial en error*, kan een grenswaarde worden gevonden waarbij automatisch de woordkoppels die deze grens overschrijden als similair kunnen worden beschouwd. In [10] wordt een grenswaarde van 0.83 vooropgesteld bij toepassing op Engelstalige tags.

Het soundex algoritme zou eveneens een waardevolle inbreng kunnen hebben bij het zoeken naar gelijke woorden. Dit algoritme detecteert namelijk de woorden die eenzelfde klank hebben[11]. Aangezien in de Nederlandse taal vaak verwarring bestaat omtrent de schrijfwijze van bepaalde woorden, zou dit algoritme eveneens een uitweg bieden. Tags zoals *product* en *produkt* zouden zo gedetecteerd kunnen worden.

Het berekenen van de co-occurrence per tag pair, laat toe te achterhalen welke woorden vaak in combinatie worden gebruikt. Indien twee tags heel veel samen worden gebruikt, dan kan dit een indicatie zijn dat er een relatie bestaat tussen beide woorden[12]. Om de co-occurrence van een tag pair te bepalen dient voor elk getagd object alle mogelijke tag koppels te worden gemaakt. Dit proces moet herhaald worden voor alle objecten om vervolgens de frequentie te kunnen bepalen voor elk tag koppel.

We kunnen concluderen dat er verschillende technieken kunnen worden ingezet om de nadelen van een folksonomie te verkleinen. Omwille van de sterke kanten van een folksonomie en het feit dat hun nadelen kunnen worden verkleind met verschillende technieken, zijn onderzoekers meer en meer de wetenschappelijke waarde ervan gaan beseffen. Er zijn twee tendenzen merkbaar. Langs de ene kant is er onderzoek die zich focust op hoe een folksonomie kan worden verrijkt met statistische technieken zoals hierboven beschreven, bestaande online web bronnen en

ontologieën[9] en anderzijds zijn er onderzoekers die folksonomieën trachten om te buigen in ontologieën[13]. In volgende paragrafen bespreken we twee papers die elk van een van deze onderzoeksrichtingen behandelen.

3. Het Verrijken van een folksonomie.

De paper geschreven door Lucia Specia en Enrico Motta[9] is volgens ons op dit ogenblik een paper die het best beschrijft hoe een folksonomie kan worden verrijkt met concepten en relaties door gebruik te maken van verschillende technieken en online bronnen. In deze paper stellen de auteurs een methodologie voor die we in volgende paragrafen zullen bespreken. De auteurs hebben hun methodologie getest op een dataset van de social bookmark manager deli.cio.us en de online foto manager Flickr.

3.1. Filter

In eerste instantie worden de irrelevante tags verwijderd: tags die beginnen met een cijfer of tags die een frequentie hebben kleiner dan 10. Vervolgens worden similaire tags gedetecteerd door middel van de Levenshtein similariteit. Hiervoor worden alle mogelijk tag koppels gemaakt. Wanneer een koppel een grenswaarde van 0.83 overschrijdt dan wordt het koppel geclusterd en wordt er gezocht naar een representatieve term in de generische thesaurus Wordnet. Indien een van deze similaire tags kan teruggevonden worden in Wordnet dan wordt deze gekozen als representatieve term voor de cluster. Er wordt echter niet in de paper gespecifieerd wat er gebeurt indien beide voorkomen in Wordnet.

3.2. Clusteren.

De co-occurrence wordt berekend voor elk tag koppel met een minimumfrequentie. De bekomen waarden en de tag koppels worden vervolgens in een matrix geplaatst waarbij elke lijn of kolom een vector voorstelt. De tag paren worden geclusterd door toepassing van de cosinus similariteit[14]. Deze similariteit wordt berekend door het product van de twee vectoren te delen door de euclidische afstand[1] tussen beide vectoren. Volgens de auteurs van deze paper is deze clusteringtechniek het meest passende in deze situatie omdat de andere clusteringtechnieken gevoeliger zijn voor variaties tussen de elementen. Deze similariteit berekent de cosinus of de hoek tussen twee vectoren en bekomt een waarde die ligt tussen 0 en 1. Hoe kleiner deze waarde hoe similairder de vectoren en hoe dichter bij 1 hoe meer verschillend de vectoren of de tag koppels zijn.

3.3. Detecteren van relaties en concepten.

1 De euclidische afstand wordt berekend door de vierkantswortel te nemen uit de som van de kwadraten van de verschillen tussen de coördinaten van de vectoren. (< http://en.wikipedia.org/wiki/Euclidean_distance > beozcht op 25 december 2007.)

In een derde stap detecteren de auteurs de relaties tussen de geclusterde tag koppels. Hiervoor maken ze gebruik van bestaande ontologieën, Wikipedia en Google.

Eerst zoeken ze naar de relatie tussen een tag koppel (vb. *university* en *course*) door gebruik te maken van bestaande ontologieën. Ontologieën hebben namelijk een veel rijkere semantiek dan folksonomieën[2]. Binnen het domein van de computerwetenschappen worden ontologieën omschreven als een middel om de communicatie te bevorderen tussen mensen en machines. Een ontologie beschrijft namelijk een gemeenschappelijk domein waarin concepten, instanties en relates gedefinieerd zijn op een informele (natuurlijke taal) of formele wijze (ontologietaal)[15]. Zo zou een ontologie van een gezin de concepten moeder, vader, zoon, dochter bevatten en de relaties “heeftmoeder”, “heeftvader”, ”heeftzoon”, “heeftdochter”, ”heeftbroer” en “heeftzus” beschrijven. Vervolgens kan deze ontologie worden aangevuld met instanties of elementen, meer bepaald met personen uit een gezin. Een aantal ontologieën zijn toegankelijk via de search engine Swoogle die RDF documenten indexeert. RDF is een van de standaard ontologietalen voorgesteld door W3C, het world wide web consortium[16]. Indien er geen resultaat wordt gevonden met de search engine Swoogle, dan gaan de auteurs ervan uit dat de tags een acroniem of misgeschreven zijn. De termen worden opgezocht in Wikipedia om te zoeken naar acrononiemen of letterwoorden. In het geval het om een acroniem gaat bijvoorbeeld de tag *nyc* dan wordt het vervangen door de volledige term, in dit geval *New York City,* en wordt de vorige stap opnieuw herhaald. In het andere geval wordt nagegaan of het woord juist geschreven is door Google te gebruiken. Google fungeert namelijk als een spellingschecker. Telkemale de gebruiker een query opgeeft maakt Google variaties op de zoektermen en vergelijkt het aantal corresponderende zoekresultaten. Indien de zoekterm van de gebruiker kleiner aantal zoekresultaten heeft dan suggereert Google een andere zoekterm.

De auteurs maken gebruik van de technieken en bronnen in slechts één manier hoewel een aantal van deze bronnen ook voor andere doeleinden en resultaten kunnen worden gebruikt zoals blijkt uit volgende paragrafen.

4. Een folksonomie, een vruchtbare bodem voor een ontologie.

Er wordt eveneens onderzoek verricht in de andere richting, namelijk het ombuigen van folksonomieën in ontologieën. Ontologieën hebben namelijk een veel rijkere semantiek en zijn één van de technologieën om van het semantische web, het web waar alle informatie interpreteerbaar is door machines, een realiteit te maken. Het bouwen van een ontologie is echter zeer arbeidsintensief en de ontwikkeling van deze ontologieën gebeurt meestal door experten. Gebruikers worden niet actief betrokken bij de ontwikkeling ervan. Daarenboven duurt het een hele poos vooraleer nieuwe woorden, relaties etc. worden opgenomen in de ontologie. Gezien

de sterke kanten van een folksonomie wordt meer en meer onderzoek verricht om deze folksonomieën om te buigen tot ontologieën[2].

In onderstaande paragraaf beschrijven we de Folksontology methodologie die is voorgesteld in een paper van Céline Van Damme, Martin Hepp en Katharina Siorpaes[13]. Deze methodologie is voorlopig nog niet getest op een data set, maar vormt wel een eerste aanzet van hoe folksonomieën kunnen omgebouwd worden tot ontologieën.

4.1. Folksontology methodologie.

Om folksonomieën te kunnen omzetten naar ontologieën wordt zowel het gebruik van verschillende technieken als online bronnen voorgesteld. Het kernidee achter de methodologie bestaat erin om zoveel mogelijk informatie te destilleren uit de beschikbare folksonomie data. Zo zijn er niet alleen actoren, tags en objecten betrokken bij het proces, maar eveneens systemen. Er zijn verschillende systemen waar folksonomieën worden gebruikt voor het categoriseren van objecten handelend over eenzelfde topic. Zo worden er op Deli.cio.us bookmarks verzameld omtrent computers, maar worden er eveneens foto's op Flickr bijgehouden over computers. Vandaar dat er verschillende technieken worden voorgesteld die het benutten van deze verschillende databronnen mogelijk maken. In tegenstelling tot vorige methodologie worden de bronnen op verschillende manieren benut en wordt eveneens de community betrokken om extra input te leveren bij het bouwen van de ontologieën.

4.1.1. Tags filteren.

Stemming algoritmes worden voorgesteld om tags te herleiden naar de stam van het woord. Vervolgens wordt de schrijfwijze van de tags geverifieerd door lexicale bronnen te gebruiken zoals de thesaurus Wordnet, een online woordenboek vb. Leo dictionaries, Google en Wikipedia. Indien de tags niet teruggevonden worden in een van deze bronnen wordt de frequentie van de tags nagegaan. Bij een lage frequentie wordt er geconcludeerd dat het om een foutief geschreven woord gaat en in het andere geval wordt het woord voorgelegd aan de community. Het kan immers gaan om een nieuwe term die nog niet is opgenomen in de lexicale bronnen.

4.1.2. Technieken

Er wordt een overzicht gegeven van de verschillende technieken die vervolgens kunnen worden toegepast op de gefilterde tags. Als eerste optie wordt het berekenen van de co-occurrence van de verschillende tag paren aangeraden. Vervolgens worden er verschillende voorstellen gedaan van hoe door middel van sociale netwerk technieken actoren kunnen worden geclustered op basis van gemeenschappelijke tags of objecten. Deze technieken kunnen namelijk worden toegepast op verschillende systemen. Zo wordt het mogelijk om meer aspecten van een

topic te belichten. Veronderstel dat verschillende systemen (vb. Deli.cio.us en Flickr) het topic wijn behandelen, dan kan het aggregeren van de data van beide systemen het mogelijk maken om alle aspecten van het topic te belichten terwijl dit met één systeem dit niet altijd mogelijk is.

4.1.3. Verrijken.

De auteurs stellen vervolgens voor om de tag clusters te verrijken met synoniemen, homoniemen, concepten en anderstalige tags te vertalen naar het Engels door middel van de lexicale bronnen (Wikipedia, Wordnet en een online woordenboek). Zo kunnen de URL pagina's van Wikipedia namelijk niet alleen worden gebruikt voor het vinden van concepten, maar ook voor het zoeken naar homoniemen dankzij de doorverwijspagina's. Doorverwijspagina's groeperen namelijk de verschillende betekenissen van een begrip. Zo beschrijft de doorverwijspagina pc op Wikipedia 8 verschillende concepten[2]. Gezien een thesaurus relaties beschrijft zoals synoniemen, homoniemen e.d. kan deze bron worden gebruikt voor het zoeken naar dergelijke relaties tussen de tag koppels. Bestaande online ontologieën worden eveneens voorgesteld als een bron voor het zoeken naar relaties tussen tags. Hierbij wordt het gebruik van mapping en matching technieken voorgesteld. Deze techieken kunnen namelijk de tags gaan matchen met de concepten uit een ontologie.

4.1.4. Community.

In het geval er geen relatie kan worden gevonden voor een tag koppel door gebruik te maken van de bestaande online ontologieën en de lexicale bronnen, dan dient dit worden voorgelegd aan de community evenals tags met een hoge frequentie die niet teruggevonden worden in de online bronnen.

5. Conclusie.

Ondanks een aantal nadelen (problemen van getal, geen synoniemen of homoniemen,...), zijn onderzoekers zijn zich meer en meer bewust geworden van de wetenschappelijke waarde die schuil gaat achter een folksonomie, een taxonomie gecreëerd door de gebruikers. In deze paper hebben we enkele technieken beschreven die door middel van een extra input van de gebruiker of op een automatische manier de nadelen van een folksonomie kunnen verkleinen. Daarnaast hebben we de twee onderzoeksrichtingen toegelicht die er momenteel heersen binnen dit domein: het verrijken van een folksonomie en het ombouwen van een folksonomie in een ontologie. We hebben deze onderzoeksrichtingen beschreven door de methodologieën voorgesteld in twee papers te bespreken. In beide papers wordt het gebruik van statistische technieken, beschikbare online bronnen en ontologieën voorgesteld. In de paper die

2 < http://nl.wikipedia.org/wiki/Pc > bezocht op 31 december 2007.

eerstegenoemde onderzoeksrichting bespreekt worden de online bronnen slechts op één manier benut terwijl bij de andere onderzoeksrichting deze bronnen voor verschillende doeleinden worden aangewend. Daarenboven wordt voorgesteld de community bij het hele proces te betrekken bij het bouwen van ontologieën.

Bibliografie

[1] <http://vanderwal.net/folksonomy.html> bezocht op 5 april 2007

[2] Van Damme, C. Informatie zoeken op het web: directories, zoekmachines, folksonomies...en ontologies. *Bladen voor de documentatie*, 2007, Nr.4, pp.1-11.

[3] Sinclair, J.; Cardrew Hall, M. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 2007, pp. 1-18.

[4] Moreville, P. *Ambient Findability*. O'Reilly, 2005, 204p.

[5] Quintarelli, E. Folksonomies: power to the people. Paper presented at the *ISKO Italy-UniMIB meeting*, June 2005. Online beschikbaar <http://www.iskoi.org/doc/folksonomies.htm>

[6] <http://en.wikipedia.org/wiki/Autocompletion> bezocht op 25 december 2007.

[7] Spyns, P.; de Moor, A.; Vandenbussche, J.; Meersman, R. How the twain meet. In *On the Move to Meaningful Internet Systems 2006: CooPIS, DOA and ODBASE,* 2006, Montpellier, France, Springer 2006, pp.738-755.

[8] van Rijsbergen, C.J.; Robertson, S.E.; Porter, M.F. New models in probabilistic information retrieval. In Information retrieval Research, S.E. Robertson, C.J. van Risjbergen and P. Williams, Eds. Butterworths, 1981, ch. 4, pp. 35-56.

[9] Levenshtein, V. Binary Code Capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 1966, Vol. 10, p.707.

[10] Holmes, D.; Mc Cabe, M. C. Improving precision and recall four soundex retrieval. In ITCC02: Proceedings of the International Conference on Information Technology: Coding and Computing. 2002, Washington, DC, USA.

[11] Schmitz, P .Inducing ontology from Flickrt ags. In Proceedings of Collaborative Web
Tagging Workshop at WWW 2006, Edinburgh, UK, 2006.

[12] Specia, L.; Motta, E. Integrating folksonomies with the semantic Web. in: E. Fraconi, M. Kifer, and W. May (Eds.): *Proceedings of the*

European Semantic Web Conference (ESWC 2007), Innsbruck, Austria. Springer 2007, pp.624-639.

[13] Van Damme, C.; Hepp M.; Siorpaes, K. FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. *Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, 2007, Innsbruck, Austria.

[14] < http://en.wikipedia.org/wiki/Jaccard_index > bezocht op 25 december 2007.

[15] Uschold, M.; Gruninger, M. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 1996, 11(2), pp. 93-155.

[16]<http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OwlVarieties> bezocht op 25 april 2007