Een Kritische Kijk op Nieuwe Classificatietechnieken

Inforum2007, 26 april 2007

Céline Van Damme
Celine.Van.Damme@vub.ac.be

Vrije Universiteit Brussel

Structuur

- Informatie overload op het web
- Informatie zoeken op het web
- Vergelijking huidige classificatietechnieken documentalist

1.Informatie overload op het web (1)

- Statische web pagina's
  - Altijd aanspreekbaar en beschikbaar
  - Indexeerbaar door meeste zoekmachines
  - Vormen het *visible web*
- Dynamische web pagina's
  - Pagina's worden gecreëerd bij opvraging (uit database) en verdwijnen daarna (vb. Online Vandale Woordenboek)
  - Niet indexeerbaar door meeste zoekmachines
  - Vormen het *dark web*

1.Informatie overload op het web(2)

- Visible Web
  - 200 miljoen pagina's (1997)[1]
  - 800 miljoen pagina's (1998) [1]
  - 11,5 miljard pagina's (januari 2005)[1]
- Dark Web
  - 500 keer visible web (2003) [2]

1. Informatie overload op het web(3)

Grote invloed social software of Web 2.0 tools

- Lage kost en lage technologiebarrière
- Internetgebruiker: geëvolueerd van een informatieconsument naar een informatiecreator
- Enkele voorbeelden:
  - Wiki → Wikipedia: meer dan 5.300.000 web pagina's [3]
  - Blogs: meer dan 71.000.000 blogs geregisteerd bij Technorati[4]

2. Informatie zoeken op het web

- Directories
- Zoekmachines & Ontologies
- Folksonomies

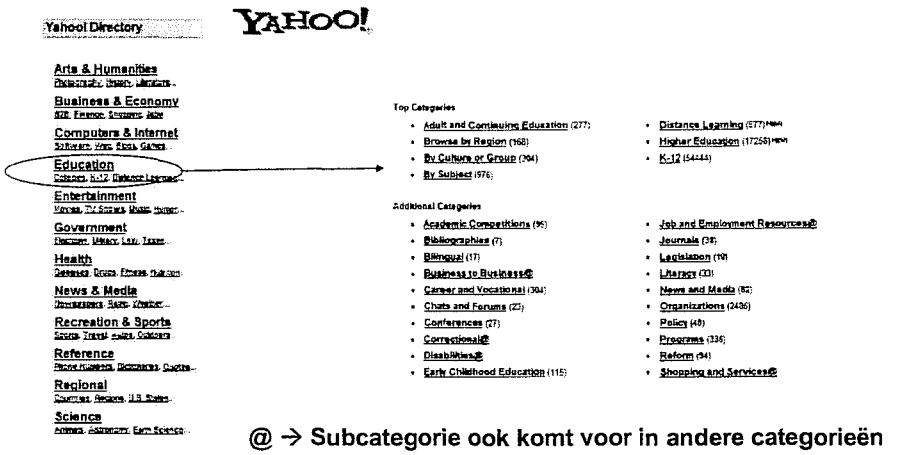
2. Informatie zoeken op het web

- ***Directories***
- Zoekmachines & Ontologies
- Folksonomies

Definitie

- Taxonomies
- Classificeren van websites in hiërarchische categorieën
- Groep experten
- Navigeren via drill down
- Vb. Yahoo Directories, Open Directory Project

Voorbeeld



@ → Subcategorie ook komt voor in andere categorieën

2. Informatie zoeken op het web

- Directories
- ***Zoekmachines & Ontologies***
- Folksonomies

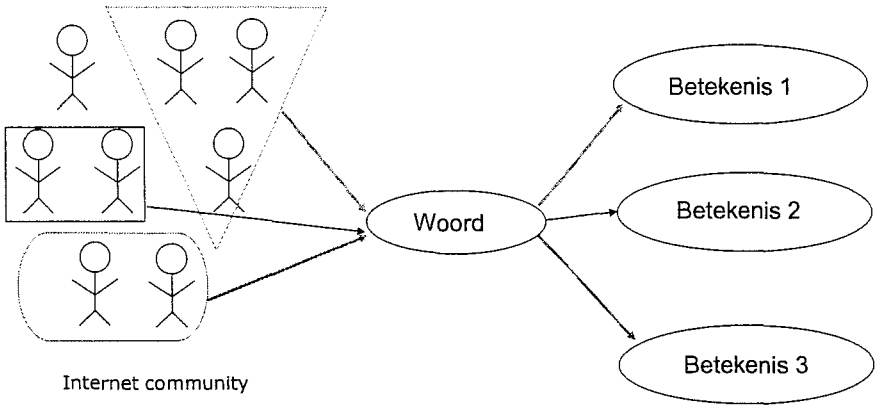
Werking Zoekmachine (1)

- Web crawlers doorzoeken het web
- Lijstje van URLs
- Kopiëren en indexeren web pagina
- Afhankelijk soort zoekmachine
- Data bewaren in database

Werking Zoekmachine (2)

- Zoekopdracht gebruiker toetsen aan database
- Genereren van een resultatenlijst

Problemen



- Zoekopdracht wordt verkeerd geïnterpreteerd
- Web pagina's worden geïndexeerd als *plain text*

Oplossing

- Meta data in web pagina plaatsen via
  - XML: <titel> abc </titel>
    - MAAR <titel> kan verschillende betekenissen hebben: titel van een boek, film, paper...

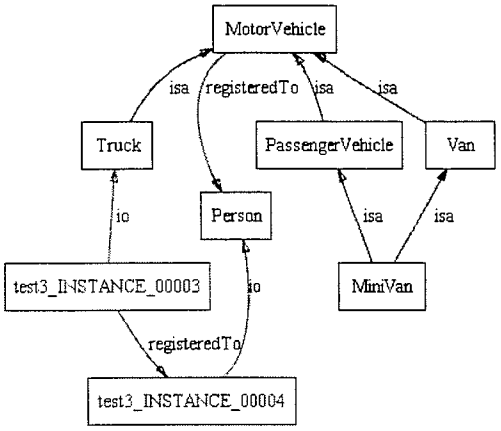
Definitie ontology

- beschrijft de natuurlijke taal van een domein
- bevat concepten en attributen (instances)
- beschrijft hun onderlinge relaties
- beschrijft hun regels
- geschreven in een formele taal: een taal begrijpbaar voor machines (RDF, OWL)

Een vergelijking...

- Controlled vocabulary
- Taxonomy = controlled vocabulary + hiërarchische relaties
- Thesaurus = taxonomie met horizontaal gerelateerde terminologie (synoniemen, antoniemen, meroniemen etc,) vb. Wordnet
- Ontology =uitgebreider dan thesaurus

Voorbeeld




```
....
<rdfs:Class rdf:about="&mv;MotorVehicle">
        <rdfs:subClassOf
rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;PassengerVehicle">
        <rdfs:subClassOf
rdf:resource="&mv;MotorVehicle"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;Person">
        <rdfs:subClassOf
rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
....
```


Semantische Web

- Belangrijke technologie voor de ontwikkeling van het semantische web

  web waar alle informatie begrijpbaar en interpreteerbaar is voor machines

- Rijker dan een taxonomie: meer relaties worden blootgelegd zoals meroniem vb. hand is deel van een arm

- Bevordert het opzoeken van informatie: zoekmachines zullen veel betere resultaten kunnen genereren aan de gebruikers

Problemen

- Ontwikkeling en onderhoud duur en arbeidsintensief
- Groep experts <----> Effectieve gebruikers
- Formele taal schrikt gebruikers af om te participeren in ontwikkeling

2. Informatie zoeken op het web

- Directories
- Zoekmachines & Ontologies
- ***Folksonomies***

Definitie folksonomy (1)

- Sociaal Classificatiesysteem
- Ontwikkelaars = gebruikers
- Gebruikers mogen hun eigen keywords of *tags* gebruiken voor het omschrijven van content:
  - volgens Amerikaanse studie: 28% internet gebruikers heeft reeds content getagd[5]
- Vergelijkbaar met keywords toegevoegd door auteur(s) aan een paper

Definitie folksonomy (2)

- Het aggregeren van alle tags = vlakke bottom-up taxonomy
- Folksonomy = folk + taxonomy (Thomas Vander Wal[6])
- Sociale navigatie: informatie vinden via personen met gelijke interesse.

Enkele voorbeelden...

slideshare

43 Things

del.icio.us

Connotea

YouTube

flickr

citeulike

BibSonomy

Technorati

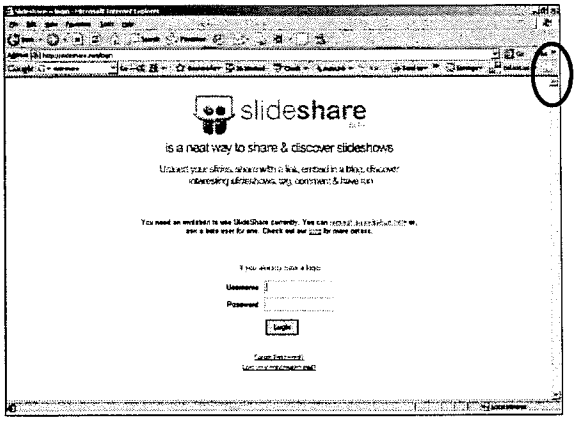
En vele andere...

Del.icio.us (1)

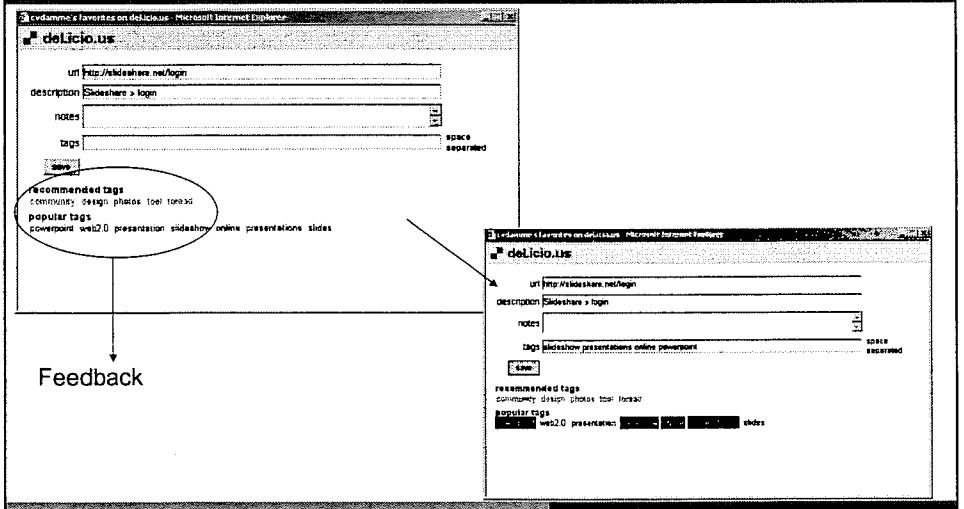
del.icio.us

- Beheren van favoriete websites of bookmarks
- Tags worden gebruikt voor het omschrijven van bookmarks
- Tags kunnen door de gebruiker worden geclusterd in bundels
- Feedback
- Elke gebruiker heeft zijn eigen account http://del.icio.us/username
- Op basis van tags of bookmarks kunnen personen met gelijke interesses elkaar terugvinden
- Knoppen in browser

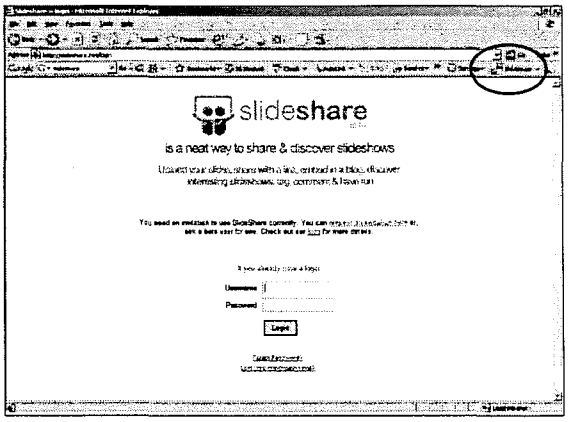
Del.icio.us (2)



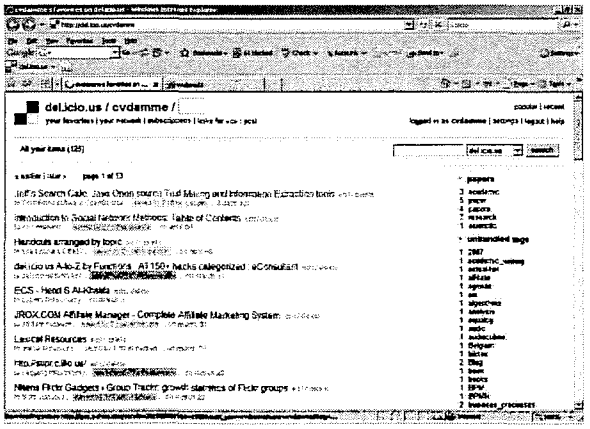
Del.icio.us (3)



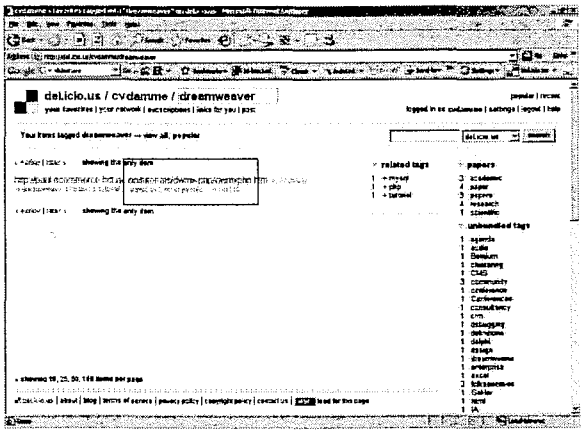
Del.icio.us (4)



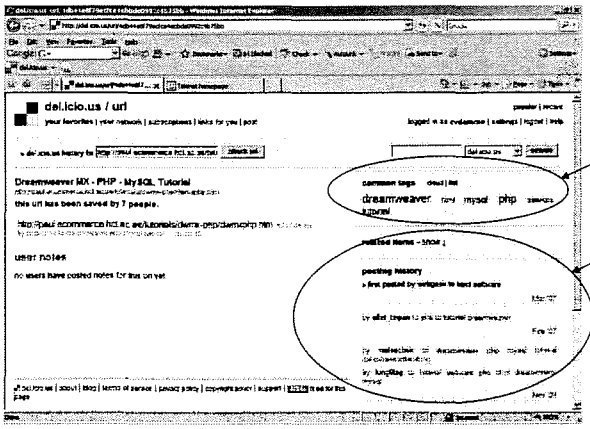
Del.icio.us (5)



Del.icio.us (6)



Del.icio.us (7)



Del.icio.us (8)

del.icio.us Cloud (popular tags)

This is a tag cloud - a list of tags where size reflects popularity.
sort: alphabetically | by #ttx

advertising **ajax** apple architecture **art** article articles audio **blog** blogging **blogs** books **business** car comics **community** computer cooking cool crafts **CSS** culture database **design** **development** diy download **education** email entertainment environment fashion flc film finance firefox **flash** fonts **food** **free** **freeware** **fun** **funny** gallery **games** **google** graphics green gtd hardware **health** history home house **howto** html humor illustration images **imported** information **inspiration** **internet** **java** **javascript** jobs learning library lifehacks **linux** **mac** magazine marketing **media** mobile money movies mp3 **music** myspace network **news** **online** **opensource** **osx** photo **photography** **photos** photoshop php podcast **politics** portfolio productivity **programming** radio rails recipe **recipes** **reference** religion research resources ruby rubyonrails **science** **search** **security** seo sga shop **shopping** slash social **software** sports tech **technology** **tips** **tools** toread **travel** **tutorial** tutorials tv twitter typography ubuntu **video** videos **web** **web2.0** **webdesign** webdev wiki **windows** wordpress work writing youtube

Enkele nadelen

- Homoniemen
- Synoniemen
- Idiosyncratisch taggen
- Meervouden
- Schrijf-en tikfouten
- Algemene versus gespecialiseerde termen

Nadelen zichtbaar in Tag Cloud

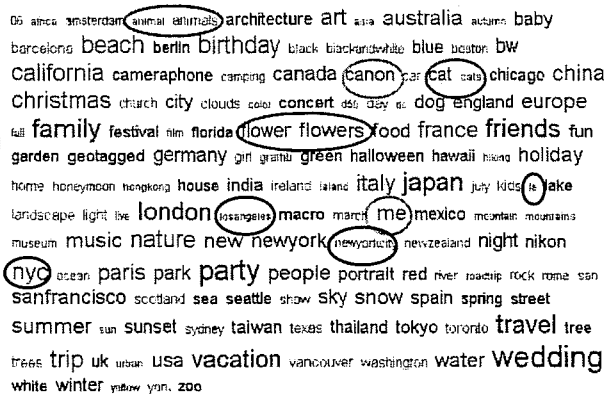
Meervouden

Homoniemen

Idiosyncratisch taggen

Synoniemen

...



flickr™

Oplossing

- Stemming algoritmes
- Clustering
- Facets
- Folksonomies + Ontologies

Voordelen

- Lage kost
- Lage cognitieve overhead
- Gebruikers = ontwikkelaars
- Nieuwe woorden worden direct opgenomen
- Gebruikers vinden hun content zeer snel terug
- Tags gecreëerd door mensen sluiten veel beter aan dan deze gecreëerd door automatische creatie[7]
- ...

3. Vergelijking huidige classificatie-technieken documentalist

- Hiërarchisch – opsommend: vb. DDC
- Analytisch-synthetisch: vb. Colon classificatie, Bliss Bibliographic classification

Dewey Decimal Classificatie (DDC)

- Melvil Dewey
- In 200.000 bibliotheken
- 10 hoofdklasses
- 1 hoofdklasse heeft 10 subklasses
- 1 subclasse heeft 10 secties
- Arabische getallen
- Updates op regelmatige basis

DDC <----> Classificatietechnieken Web

- Parallellismen met de directories op www, alleen zijn het aantal categorieën veel kleiner (Open Directory Project 500.000 [8])
- Enkel hiërarchische relaties
- Een boek kan maar op 1 plaats voorkomen
- Het wordt ge-update door een commissie: mist flexibiliteit van folksonomies: gebruikers hebben geen inspraak

Colon Classificatie (CC)

- S.R. Ranganathan
- Reactie op beperking van hiërarchische en opsommende classificatietechnieken
- Facet classificatie: alle aspecten van een domein worden verzameld in een soort *clusters* of facets. De facets worden gebruikt om de boeken te beschrijven
- Personality Matter Energy Space Time

Bliss Bibliographic Classification (BCC)

- Bouwt verder op werk Ranganathan
- Facets:
  - Thing
  - Kind
  - Part
  - Property
  - Material
  - Process
  - Operation
  - Patient
  - Product
  - By product
  - Agent
  - Space
  - Time

CC & BBC <----> Classificatietechnieken Web

- Facets zijn voorafbepaald
- Opportuniteit folksonomies
- Creatie van facets = doelstelling FaceTag

Referenties

[1] A. Gulli and A. Signorini. (2005) The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan, ACM Press.

[2] P. Lyman,H. R. Varian,K. Searingen,P. Charles, N. Good, L. L. Jordan, and J. Pal. (2003) How much information? Online beschikbaar http://www.sims.berkeley.edu/research/projects/how-much-info-2003

[3] Wikipedia Foundation: About Wikipedia. 2007 Online beschikbaar op http://en.wikipedia.org/wiki/Wikipedia:About#Contributing_to_Wikipedia

[4] Technorati http://technorati.com/

[5] PEW internet & american Life Project (2005) Online News and User-generated Content Dec.2005 http://www.pewinternet.org/dataset_display.asp?r=55

[6] Vander Wal, T. (2004). Folksonomy. http://vanderwal.net/folksonomy.html

[7] Al-Khalifa, H. S. and Davis, H. C. (2007) Exploring The Value Of Folksonomies For Creating Semantic Metadata. International Journal on Semantic Web and Information Systems (IJSWIS) 3(1) pp. 13-39

[8] SIEVERTS, Eric. (2004). Inhoudelijk toegankelijk maken van hybride bibliotheekcollecties. Paper Koninklijke bibliotheek Den Haag. 50 p.