

# *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 Vandalen Woordenboek)
  - Niet indexeerbaar door meeste zoekmachines
  - Vormen het *dark web*

## 1. Informatie overload op het web(2)

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

## 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 geregistreerd bij Technorati[4]

## 2. Informatie zoeken op het web

- Directories
- Zoekmachines & Ontologies
- Folksonomies

## 2. Informatie zoeken op het web

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

Inforum2007 Celine van Damme  
15-04-07 Pag. 7

## Definitie

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

Inforum2007 Celine van Damme  
15-04-07 Pag. 8

# Voorbeeld

Yahoo! Directory

**YAHOO!**

**Arts & Humanities**  
Drama, Music, Movies, TV Shows

**Business & Economy**  
Art, Finance, Investors, Jobs

**Computers & Internet**  
Software, Tools, Sites, Games

**Education**  
Colleges, K-12, Distance Learning

**Entertainment**  
Books, TV Shows, Music, Comics

**Government**  
Business, History, Law, Taxes

**Health**  
Dietetics, Drugs, Fitness, Nutrition

**News & Media**  
Journals, Blogs, Websites

**Recreation & Sports**  
Games, Travel, Sports, Outdoors

**Reference**  
Biographies, Dictionaries, Quizzes

**Regional**  
Countries, Regions, U.S. States

**Science**  
Animals, Astronomy, Earth Sciences

**Top Categories**

- **Adult and Continuing Education** (277)
- **Browse by Region** (168)
- **By Culture or Group** (704)
- **By Subject** (976)
- **Distance Learning** (577)
- **Higher Education** (17256)
- **K-12** (6444)

**Additional Categories**

- **Academic Competitions** (96)
- **Bibliographies** (7)
- **Bilingual** (17)
- **Business to Business**
- **Career and Vocational** (304)
- **Chats and Forums** (22)
- **Conferences** (7)
- **Correctional**
- **Disabilities**
- **Early Childhood Education** (115)
- **Job and Employment Resources**
- **Journals** (34)
- **Legislation** (10)
- **Literary** (33)
- **News and Media** (65)
- **Organizations** (2436)
- **Policy** (48)
- **Programs** (136)
- **Reform** (94)
- **Shopping and Services**

@ → Subcategorie ook komt voor in andere categorieën

Inform2007 Celine Van Damme  
15-04-07 Pag. 9

## 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

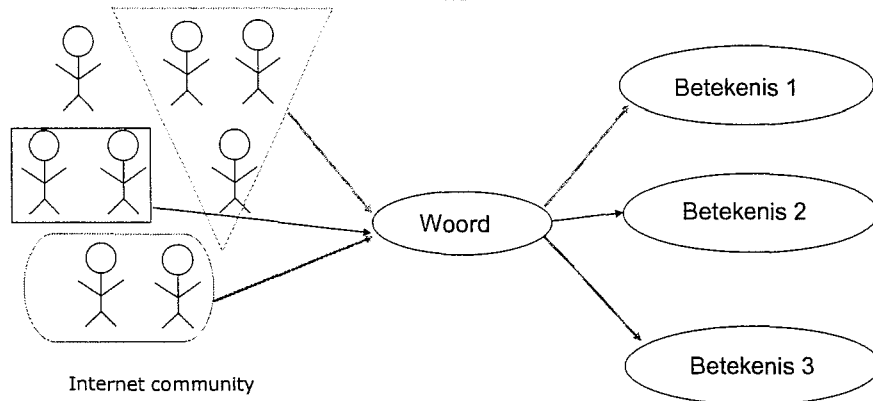
Inforum2007 Celine Van Damme  
15-04-07 Pag. 11

## Werking Zoekmachine (2)

- Zoekopdracht gebruiker toetsen aan database
- Genereren van een resultatenlijst

Inforum2007 Celine Van Damme  
15-04-07 Pag. 12

# Problemen



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

Informatie 2007 Celine van Damme  
15-04-07 Pag. 15

# Oplossing

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

Informatie 2007 Celine van Damme  
15-04-07 Pag. 14

## 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 begripbaar voor machines (RDF, OWL)

Inforum2007 Celine Van Damme  
15-04-07 Pag. 15

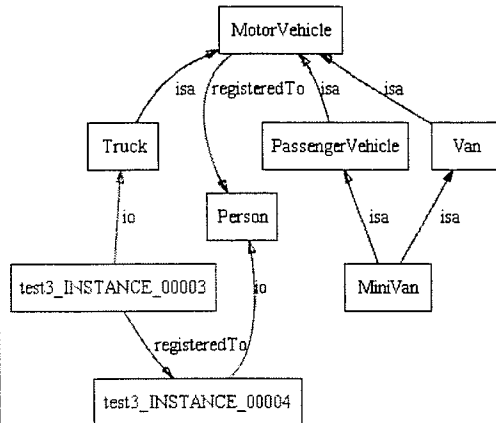
## Een vergelijking...

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

Inforum2007 Celine Van Damme  
15-04-07 Pag. 15



## Voorbeeld



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

Inforum2007 Celine van Damme  
15-04-07 Pag. 17

## Semantische Web

- Belangrijke technologie voor de ontwikkeling van het semantische web  
*web waar alle informatie begripbaar 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

Inforum2007 Celine van Damme  
15-04-07 Pag. 18

# Problemen

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

Inforum2007 Celine van Damme  
15-04-07 Pag. 19

## 2. Informatie zoeken op het web

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

Inforum2007 Celine van Damme  
15-04-07 Pag. 20

## 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<sup>[5]</sup>
- Vergelijkbaar met keywords toegevoegd door auteur(s) aan een paper

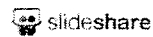
Inform2007 Celine Van Damme  
15-04-07 Pag.21

## Definitie folksonomy (2)

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

Inform2007 Celine Van Damme  
15-04-07 Pag.22

## Enkele voorbeelden...



43 Things



citeulike

BibSonomy



En vele andere...

Inform2007 Ceine Van Damme  
15-04-07 pag. 23

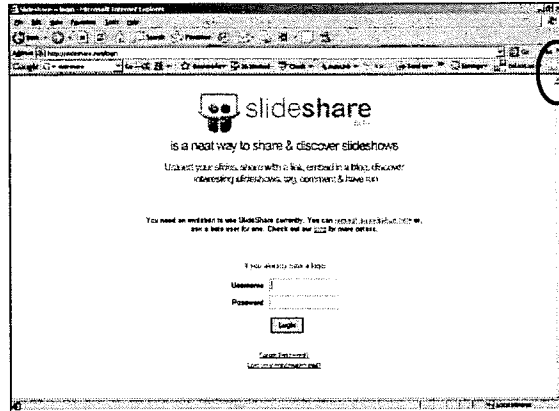
## Del.icio.us (1)



- 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

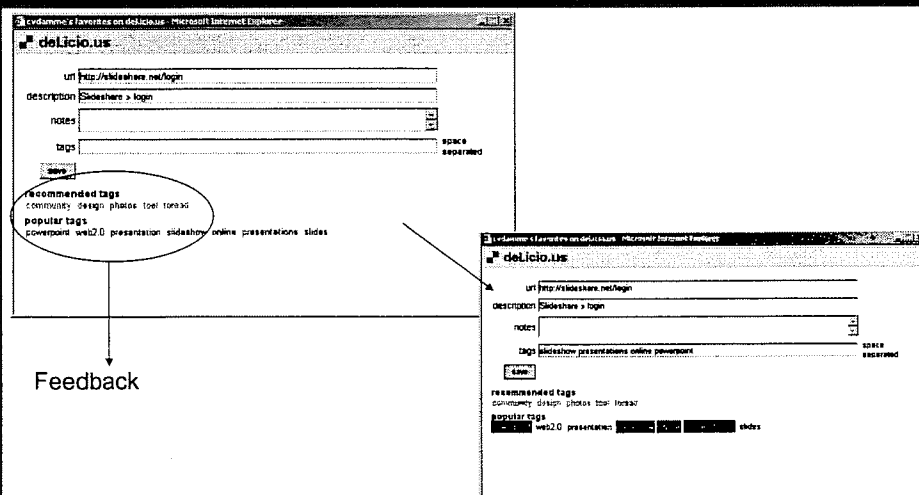
Inform2007 Ceine Van Damme  
15-04-07 pag. 24

## Del.icio.us (2)



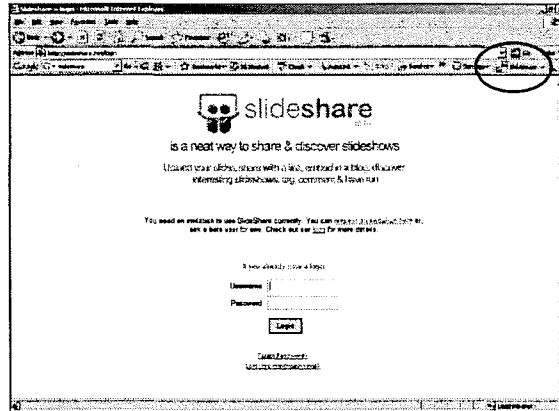
Infernum2007 Celine Van Damme  
15-04-07 Page 25

## Del.icio.us (3)



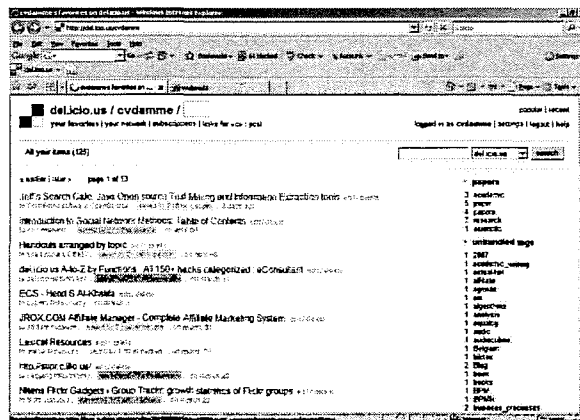
Infernum2007 Celine Van Damme  
15-04-07 Page 25

# Del.icio.us (4)



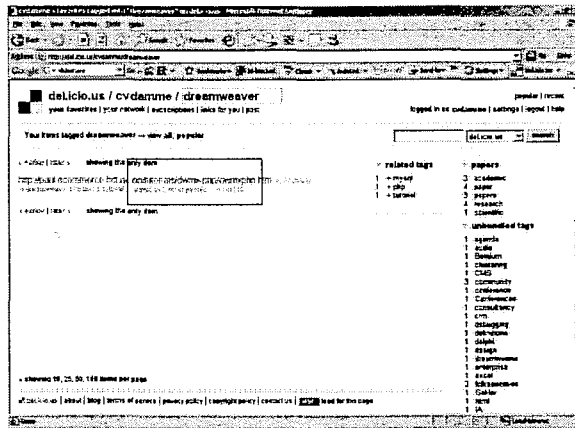
Inform 2007 Céline Van Damme  
15-04-07 Page 22

# Del.icio.us (5)

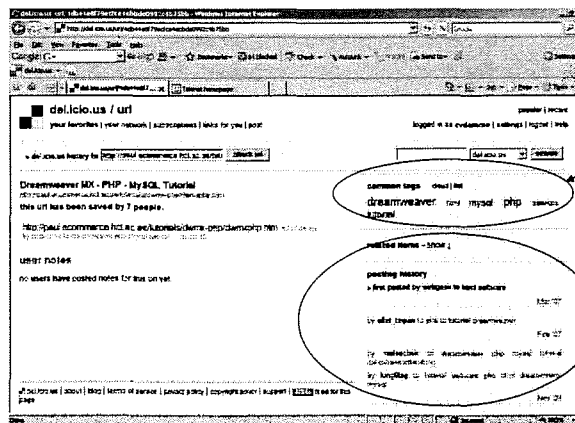


Inform 2007 Céline Van Damme  
15-04-07 Page 25

# 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: rtrn

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 fic film finance firefox flash fonts food free freeware fun funny gallery games google  
graphics green gtk hardware health history home house howto html humor illustration images imported information  
inspiration internet java javascript jobs learning library lifetacks linux mac magazine marketing media mobile money  
movies mp3 music myspace network news online opensource osx photo photography photos photoshop sfp  
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 wkid windows wordpress work writing youtube

Inforum2007 Celine Van Damme  
15-04-07 pag 21

## Enkele nadelen

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

Inforum2007 Celine Van Damme  
15-04-07 pag 22



# Nadelen zichtbaar in Tag Cloud

Meervouden

Homoniemen

Idiosyncratisch taggen

Synoniemen

...

flickr

Os aino amsterdam animal animals architecture art asia australia autumn baby  
barcelona beach berlin birthday black blackandwhite blue ucator bw  
california cameraphone camping canada canon car cat cats chicago china  
christmas church city clouds color concert dslr day de dog england europe  
fall family festival film florida flower flowers food france friends fun  
garden geotagged germany gift granh green halloween hawaii hiana holiday  
home honeymoon hongkong house india ireland itane italy japan july kids lake  
landscape light live london losangeles macro march me mexico mountain mountains  
museum music nature new newyork newyorkers newzealand night nikon  
nyc ocean paris park party people portrait red river roadrip rock rome san  
sanfrancisco scotland sea seattle show sky snow spain spring street  
summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree  
trees trip uk usaa usa vacation vancouver washington water wedding  
white winter yellow ypn zoo

Inforum2007 Celine van Damme  
15-04-07 pag. 33

## Oplossing

- Stemming algoritmes
- Clustering
- Facets
- Folksonomies + Ontologies

Inforum2007 Celine van Damme  
15-04-07 pag. 34

## 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<sup>[7]</sup>
- ...

Inforum2007 Celine Van Damme  
15-04-07 Pag. 35

## 3. Vergelijking huidige classificatie-technieken documentalist

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

Inforum2007 Celine Van Damme  
15-04-07 Pag. 35

## Dewey Decimal Classificatie (DDC)

- Melvil Dewey
- In 200.000 bibliotheken
- 10 hoofdklassen
- 1 hoofdklasse heeft 10 subklassen
- 1 subklasse heeft 10 secties
- Arabische getallen
- Updates op regelmatige basis

Inforum2017 Celine Van Damme  
15-04-07 Pag. 37

## 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

Inforum2017 Celine Van Damme  
15-04-07 Pag. 38

## 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

Infodan2007 Celine Van Damme  
15-04-07 Pag.39

## Bliss Bibliographic Classification (BCC)

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

Infodan2007 Celine Van Damme  
15-04-07 Pag.40

## CC & BBC <----> Classificatietechnieken Web

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

Inform2007 Celine Van Damme  
15-04-07 Pag. 41

## 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](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?tr=55](http://www.pewinternet.org/dataset_display.asp?tr=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.

Inform2007 Celine Van Damme  
15-04-07 Pag. 42