Free text of gecontroleerd vocabulaire: een dilemma

Dr. Gerhard J.A. Riesthuis
Universiteit van Amsterdam
17 februari 2004

Een oude discussie

- In de jaren dertig van de 20ste eeuw
  - Is onderwerpsontsluiting nodig?
  - Is een trefwoordencatalogus beter dan een systematische catalogus?

Onderwerpsontsluiting nodig?
Neen

- *Het is zeer duur*
- *Het is overbodig*
  - *Het kan ook met titelwoorden*
  - *Er bestaan vele bibliografiëen*

Onderwerpsontsluiting nodig?
Ja

- De omweg via bibliografieën is te omslachtig en verschijnen te laat
- Bibliografieën bevatten niet altijd ook boeken
- Systematische ontsluiting nodig voor overzicht van collecties
- Bibliotheek kan onderwerpsontsluiting afstemmen op eigen gebruikers

RESULTAAT VAN DISCUSSIE

- Marburg stopte met onderwerpsontsluiting
- De andere UB's gingen door met een onderwerpscatalogus
- Wel verschuiving van systematische naar trefwoordcatalogi

Waarom die discussie toen?

Documentatiebeweging is volwassen

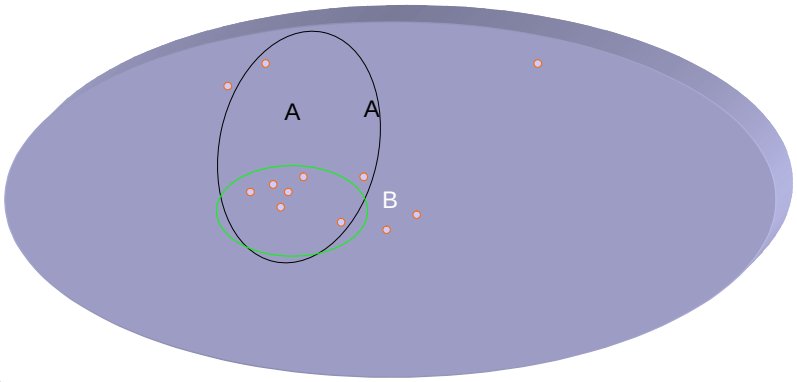
- Veel bibliografieën beschikbaar
- Verschuiving van boeken naar tijdschriftartikelen
- Voor literatuuronderzoek studie van bibliografieën nodig (voor artikelen)

En toen?

- 50 jaar later stopte ook de Universiteitsbibliotheek van Gent met onderwerpsontsluiting
- Argumenten
  - Bibliografieën
  - Titelwoorden
- Maar: besluit wordt nu betreurd

Hoe werkt zoeken?

Collectie met gezochte documenten.



In woorden:

- Zoeken betekent het vormen van een deelverzameling waarin het aandeel van relevante documenten groter is dan in de verzameling als geheel
- Zonder te veel relevante documenten te verliezen
- Ideaal: deelverzameling met alle relevante en geen niet-relevante documenten

Vele wegen naar het doel

- Onderwerpsontsluiting met gecontroleerde vocabulaires
- Titelwoorden
- Vrije tekst [tekst digitaal beschikbaar!]
- Andere
  - Namen van uitgevers
  - Titels van reeksen
  - Namen van auteurs, redacteuren, enz.

Kosten

- Relatief duur: Onderwerpsontsluiting
- Relatief goedkoop: De overige methoden
- Vragen:
  - Zijn de resultaten van onderwerpsontsluiting zoveel beter dat de extra kosten verantwoord zijn?
  - Kunnen de kosten verlaagd worden zonder (veel) kwaliteitsverlies?

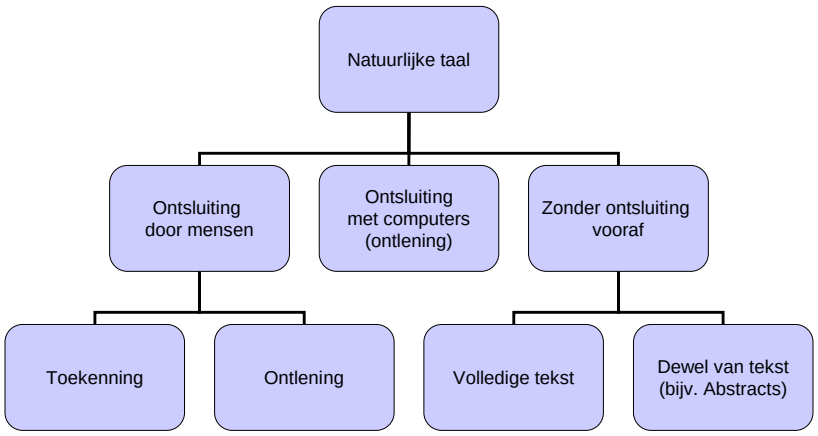
Gecontroleerde vocabulaires

- Classificaties
  - Enumeratieve classificaties
  - Facetclassificaties (analytisch-synthetisch)
- Woordsystemen
  - Trefwoordsystemen
  - Thesauri

Classificaties à ß Woordsystemen

- Classificaties
  - geven overzicht
  - relatief moeilijk bij het zoeken
  - relatief betere resultaten bij ontsluiten
- Woordsystemen
  - relatief makkelijk bij het zoeken,
  - goede resultaten bij zoeken op “namen”, maar slecht bij vage onderwerpen
  - relatief slechtere resultaten bij ontsluiten

Vrije tekst



Voordelen van vrije tekst

- Een betere specificiteit in de retrieval
  - Voorbeeld: “geluidshinder door Schiphol in Heiloo”
  - UDC: 628.517.2 : 629.73 (492.62) [Geluidshinder door luchtvaart in Noord-Holland]

Voordelen van controle

- Reductie van semantische ambiguïteit
  - Door controle van homografen
- Bevordering van consistentie bij het weergeven van een onderwerp
  - Door controle van synoniemen
- Het uitvoeren van veelomvattende zoekacties
  - Door de semantische relaties tussen termen

Kosten

- Bij gecontroleerde vocabulaires
  - bij de input **à** de indexer lost de problemen op
- Bij vrije tekst
  - bij de output **à** de zoeker lost de problemen op

Twee vragen

1. Kan een menselijke zoeker de problemen oplossen?
2. Kan een zoekprogramma de problemen oplossen?

Maar eerst over de problemen …

Het probleem van homografen

- bot (been)
- bot (vis)
- bot (stomp)
- bot (brutaal)

- Tamelijk triviaal probleem in relatief kleine, specialistische domeinen
- Minder triviaal in een universeel domein zoals het Internet

Het probleem van synoniemen

- Echte synoniemen
  - Chemie
  - Scheikunde

- Ook tameijk triviaal probleem: gebruik een woordenboek
- Pseudo-synoniemen: woorden en uitdrukkingen die ongeveer hetzelfde betekenen vormen het echte probleem

Pseudo-synoniemen

- Komen in vele gedaanten voor
  - Peuter, kleuter, klein kind, vier-jarigen, kinderen uit groep één
  - Toneel, toneelspel, toneelstuk, toneelvoorstelling, toneelspeler
- Vrije tekst: betere specificiteit
- Controle: makkelijker “alles” te vinden (betere recall)
- Ook met vrije-tekstsystemen kunnen goede zoekers redelijke recall bereiken (trunceren, ‘or’)
- Vereist kennis van het betreffende domein

Het probleem van de veelomvattende zoekacties

“Regelingen met betrekking tot het houden van huisdieren in Nederlandse gemeenten”

- Alle gemeenten …
- Alle mogelijke dieren die als huisdier worden gehouden …
- Mogelijke plossing een “explode command”
- Vereist echter controle van (pseudo-) synoniemen en semantische relaties

Conclusies

- Controle speelt vooral een rol voor verbetering van de recall [“het vangen van zoveel mogelijk relevante documenten uit de verzameling relevante documenten”]
- Vrije-tekstzoeken waardevol voor vragen met een hoge specificiteit

Hoe te controleren?

- Klassieke informatietalen (met name facetclassificaties en thesauri)
- Post-controlled vocabulaires

Thesauri

- Bij het zoeken wordt een synoniem vervangen door de “standaardterm”
- Relaties meestal beperkt tot
  - Relaties die synoniemen aangeven (UF – USE)
    - Meestal niet tussen samenstellingen en de afzonderlijke termen – Kinderkleding USE Kinderen + Kleding
    - Dit type relaties is belangrijk voor het Nederlands
  - Hierarchische relaties
  - Overige, niet nader gespecificeerde relaties
- Moet bijgehouden worden

Post-controlled thesauri

- In principe niet meer dan verzameling van groepen termen met dezelfde of nauw verwante betekenis
- Vaak ook hierarchische relaties
- Bij het zoeken wordt de term aangevuld met de andere termen uit de groep waartoe de zoekterm behoort
- Bijhouden op basis van context in de gevonden documenten

Algemene conclusies

- Voor goede retrieval nodig:
  - Vrije-tekstzoekmachine
  - Gecontroleerd vocabulaire
    - Kan zowel pre- als post-controlled zijn
    - Bijhouden is een probleem, speciaal in gebieden met snel veranderende vocabulaire
  - Bijhouden met behulp van computers