

# Automatische Document Klassificatie

C.H.A. Koster  
Kath. Universiteit Nijmegen



PEKING  
*Text Classification*

# overzicht

- trends in Information Retrieval?
- nieuwe IR toepassingen
- het PEKING project
- resultaten

# trends in Information Retrieval?

- van databases naar netwerken
- van keywords naar taal
- van queries naar profielen
- nieuwe IR toepassingen.

# van keywords naar taal

## Informatiebehoefte uitdrukken

- in enkele woorden?
- samenhangende uitdrukkingen en vragende zinnen
- in je eigen taal

## Resultaten krijgen

- als een lijst van documenten?
- in de vorm van relevante passages
- ook in andere talen.

# van queries naar profielen

- query bestaat uit enkele woorden

**trein boeken**

**boek een duitse nachttrein**

**boek een trein van Amsterdam naar Trondheim**

# van queries naar profielen

- query bestaat uit enkele woorden
- profiel bestaat uit queries met gewichten

protective	0.137295	collective	-0.017039
safety	0.073049	payment	-0.015385
hazards	0.063303	annual	-0.013483
dangerous	0.060150	agreement	-0.011570
hygiene	0.054662	wage	-0.010222

# van queries naar profielen

- query bestaat uit enkele woorden
- profiel bestaat uit queries met gewichten
- automatische constructie van profielen?

# lerende document klassificatie

- iedere klasse wordt gedefinieerd door een aantal voorbeelden (en tegenvoorbeelden)
- daaruit worden profielen geleerd
- nieuwe documenten worden toegewezen aan de klasse(n) met het meest daarop lijkende profiel.

# nieuwe IR toepassingen

- **pre-classificatie van patent aanvragen** – efficiënt gebruik van mensentijd
- **routing van documenten** – vult het gat tussen OCR en WorkFlow
- **Email routing** – houdt uw Email folders in orde
- **News monitoring** – zie alleen wat je wilt zien
- **Narrowcasting** – ontvang alleen wat je wilt zien
- **Spam filtering** – ongewenste Emails afwimpelen.

toepassingen van profielen en klassificatie.

# case: pre-klassificatie (1)

- handmatige klassificatie

klassificatie-boom

op den duur niet meer te handhaven

iemand verzint een betere klassificatie

maar hoe voer je deze in?

# case: pre-klassificatie (1)

- handmatige klassificatie
- automatische klassificatie

leert een klasse-profiel uit voorbeeld-documenten  
16000 pagina's geklassificeerd binnen 2 minuten  
maar de bestaande documenten zijn vaak slechte  
voorbeelden

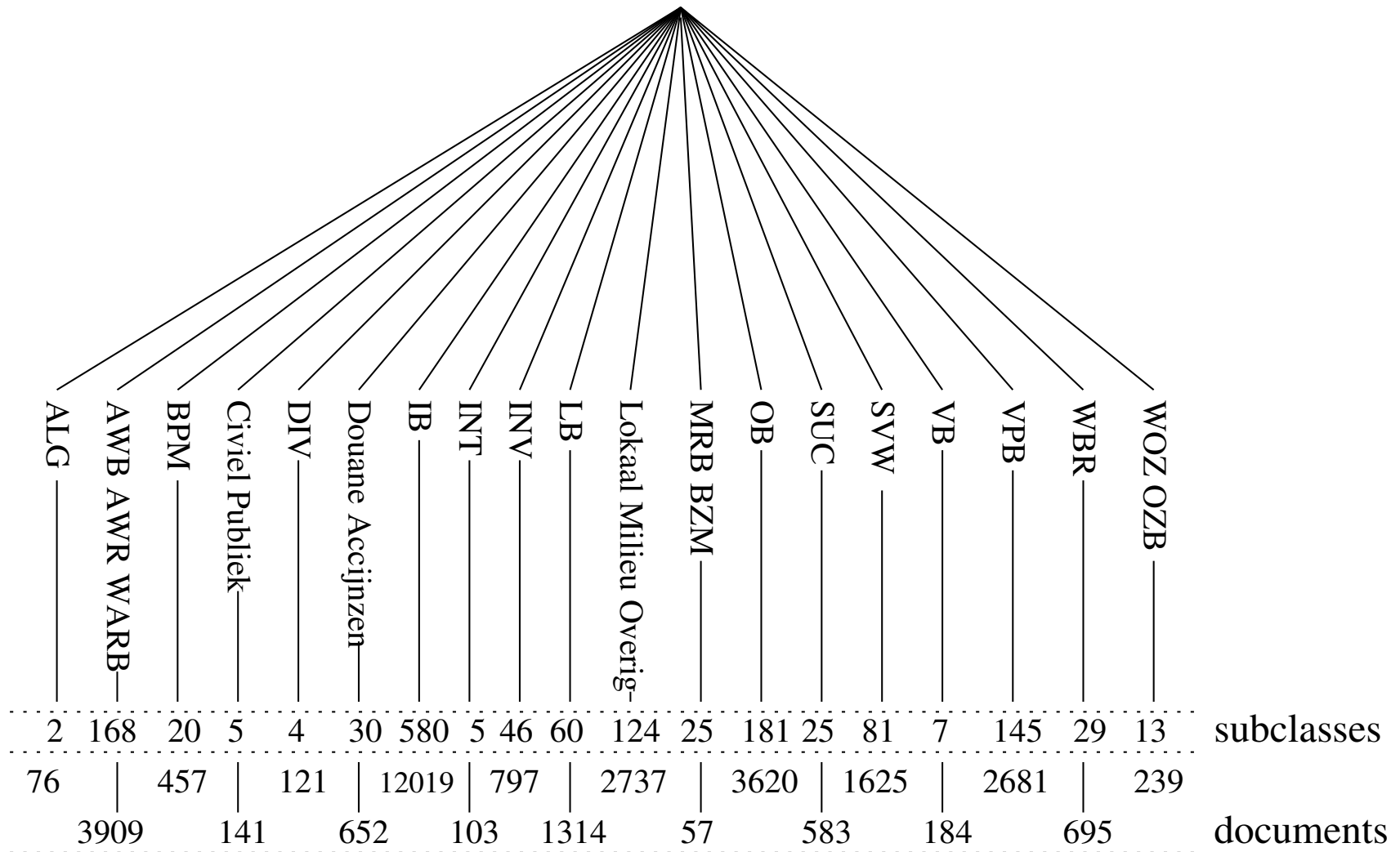
# case: pre-klassificatie (1)

- handmatige klassificatie
- automatische klassificatie
- het bootstrap probleem.

hoe kom je van de handmatige naar de automatische toestand?

ondersteuning nodig voor het herzien van de klassen en het wieden van misplaatste documenten.

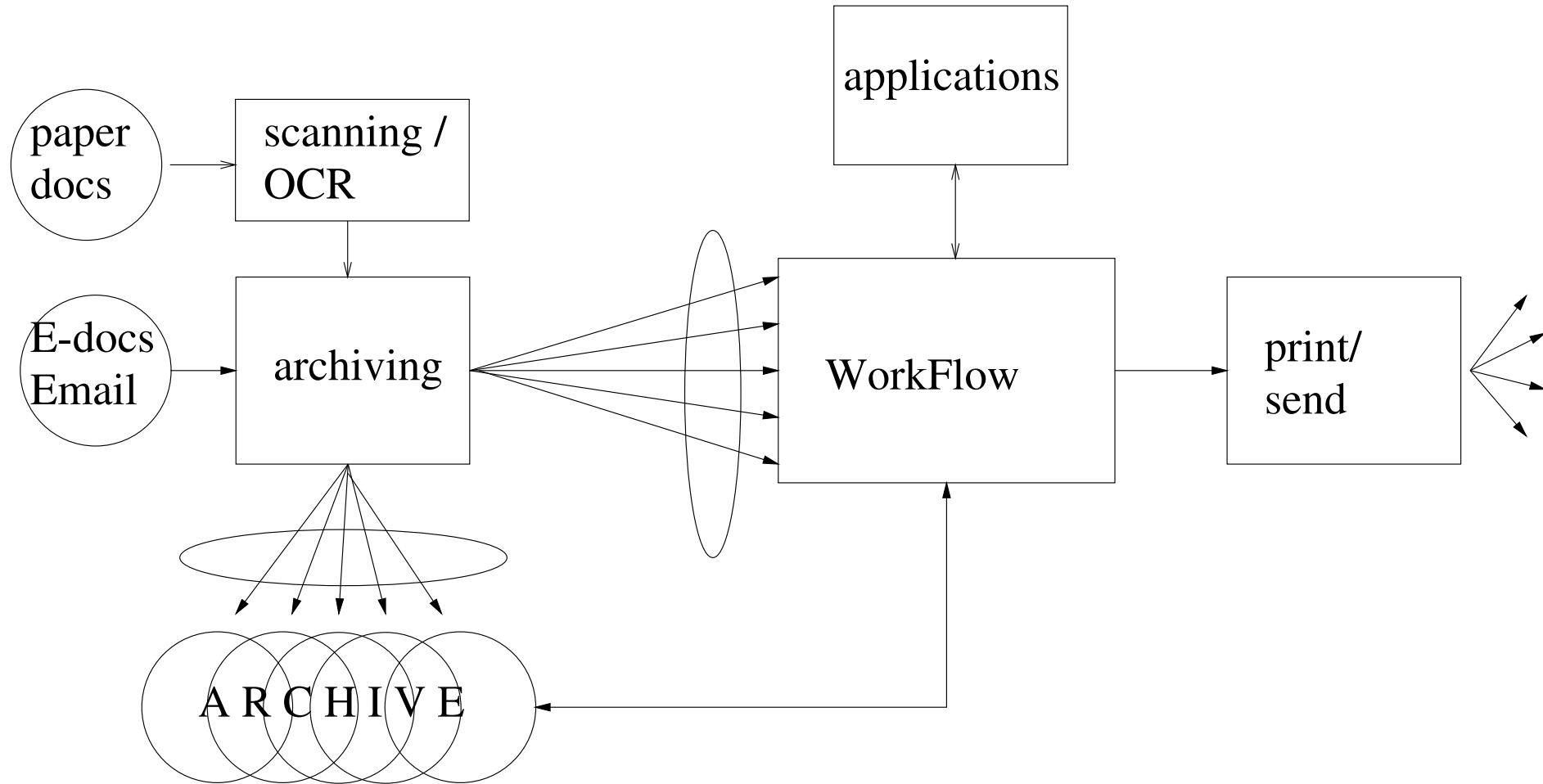
# case: pre-klassificatie (2)



## case: pre-klassificatie (3)

- eerst was de klant overtuigd van de kwaliteit van zijn klassificatieproces
- trainen op bestaande documenten gaf slechts 60% nauwkeurigheid
- uit blind onderzoek bleek dat de klanten de bestaande klassificatie vaak ook onjuist vonden
- herziening van de klassifikatie-boom en automatisch verwijderen van evident verkeerd geklassificeerde documenten gaf 80% nauwkeurigheid
- gebruikers-tevredenheids onderzoek loopt nog.

# klassificatie in de document cyclus



# het PEKING project (deelnemers)

- Europees 5e kader project (IST-25338) 2001/02
- industriële partners:  
META4 R&D, Spanje (coordinator); Quinary, Italië;  
Edmond R&D, Nederland
- academische partners:  
GilcUB, Univ. van Barcelona, Spanje; Univ. van  
Nijmegen (KUN), Nederland.
- gebruikers:  
CRF-FIAT, Italië; CINDOC, Spanje; Fiscaal up to Date,  
Nederland.

# het PEKING project (doelen)

- PEople and Knowledge INformation Gathering
- supervised and unsupervised classification and (cross-lingual) matching of documents in organizations
- research van de KUN (met gilcUB)
  - cross-lingual document classification
  - better linguistic techniques
  - better classification algorithms.
- de nederlandse driehoek (met Edmond en Fiscaal): compleet klassificatie systeem voor Fiscaal up to Date.

# resultaten

- het PEKING project is door de EG als zeer succesvol beoordeeld

# resultaten

- het PEKING project is door de EG als zeer succesvol beoordeeld
- uitstekende methoden voor cross-lingual Text Categorization zonder de documenten te vertalen zijn gevonden

# resultaten

- het PEKING project is door de EG als zeer succesvol beoordeeld
- uitstekende methoden voor cross-lingual Text Categorization zonder de documenten te vertalen zijn gevonden
- de applicatie voor Fiscaal upto Date werkt

# resultaten

- het PEKING project is door de EG als zeer succesvol beoordeeld
- uitstekende methoden voor cross-lingual Text Categorization zonder de documenten te vertalen zijn gevonden
- de applicatie voor Fiscaal upto Date werkt
- het classificatiesysteem PEKING/LCS is marktrijs

# resultaten

- het PEKING project is door de EG als zeer succesvol beoordeeld
- uitstekende methoden voor cross-lingual Text Categorization zonder de documenten te vertalen zijn gevonden
- de applicatie voor Fiscaal upto Date werkt
- het klassificatiesysteem PEKING/LCS is marktrijs
- snel, robuust, nauwkeurig en als enige voorzien van software voor de bootstrap van handmatige naar automatische klassificatie.

[www.edmond.nl](http://www.edmond.nl)

[www.cs.kun.nl](http://www.cs.kun.nl)