Contextual and contentual meta data

Cornelis H.A. Koster, University of Nijmegen

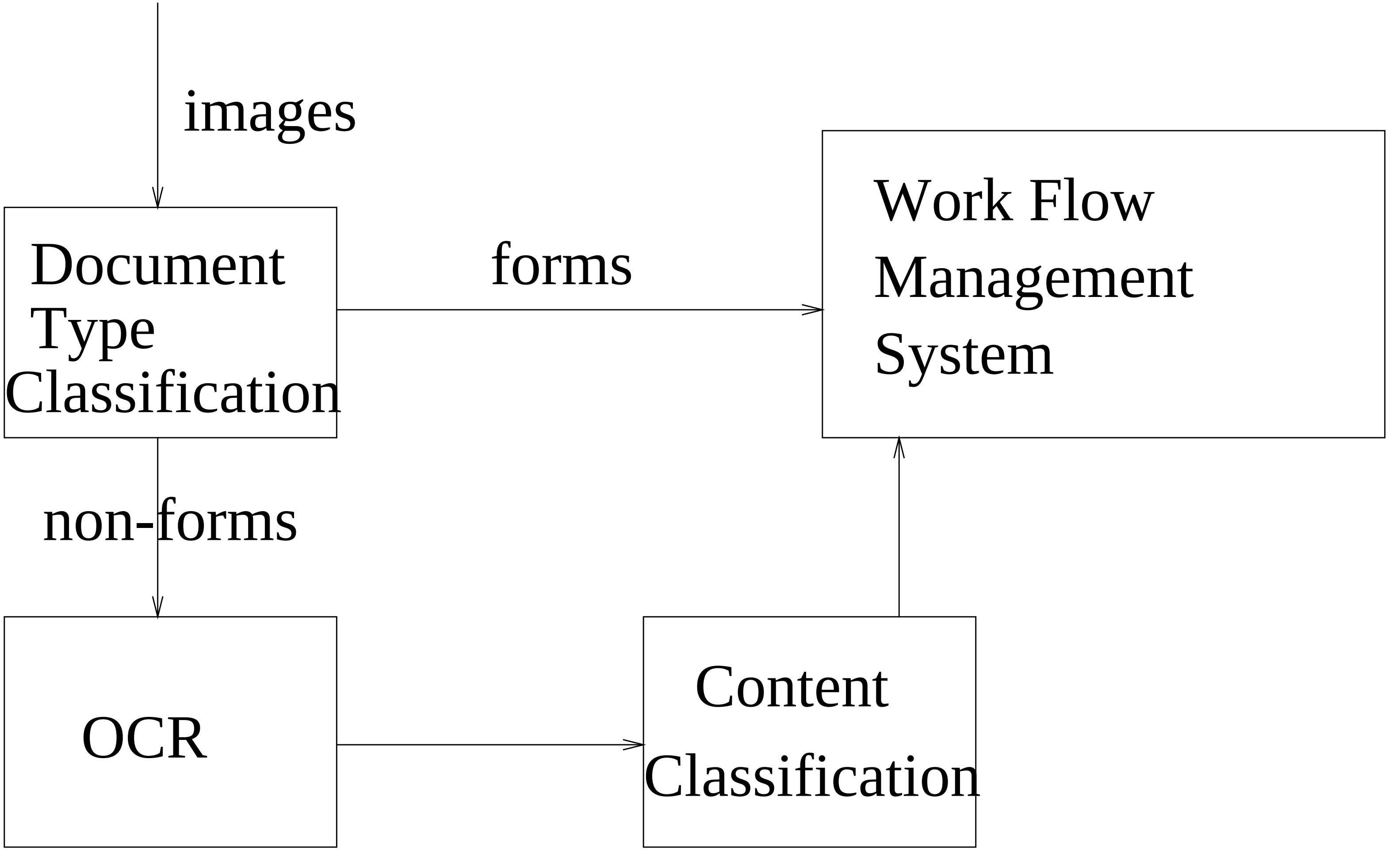
PEKING

Text Classification

Overview

- Meta data: yes
- Meta data: no
- Meta data: later
- Conclusions.

Example 1: Routing



Meta data: yes

- meta data = data “besides" the data
  - greek $\mu\epsilon\tau\alpha$ (= after, besides, with)
  - identification, source, type, ...
- don’t throw away meta data
  - known at the construction of the data
  - which you may well need later, and
  - which you can’t reconstruct later.

Example 2: Thesaurus

- classical **thesaurus** of indexing terms

  assigned terms versus extracted terms

  ontology with spec/gen/syn relations

Example 2: Thesaurus

- classical **thesaurus** of indexing terms
- how do you maintain a thesaurus?

vague boundaries, classification error

drifting terminology, new or obsolete terms

Example 2: Thesaurus

- classical **thesaurus** of indexing terms
- how do you maintain a thesaurus?
- how do you change the thesaurus?

while keeping the meta data up-to date

Example 2: Thesaurus

- classical **thesaurus** of indexing terms
- how do you maintain a thesaurus?
- how do you change the thesaurus?
- how do you introduce a new thesaurus?

Meta data: no

- meta data bring back the inflexibility of the thesaurus
- you are trying to foresee all possible later use
- maybe you will never use (some of) the data
- what do you do if you have forgotten certain meta data?
- how do you address a changing information need?
- can you assign new meta data to the already collected data?

Meta data: later

- some meta data are harder to obtain than others

Meta data: later

- some meta data are harder to obtain than others
- there are different types of meta data:

Meta data: later

- some meta data are harder to obtain than others
- there are different types of meta data:
- contextual meta data

easy at the creation of the document

hard or impossible later

Meta data: later

- some meta data are harder to obtain than others
- there are different types of meta data:
- contextual meta data
- contentual meta data

hard if it needs detective work

can just as well be derived later

Deriving contentual meta data

- assigned keywords

  manual interpretation of document contents expensive

Deriving contentual meta data

- assigned keywords
- extracted keywords

provided they’re there to be extracted

Deriving contentual meta data

- assigned keywords
- extracted keywords
- automatic techniques are available for deriving contentual meta data from documents

  automatic classification
  term extraction
  full-text data mining

Deriving contentual meta data

- assigned keywords
- extracted keywords
- automatic techniques are available for deriving contentual meta data from documents
- store the full text!

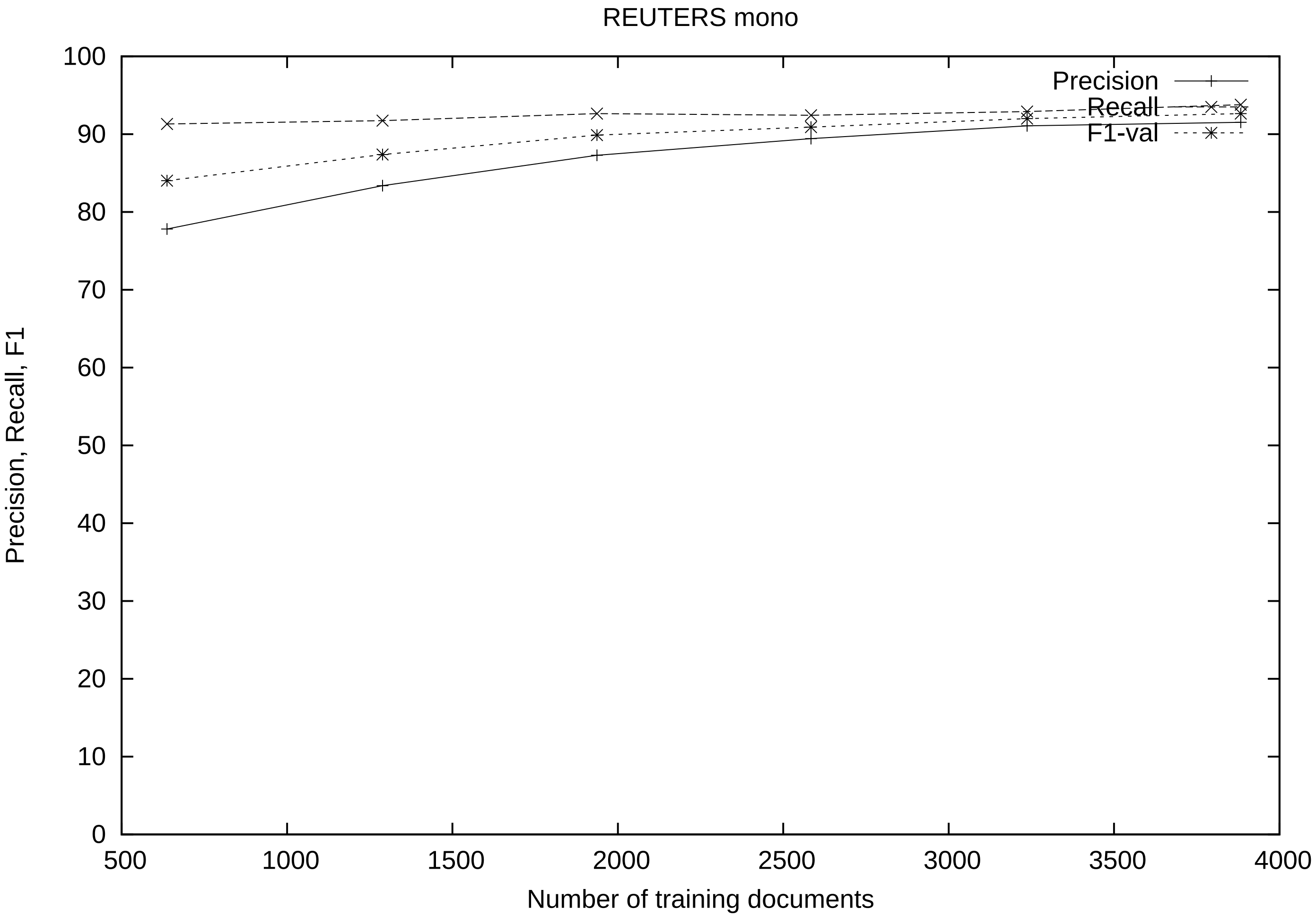
Automatic Document Classification

Given a set of document classes C and example documents for each class, construct a classifier which, given a document d, finds the class(es) to which d is most similar.

- **mono classification**: each document belongs to precisely one class out of n classes
- **multi classification**: each document belongs to zero or more classes out of n
- **hierarchical classification**: mono- or multi-classification in which the classes are arranged in a hierarchy.

Example 3: newspaper classification

- Reuters mono subset, 56 classes, 9000 documents



Term extraction

- state-of-the-art: Regular Expression matching
- grep, awk, perl
- future: grammar-based pattern matching
  - recognizing more complicated patterns
  - exploiting the structure of the language
  - expressions of place, time, syntactic roles.

Full-text mining

- still in infancy
- interactive construction of search profile
- on very large collections
- using strong linguistic techniques
- backed-up by strong statistical techniques.

- see e.g. my research projects: MINIT, BioMine
- a new search engine for a new way of searching.

Conclusions

- adding meta data which are not used is a waste
- retrospective change of meta data is practically impossible
- ==> add only contextual meta data
- use classification, extraction and text mining to derive contentual meta data.

`http://www.cs.kun.nl/peking/`