

From key-words to key-phrases

C.H.A. Koster
University of Nijmegen
The Netherlands
`kees@cs.kun.nl`

ICT-kenniscongres
6 September 2001

- From Retrieval to Classification
- Keyword-based retrieval
- Phrase-based retrieval
- Noun phrases as terms
- Verb phrases as terms
- Linguistics for Information Retrieval
- The PEKING project

Traditional Information Retrieval

Given a collection of natural language **documents** and a **query** in natural language, find the document(s) (best) matching that query.

Less traditional Information Retrieval

- mail filtering
- mail routing
- news monitoring
- narrowcasting
- contents classification.

The changing face of Information Retrieval

- from retrieval to classification
- from keywords to key phrases
- towards novel applications.

1 From Retrieval to Classification

Document Classification

Given example documents for n document classes, construct for each a class a profile P_i , such that given a document d with the profile p , the class(es) are found whose profile is most similar to p .

Automatic learning

- in the training phase, documents labeled with a class are presented to the system
- from each document, a **document profile** is derived (a weighted set of terms)
- from these, the system builds a classifier for each class, in the form of a **class profile**
- in the test phase or production phase the class profiles are used to assign new documents to a class.

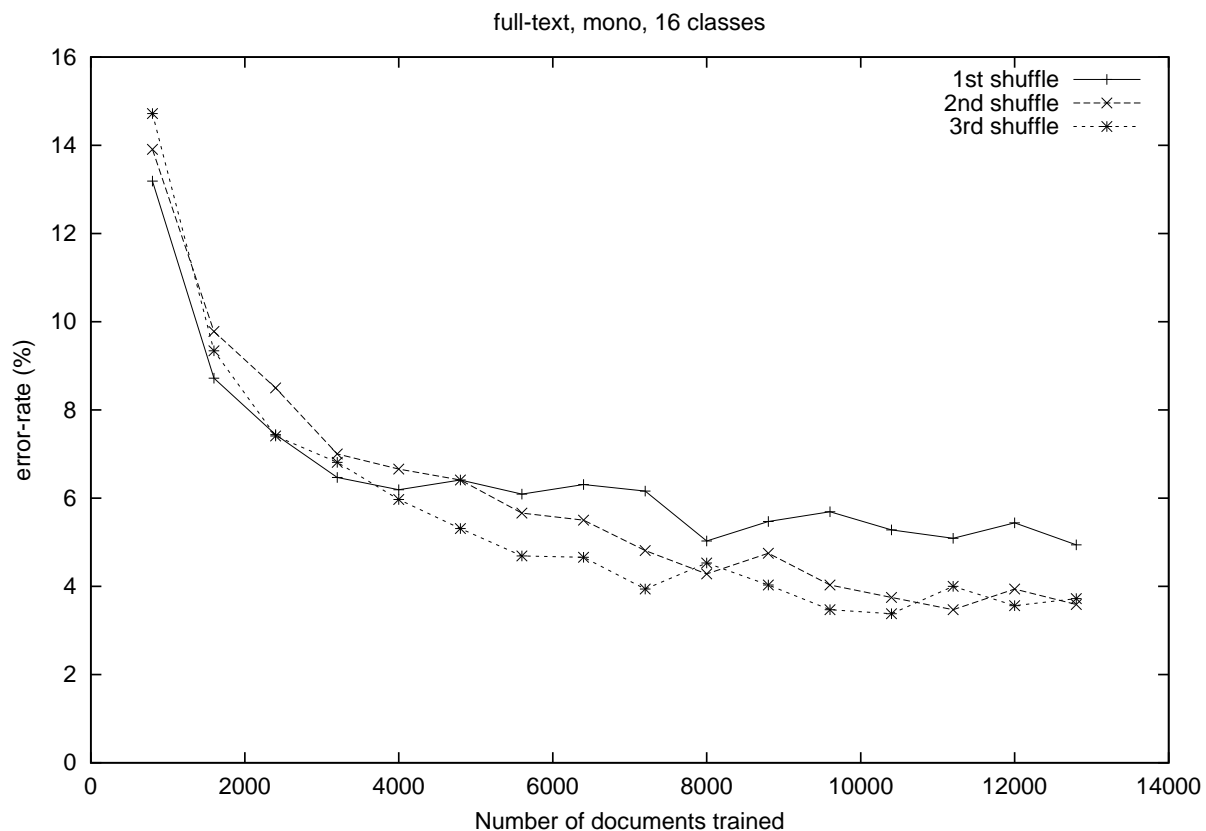
1.1 Example: classifying patent applications

- many learning and classification algorithms available
- Rocchio, Simple Bayesian, Support Vector Machines, Winnow ...
- can classify anything

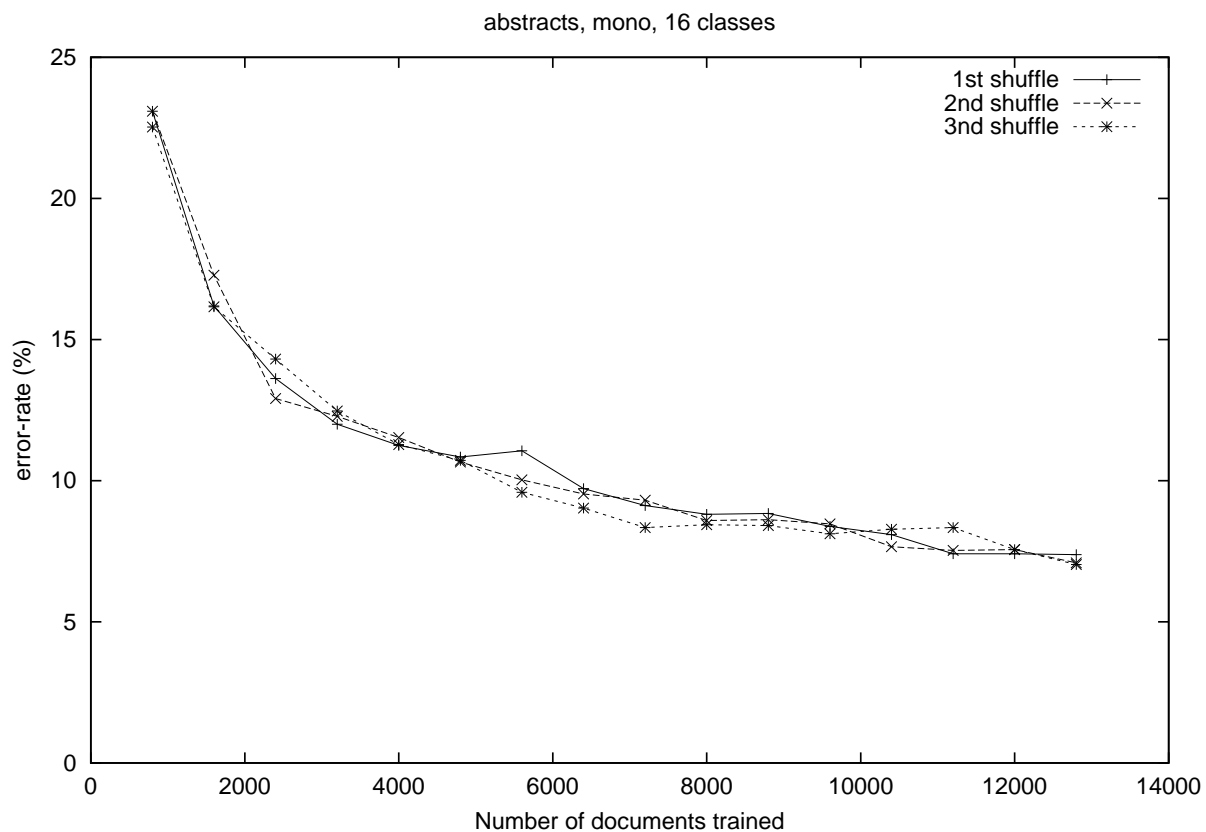
- **A technology in search of applications!**

- KUN studies the classification of large collections
of large documents
- experiment with 16000 patent applications in 16 categories
- split 80/20 into train and test set
- three different random shuffles.

- full-text documents (about 10000 words each)



- abstracts (about 500 words each)



2 Keyword-based retrieval

The **keyword hypothesis**

document = set of words

query = set of words

If the word x occurs in the document then the document is about x .

precision:

$$P = \frac{\#documents\ retrieved\ and\ relevant}{\#documents\ retrieved}$$

recall:

$$R = \frac{\#documents\ retrieved\ and\ relevant}{\#documents\ relevant\ and\ in\ collection}$$

- recall = precision = 50% is considered good
- but that means that 50% of the documents retrieved are irrelevant, and only half of the relevant documents is found
- luckily recall is hard to measure in traditional retrieval ...
- classification provides ideal testbed for new techniques, because recall is measurable using pre-classified documents.

Problems of keyword-based retrieval

- document and query are just isolated words, the connections between the words are lost
the Hilary Clinton Health Care Bill Proposal
- one word may have different meanings (**polysemy**)
- different words may have the same meaning (**synonymy**)
but also hyper/hyponymy; mero/holonymy; antonymy
and synonymy is often partial (context-dependent)
- a word may appear in different forms (**morphological variation**)
in English and Dutch mostly just singular/plural, but
much more in inflected languages (Russian, Greek, Arabic).

- for stylistic reasons, people vary their choice of words
air pollution
pollution of the air
- for precision, people use **phrases** rather than single **words**
- people usually don't talk in isolated words except when searching the Internet.

3 Phrase-based retrieval

document = set of terms

query = set of terms

If the term x occurs in the document then the document is about x .

- a term may be a (key) word or a phrase
- still disregarding discourse structure.

3.1 Noun phrases as terms

- noun phrase serves as a reference to or description of a complicated concept
- is roughly that which can occur as the subject or object of a sentence
 - software engineering conference
 - the air pollution in industrial activity zones
 - a threshold, which should not be exceeded
- a noun made more precise by
 - premodifiers (adjectives, nouns, complete noun phrases) and
 - post modifiers (preposition phrases, relative phrases, etc).

3.2 Phrases and frames

- the “essence” of a phrase can be represented by a (recursive) head-modifier structure (**frame**):

```
[pollution, air]
[pollution, water]
[engineering, software]
[conference, [engineering, software]]
[[proposal, [bill, [care, health]]], [clinton, hillary]]
```

3.3 Normalization

- use of noun phrases brings little improvement:
 - it raises precision but lowers recall
- people avoid repeating a phrase literally
- **normalization** needed to map equivalent phrases onto the same frame
- undoing syntactic variation
 - pollution of the air \Rightarrow [pollution, air]
- undoing morphological variation
 - car doors \Rightarrow [door, car]

3.4 Verb phrases as terms

- the verb phrase describes a situation or process
- consists of a (main) verb with its auxiliaries and a number of complements, each fulfilling a certain role
- abstract structure

verbal kernel *subject* [*object*] [*indirect object*] ...

- nounphrases occur as (or in) its complements. Some examples:

I subject
would have liked to hit verbal kernel
the man object
with a hammer. prep. complement (instrumental)

my husband subject
does not understand verbal kernel
me. object

I subject
owe verbal kernel
you indirect object
twenty dollars. object

3.5 Normalizing transformations on the VP

- idea is:

to map obviously equivalent frames on each other

not to lose information obviously relevant for retrieval purposes.

- representation by nested frames:

[subject, [verb, object | other complements]]

- elimination of irrelevant elements, like adverbs.
- bringing passive VP's into active form
- elimination of auxiliaries for time and modality
- morphological normalization of main verb form
- unnesting of the complements, reducing each NP complement to its head plus a separate NP frame.

Examples:

[he, [visit, [conference, [engineering, software]]]

[[doctor, [stink, of gin]], come in | [lie, in corner]]

3.6 Unnesting

- a composed frame [c, [b, a]] will be decomposed into two frames [b, a] and [c, b]
- using the head b as an *abstraction* for [b, a] in the second one
- applied recursively to the head and modifiers of a frame
- result: a frame is rewritten into a collection of terms = frames without nesting
- example:

```
{[visit, conference],  
 [conference, engineering],  
 [engineering, software]}
```

- example:

```
{[doctor, stink],  
 [stink, of gin],  
 [doctor, come in],  
 [doctor, lie]  
 [lie, in corner]}
```

- example:

```
{[clinton, hillary],  
 [care, health],  
 [bill, care],  
 [proposal, bill],  
 [proposal, clinton]}
```

which does not include [clinton, bill].

4 Linguistics for Information Retrieval

- Linguistics has much to offer to modern IR
- but only the simplest gifts are accepted:

low technological investment	
in current use	unusual or unexplored
stoplists stemming lemmatization noun phrases	collocations verb phrases syntactic normalization fuzzy semantic matching semantic interpretation
high technological investment	

- simple linguistic techniques have brought modest rewards
- point of diminishing returns reached
- for a long time, better statistical techniques reaped more progress
- again, point of diminishing returns reached
- high-tech linguistics needs large investments
- so why bother?

Missing linguistic resources:

- technology for parsing at the speed of the Internet
- special grammars for Information Retrieval
- analyzed corpora (**treebanks**) providing statistical data
- lexica of collocations and phrases (including synsets)
- **all of these in the public domain.**

5 The PEKING project

- “PEople and Knowledge INformation Gathering”
- 5th framework project
- supervised and unsupervised classification and (cross-lingual) matching of documents in organizations.

the consortium:

1. META4 R&D (coordinator)
2. Univ. of Barcelona (GilCUB) and CINDOC in Spain
3. Quinary and CRF-FIAT in Italy
4. Univ. of Nijmegen (KUN), Edmond bv and Fiscaal up to Date in The Netherlands

goal of the project:

to develop a practical classification system for companies which make a living by providing value-added access to a database of documents

- hierarchical mono-classification by subject
“what is the logical place to store this document in a tree of documents?”
- orthogonal multi-classification on index terms
“which index terms are applicable to this document?”
- additional hypertext structure
- bootstrap from current manual system to semi-automatic system.

key questions on the application side:

- Can an automatically learning system be made to provide a hierarchical classification which is good enough for the users?
- Can the consistency and quality of automatic classification approximate the experience and insight of experts performing manual classification?
- Can a semi-automatic system provide an economically attractive solution to the disclosure problems of firms like FISCAAL?

technical problems to be solved:

- learning reliably from unreliably classified documents
- exploiting the notion of uncertainty in improving classification results
- developing special IR grammars for English, Dutch, Spanish, Italian
- deriving normalized phrasal representations from documents, and
- using those phrasal representations in conjunction with statistical learning methods to increase precision in learning.

best statistical learning techniques

+ best linguistical document representation

= best document classification system