7. Besondere Datentypen und Anwendungen

Inhalt dieses Kapitels

7.1 Temporal Data Mining

Problemstellung, Sequential Patterns, Modifikation des Apriori-Algorithmus

7.2 Spatial Data Mining

Aufgaben und Probleme, typische Methoden, räumliche Charakterisierung und Trenderkennung

7.3 Text- und Web-Mining

Aufgaben und Probleme, Clustering von Web/Text-Dokumenten, Suchmaschine mit Berücksichtigung der Linkstruktur

7.4 Lernen von Ontologien

7.5 Text Klassifikatio/Clustern mit Hintergrundwissen

7.1 Temporal Data Mining

Problemstellung

- Analyse von zeitbezogenen Daten
- Anwendungen

  Finanzen: Aktienkurse, Inflationsraten, . . .

  Medizin: Blutdruck, . . .

  Meteorologie: Niederschläge, Temperaturen, . . .
- ausgezeichnetes Attribut:

  Punkte oder Abschnitte in einem zeitlichen Bezugssystem

  ➡ impliziert zeitliche Ordnung der Datensätze

7.1 Temporal Data Mining

Problemstellung

- zwei Arten von Methoden

  Analyse zeitlicher Zusammenhänge *innerhalb* einzelner Abläufe

  Analyse zeitlicher Zusammenhänge *zwischen* verschiedenen Abläufen

- Besonderheit des Temporal Data Mining

komplexe zeitliche Relationen zwischen Zeitpunkten und Zeitintervallen: „während“, „überschneidend“, „direkt aufeinanderfolgend“ . . .

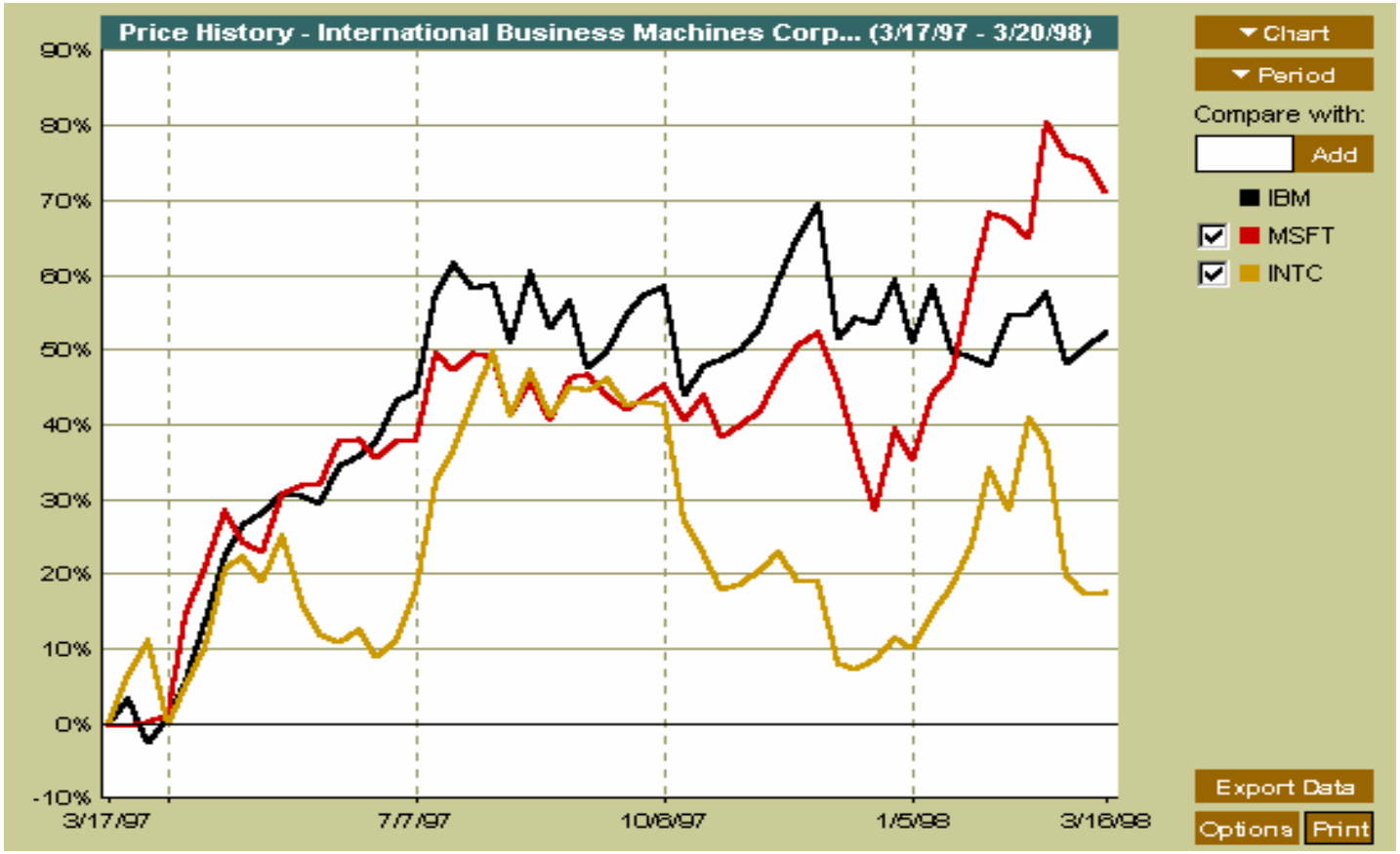
neue Typen interessanter Regeln

⇒

zusätzliche Komplexität der Algorithmen

7.1 Zeitreihen -Analyse

Beispiel



7.1 Zeitreihen-Analyse

Komponenten von Zeitreihen [Fahrmeier et al.1999]

Trendkomponente

langfristige systematische Veränderung

Konjunkturkomponente

Verlauf von Konjunkturzyklen

Saisonalkomponente

jahreszeitlich bedingte Schwankungen

Restkomponente

Irreguläre Veränderungen, zufällig, relativ gering

7.1 Zeitreihen-Analyse

Methoden [Fahrmeier et al.1999]

Globale Regression

- Auswahl eines Funktionstyps
- Schätzung der unbekannten Parameter mit Hilfe der Methode der kleinsten Fehlerquadrate

globaler Trend häufig zu grob

Lokale Methoden

- gleitender Durchschnitt (Moving Window)

  Glättung
- lokale Regression

  Regressionsfunktion für Umgebung des jeweiligen Punkts

7.1 Sequential Patterns

Idee

- nicht einzelne Transaktionen, sondern Mengen von zusammengehörigen und zeitlich geordneten Sequenzen von Transaktionen
- häufige Sequenz:

  viele Kunden, die zu einem Zeitpunkt Produkte *A*, *B*, *C* eingekauft haben, haben zu einem späteren Zeitpunkt auch die Produkte *D*, *E* und *F* gekauft

  „5% aller Kunden haben zuerst das Buch *Solaris*, danach das Buch *Transfer* und dann *Der Futurologische Kongreß* gekauft."

- Anwendung

  Kunde hat schon *Solaris* gekauft, bestellt jetzt *Transfer*:

  ⇨ empfehle *Der Futurologische Kongreß*

7.2 Spatial Data Mining

Problemstellung

- Analyse von raumbezogenen Daten
- ausgezeichnetes Attribut:

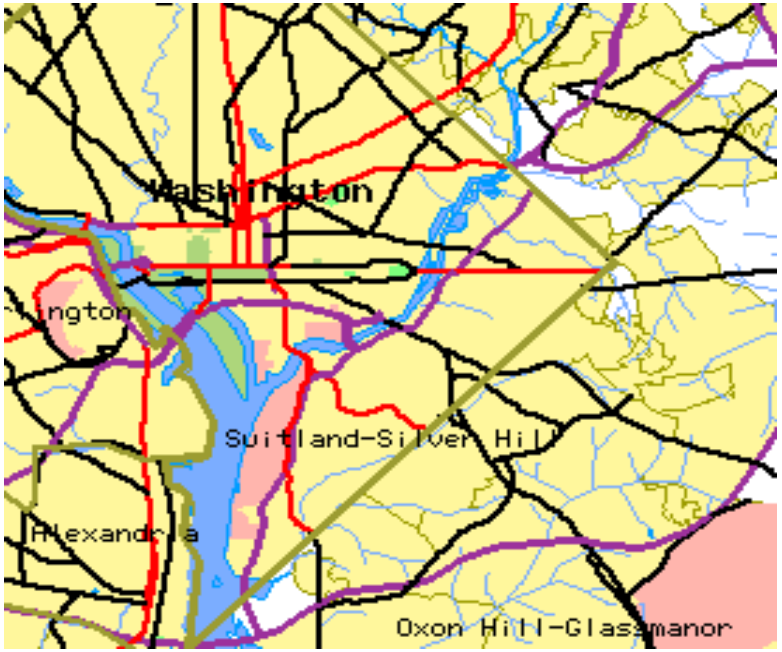
  Lage und Ausdehnung in einem
  2- oder 3-dimensionalen Raum

  ➡ Punkte, Linien, Polygone, Polyeder

- Anwendungen

  Geographie: Topologische Karten,
  Thematische Karten, . . .

  Biologie: Proteine, . .



7.2 Spatial Data Mining

Problemstellung

- Aufgaben

  Analyse von *einzelnen* räumlichen Verteilungen bestimmter Attribute

  Analyse von Abhängigkeiten *zwischen* räumlichen Verteilungen von Attributen

- Anwendungen

  Geo-Marketing

  Verkehrssteuerung

  Umweltschutz . . .

- Besonderheit des Spatial Data Mining

  Attribute von Nachbarn beeinflussen ein gegebenes Objekt

  Einfluß hängt ab von räumlichen Nachbarschaftsbeziehungen

7.3 Text- und Web-Mining

Problemstellung

- Analyse von Text- und Hypertext-Daten sowie ihrer Benutzung
- Anwendungen

  elektronische Mails einer Firma

  Newsgroup-Artikel

  Webseiten aus dem Internet oder dem Intranet einer Firma
- Text- und Hypertext-Daten

  Text

  Präsentation

  Inhalt

  Hyper-Links

7.3 Text- und Web-Mining

Problemstellung

- Text

  Transformation eines Dokuments D in Vektor $r(D) = (h_1, ..., h_d)$

  $h_i \geq 0$: die Häufigkeit des Terms t_i in D

  Reduktion der Anzahl der Terme

  Stop-Listen, Stemming, Entfernen besonders häufiger bzw. seltener Terme

- Präsentation (HTML)

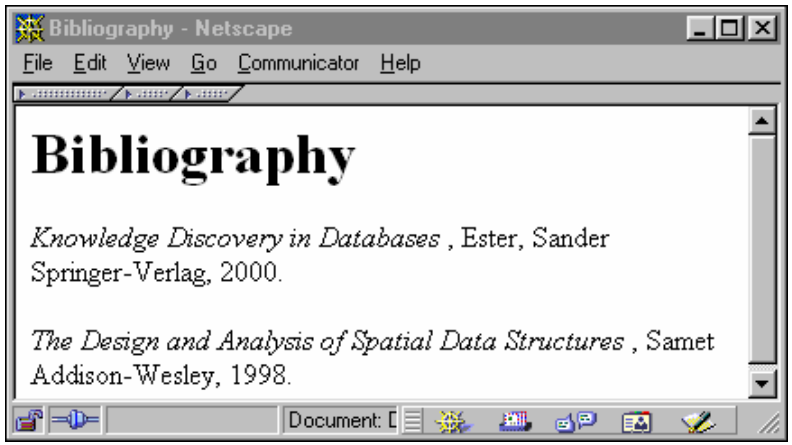
<h1> Bibliography </h1>

 Knowledge Discovery in Databases,

Ester, Sander

Springer-Verlag, 2000.

. . .



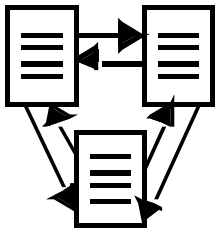
7.3 Text- und Web-Mining

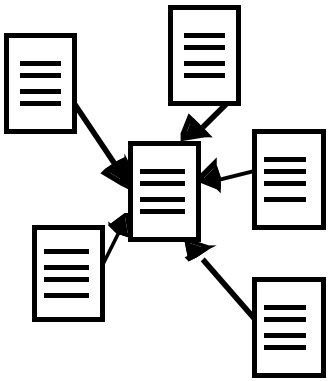
Problemstellung

- Inhalt (XML)

```
<bibliography> <book> <title> Knowledge Discovery in Databases </title>
  <author> Ester </author>  <author> Sander </author>
  <publisher> Springer-Verlag </publisher>
  <year> 2000 </year>
  </book>
  . . .
</bibliography>
```

- Hyper-Links





7.3 Text- und Web-Mining

Problemstellung

- Aufgaben

  Analyse von *Inhalt* und *Struktur* von Hypertext-Dokumenten

  Analyse der *Link-Struktur* einer Menge von Hypertext-Dokumenten

  Analyse der *Benutzung* einer Menge von Hypertext-Dokumenten

- Besonderheit des Text- und Web-Mining

  ➡ Diversität des Vokabulars, z.B. verschiedene Sprachen

  Vagheit der Texte

  Unterschiedliche Qualität der Texte

  ➡ Link-Struktur

7.3 Clustering der Antwortmengen von Suchmaschinen

Motivation

- Ergebnisse von Web-Suchmaschinen

  im allgemeinen in Form einer Liste

- Probleme

  Antwortlisten typischerweise sehr lang

  viele Terme treten in ganz verschiedenen Kontexten auf

  sehr unübersichtliche Darstellung

z.B. „Cluster“: Datenanalyse, Rechnernetze, Astronomie, . . .

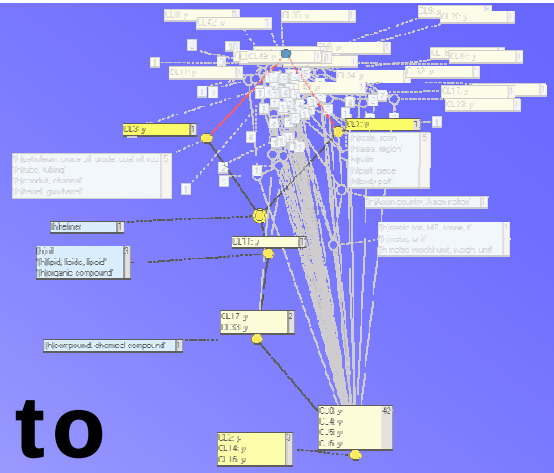
- Ziel

  Clustering der Antwortmengen nach Kontexten

  Browsen des Clustering statt der Antwortliste

CLUSTERCOUNT	60
EXAMPLE	100
MINCOUNT	15

Mittelwert - PURITY				WEIGHT PRUNE					
				tfidf			without		
ONTO	HYPDEP	HYPDI	HYPINT	30	5	0	30	5	0
false	0	context	add	0,570	0,547	0,526	0,470	0,461	0,462
true	0	context	repl	0,572	0,552	0,526	0,462	0,467	0,459
			add	0,585	0,560	0,546	0,483	0,484	0,483
			only	0,603	0,581	0,571	0,485	0,478	0,481
		first	repl	0,562	0,550	0,525	0,469	0,469	0,466
			add	0,590	0,563	0,551	0,488	0,482	0,484
			only	0,599	0,584	0,569	0,483	0,487	0,477
		all	repl	0,512	0,504	0,488	0,472	0,469	0,472
			add	0,533	0,530	0,525	0,481	0,480	0,484
			only	0,500	0,503	0,497	0,455	0,456	0,454
	5	context	repl	0,602	0,595	0,578	0,404	0,396	0,398
			add	0,618	0,598	0,586	0,423	0,413	0,407
			only	0,593	0,594	0,588	0,392	0,387	0,390
		first	repl	0,602	0,602	0,584	0,411	0,400	0,401
			add	0,616	0,600	0,590	0,423	0,417	0,411
			only	0,593	0,593	0,585	0,394	0,389	0,390
		all	repl	0,522	0,521	0,519	0,318	0,315	0,308
			add	0,542	0,537	0,531	0,324	0,318	0,309
			only	0,520	0,513	0,513	0,315	0,311	0,308



Using Ontologies to Improve the Text Custering and Classification Task

Andreas Hotho

UNIKASSEL VERSITÄT

FACHBEREICH MATHEMATIK / INFORMATIK
Fachgebiet Wissensverarbeitung
STIFTUNGSPROFESSUR DER GEMEINNÜTZIGEN HERTIE-STIFTUNG

Joint work with:
- **Stephan Bloehdorn**
- **Steffen Staab**
- **Gerd Stumme**

Motivation

requirements on the cluster methods

Efficient

- Results should also be available on large data sets or on ad-hoc collect e.g. from search engines

Effective

- Cluster result must be correct

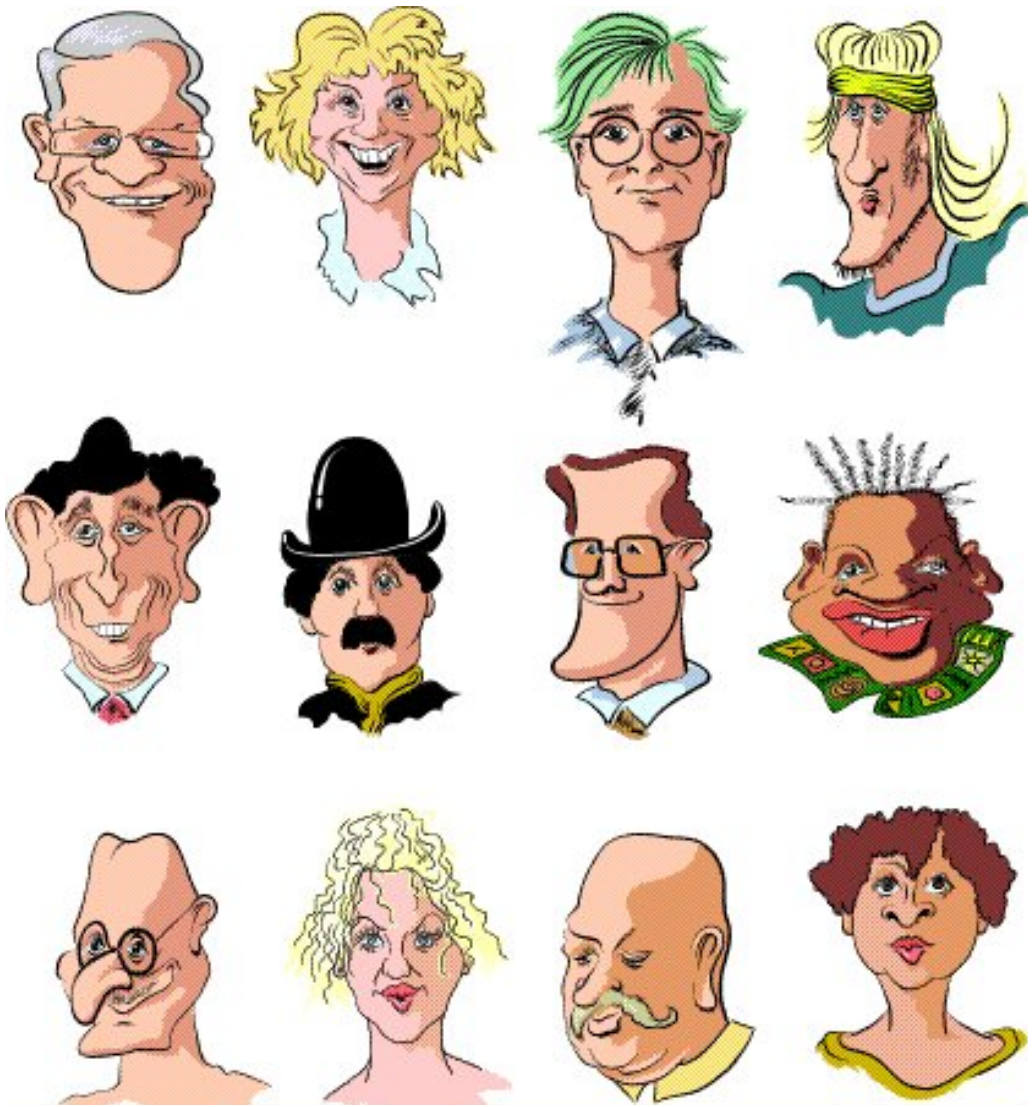
Problem of explanatory power

- Results of the clustering process must be understandable

User interaction und subjectivity

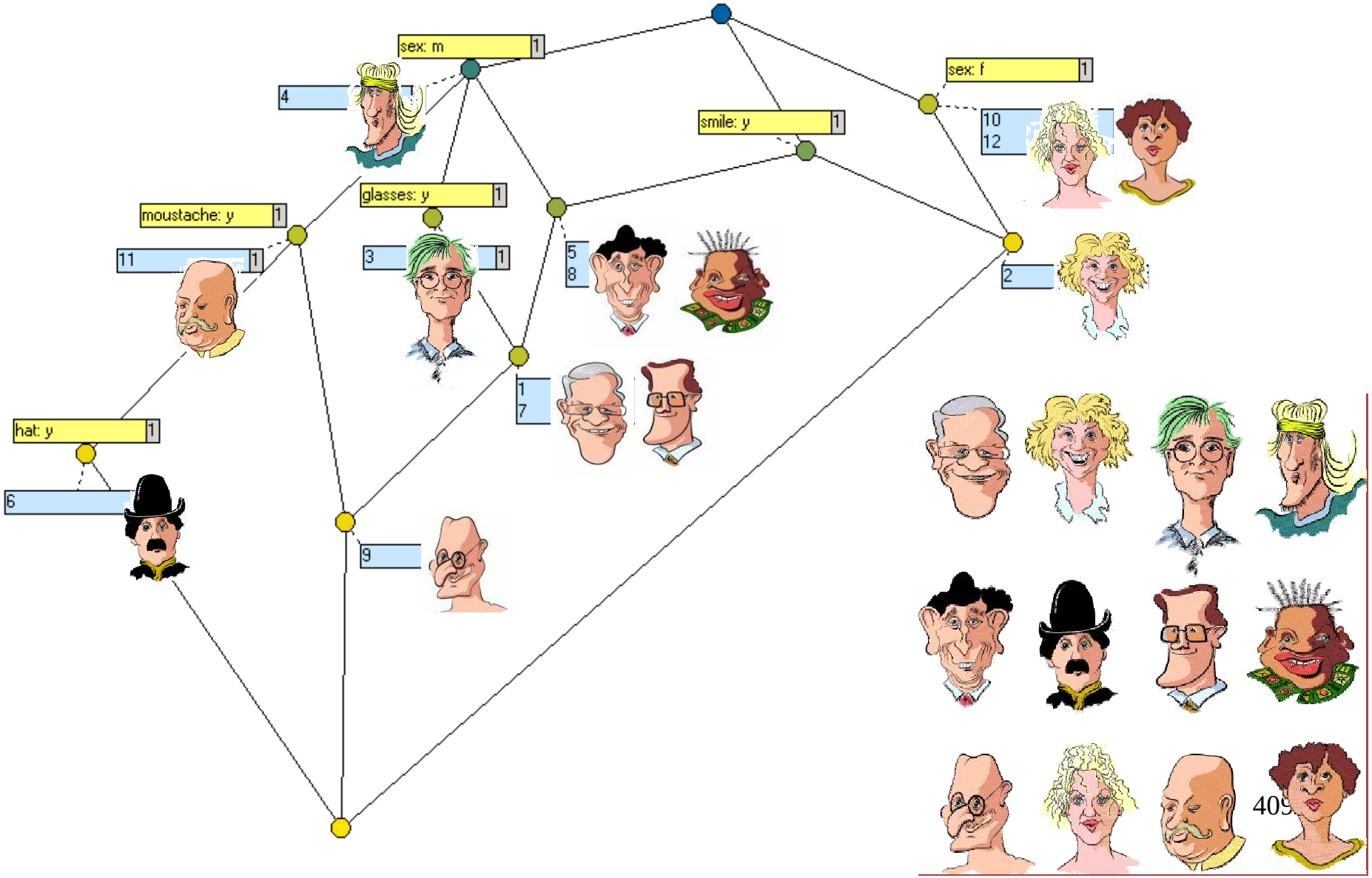
- User has his own imagining of the clustering goal and want integrate this in the cluster process

Introduction Clustering



case	sex	glasses	moustache	smile	hat
1	m	y	n	y	n
2	f	n	n	y	n
3	m	y	n	n	n
4	m	n	n	n	n
5	m	n	n	y?	n
6	m	n	y	n	y
7	m	y	n	y	n
8	m	n	n	y	n
9	m	y	y	y	n
10	f	n	n	n	n
11	m	n	y	n	n
12	f	n	n	n	n

Introduction Formal Concept Analysis



Questions…

- What ist the optimal feature representation for text documents ?
- More precisely: which representation optimally mirrors the semantic similarity of text documents in the feature space ?
- Tasks:
  - group semantically similar text documents (text clustering)
  - classify unseen text documents against classes of known text documents based on semantic similarity (text classification)
- Can formal semantic structures like ontologies support this task ?
- Can ontology learning techniques produce competitive results in this context ?

Text Clustering & Classification Task

Datasets:

Reuters-21578

- Documents about finance from 1987

OHSUMED Corpus

- Titles and Abstracts of medical journal

FAODOC Corpus

- Documents about agricultural information

Given a set of training documents, annotated with one or more categories, learn to automatically annotate previously unseen documents.

Datasets

Datasets:

Reuters-21578

- Documents about finance from 1987
- 9603 training documents and 3299 test documents (ModApte Split)
- Binary Classification on Top 50 classes.

OHSUMED Corpus

- OHSUMED (TREC-9), titles and abstracts from medical journals, 1987
- 36369 training documents and 18341 test documents
- Binary Classification on Top 50 classes (MeSH classifications).

FAODOC Corpus

- Documents about agricultural information
- 1501 docs within 21 categories

Text Classification Approaches

Documents

Dok 17892 crude
==========
Oman has granted term crude oil customers retroactive discounts from official prices of 30 to 38 cents per barrel on liftings made during February, March and April, the weekly newsletter Middle East Economic Survey (MEES) said. MEES said the price adjustments, arrived at through negotiations between the Omani oil ministry and companies concerned, are designed to compensate for the difference between market-related prices and the official 17.63 dlrs per barrel adopted OPEC Oman since February. REUTER

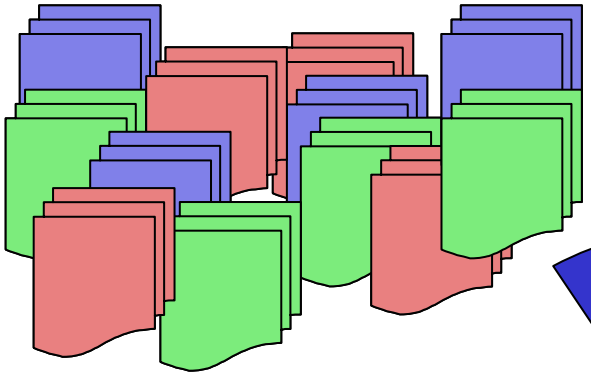
Bag of Words

Oman	2
has	1
granted	1
term	1
crude	1
oil	2
customers	1
retroactive	1
discounts	1
...	...

Further preprocessing steps
-Stopwords
-Stemming

Text Clustering & Classification Approaches

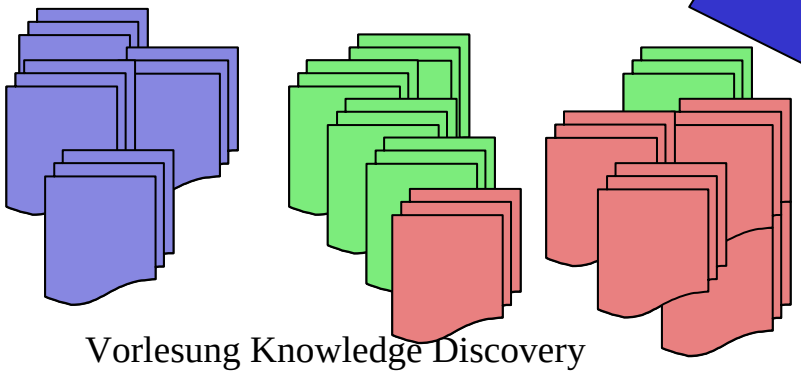
Documents



Bag of Words

	oman	has	granded	…
Obj1	2	2	1	…
Obj2	1	1	0	…
Obj3	0	0	2	…
Obj4	0	0	2	…

background knowledge



clustering/ classification algorithm

Bi-Partitioning K-Means

Input: Set of documents D, number of clusters k

Output: k cluster that exhaustively partition D

Initialize: $P^* = \{D\}$

Outer Loop:

Repeat k-1 times: **Bi-Partition** the largest cluster $E \in P^*$

Bi-Partitioning K-Means

Input: Set of documents D, number of clusters k

Output: k cluster that exhaustively partition D

Initialize: $P^* = \{D\}$

Outer loop:

Repeat k-1 times: **Bi-Partition** the largest cluster $E \in P^*$

Inner loop:

- Randomly initialize two documents from E to become e_1, e_2
- **Repeat** until convergence is reached
  - Assign each **document** from E to the **nearest** of the two $\boldsymbol{e_i}$; thus split E into E_1, E_2
  - **Re-compute** $\boldsymbol{e_1, e_2}$ to become the centroids of the document representations assigned to them
- $P^* := (P^* \setminus E) \cup \{E_1, E_2\}$

AdaBoost

- Boosting is a relatively young and very successful machine learning technique.

- Boosting algorithms build so called **ensemble classifiers** (meta classifiers):

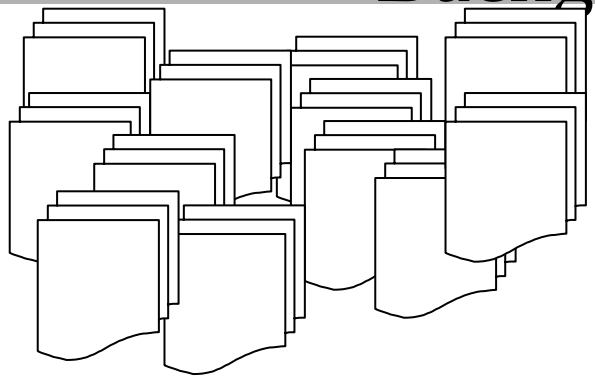
  1. Build many very simple “weak” classifiers.
  2. Combine weak learners in an additive model:

$$\hat{f}(x) = \mathsf{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

AdaBoost

- AdaBoost maintains weights D_t over the training instances.
- At each iteration t: choose a base classifier h_t that performs best on *weighted* training instances.
- Calculate weight parameter α_t based on performance base classifier. Higher errors lead to smaller weights and smaller errors lead to higher weights.
- Weight update increases (decreases) weights for wrongly (correctly) classified instances.
- Thereby, AdaBoost “is focusing in” on “hard” training instances.

Text Clustering with Background Knowledge



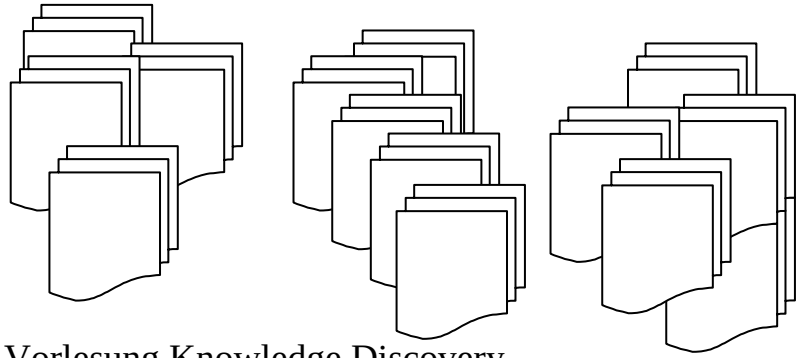
Reuters data set for our studies (min15 max100)

- choose a representation
- similarity measure
- clustering algorithm

bag of terms (details on the next slide)

cosine as similarity measure

Bi-Section is a version of KMeans



Preprocessing steps

– build a bag of words model

docid	term1	term2	term3	...
doc1	0	0	1	
doc2	2	3	1	
doc3	10	0	0	
doc4	2	23	0	
...				

– extract word counts (term frequencies)
– remove stopwords
– pruning: drop words with less than e.g. 30 occurrences
– weighting of document vectors with tfidf

(term frequency - inverted document frequency)

$$tfidf(d,t) = \log(tf(d,t)+1) * \log\left(\frac{|D|}{df(t)}\right)$$

$|D|$ no. of documents d
$df(t)$ no. of documents d which contain term t

The Bag-of-Words-Model – the Classical Approach

- The bag-of-words-model is the standard feature representation for content-based text mining.
  - Hypothesis: patterns in terminology reflect patterns in conceptualizations.
  - Steps: chunking, stemming, stop words, weighting… go !
  - Good statistical properties.

[Salton 1989]

- Some known deficiencies:
  - collocations (multi word expressions),
  - synonymous terminology,
  - polysemous terminology,

    and
  - varying degrees of specificity / generalization.

Alternative: Conceptual Document Representation

- Enhancing the bag-of-words representation with conceptual features from ontologies improves text clustering and classification.
  - Steps: collocation detection, morphological transformations, concept retrieval.
  - Hard problem: word sense disambiguation (if necessary); simple strategies used.
  - Mostly synonymy and collocations effects.
- Carefully generalizing concepts improves results much further.
  - "Generalizing": moving upwards in the ontologies' taxonomy.
- "Concepts Only" strategy is competitive but still worse than bag-of-words.
- "Hybrid" strategy outperforms bag-of-words significantly.

[Hotho et al. SWW@SIGIR 2003]
[Bloehdorn & Hotho ICDM 2004]

Limitations of the Bag-Of-Words Model

Thus, algorithms can only detect patterns in *terminology* -- *conceptual patterns* are ignored.

Specifically, such systems fail to cope with:

1. Multi Word Expressions: **European Union** vs. **Union**,
2. Synonymous Terminology: **Tungsten** vs. **Wolfram**,
3. Polysemous Terminology: **nut**
4. Generalizations: **beef** vs. **pork**

“Wolfram”

Our Approach

3. Polysemous Ter

- If we enhance the bag-of-words document representation with appropriate ontology concepts, this should improve classification by addressing issues 1-3.

4. Generalizations:

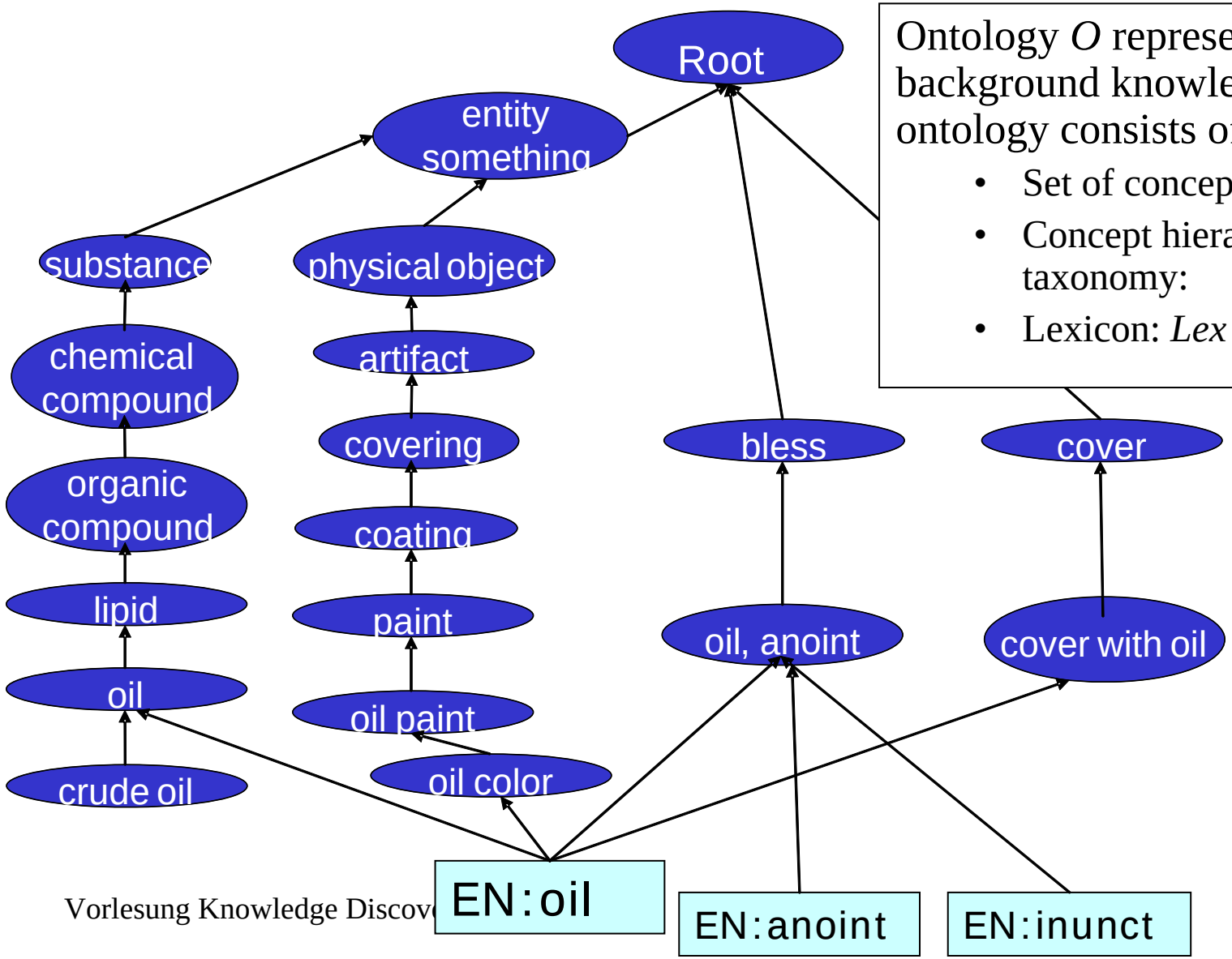
- If we carefully generalize these concepts, this should improve classification even more by addressing issue 4.

Conceptual Document Representation

Overview

- Motivation
- Current Approach for Text Classification
- **Conceptual Document Representation**
- Evaluation
- Conclusion and Outlook

Ontology

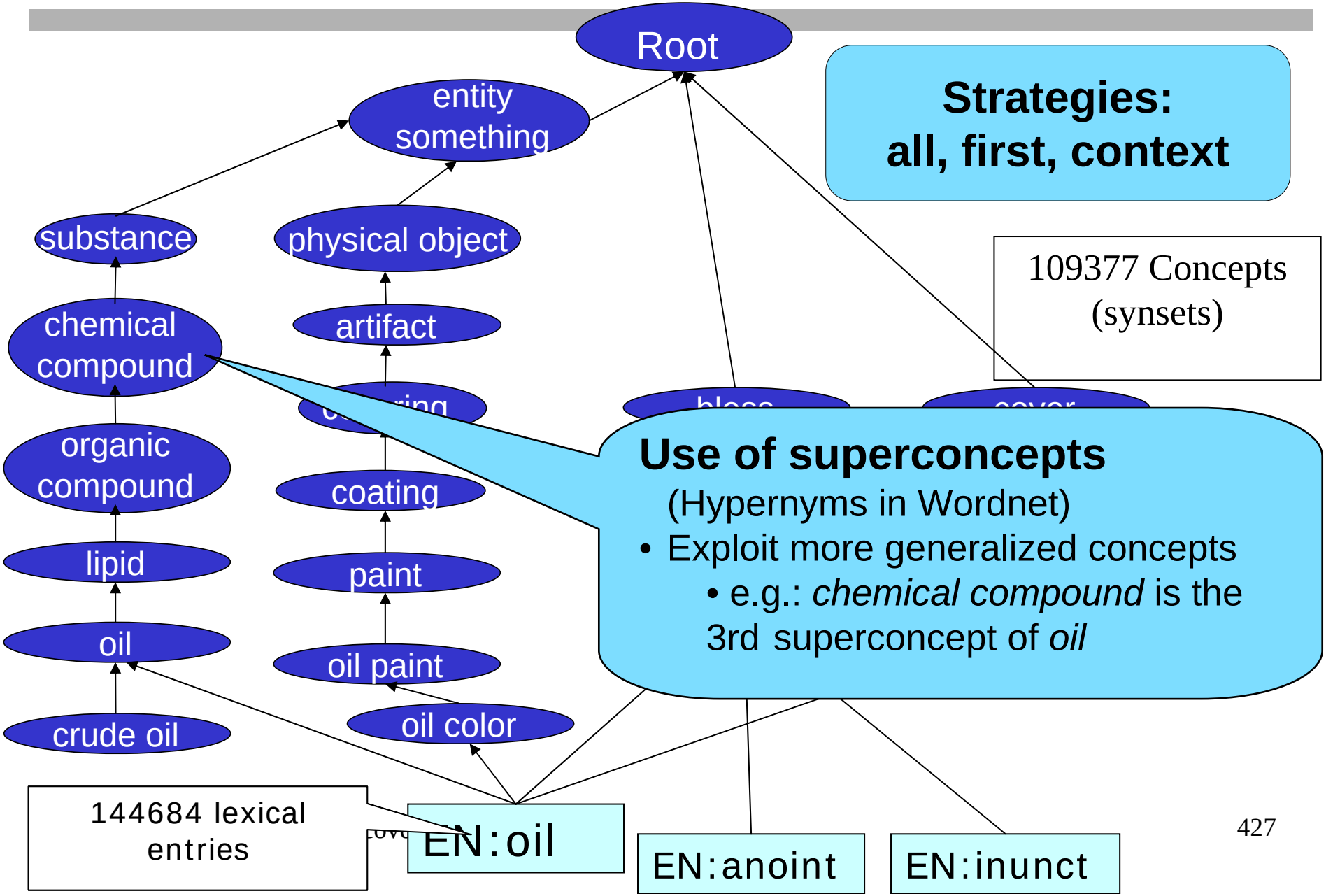


Ontology *O* represents the background knowledge core ontology consists of:

- Set of concepts: $\mathcal{C}$
- Concept hierarchy or taxonomy:
- Lexicon: *Lex*

Word Sense Disambiguation

WordNet as an example



Ontology-based representation

Oman	1	Oman	1	Oman	1
has	1	granted	1	granted	1
granted	1	term	1	term	1
term	1	(C) term	1	(C) term	1
crude	1	crude	1	crude	1
oil	1	(C) crude	1	(C) crude	1
customers	1	oil	1	oil	1
retroactive	1	(C) oil	1	(C) oil	1
discounts	1	customer	1	(C) lipid	1
...	...	(C) customer	1	(C) compound	1
		...	...	...	...
1		2		3	

strategy: add

Conceptual Document Representation

Simple Strategy to map terms to concept is to map every single term to a concept

Advanced version of mapping terms to concepts requires the following steps

Detecting the appropriate a set of concepts from an Ontology ($\mathcal{O}$, *Lex*) requires multiple steps:

1. Candidate Term Detection
2. Morphological Transformations
3. Word Sense Disambiguation
4. Generalization

Conceptual Document Representation
Candidate Term Detection

- Querying the lexicon directly for each single word will not do the trick! (Remember the multi-word expressions!)

  Solution:

  Move a window of maximum size over the text, decrease window size if unsuccessful before moving on.

- But querying the lexicon for *any* candidate term window produces much overhead !

  Solution:

  Avoid unnecessary lexicon queries by matching POS tags in the window against appropriately defined syntactical patterns (e.g. noun phrases).

Evaluation of Text Document Clustering

given classes $\mathcal{L}^*$

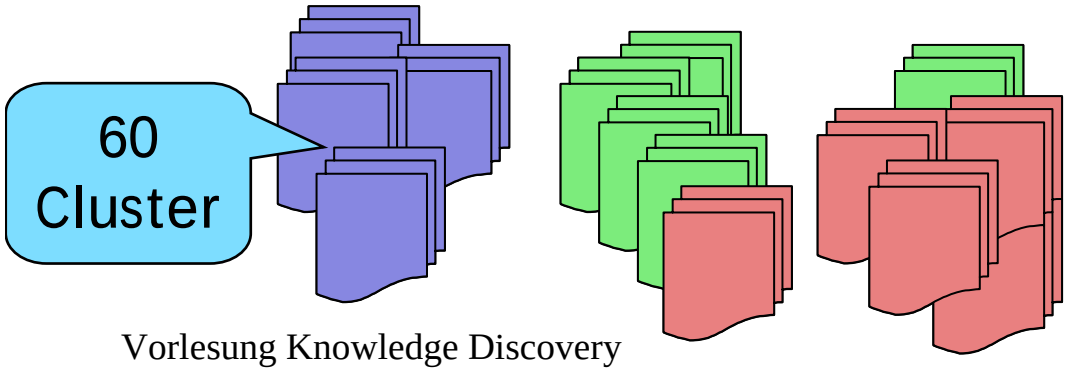
Goal:
Cluster should be as similar as possible to the given classes

compare clustering $\mathcal{P}^*$ of document set D
with given classes $\mathcal{L}^*$

$$\mathbf{Precision}(P,L) := \frac{|P \cap L|}{|P|}$$

$$\mathbf{Purity}(\mathcal{P}^*,\mathcal{L}^*) := \sum_{P \in \mathcal{P}^*} \frac{|P|}{|D|} \max_{L \in \mathcal{L}^*} \mathbf{Precision}(P,L)$$

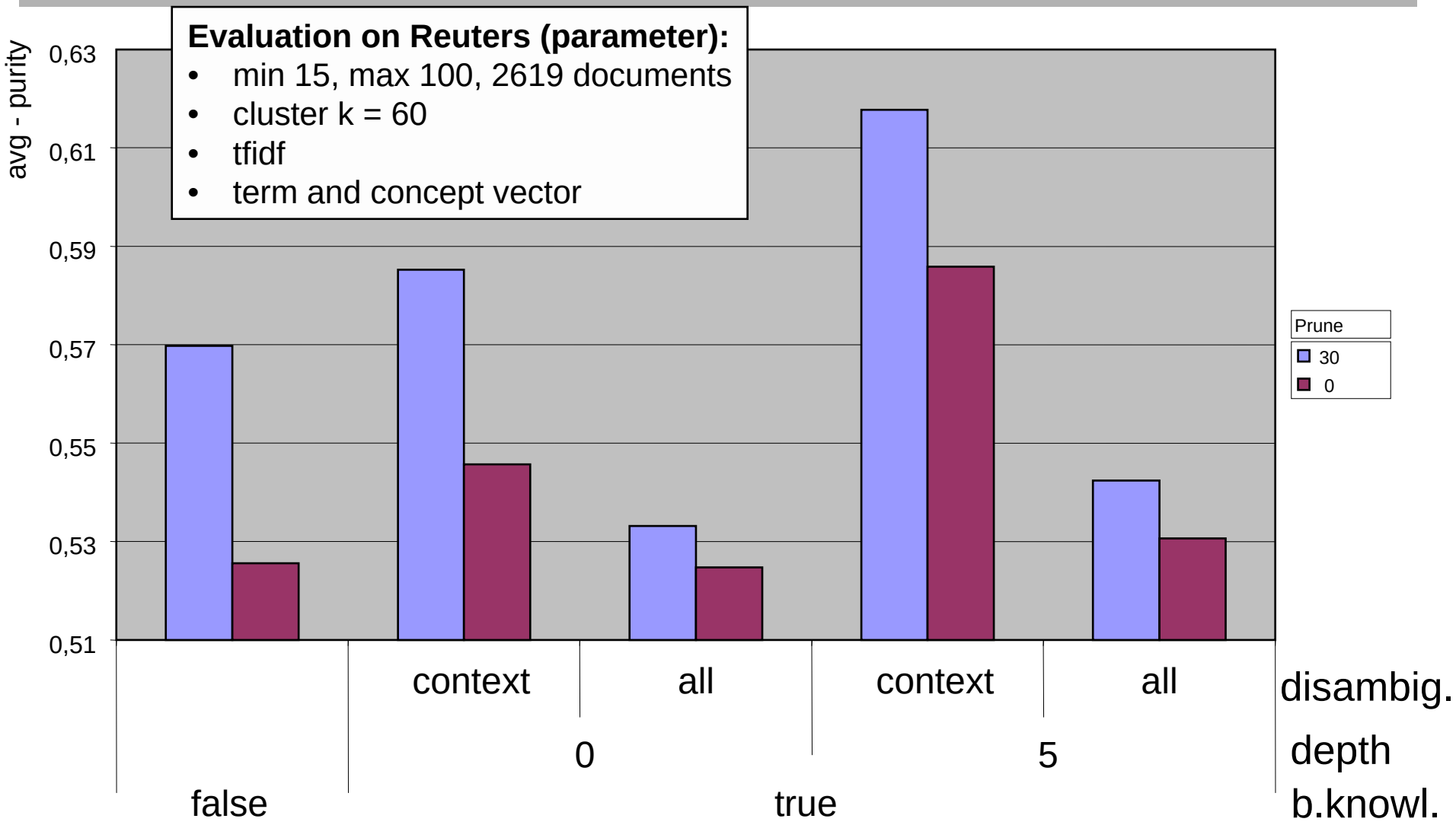
$$\mathbf{InvPrty}(\mathcal{P}^*,\mathcal{L}^*) := \sum_{L \in \mathcal{L}^*} \frac{|L|}{|D|} \max_{P \in \mathcal{P}^*} \mathbf{Precision}(L,P)$$



Clustering $\mathcal{P}^*$

Evaluation Metrics: Text Classification

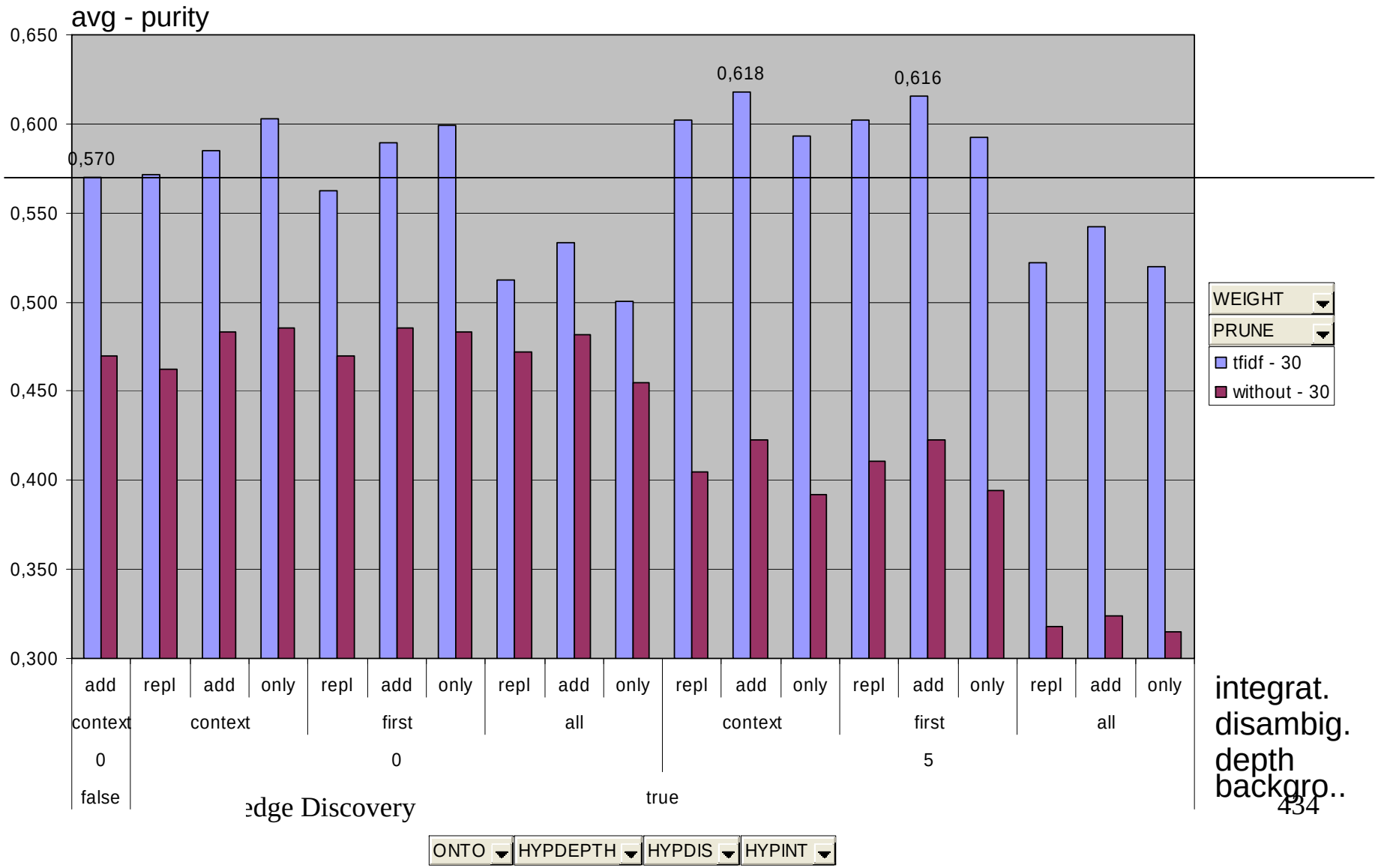
Evaluation of Text Clustering



Evaluation parameter

- min 15, max 100, 2619 documents
- cluster k = 60

on of Text Clustering



integrat.
disambig.
depth
backgro..

edge Discovery

ONTO HYPDEPTH HYPDIS HYPINT

Evaluation of Text Clustering

Evaluation parameter

- min 15, max 100, 2619 Dokumente
- Cluster k = 60
- Disamb = context
- Prune = 30

Backgr.	depth	integr.	Mean - PURITY		Mean - INVPURITY	
false			0,570	±0,019	0,479	±0,016
true	0	add	0,585	±0,014	0,492	±0,017
		only	0,603	±0,019	0,504	±0,021
	5	add	0,618	±0,015	0,514	±0,019
		only	0,593	±0,01	0,500	±0,016

Variance analysis of the Reuters classes

Idea:

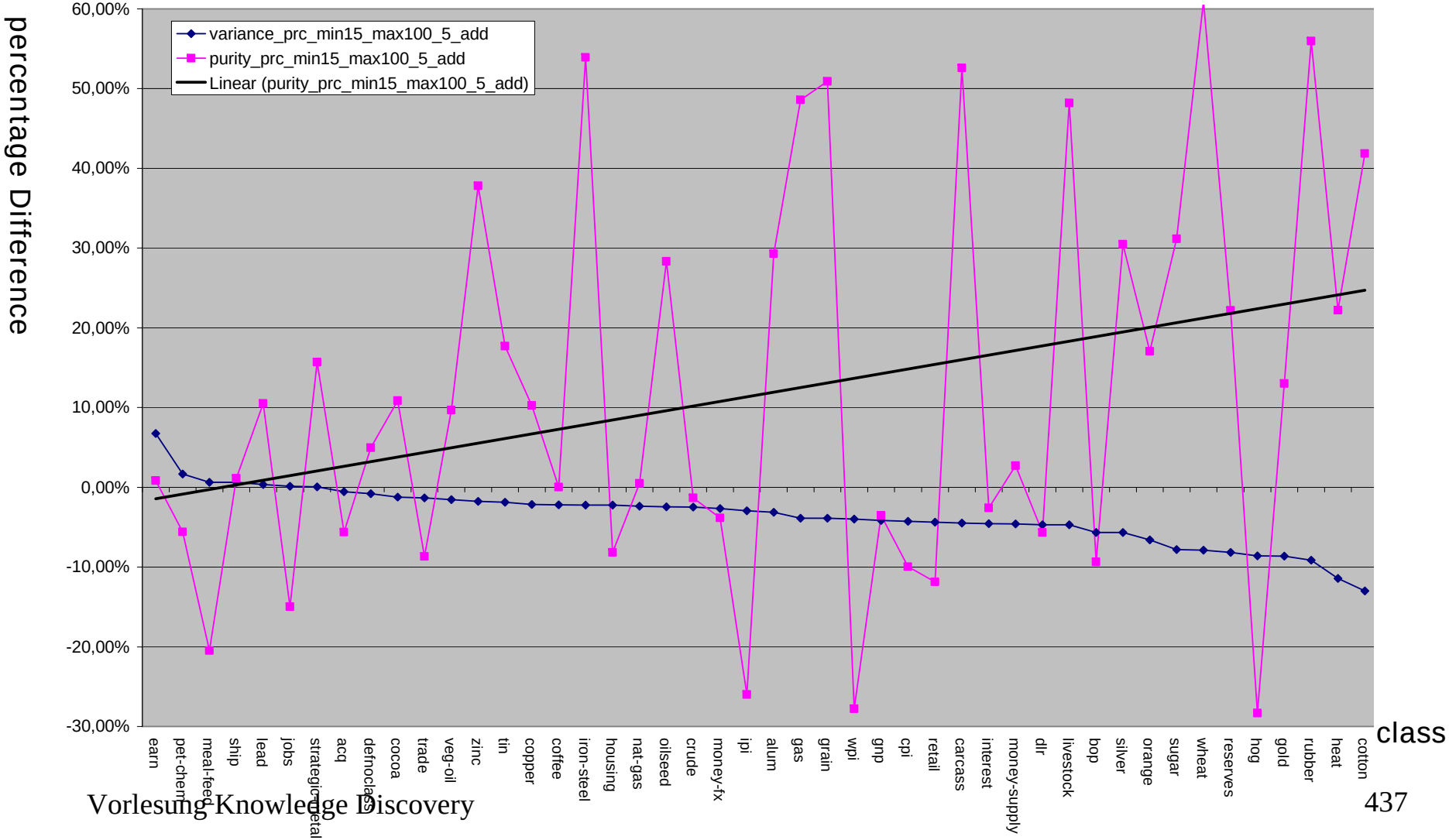
- Ideally documents of one class should have the same representation ➔ variance = 0
- representation of the documents is changed, variance will also change

Analysis:

- Compare the variance of the classes of both Representations (with and without ontology)
- Compare the purity per class

Variance analysis

Variance and purity per class for PRC-min15-max100

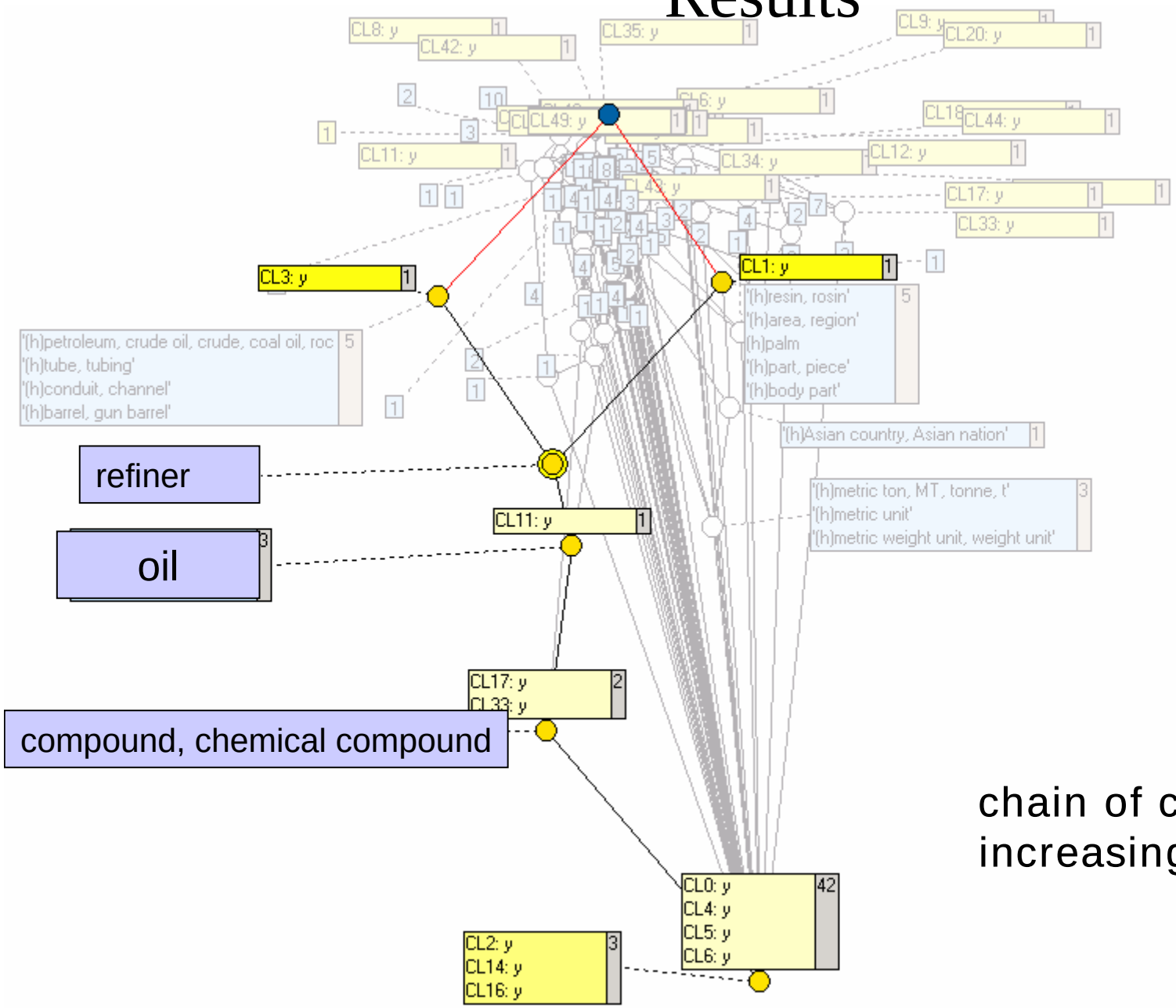


Extracted Word description

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
amount	0,12	depository financial institu	0,09	loss	0,34	Irani, Iranian, Persian'	0,14	indebtedness, liability, fin	0,12
billion, one million million,	0,11	financial institution, financ	0,09	failure	0,33	Iran, Islamic Republic of	0,13	obligation	0,12
large integer'	0,11	rate, charge per unit'	0,09	nonaccomplishment, nona	0,32	gulf	0,13	debt	0,12
integer, whole number'	0,11	charge	0,09	Connecticut, Nutmeg Sta	0,28	vessel, watercraft'	0,12	written agreement'	0,1
insufficiency, inadequacy	0,1	institution, establishment'	0,09	ten, 10, X, tenner, decade	0,24	ship	0,12	agreement, understanding	0,08
deficit, shortage, shortfall	0,1	loss	0,08	American state'	0,23	craft	0,12	creditor	0,08
number	0,09	monetary unit'	0,07	state, province'	0,22	Asian, Asiatic'	0,11	lender, loaner'	0,08
excess, surplus, surplusa	0,09	central, telephone excham	0,07	system, unit'	0,19	person of color, person of	0,10	statement	0,07
overabundance, overmuc	0,09	financial loss'	0,06	network, net, mesh, mesh	0,19	Asian country, Asian natio	0,10	billion, one million million,	0,06
abundance, copiousness,	0,09	outgo, expenditure, outlay	0,06	September, Sep, Sept'	0,18	oil tanker, oiler, tanker, ta	0,10	large integer'	0,05

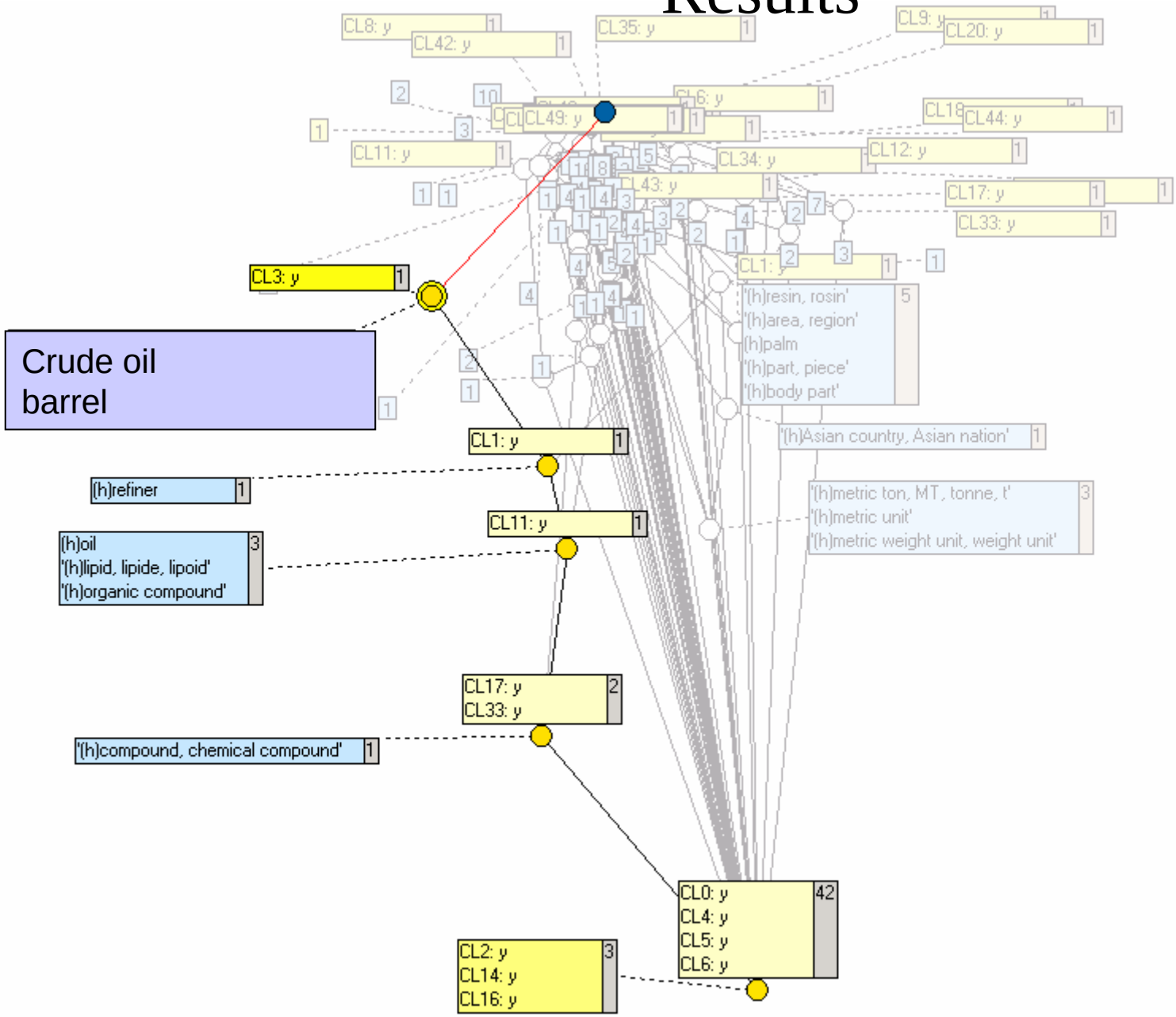
Cluster 5		Cluster 6		Cluster 7		Cluster 8		Cluster 9	
text, textual matter'	0,15	loss	0,34	gross sales, gross revenu	0,11	tender, legal tender'	0,15	metric weight unit, weight	0,15
matter	0,15	failure	0,33	sum, sum of money, amo	0,09	offer, offering'	0,14	metric ton, MT, tonne, t'	0,15
letter, missive'	0,15	nonaccomplishment, nona	0,32	income	0,09	medium of exchange, mon	0,11	mass unit'	0,14
sign, mark'	0,13	common fraction, simple f	0,22	financial gain'	0,09	speech act'	0,1	palm, thenar'	0,14
clue, clew, cue'	0,13	fraction	0,22	gain	0,09	indicator	0,1	area, region'	0,12
purpose, intent, intention	0,11	rational number'	0,22	enterprise	0,05	standard, criterion, measu	0,1	unit of measurement, unit	0,10
evidence	0,11	real number, real'	0,22	business, concern, busine	0,05	reference point, point of r	0,09	organic compound'	0,10
indication, indicant'	0,11	complex number, complex	0,22	assets	0,05	signal, signaling, sign'	0,08	oil	0,10
goal, end'	0,1	one-half, half'	0,22	division	0,05	acquisition	0,06	lipid, lipide, lipoid'	0,10
writing, written material, p	0,07	revolutions per minute, rp	0,22	army unit'	0,05	giant	0,06	compound, chemical com	0,08

Results

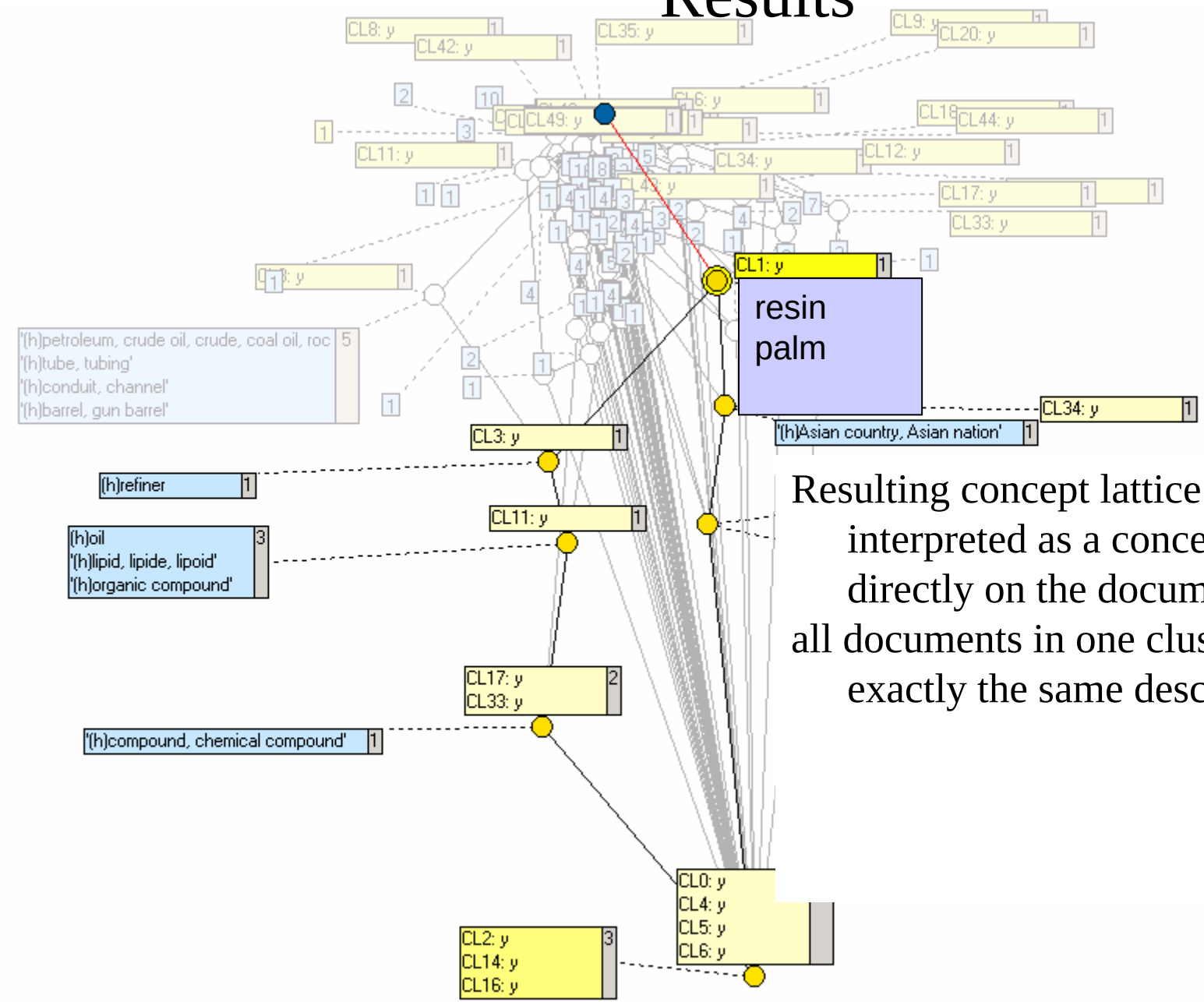


chain of concepts with increasing specificity

Results


Crude oil barrel
CL3: y
CL1: y
(h)refiner
CL11: y
(h)oil
'(h)lipid, lipide, lipoid'
'(h)organic compound'
CL17: y
CL33: y
'(h)compound, chemical compound'
CL0: y
CL4: y
CL5: y
CL6: y
CL2: y
CL14: y
CL16: y
'(h)resin, rosin'
'(h)area, region'
(h)palm
'(h)part, piece'
'(h)body part'
'(h)Asian country, Asian nation'
'(h)metric ton, MT, tonne, t'
'(h)metric unit'
'(h)metric weight unit, weight unit'

Results



Resulting concept lattice can also be interpreted as a concept hierarchy directly on the documents

all documents in one cluster obtain exactly the same description

Evaluation: Reuters Results Classification

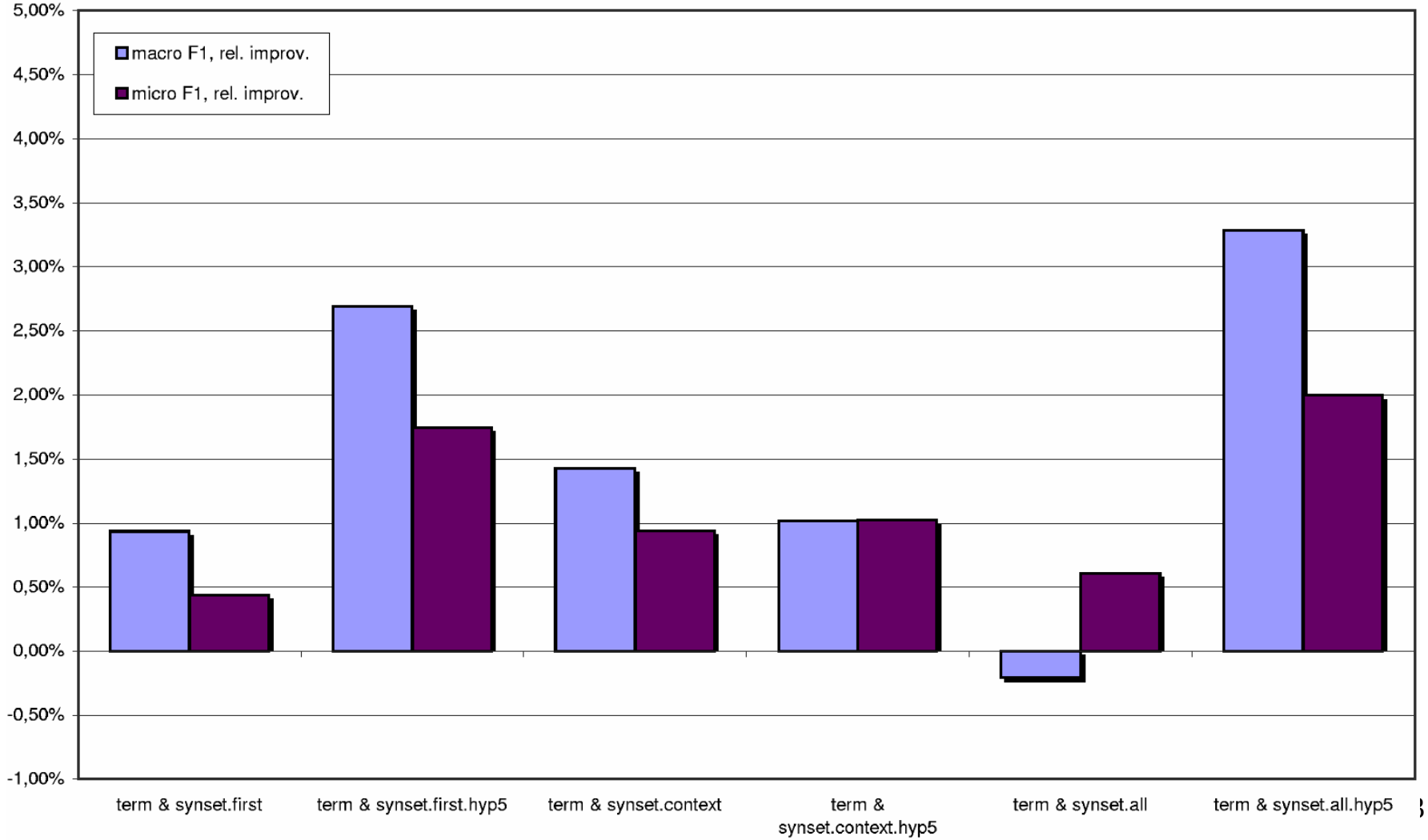
- Top 50 reuters classes with 17525 term stems/10259 – 27236 synset features

Feature Type	Error	macro-averaged Prec	Rec	F_1	BEP
term	00.65	80.59	66.30	72.75	74.29
term & synset.first	00.64	80.66	67.39	73.43	**75.08**
term & synset.first.hyp3	**00.62**	79.71	67.76	73.25	74.74
term & synset.first.hyp5	**00.60**	80.67	**69.57**	**74.71**	74.84
term & synset.first.hyp10	**00.62**	80.43	**68.40**	**73.93**	**75.58**
term & synset.context	00.63	79.96	**68.51**	73.79	74.46
term & synset.context.hyp5	**00.62**	79.48	68.34	73.49	74.71
term & synset.all	00.64	80.02	66.44	72.60	73.62
term & synset.all.hyp5	**00.59**	**83.76**	68.12	**75.14**	**75.55**

Feature Type	Error	micro-averaged Prec	Rec	F_1	BEP
term	00.65	**89.12**	79.82	84.21	85.77
term & synset.first	00.64	88.75	80.79	84.58	85.97
term & synset.first.hyp3	**00.62**	**89.12**	81.40	85.09	**86.25**
term & synset.first.hyp5	**00.60**	**89.16**	**82.46**	**85.68**	85.91
term & synset.first.hyp10	**00.62**	88.78	**81.74**	**85.11**	**86.14**
term & synset.context	00.63	88.86	81.46	85.00	85.91
term & synset.context.hyp5	**00.62**	89.09	81.40	85.07	**85.97**
term & synset.all	00.64	88.82	80.99	84.72	85.69
term & synset.all.hyp5	**00.59**	**89.92**	**82.21**	**85.89**	**86.44**

Evaluation: Reuters Results

Relative improvement on the top 50 classes



Evaluation: OHSUMED Results

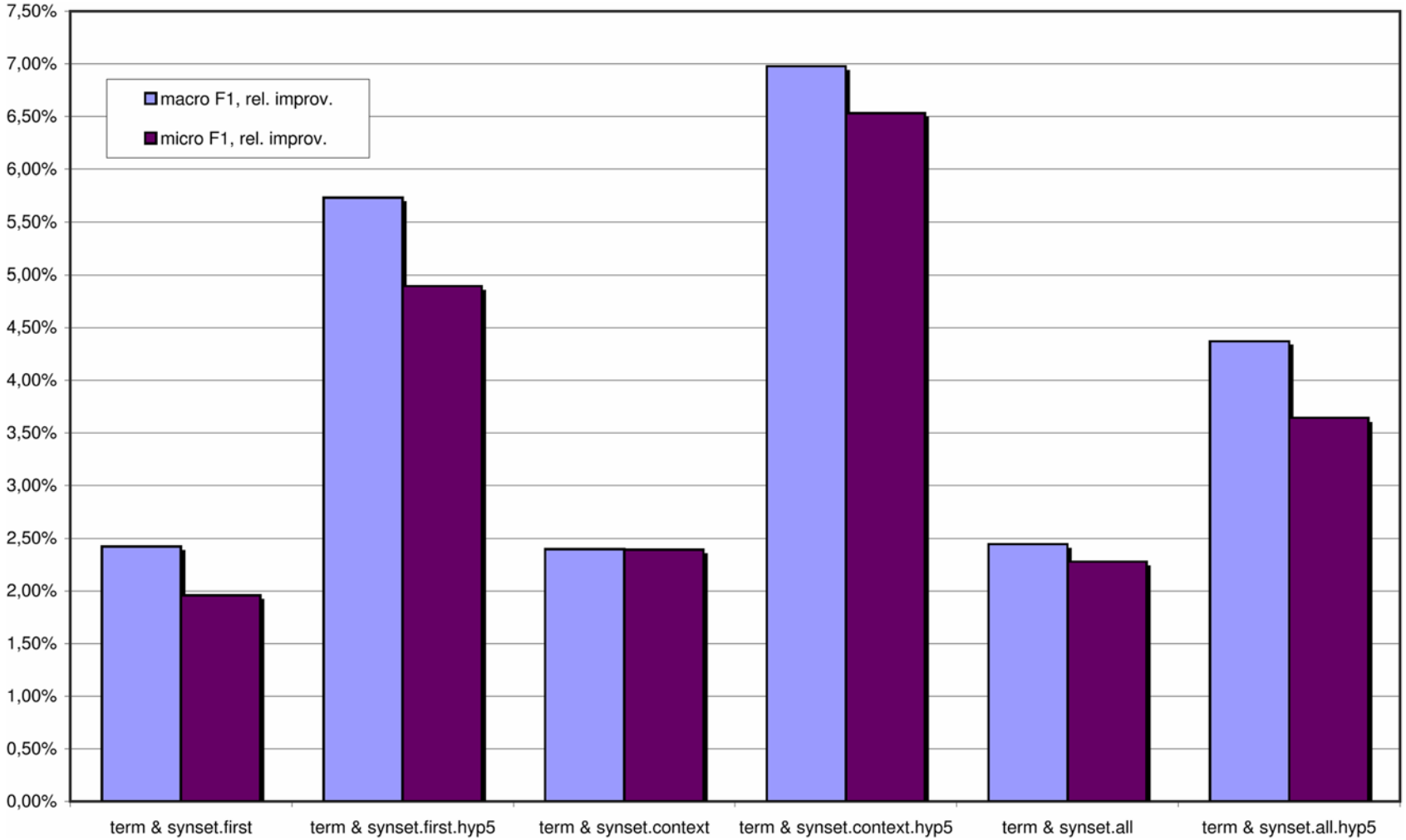
Top 50 classes with WordNet

Feature Type	Error	macro-averaged Prec	Rec	F_1	BEP
term	00.53	52.60	35.74	42.56	45.68
term & synset.first	**00.52**	53.08	36.98	43.59	46.46
term & synset.first.hyp5	**00.52**	**53.82**	**38.66**	**45.00**	**48.01**
term & synset.context	**00.52**	52.83	37.09	43.58	**46.88**
term & synset.context.hyp5	**00.51**	**54.55**	**39.06**	**45.53**	**48.10**
term & synset.all	**00.52**	52.89	37.09	43.60	46.82
term & synset.all.hyp5	**00.52**	**53.33**	**38.24**	**44.42**	46.73

Feature Type	Error	micro-averaged Prec	Rec	F_1	BEP
term	00.53	55.77	36.25	43.94	46.17
term & synset.first	**00.52**	56.07	37.30	44.80	47.01
term & synset.first.hyp5	**00.52**	**56.84**	**38.76**	**46.09**	**48.31**
term & synset.context	**00.52**	**56.30**	37.46	44.99	**47.34**
term & synset.context.hyp5	**00.51**	**58.10**	**39.18**	**46.81**	**48.45**
term & synset.all	**00.52**	56.19	37.44	44.94	47.32
term & synset.all.hyp5	**00.52**	**56.29**	**38.24**	**45.54**	46.73

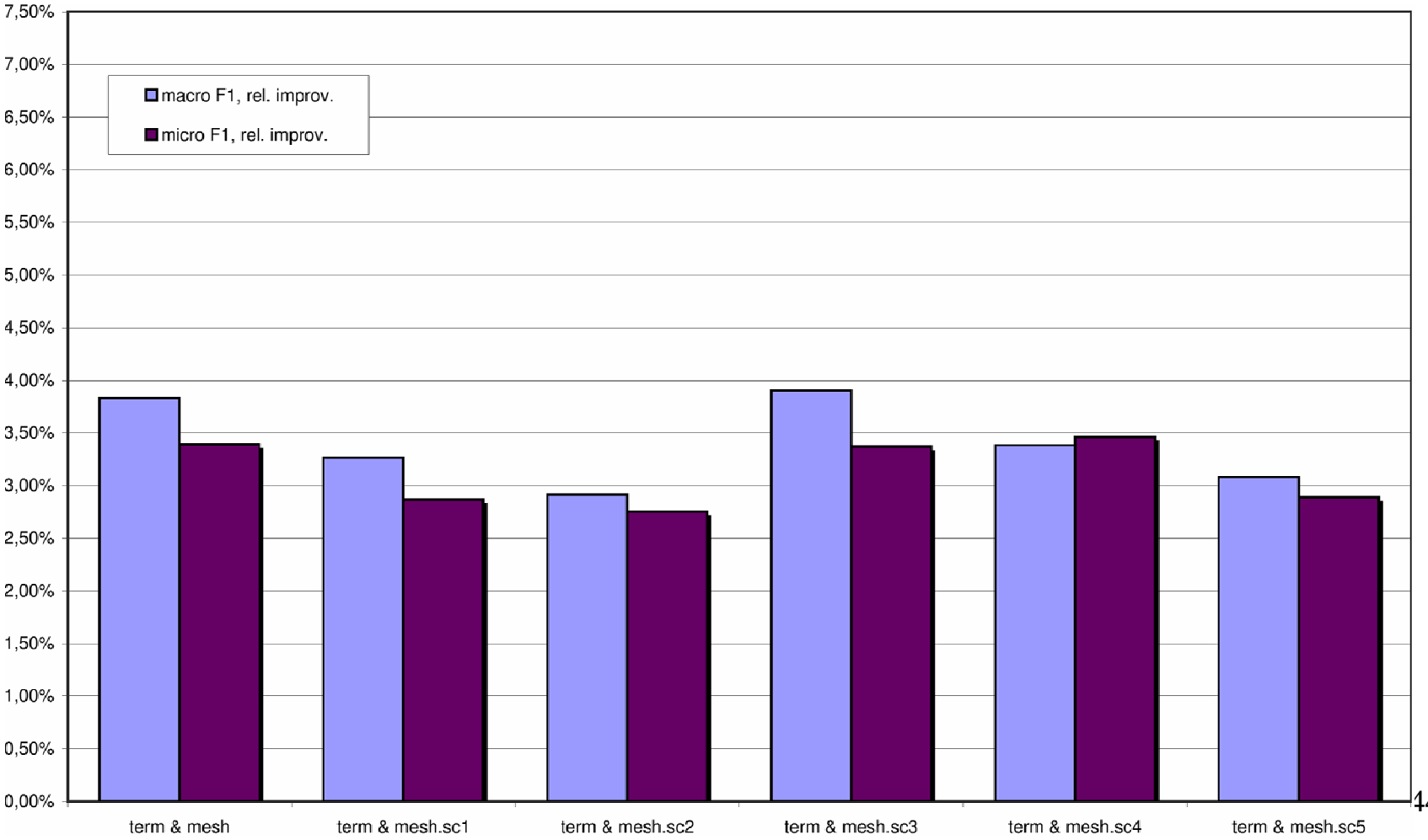
Evaluation: OHSUMED Results

Relative improvement on the top 50 classes with WordNet



Evaluation: OHSUMED Results

Relative improvement on the top 50 classes with Mesh Ontology (~ 22.000 Concepts, all Strategy)

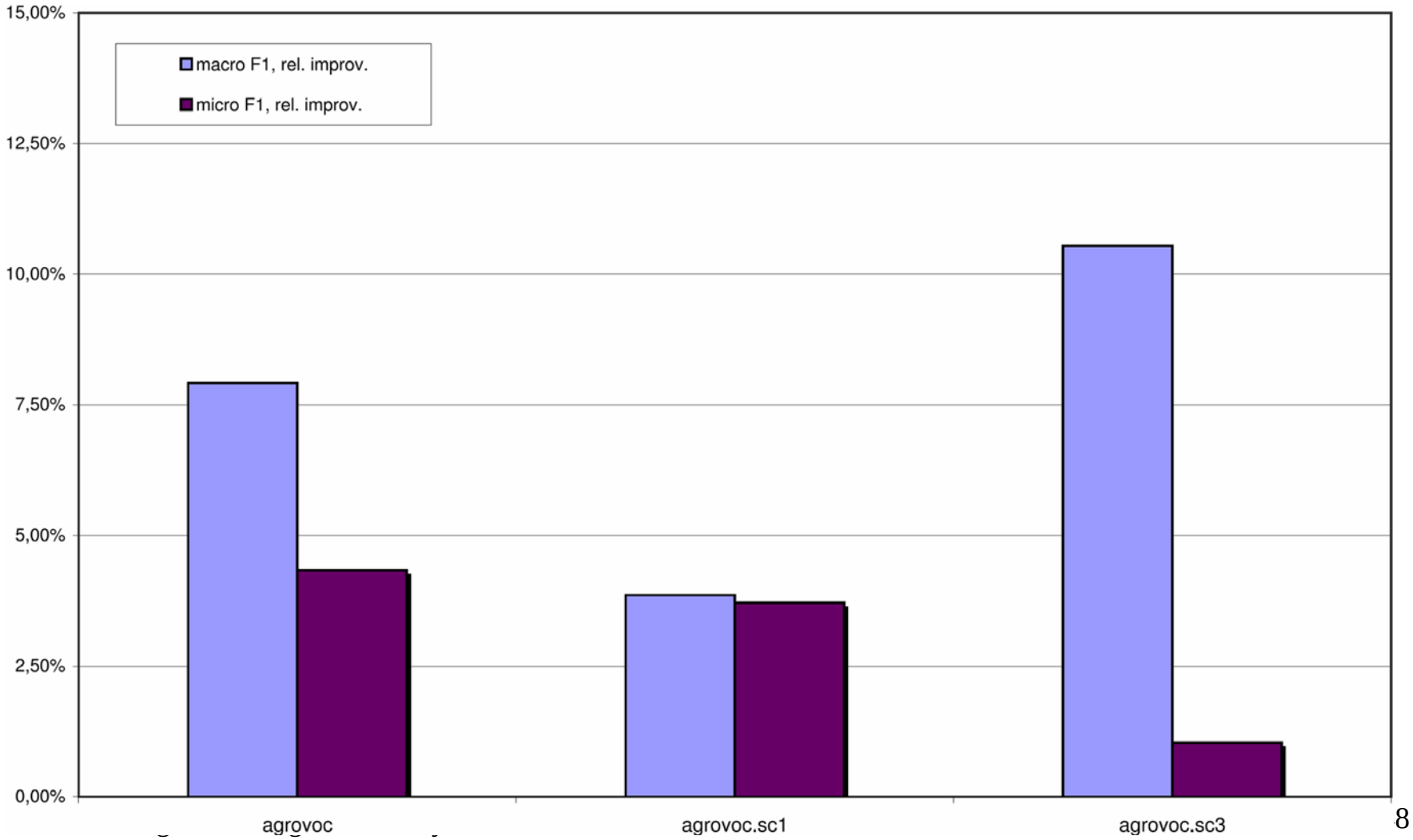


Evaluation: FAODOC Results

Feature Type	Error	macro-averaged Prec	Rec	F_1	BEP
term	06.87	45.47	27.11	33.97	36.93
term & agrovoc	**06.66**	**50.96**	28.63	36.66	39.84
term & agrovoc.sc1	06.76	49.26	27.48	35.28	39.40
term & agrovoc.sc3	06.79	49.08	**30.41**	**37.55**	**41.69**

Feature Type	Error	micro-averaged Prec	Rec	F_1	BEP
term	06.87	50.44	31.22	38.57	44.29
term & agrovoc	**06.66**	**52.91**	32.46	**40.24**	**48.01**
term & agrovoc.sc1	06.76	51.75	**32.60**	40.00	46.77
term & agrovoc.sc3	06.79	51.47	31.36	38.97	47.73

Evaluation: FAODOC Results



Overview

- Motivation
- Current Approach for Text Classification
- Conceptual Document Representation
- Evaluation
- **Conclusion and Outlook**

Conclusion and Outlook

- Successful integration of conceptual features to improve classification performance
- Gereralization does improve classification results in most cases

Conclusion and Outlook

- Advanced Generalization Strategies
- Development of additional weak learner plugins that exploit ontologies more directly
- Heuristics for efficient handling of continuous feature values like TFIDF in AdaBoost
- Multilingual Text Classification