UNIKASSEL VERSITÄT

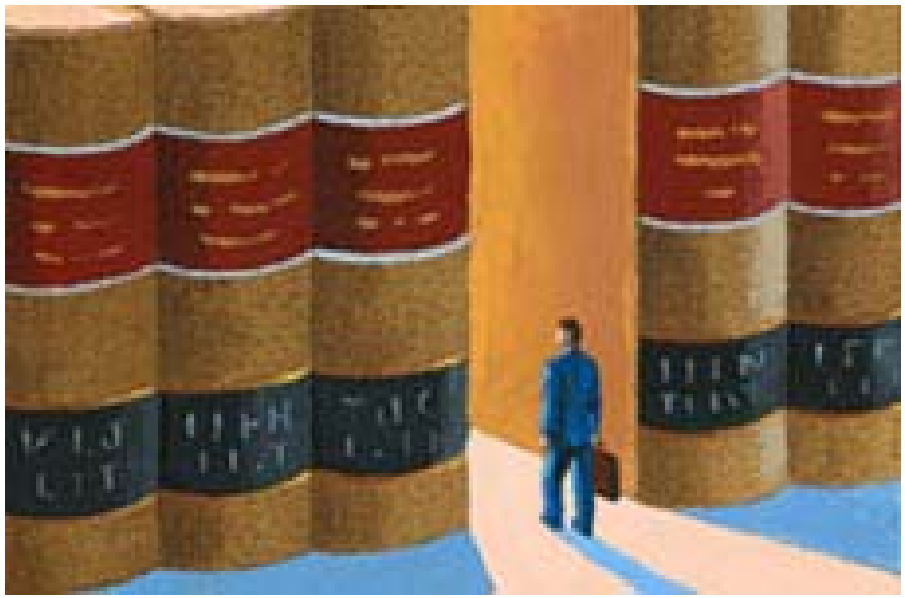
ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Some Applications of

Conceptual Knowledge Processing

Prof. Dr. Gerd Stumme

Introduction

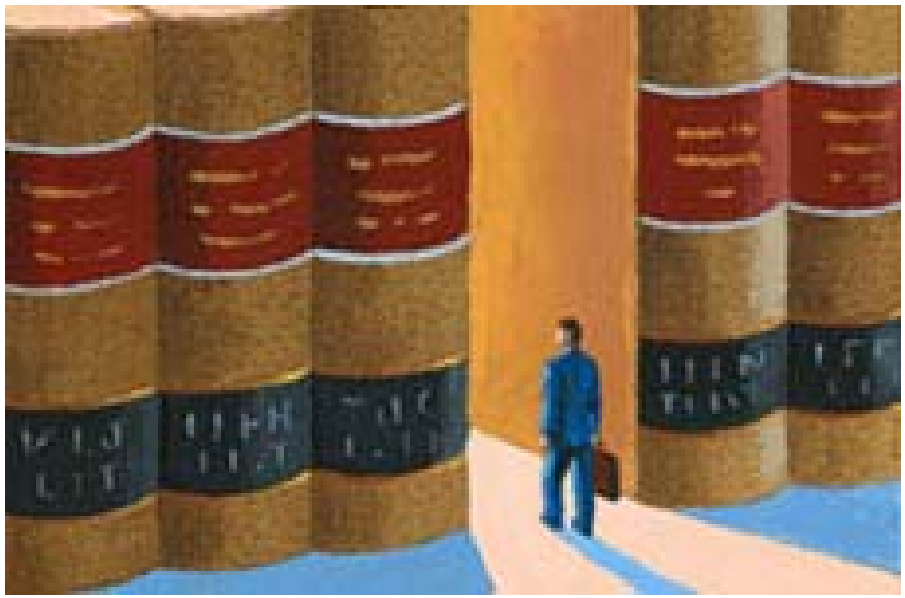


„ An Ontology is a formal and explicit specification of a shared conceptualisation of a domain of interest.“ T. Gruber, 1993.

There are many formalisms, ranging from light-weight to heavy-weight.

Examples: Concept Hierarchies, Thesauri, XML, RDF, OWL.

Introduction



„ An Ontology is a formal specification of a shared conceptualisation of a domain of interest.“ T. Gruber, 1993.

There are many formalisms, ranging from light-weight to heavy-weight.

Examples: Concept Hierarchies, Thesauri, XML, RDF, OWL.

Well, so far so good. But what are they good for?

Introduction

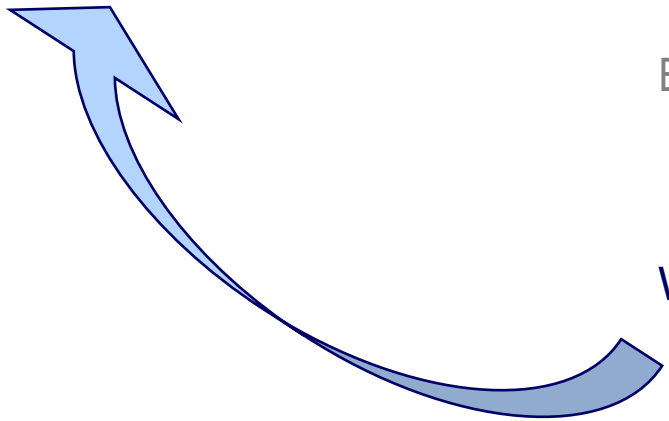
Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge

„ An Ontology is a formal specification of a shared conceptualisation of a domain of interest.“ T. Gruber, 1993.

There are many formalisms, ranging from light-weight to heavy-weight.

Examples: Concept Hierarchies, Thesauri, XML, RDF, OWL.



Well, so far so good. But what are they good for?

Overview

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge

- Combining the above

This talk gives some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems

- Courseware Watchdog

Overview

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge

- Combining the above

This talk gives some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems

- Courseware Watchdog

with A. Hotho (Karlsruhe/Kassel)

Text Clustering with Background Knowledge

Task:

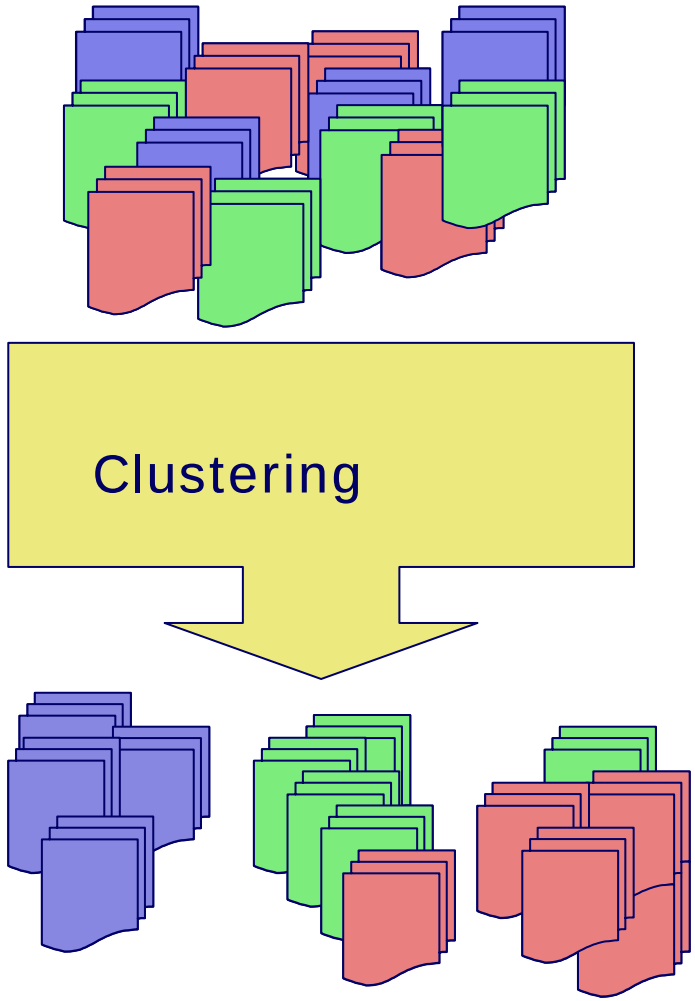
Discovering structure in text collections by grouping together similar documents.

Test Data:

(A subset of) 21578 Reuters News

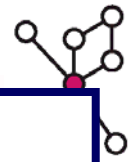
Problem:

1. Overlapping clusters should be allowed.
2. A conceptual description of the clusters is required.
3. The method should be computationally effective.



Can Background Knowledge – in form of a thesaurus – improve the result?

Text Clustering with Background Knowledge



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 price of 17.63 dlrs per barrel adopted by non-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
...	...

WordNet 2.0 Search

Search word: Find senses

Results for "Hypernyms (this is a kind of...)" search of noun "oil"

3 senses of oil

Sense 1
oil -- (a slippery or viscous liquid or liquefiable substance not miscible with water)
=> lipid, lipide, lipoid -- (an oily organic compound insoluble in water but soluble in organic solvents; essential structural component of living cells (along with proteins and carbohydrates))
=> macromolecule, supermolecule -- (any very large complex molecule; found only in plants and animals)
=> molecule -- ((physics and chemistry) the simplest structural unit of an element or compound)
=> unit, building block -- (a single undivided natural thing occurring in the composition of something else; "units of nucleic acids")
=> thing -- (a separate and self-contained entity)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> organic compound -- (any compound of carbon and another element or a radical)
=> compound, chemical compound -- ((chemistry) a substance formed by chemical union of two or more elements or ingredients in definite proportion by weight)
=> substance, matter -- (that which has mass and occupies space; "an atom is the smallest indivisible unit of matter")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Sense 2
oil, oil color -- (oil paint used by an artist)
=> oil paint -- (paint in which a drying oil is the vehicle)
=> paint -- (a substance used as a coating to protect or decorate a surface (especially a mixture of pigment suspended in a liquid); dries to form a hard coating)
=> coating, coat -- (a thin layer covering something; "a second coat of paint")
=> covering -- (an artifact that covers something else (usually to protect or shelter or conceal it))
=> artifact, artefact -- (a man-made object taken as a whole)
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> whole, whole thing, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> coloring material, colouring material, color, colour -- (any material used for its color; "she used a different color for the trim")
=> material, stuff -- (the tangible substance that goes into the makeup of a physical object; "coal is a hard black material"; "wheat is the stuff they use to make bread")
=> substance, matter -- (that which has mass and occupies space; "an atom is the smallest indivisible unit of matter")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Sense 3
vegetable oil, oil -- (any of a group of liquid edible fats that are obtained from plants)
=> edible fat -- (oily or greasy matter making up the bulk of fatty tissue in animals and in seeds and other plant tissue)

Text Clustering with Background Knowledge

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 price of 17.63 dlrs per barrel adopted by non-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
...	...
chem. comp.	2

Adding of superconcepts of WordNet

Two-Step Clustering Approach:

- **First Step:**
  - with standard algorithm “Bisection k-Means”
  - reduces efficiently the number of objects
- **Second Step:**
  - with Formal Concept Analysis
  - provides descriptions of the clusters
  - and allows for multiple inheritance

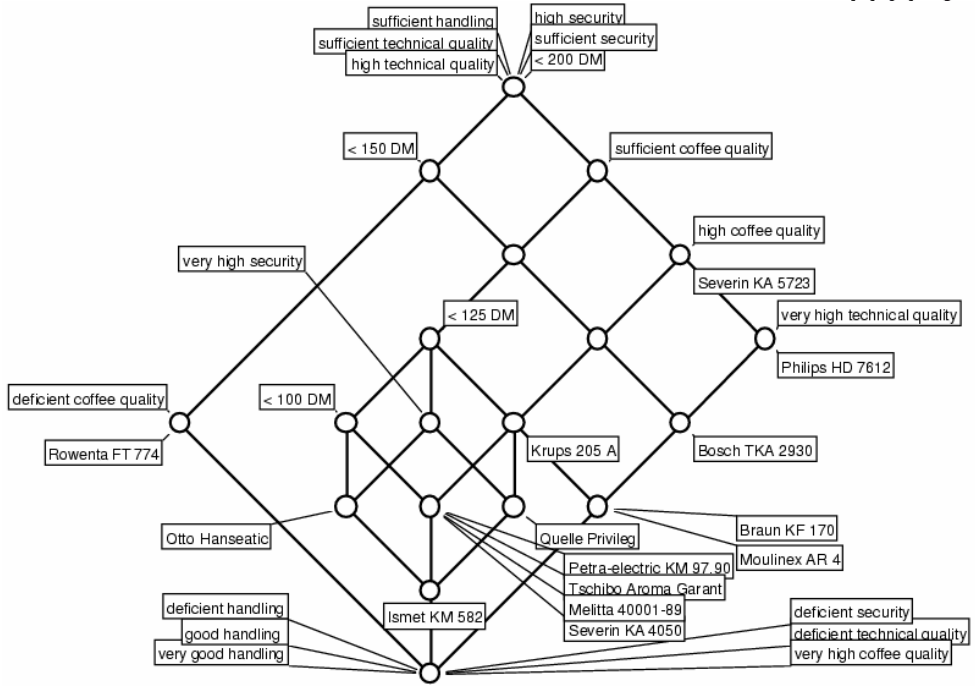
Excursion: Formal Concept Analysis

Formal Concept Analysis (FCA) arose in the 1980ies as a mathematical theory, formalizing the concept „concept“.

Since then, FCA has increasingly been applied in computer science, esp. for

- data analysis,
- knowledge discovery,
- and software engineering.

FCA derives **concept hierarchies** from data tables, and provides means for their **visualisation**.



STIFTUNG WARENTEST test KOMPASS — KAFFEEMASCHINEN MIT WARMHALTEKANNE (8 bis 10 Tassen) — test Ausgabe 12/98

	Mittlerer Preis in DM ca.	Preis für Ersatzkanne/ Glaseinsatz in DM ca.	Kaffeequalität	Technische Prüfung	Sicherheit	Handhabung	test-Qualitätsurteil
Gewichtung			**35 %**	**30 %**	**10 %**	**25 %**	
Neckermann Best.-Nr. 8628/409	40,–	35,–¹) / ❑	baugl. mit Otto Hanseatic Best.-Nr. 4327357				**zufriedenst.**
Otto Hanseatic Best.-Nr. 4327357	40,–	30,–²) / ❑	**O**	**+**	**++**	**O**	**zufriedenst.**
Quelle Privileg Best.-Nr. 7030720	40,–	24,50 / 17,50	baugl. mit Otto Hanseatic Best.-Nr. 4327357				**zufriedenst.**
Severin KA 9660	50,–	35,– / 23,–	baugl. mit Otto Hanseatic Best.-Nr. 4327357				**zufriedenst.**
Severin KA 4050	80,–	50,– / ❑	**+**	**+**	**+**	**O**	**gut**
Tchibo Aroma Garant Art.-Nr. 48469	80,–	27,50 / 19,50	**+**	**+**	**+**	**O**	**gut**
Ismet KM 582 starlight	84,–	47,– / 14,–	**+**	**+**	**++**	**O**	**gut**

Excursion: Formal Concept Analysis

Def.: A **formal context** is a triple (G,M,I), where

- G is a set of objects,
- M a set of attributes
- and I a relation between G and M.

- $(g,m) \in I$ is read as „object g has attribute m“.

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Excursion: Formal Concept Analysis

Def.: A

formal concept

is a pair (A,B) with

- $A \subseteq G$ and $B \subseteq M$,
- A and B are maximal with $A \times B \subseteq I$.

A is the **extent** and B the **intent** of the concept.

intent

extent

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Excursion: Formal Concept Analysis

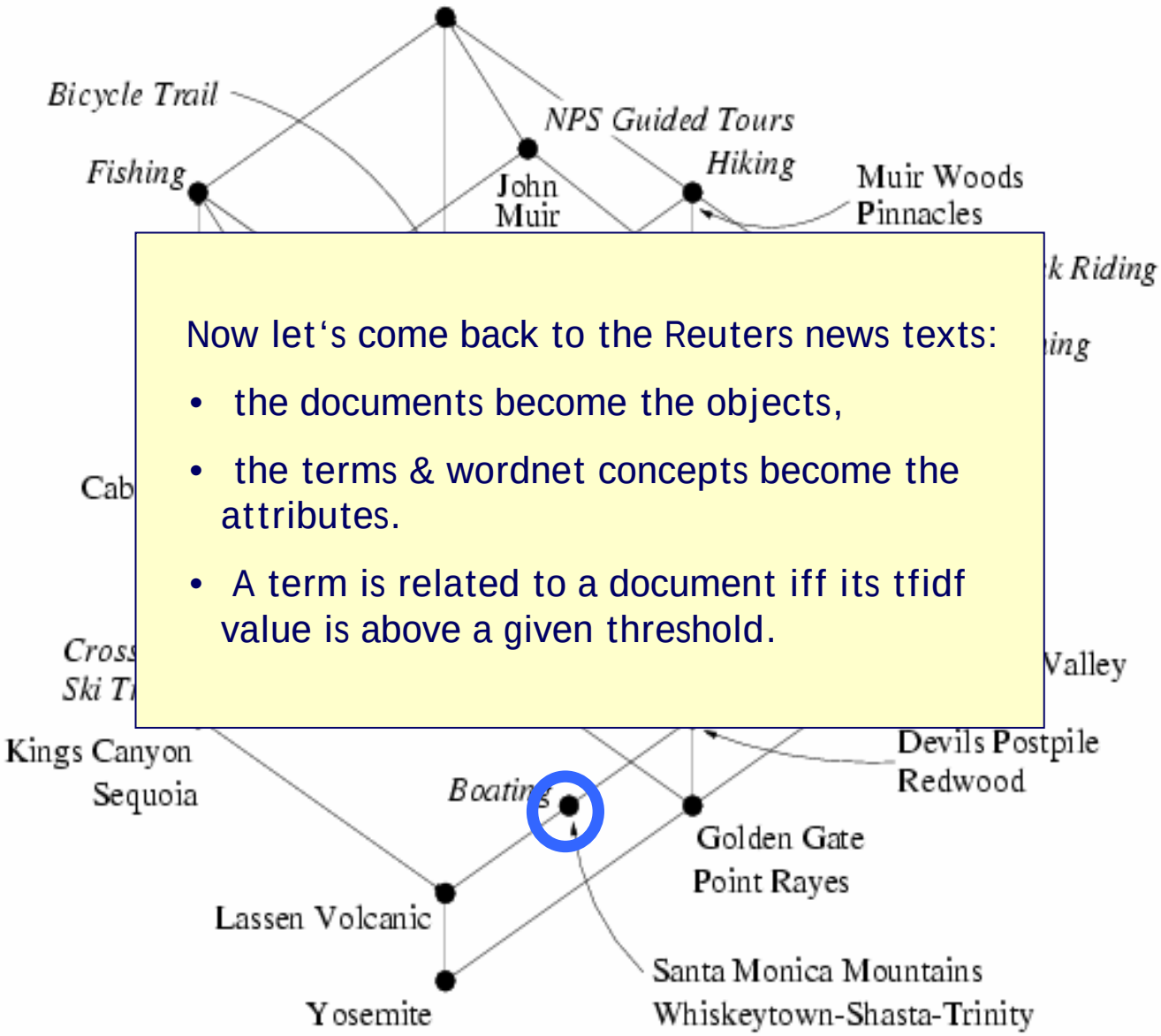
The blue concept is a **subconcept** of the yellow one, because the blue extent is contained in the yellow one.

($\Leftrightarrow$ the blue intent comprises the yellow one)

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Excursion: Formal Concept Analysis

The **concept lattice** of the formal context.



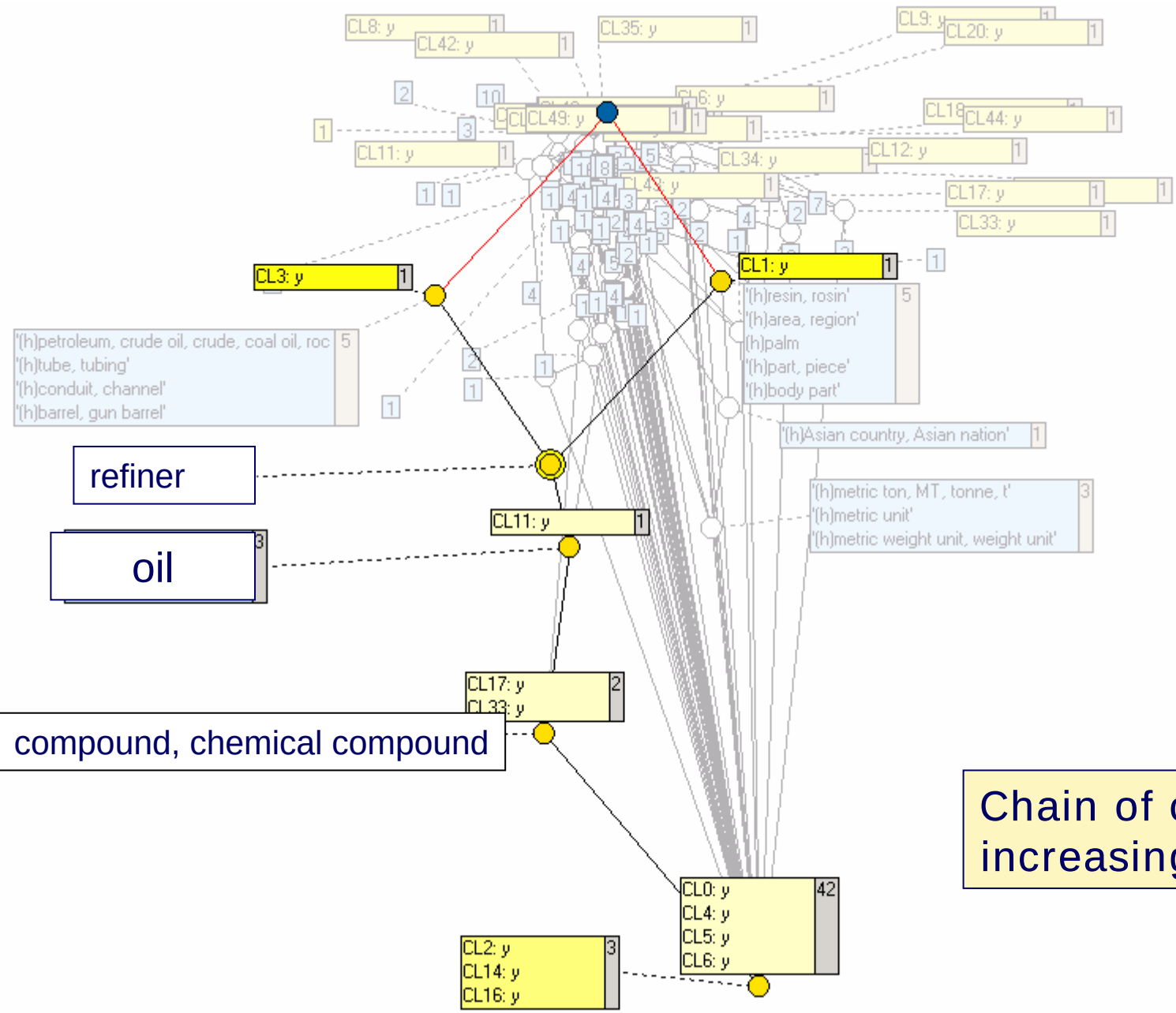
Now let's come back to the Reuters news texts:

- the documents become the objects,
- the terms & wordnet concepts become the attributes.
- A term is related to a document iff its tfidf value is above a given threshold.

National Parks in California

	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Text Clustering with Background Knowledge



Chain of concepts with increasing specificity.

Overview

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge

- Combining the above

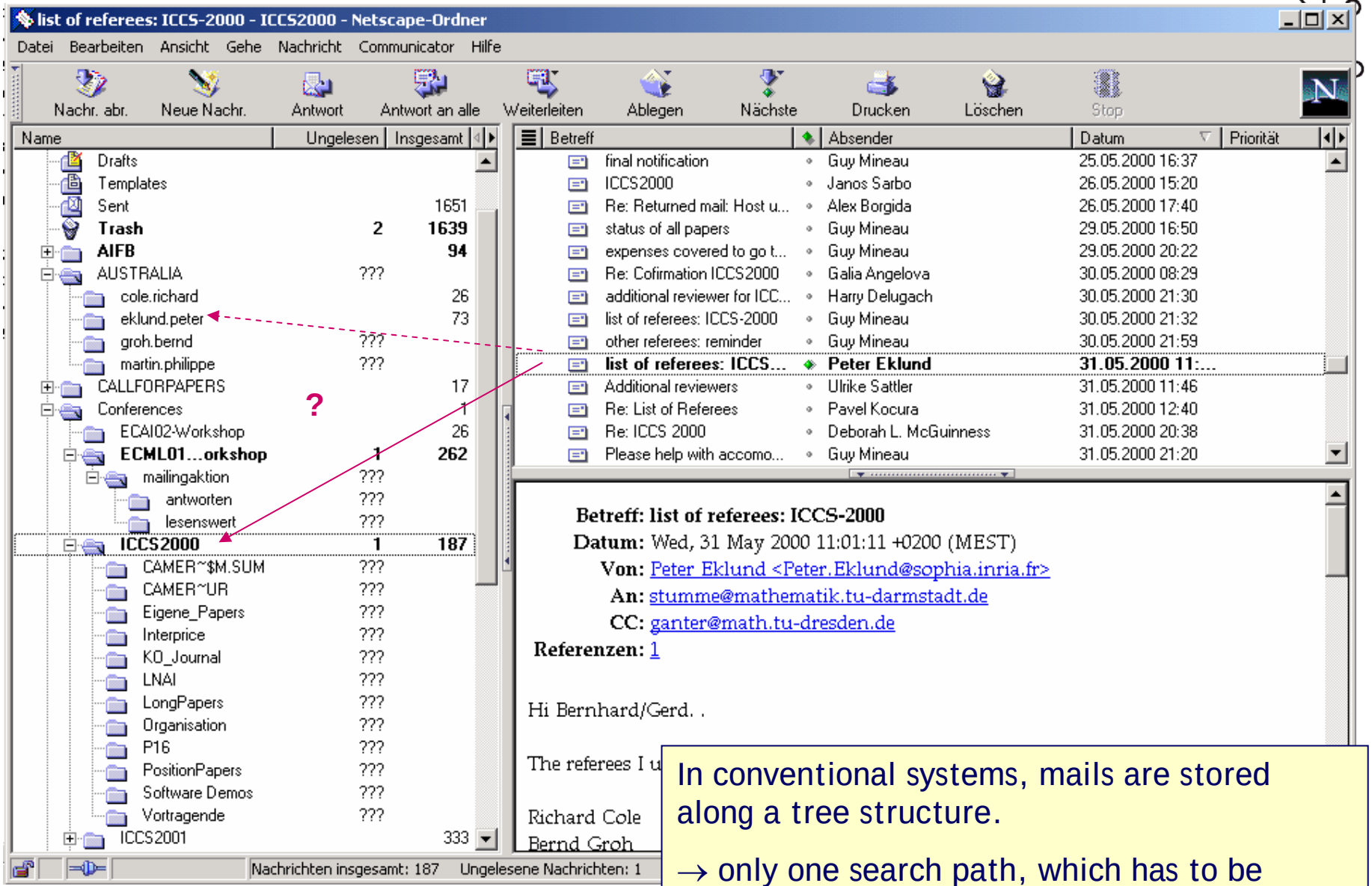
This talk gives some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems

- Courseware Watchdog

with R. Cole, P. Eklund (Australia)

Structuring Email Collections



In conventional systems, mails are stored along a tree structure.

→ only one search path, which has to be defined at storage time.

Structuring Email Collections

Our approach: FCA allows for multiple search paths and different conceptual views.

Storing emails in a formal context (G,M,I):

- G: set of all emails
- M: set of all catchwords
- $(g,m) \in I$ is read „email g has catchword m“

The relation I

- is composed automatically by Information Extraction:
  (from p.eklund@ subject kvo subject meeting)
- and can be modified by the user.

Folders correspond to formal concepts.

A partial order on M allows for inheritance of catchwords:

- ‚from Peter‘ $\rightarrow$ ‚from Australia‘

Conceptual scales allow for different views on the data.

Structuring Email Collections

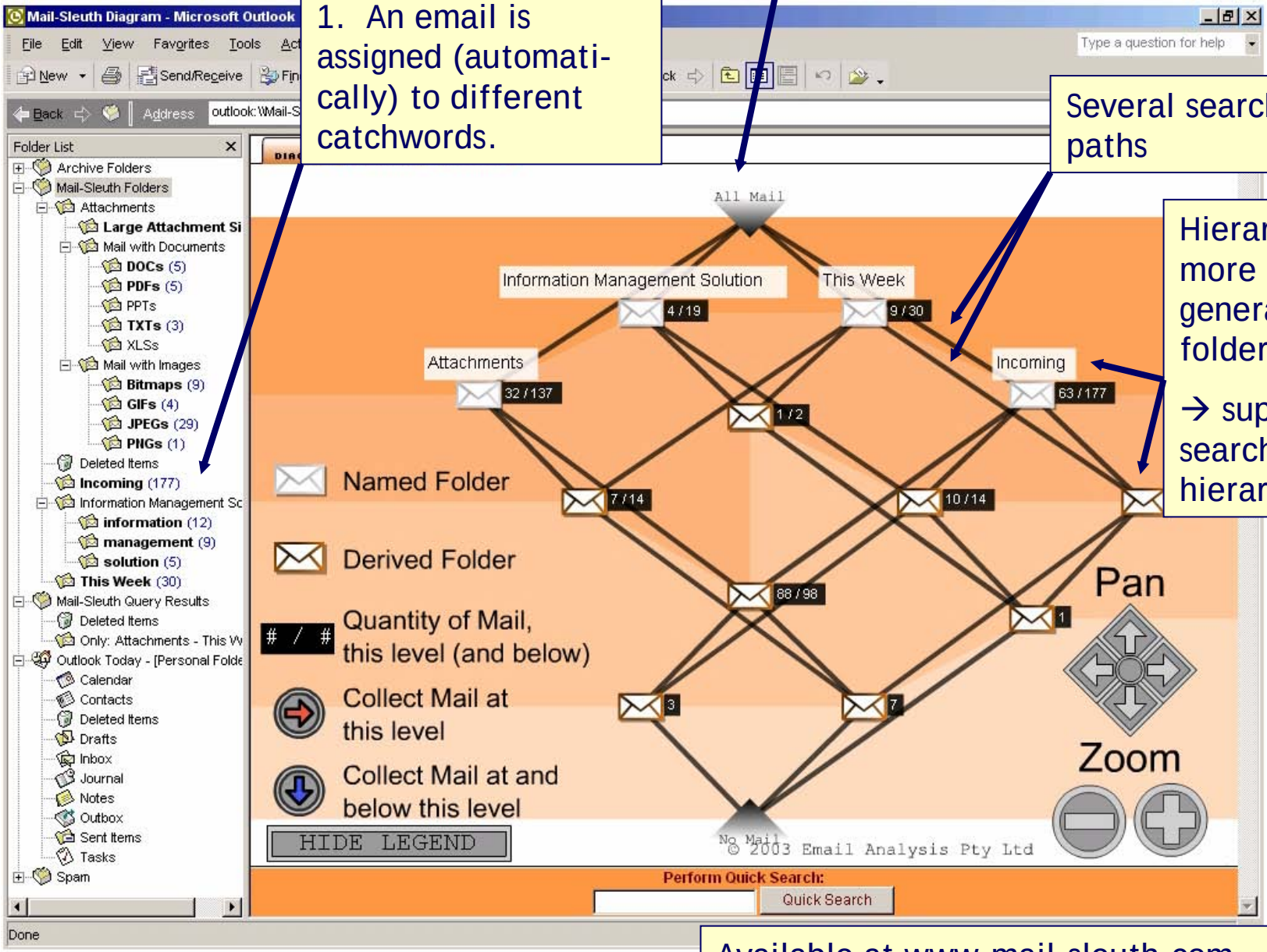
2. Computation of the concept lattice.

1. An email is assigned (automatically) to different catchwords.

Several search paths

Hierarchy of more general/specific folders.

→ supports search along hierarchy



Available at www.mail-sleuth.com

Overview

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge
- Combining the above

This talk gives some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems
- Courseware Watchdog

with Ch. Schmitz (Karlsruhe/Kassel)

Semantic Routing in P2P Networks

Peer to Peer Networks support resource sharing without central server.

Problem: How to organise the communication between the peers efficiently?

- The number of communication partners should not be too low nor too high.

Database oriented solution: Distributed Hashtables. The hash function decides which document to store where…

Our approach is inspired by "Small Worlds", known from

- Sociometry
- Biology
- Bibliometry.

→ Can peers cluster themselves into Small Worlds using local knowledge only?

- Peers in one cluster need to be semantically close, but they should also provide relations to communities further apart.

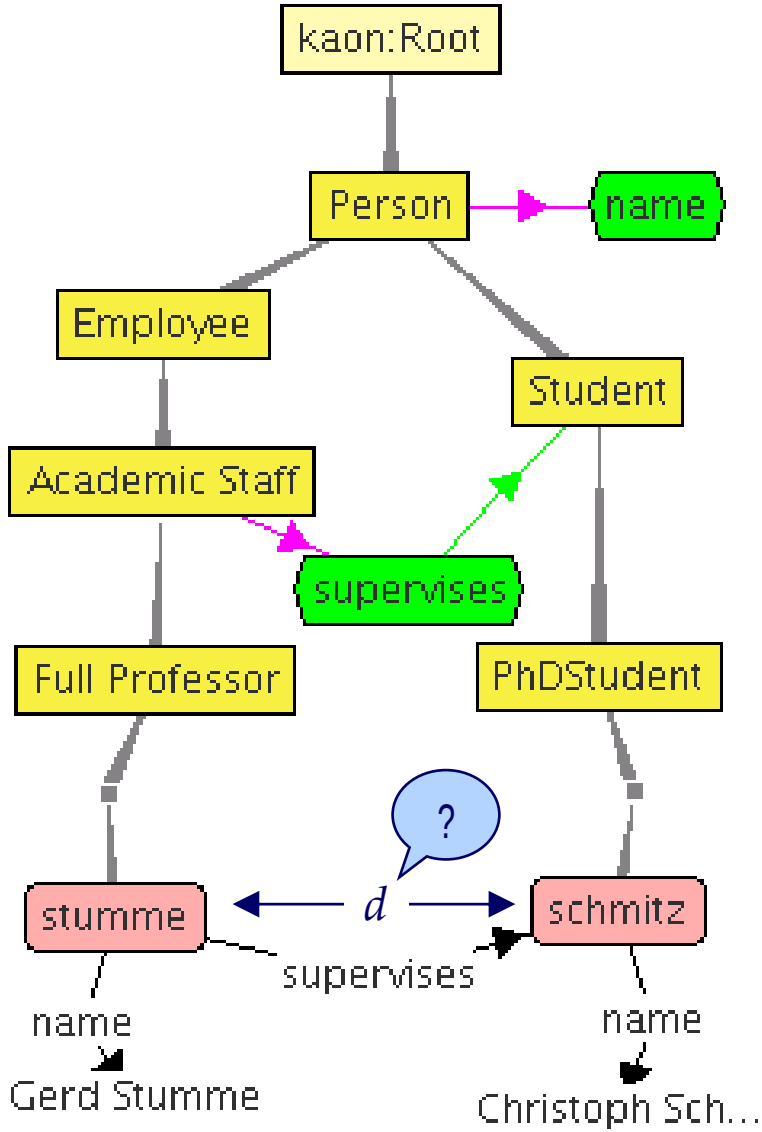
→ Can this improve the routing?

→ Are ontologies useful for this?

Semantic Routing in P2P Networks

Peers

- contain content (sic)
- which is described by an ontology.
- Peers know something about the knowledge of other peers
- and can query them.
- We try to measure the similarity of contents using the ontology.



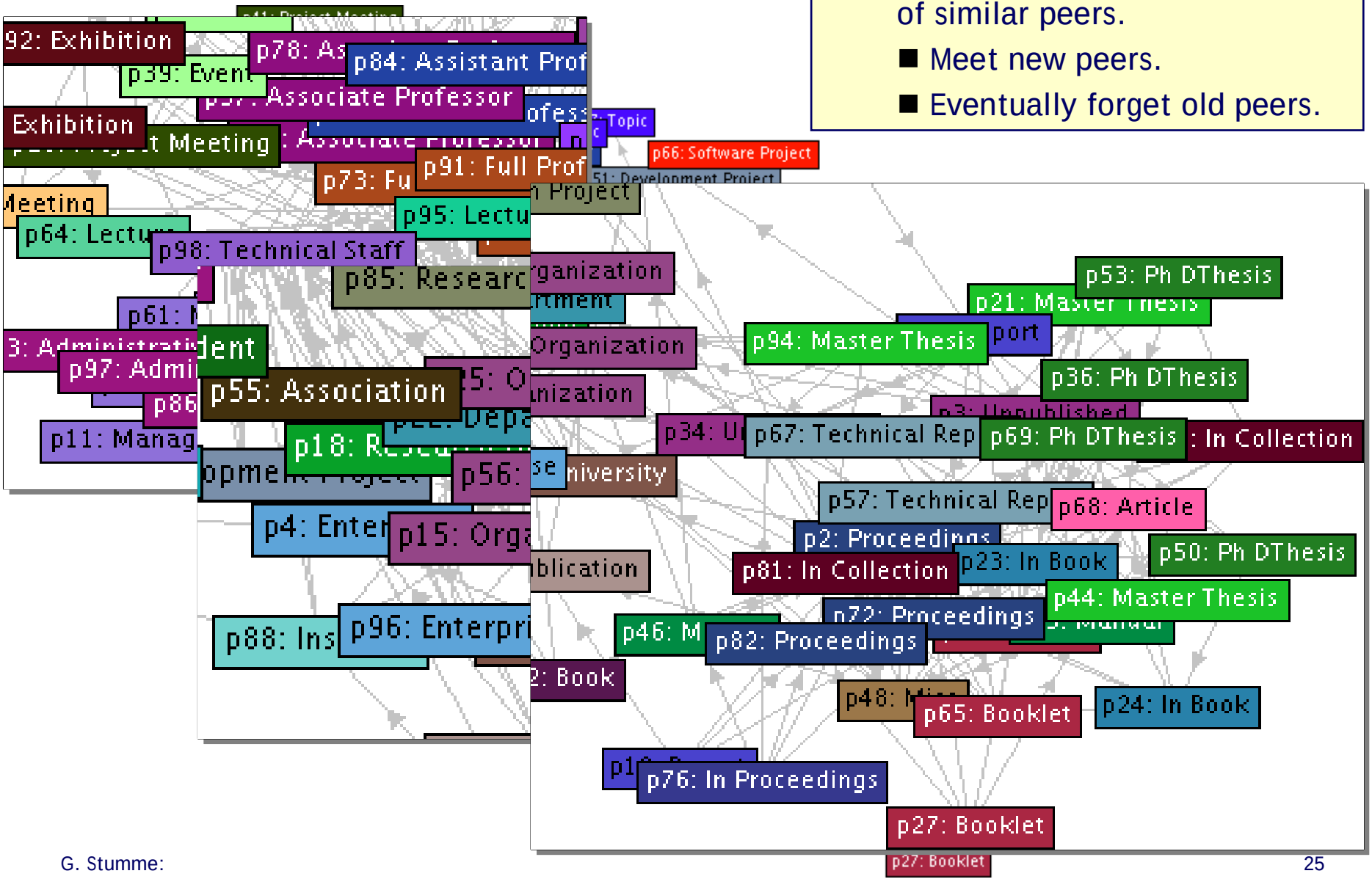
Ontology:

- Concepts
- Instances
- Relations
- Attributes

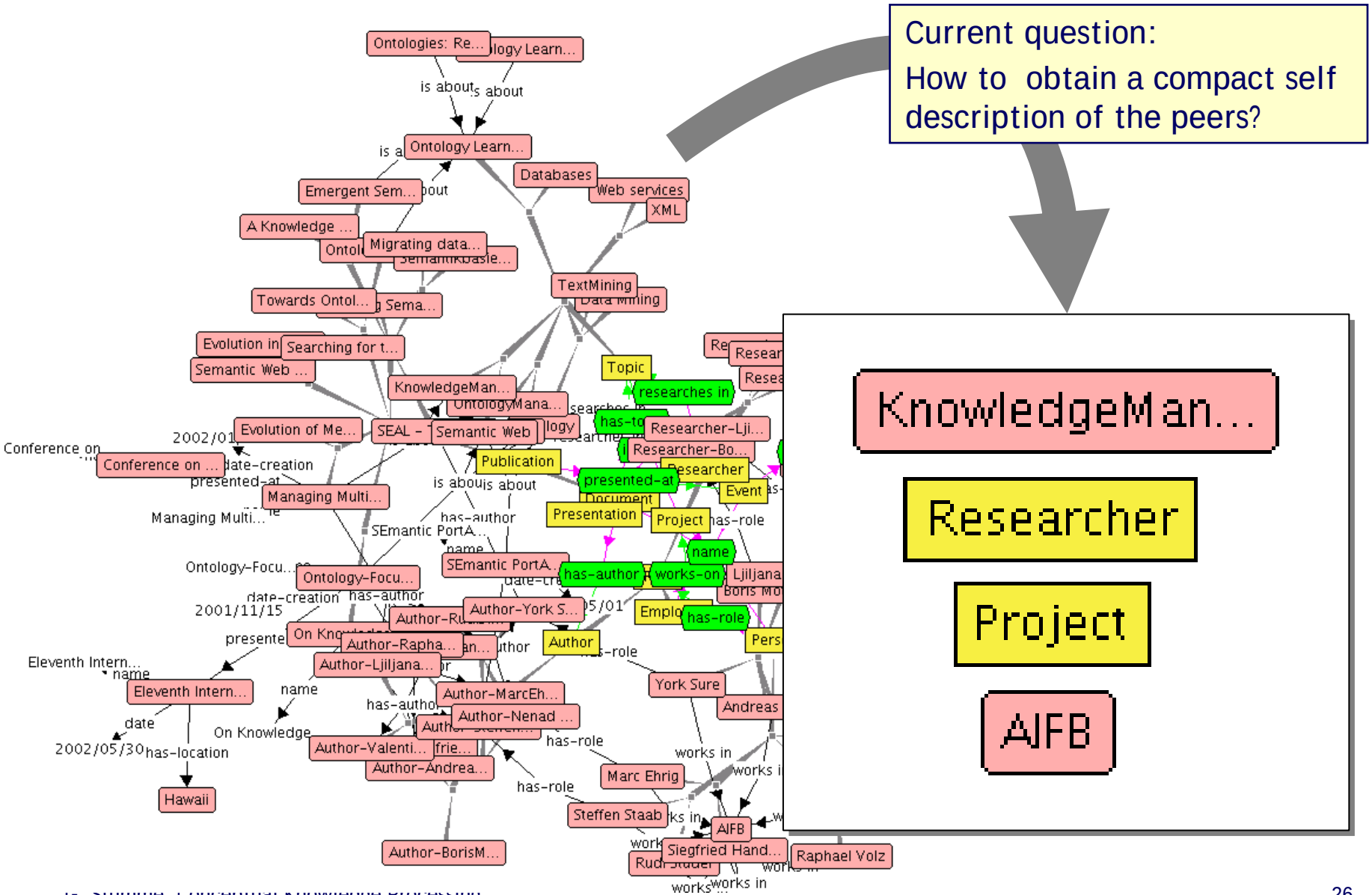
Semantic Routing in P2P Networks

Strategies for self organisation:

- Go to a dense neighborhood of similar peers.
- Meet new peers.
- Eventually forget old peers.



Semantic Routing in P2P Networks



Overview

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge
- Combining the above

This talk gives some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems
- Courseware Watchdog

with Ch. Schmitz, J. Tane (Karlsruhe/Kassel)

Courseware Watchdog

Scenario:

- Accessing and structuring distributed resources, e.g. of lecture notes of all German computer science departments.

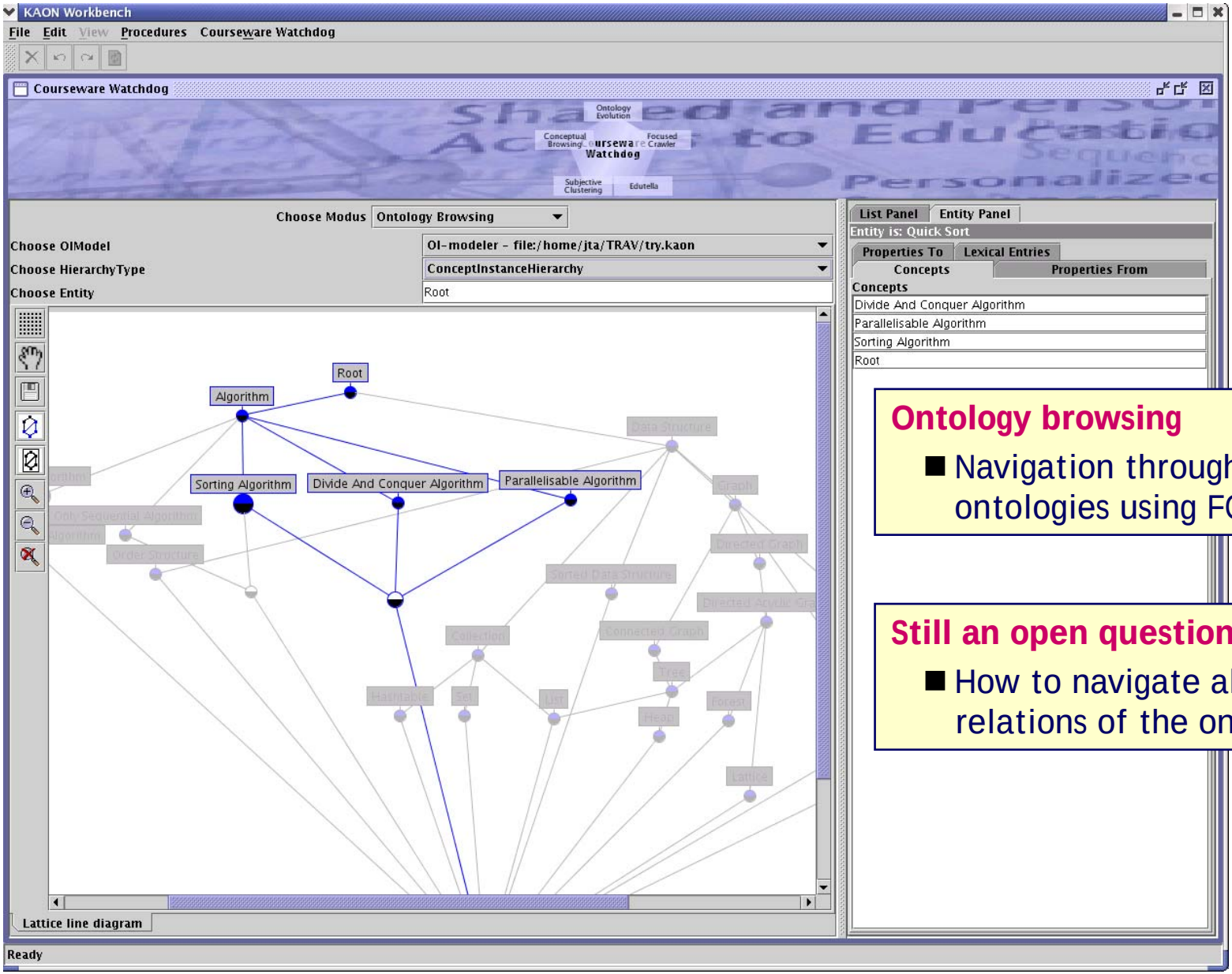
- Descriptions of resources are complex.
- The resources are distributed:
  - in different documents,
  - at different locations,
  - among different users.

Solution: Courseware Watchdog

- **Information sources:**
  - WWW via a focused crawler
  - Peer to peer network Edutella

- **User interface**
  - Conceptual browsing

Courseware Watchdog



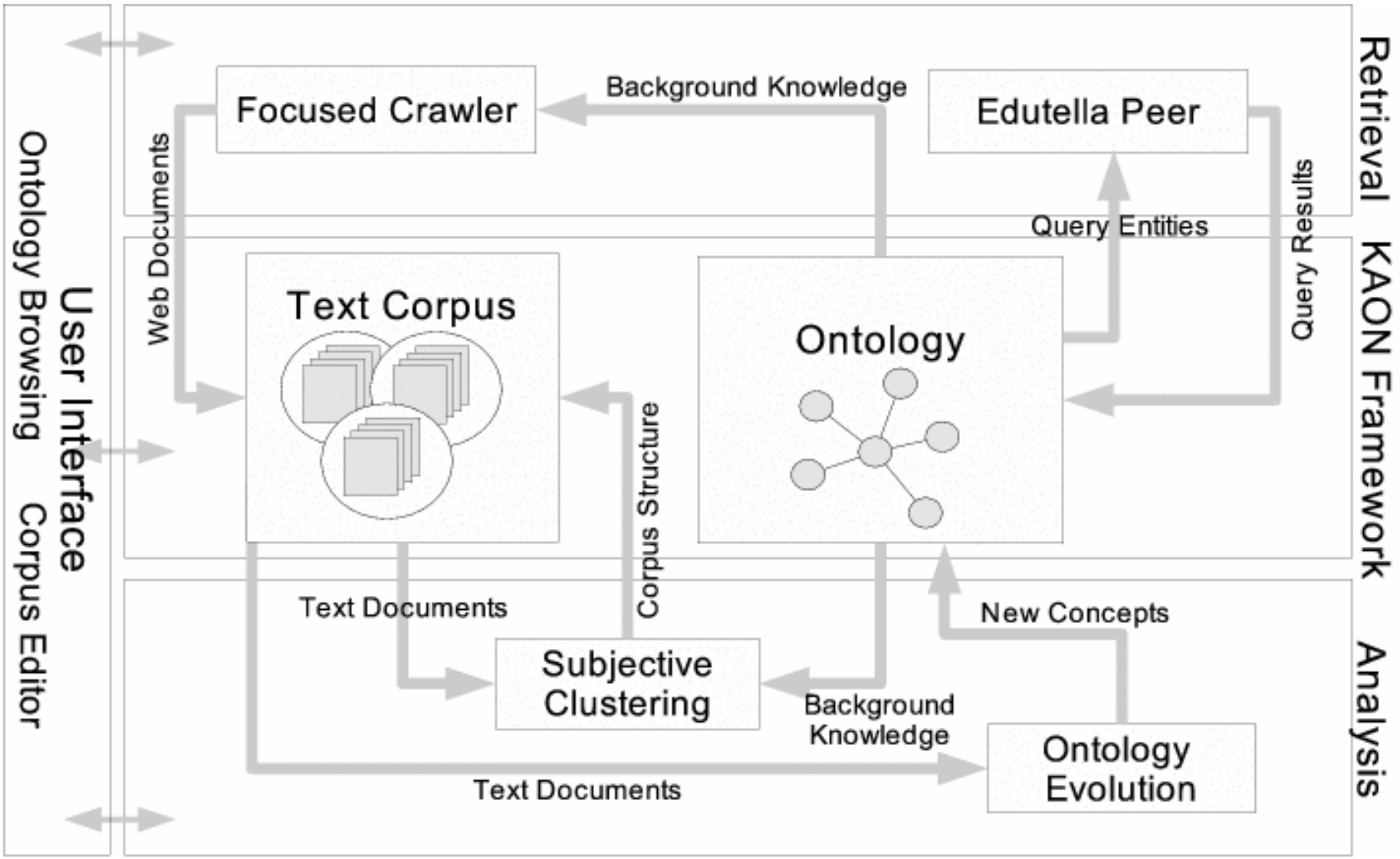
Ontology browsing

- Navigation through ontologies using FCA.

Still an open question:

- How to navigate along the relations of the ontology?

Courseware Watchdog



The End

Ontologies support a.o. the following tasks of knowledge management:

- Acquiring Knowledge
- Organizing Knowledge
- Retrieving Knowledge

- Combining the above

This talk *gave* some example applications:

- Text Clustering
- Conceptual Email Manager
- Semantic Routing in P2P Systems

- Courseware Watchdog

with R. Cole, P. Eklund, A. Hotho, Ch. Schmitz, J. Tane