Reducing the Representation Complexity of Lattice-based Taxonomies

Sergei Obiedkov

Higher School of Economics, Moscow, Russia

Taxonomies of epistemic communities

Problem

Describe the structure of an *epistemic community*, e.g. of a scientific community.

Epistemic communities

- Groups of agents sharing a common set of subjects, concepts, notions, issues; a common goal of knowledge creation—Haas (1992)
- Consist of minor subcommunities, possibly not disjoint
- Citations, co-authorship, terminology...
- Social and semantic aspects

Taxonomies of epistemic communities

Problem

Describe the structure of an *epistemic community*, e.g. of a scientific community.

Epistemic communities

- Groups of agents sharing a common set of subjects, concepts, notions, issues; a common goal of knowledge creation—Haas (1992)
- Consist of minor subcommunities, possibly not disjoint
- Citations, co-authorship, terminology...
- Social and semantic aspects

Taxonomies of epistemic communities

Problem

Describe the structure of an *epistemic community*, e.g. of a scientific community.

Epistemic communities

- Groups of agents sharing a common set of subjects, concepts, notions, issues; a common goal of knowledge creation—Haas (1992)
- Consist of minor subcommunities, possibly not disjoint
- Citations, co-authorship, terminology...
- Social and semantic aspects

Taxonomies of epistemic communities

Problem

Describe the structure of an *epistemic community*, e.g. of a scientific community.

Epistemic communities

- Groups of agents sharing a common set of subjects, concepts, notions, issues; a common goal of knowledge creation—Haas (1992)
- Consist of minor subcommunities, possibly not disjoint
- Citations, co-authorship, terminology...
- Social and semantic aspects

Taxonomies of epistemic communities

Problem

Describe the structure of an *epistemic community*, e.g. of a scientific community.

Epistemic communities

- Groups of agents sharing a common set of subjects, concepts, notions, issues; a common goal of knowledge creation—Haas (1992)
- Consist of minor subcommunities, possibly not disjoint
- Citations, co-authorship, terminology...
- Social and semantic aspects

Concept lattices
Formal Concept Analysis (Ganter, Wille 1999)

A formal context $\mathbb{K} = (G, M, I)$

- set of objects G
- set of attributes M
- objects have attributes: relation $I \subseteq G \times M$

Derivation operators

For $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

Concept lattices
Formal Concept Analysis (Ganter, Wille 1999)

A formal context $\mathbb{K} = (G, M, I)$

- set of objects G
- set of attributes M
- objects have attributes: relation $I \subseteq G \times M$

Derivation operators

For $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

Concept lattices

Derivation operators

For $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

A formal concept (A, B)

- $A \subseteq G$
- $B \subseteq M$
- $A' = B$
- $B' = A$

A is the concept extent, B is the concept intent

The set of concepts of context $\mathbb{K}$ forms a lattice $\underline{\mathfrak{B}}(\mathbb{K})$

Concept lattices

Derivation operators

For $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

A formal concept (A, B)

- $A \subseteq G$
- $B \subseteq M$
- $A' = B$
- $B' = A$

A is the concept extent, B is the concept intent

The set of concepts of context $\mathbb{K}$ forms a lattice $\underline{\mathfrak{B}}(\mathbb{K})$

Concept lattices

Derivation operators

For $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

A formal concept (A, B)

- $A \subseteq G$
- $B \subseteq M$
- $A' = B$
- $B' = A$

A is the concept extent, B is the concept intent
The set of concepts of context $\mathbb{K}$ forms a lattice $\underline{\mathfrak{B}}(\mathbb{K})$

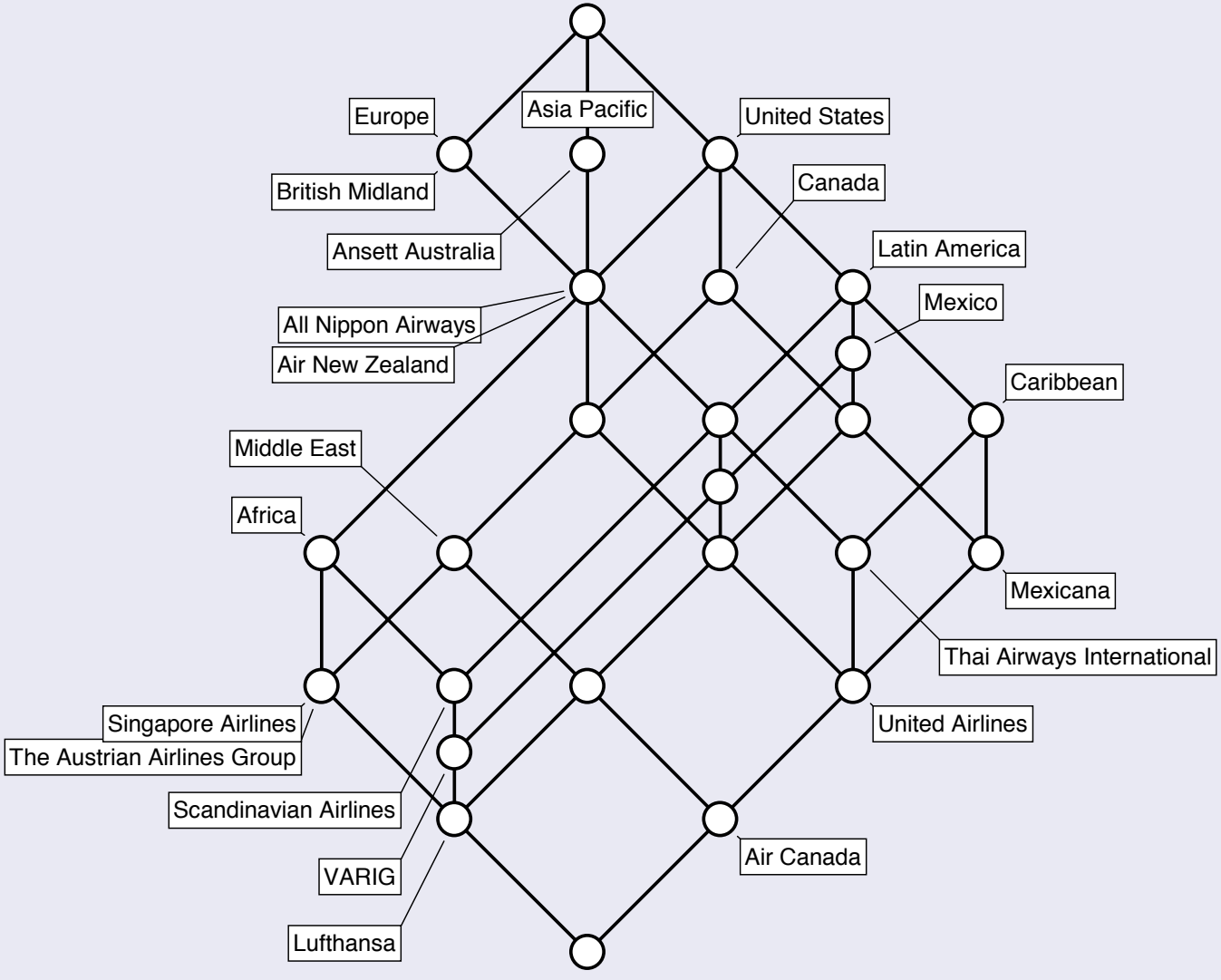
Example

Destinations of the Star Alliance Members

	Latin America	Europe	Canada	Asia Pacific	Middle East	Africa	Mexico	Caribbean	United States
Air Canada	X	X	X	X	X		X	X	X
Air New Zealand		X		X					X
All Nippon Airways		X		X					X
Ansett Australia				X					
The Austrian Airlines Group		X	X	X	X	X			X
British Midland		X							
Lufthansa	X	X	X	X	X	X	X		X
Mexicana	X		X				X	X	X
Scandinavian Airlines	X	X		X		X			X
Singapore Airlines		X	X	X	X	X			X
Thai Airways International	X	X		X				X	X
United Airlines	X	X	X	X			X	X	X
VARIG	X	X		X		X	X		X

Example

Destinations of the Star Alliance Members



Lattice taxonomies of epistemic communities

- Formal context over the set of authors G and the set of terms M
- The intent of a concept describes common interests of authors from the concept extent
- *An epistemic community* is a formal concept
- The concept lattice structures subcommunities of the subject domain under consideration

Example

Researchers that study zebrafish

`Medline` database, annotations that contain the word "zebrafish", 1998-2003, a random sample: 250 authors, 18 words

Experts' description of the domain

1. Biochemical signaling mechanisms, involving metabolic pathways and receptors.
2. Comparative studies.
3. Brain, nervous system.

Example

Researchers that study zebrafish

`Medline` database, annotations that contain the word "zebrafish", 1998-2003, a random sample: 250 authors, 18 words

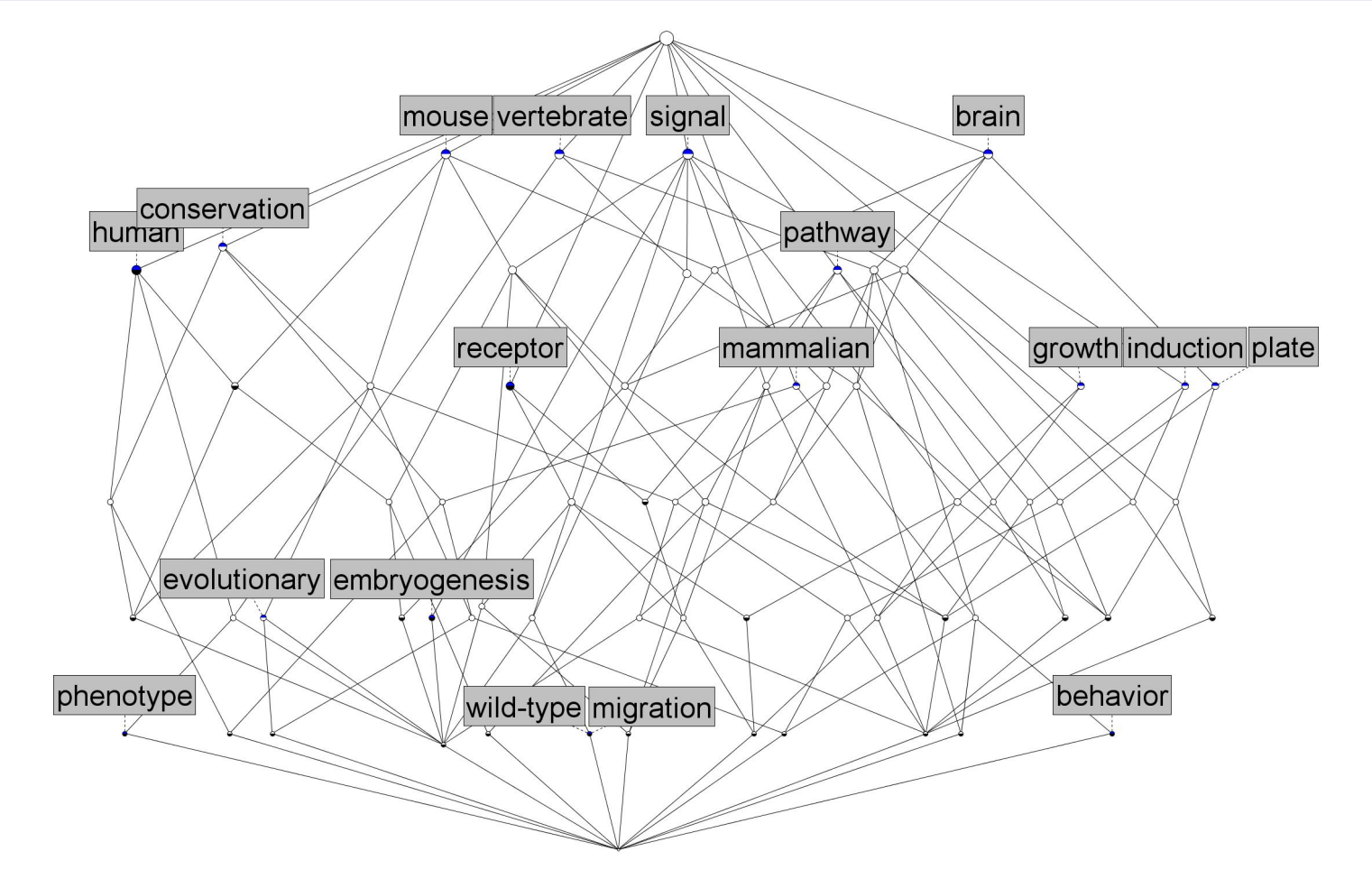
Experts' description of the domain

1. Biochemical signaling mechanisms, involving metabolic pathways and receptors.
2. Comparative studies.
3. Brain, nervous system.

Example: a concept lattice

for a sample of 25 authors

69 concepts



Lattice pruning

Large lattices are hard to

- compute
- store
- interprete

Some concepts of the lattice

- are due to noise in data
- reflect noninteresting details

Solutions

- keep an upper part of the lattice (large communities)
- discard “irrelevant” concepts
- use several representation levels

Lattice pruning

Large lattices are hard to

- compute
- store
- interprete

Some concepts of the lattice

- are due to noise in data
- reflect noninteresting details

Solutions

- keep an upper part of the lattice (large communities)
- discard “irrelevant” concepts
- use several representation levels

Lattice pruning

Large lattices are hard to

- compute
- store
- interprete

Some concepts of the lattice

- are due to noise in data
- reflect noninteresting details

Solutions

- keep an upper part of the lattice (large communities)
- discard “irrelevant” concepts
- use several representation levels

Stability-based pruning

Stability indices

- How much the intent of a formal concept (A, B) depends on particular objects from the extent:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- ...or how likely it is that the intent is retained upon deletion of an arbitrary number of objects from the context:

$$\sigma(A, B) = \frac{|\{\mathbb{K}_H \mid H \subseteq G \text{ and } B = B^{I_H I_H}\}|}{2^{|G|}}$$

Stability-based pruning

Stability indices

- How much the intent of a formal concept (A, B) depends on particular objects from the extent:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- ...or how likely it is that the intent is retained upon deletion of an arbitrary number of objects from the context:

$$\sigma(A, B) = \frac{|\{\mathbb{K}_H \mid H \subseteq G \text{ and } B = B^{I_H I_H}\}|}{2^{|G|}}$$

Computing stability

Concepts := $\underline{\mathfrak{B}}(\mathbb{K})$
for each (A, B) in Concepts
  Count[(A, B)] := the number of lower neighbors of (A, B)
  Subsets[(A, B)] := $2^{|A|}$
end for
while Concepts is not empty
  let (C, D) be any concept from Concepts with Count[(C, D)] = 0
  Stability[(C, D)] := Subsets[(C, D)] / $2^{|C|}$
  remove (C, D) from Concepts
  for each (A, B) $>$ (C, D)
    Subsets[(A, B)] := Subsets[(A, B)] $-$ Subsets[(C, D)]
    if (A, B) $\succ$ (C, D)
      Count[(A, B)] := Count[(A, B)] $-$ 1
    end if
  end for
end while
return Stability

Stability: Properties useful for approximate computation

- Given a concept (A, B) of a context (G, M, I), if there is a set $A_1 \subset A$ such that $A_1' \neq B$, then $\sigma(A, B) \leq 1 - 1/2^{|A \setminus A_1|}$.
- Given a concept (A, B) of a context (G, M, I), if there are two sets $A_1, A_2 \subset A$ such that $|A_1| = |A_2|$, $A_1 \neq A_2$, and $A_1', A_2' \neq B$, then $\sigma(A, B) \leq 1 - \frac{3}{2^{|A \setminus A_1|+1}}$.

Stability: Addition of new objects to the context

- Given a concept (A, B) of a context (G, M, I), if a new object g is added to form context $(G \cup \{g\}, M, J)$ (such that $(G \times M) \cap J = I$), then
  1. For an old concept (A, B), we have $\sigma_J(A, B) = \sigma_I(A, B)$.
  2. For a modified concept $(A \cup \{g\}, B)$, we have

$$\sigma_I(A, B) \leq \sigma_J(A \cup \{g\}, B) \leq 1/2 + \sigma_I(A, B)/2.$$

  3. For a new concept (A, B), we have

$$\sigma_J(A, B) \begin{cases} = 1/2, & \text{if } B = \{g\}'; \\ < 1/2, & \text{otherwise.} \end{cases}$$

Stability-based pruning

Lattice pruning

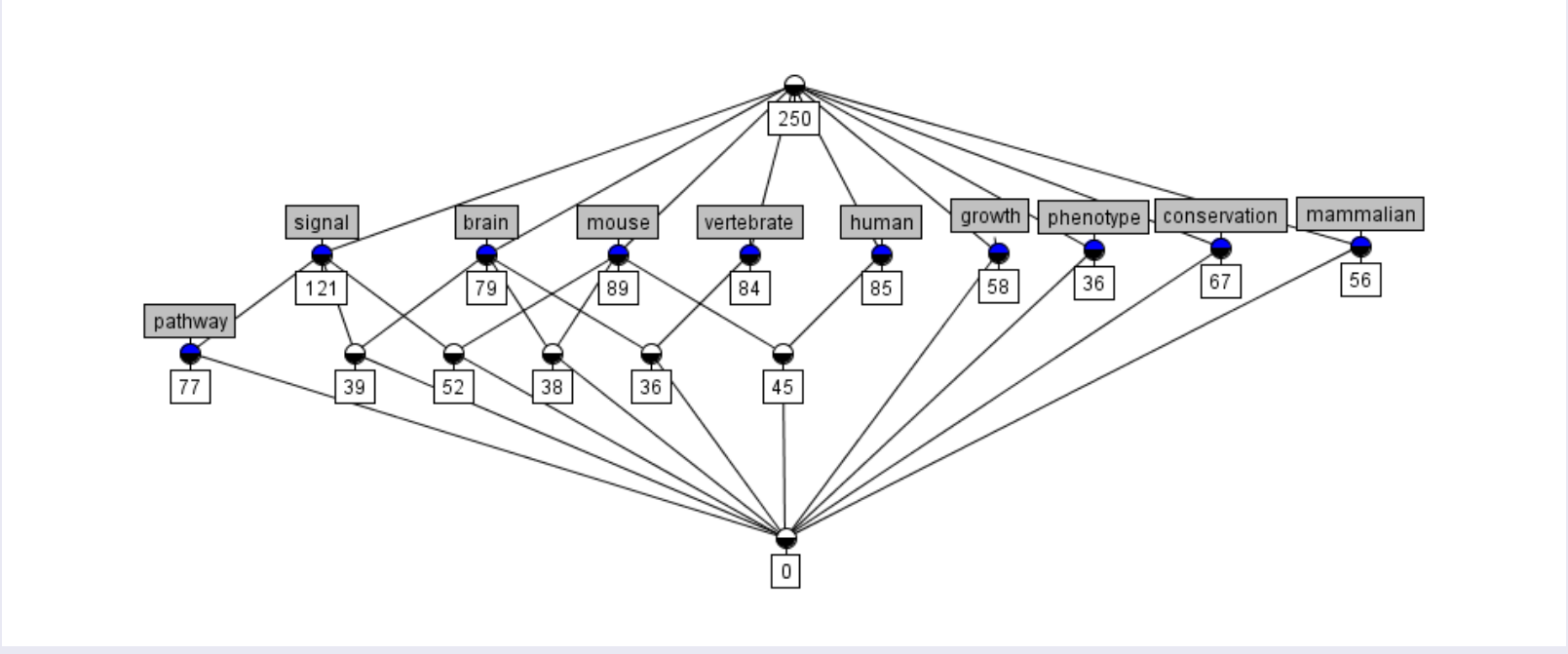
First, we simply delete concepts with the stability index below a threshold

Stability-based pruning

Lattice pruning

First, we simply delete concepts with the stability index below a threshold

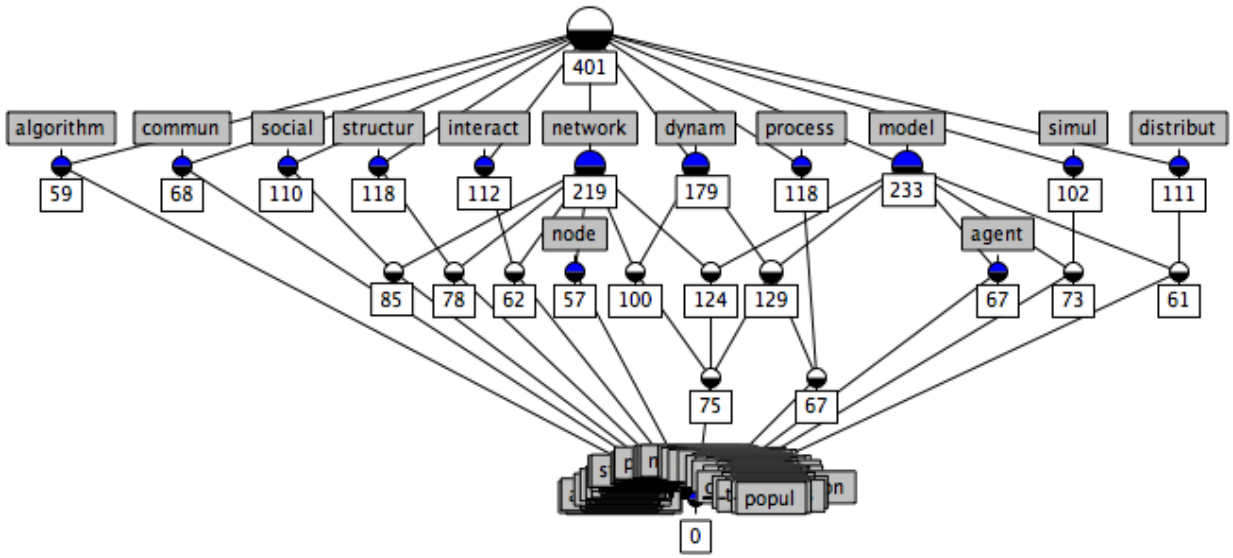
A pruned lattice: from 1100+ to 25 concepts



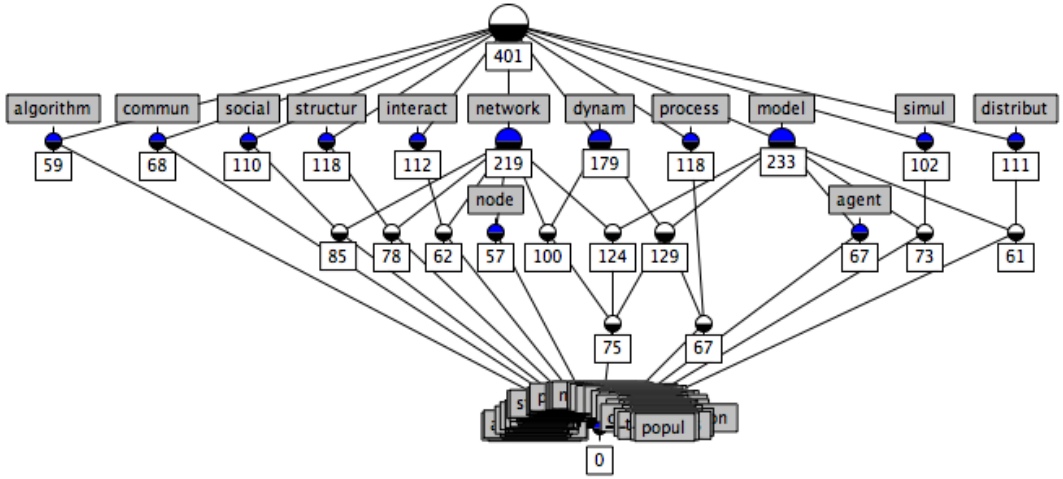
Another example: ECCS 2006

Abstracts of papers presented at the European Conference on Complex Systems

- G is the set of authors: $|G| = 401$
- M is the set of terms: $|M| = 109$
- The concept lattice contains 6011 concepts
- Select the 25 most stable concepts

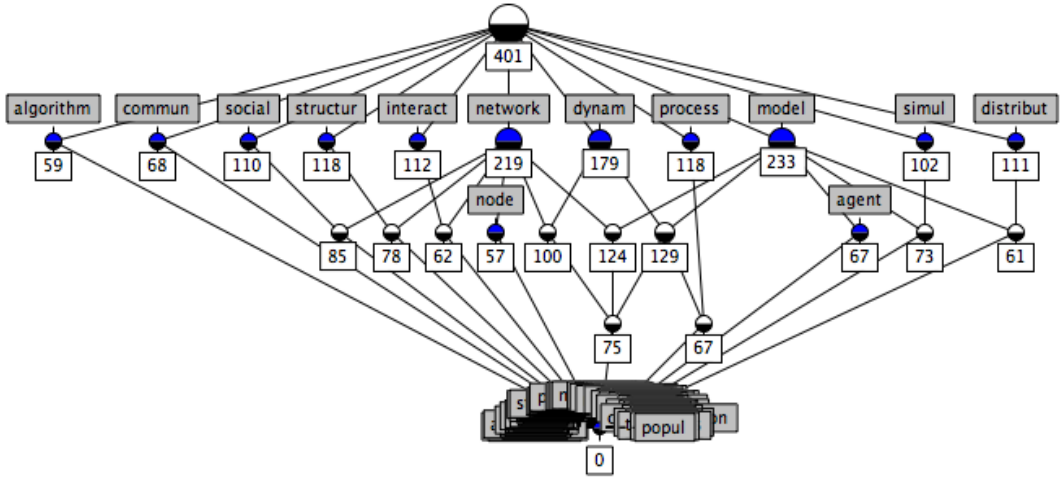


ECCS



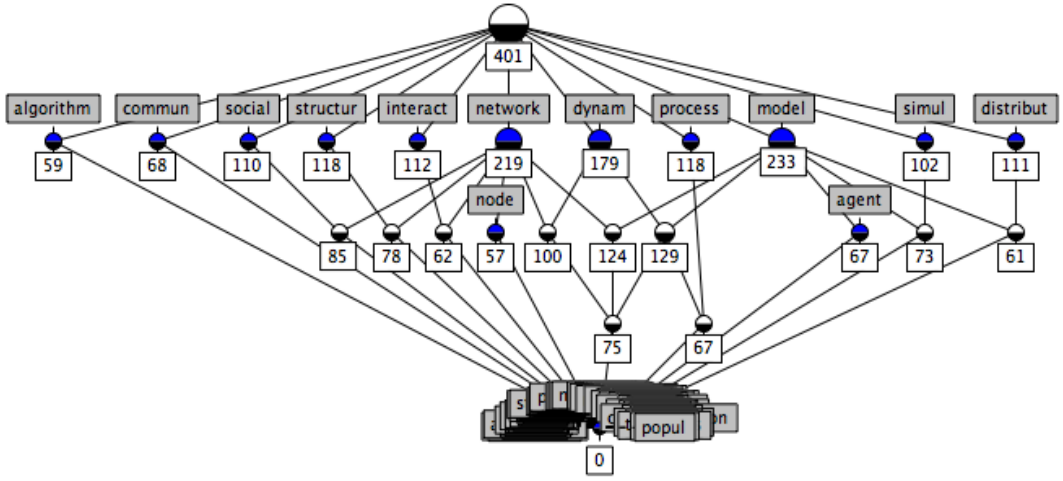
- “network” is a central issue
  - “social network” (agent-based networks)
  - “structure network” (topological issues)
  - “interact network” (networks as representation of interactions)
  - “node network” (a node being a basic unit)
  - “dynamics network” (evolution of networks)
  - “model network” (modeling of networks).

ECCS



- “model” is an important topic
  - related to “agents”, “simulation”, and “dynamics”
  - “network dynamics model”: scientists interested in the modeling of network dynamics (morphogenesis)
  - “model distribut”: the use of models to reconstruct distributions of any kind
  - “model dynam process”: the modeling of dynamical processes

ECCS



- Isolated topics: minor fields focused on particular issues
  - “algorithm”: introduction and use of novel and general algorithms to achieve empirical measurements in a variety of cases
  - “community”: community and cluster detection

Nested line diagrams

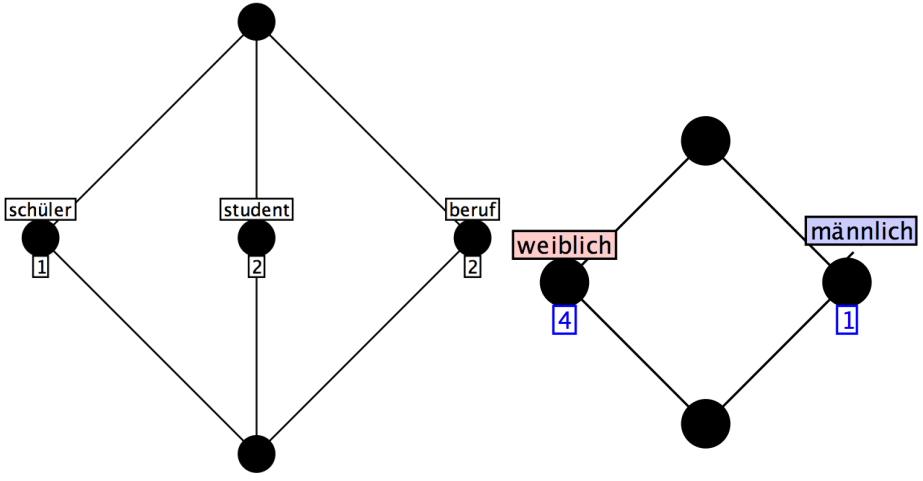
Nested line diagrams

- We partition the set of attributes into two parts
- *Internal and external parts of the lattice*: The concept (A, B) "inside" the concept (C, D) corresponds to $(A \cap C, B \cup D)$

Nested line diagrams

Nested line diagrams

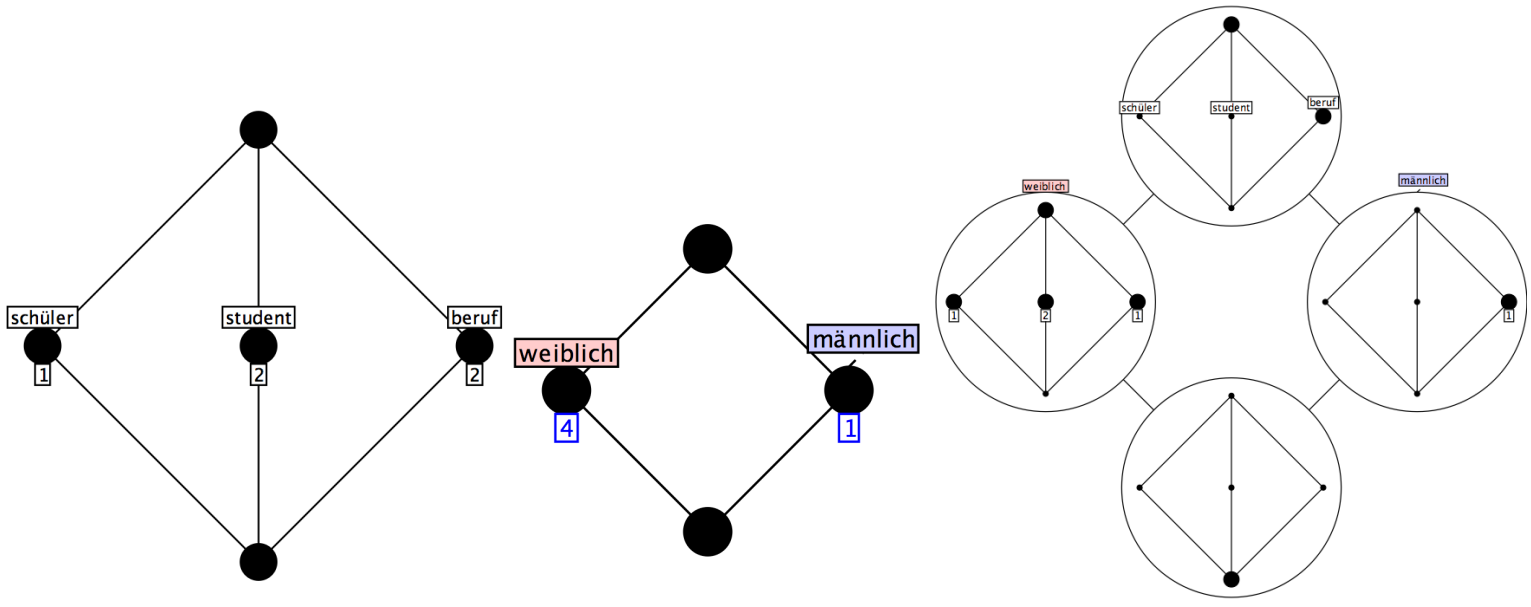
- We partition the set of attributes into two parts
- *Internal and external parts of the lattice*: The concept (A, B) "inside" the concept (C, D) corresponds to $(A \cap C, B \cup D)$



Nested line diagrams

Nested line diagrams

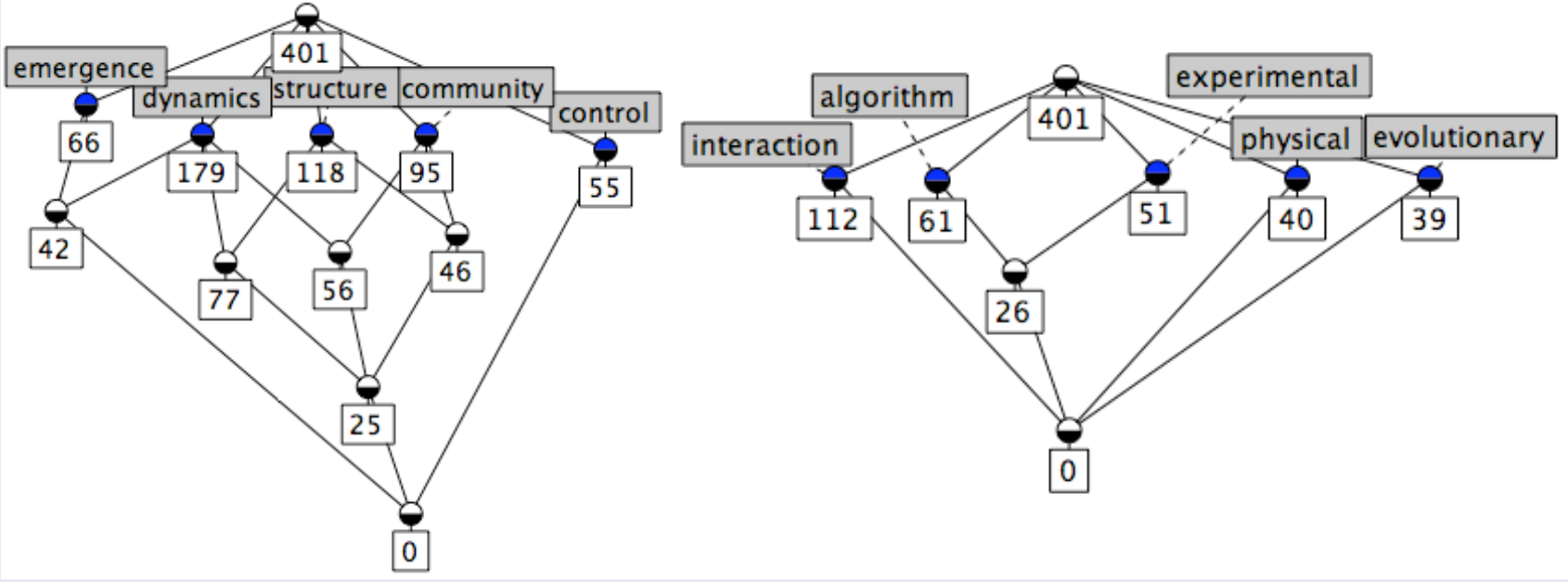
- We partition the set of attributes into two parts
- *Internal and external parts of the lattice*: The concept (A, B) "inside" the concept (C, D) corresponds to $(A \cap C, B \cup D)$



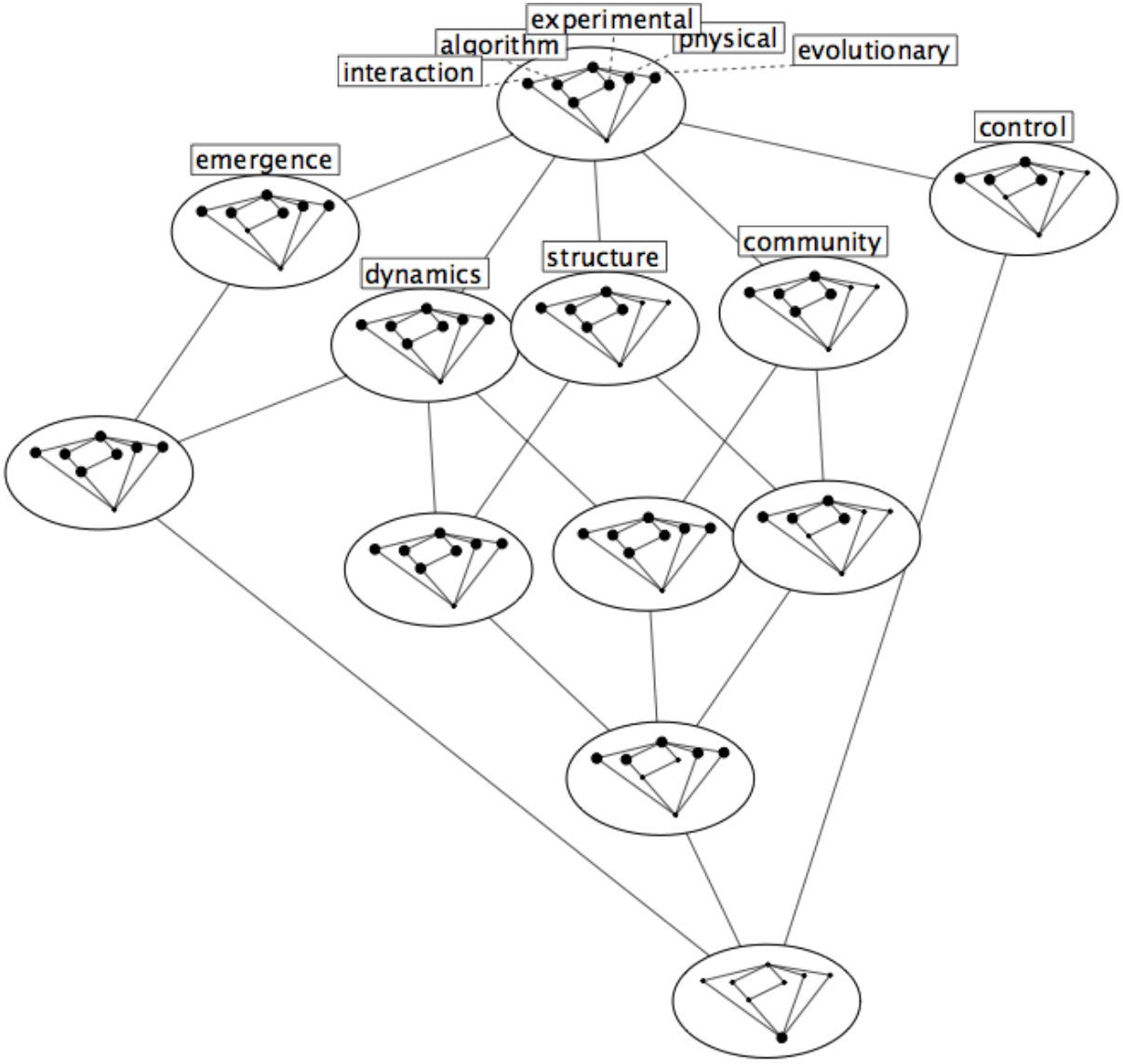
Combining nested line diagrams and stability-based pruning

We prune external and internal lattices using stability index

Pruned outer and inner lattices: topics vs. methods

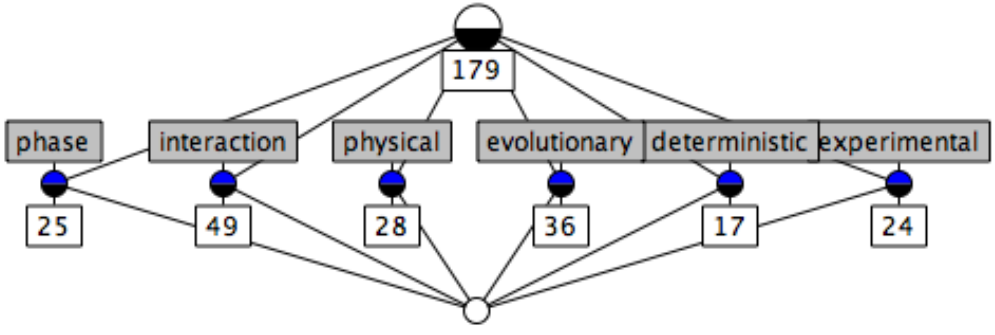


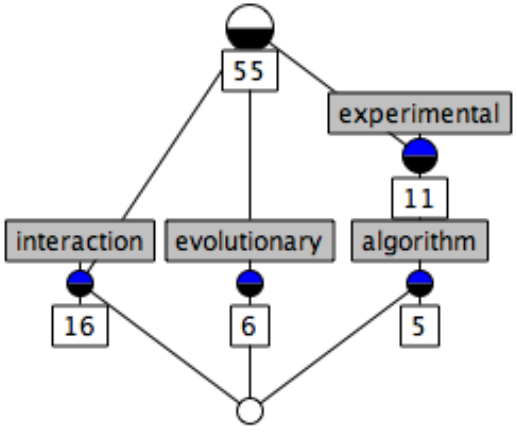
Combining nested line diagrams and stability-based pruning



Nesting vs. zooming

- The stabilized nested line diagram shows how major methods are distributed among major topics
- Zooming: To see the most important methods for each topic, construct the lattices of methods for the populations (extents) of individual topics
- “dynamics” vs. “control”





Further Work

1. Variants of stability indices
2. Pruning strategies
3. Improved embedding
4. Modeling dynamics
5. Better linguistic processing

Further Work

Variants of stability

- Intensional stability:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- Extensional stability:

$$\sigma_e(A, B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}}$$

- Generalized stability ($A_H = A \cap H$ and $B_N = B \cap N$):

$$\frac{|\{(H, N, J) \mid H \subseteq G, N \subseteq M, J = I \cap (H \times N), A_H^J = B_N, B_N^J = A_H\}|}{2^{|G|+|M|}}$$

- Approximate computation of stability

Further Work

Variants of stability

- Intensional stability:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- Extensional stability:

$$\sigma_e(A, B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}}$$

- Generalized stability ($A_H = A \cap H$ and $B_N = B \cap N$):

$$\frac{|\{(H, N, J) \mid H \subseteq G, N \subseteq M, J = I \cap (H \times N), A_H^J = B_N, B_N^J = A_H\}|}{2^{|G|+|M|}}$$

- Approximate computation of stability

Further Work

Variants of stability

- Intensional stability:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- Extensional stability:

$$\sigma_e(A, B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}}$$

- Generalized stability ($A_H = A \cap H$ and $B_N = B \cap N$):

$$\frac{|\{(H, N, J) \mid H \subseteq G, N \subseteq M, J = I \cap (H \times N), A_H^J = B_N, B_N^J = A_H\}|}{2^{|G|+|M|}}$$

- Approximate computation of stability

Further Work

Variants of stability

- Intensional stability:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

- Extensional stability:

$$\sigma_e(A, B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}}$$

- Generalized stability ($A_H = A \cap H$ and $B_N = B \cap N$):

$$\frac{|\{(H, N, J) \mid H \subseteq G, N \subseteq M, J = I \cap (H \times N), A_H^J = B_N, B_N^J = A_H\}|}{2^{|G|+|M|}}$$

- Approximate computation of stability

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes ("humans and mice are mammals")
- Add intersections of stable intents
- Delete some stable intents
- "Merge" a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \to D$ to the set of context implications
- "Fault-tolerant" approximate concepts

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes (“humans and mice are mammals”)
- Add intersections of stable intents
- Delete some stable intents
- “Merge” a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \to D$ to the set of context implications
- “Fault-tolerant” approximate concepts

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes (“humans and mice are mammals”)
- Add intersections of stable intents
- Delete some stable intents
- “Merge” a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \to D$ to the set of context implications
- “Fault-tolerant” approximate concepts

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes (“humans and mice are mammals”)
- Add intersections of stable intents
- Delete some stable intents
- “Merge” a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \to D$ to the set of context implications
- “Fault-tolerant” approximate concepts

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes (“humans and mice are mammals”)
- Add intersections of stable intents
- Delete some stable intents
- “Merge” a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \rightarrow D$ to the set of context implications
- “Fault-tolerant” approximate concepts

Further Work

Pruning strategies

- Monotone criteria (sizes of intent/extent)
- Additional hierarchy on attributes (“humans and mice are mammals”)
- Add intersections of stable intents
- Delete some stable intents
- “Merge” a non-stable concept (A, B) and its descendant (C, D)
  - Modify the context: $I \Rightarrow I \cup A \times D$
  - Modify the lattice: introduce an implication $B \rightarrow D$ to the set of context implications
- “Fault-tolerant” approximate concepts

Further Work

Nested line diagrams

- Several nesting levels
- Methods of attribute partition
  Word types: verb, noun, adjective;
  method, object, attribute
- Interactive software instead of a static picture

Further Work

Nested line diagrams

- Several nesting levels
- Methods of attribute partition
  Word types: verb, noun, adjective;
  method, object, attribute
- Interactive software instead of a static picture

Further Work

Nested line diagrams

- Several nesting levels
- Methods of attribute partition
  Word types: verb, noun, adjective;
  method, object, attribute
- Interactive software instead of a static picture

Further Work

Modeling dynamics

- Static approach: comparing taxonomies that correspond to different time periods
- Dynamic approach: each elementary change in the data base should be mapped to a modification of a change in community representation

Further Work

Modeling dynamics

- Static approach: comparing taxonomies that correspond to different time periods
- Dynamic approach: each elementary change in the data base should be mapped to a modification of a change in community representation

Further Work

Better linguistic processing

- Dealing with homonyms and synonyms
- Taking into account domain-specific associations between words
- Taking context into consideration
- Using phrases instead of words

Further Work

Other data

- French political blogs
  - words used in blogs
  - which bloggers list which bloggers as their favorites
  - which bloggers cite which bloggers in their posts
  - which bloggers comment which bloggers