Reducing the Representation Complexity of Lattice-Based Taxonomies

Sergei Kuznetsov[1], Sergei Obiedkov[1,2], and Camille Roth[3,4]

[1] Department of Applied Mathematics, Higher School of Economics, Moscow, Russia
[2] Moscow Institute of Physics and Technology, Moscow, Russia
[3] European Center for Living Technology, Venice, Italy
[4] Department of Sociology, University of Surrey, Guildford, UK
skuznetsov@yandex.ru, sergei.obj@gmail.com, camille.roth@polytechnique.edu

Abstract. Representing concept lattices constructed from large contexts often results in heavy, complex diagrams that can be impractical to handle and, eventually, to make sense of. In this respect, many concepts could allegedly be dropped from the lattice without impairing its relevance towards a taxonomy description task at a certain level of detail. We propose a method where the notion of stability is introduced to select potentially more pertinent concepts. We present some theoretical properties of stability and discuss several use cases where taxonomy building is an issue.

1 Introduction

Formal Concept Analysis (FCA) is generally an appropriate framework for building categories defined as object sets sharing some attributes, irrespectively of a particular domain of application. In this framework, categories are called "formal concepts" each concept being a pair of an object set and an attribute set such that every attribute holds for every object. This presents a convincing formal model of the philosophical notion of a "concept" characterized extensionally by the set of entities it covers and intensionally by the set of properties they have in common [1]. Formal concepts, in turn, can be gathered in a lattice structure, thus providing an overlapping taxonomy for the underlying categories.

Besides, traditional lattice operations translate properly in taxonomical and categorical terms: on the one hand, the meet of two categories is a sub-category holding objects belonging to both categories, along with their associated shared attributes; on the other hand, the join of two categories is the super-category defined by attributes shared by both categories, and the associated objects.

While formal concept lattices are theoretically robust, in practice, one often has to face huge structures containing a prohibitive number of categories, even for rather small datasets. Even if navigation in the structure is possible despite large sizes [2], readability generally remains a problem as, computational issues set apart, "even carefully constructed line diagrams lose their readability from a certain size up" [3] (p. 75).

U. Priss, S. Polovina, and R. Hill (Eds.): ICCS 2007, LNAI 4604, pp. 241–254, 2007.

Solutions may consist in representing only the most meaningful portions of the lattice by assuming that some concepts are likely to be less relevant than others from the standpoint of taxonomy description [4,5,6]. To this end, we need to filter out nodes that do not satisfy specified constraints of a certain kind. In this paper, we develop one such pruning technique. In particular, the notion of stability introduced in [7,8] to discriminate irrelevant nodes seems to be particularly appropriate and was fruitfully used in a previous attempt to prune concept lattices in the practical case of epistemic community representation [5].

Here, we apply the method to larger datasets and other domains—other kinds of epistemic communities, but also other kinds of data—as well as address the dynamic description of the resulting reduced structures. While stability was satisfactorily applied to a small sub-context consisting of agents using particular notions, thus yielding meaningful taxonomies, it was unclear whether it could be possible to go further in other domains and with much larger contexts.

2 Formal Framework

Before proceeding, we briefly recall the FCA terminology [3]. Given a *(formal) context* $\mathbb{K} = (G, M, I)$, where G is called a set of *objects*, M is called a set of *attributes*, and the binary relation $I \subseteq G \times M$ specifies which objects have which attributes, the derivation operators $(\cdot)^I$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A^I = \{m \in M \mid \forall g \in A : gIm\};$$

$$B^I = \{g \in G \mid \forall m \in B : gIm\}.$$

Put differently, A^I is the set of attributes common to all objects of A and B^I is the set of objects sharing all attributes of B.

If this does not result in ambiguity, $(\cdot)'$ is used instead of $(\cdot)^I$. The double application of $(\cdot)'$ is a closure operator, i.e., $(\cdot)''$ is extensive, idempotent, and monotonous. Therefore, sets A'' and B'' are said to be *closed.*

A *(formal) concept* of the context (G, M, I) is a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A = B'$, and $B = A'$. In this case, we also have $A = A''$ and $B = B''$. The set A is called the *extent* and B is called the *intent* of the concept (A, B). In categorical terms, (A, B) is equivalently defined by its objects A or its attributes B.

A concept (A, B) is a *subconcept* of (C, D) if $A \subseteq C$ (equivalently, $D \subseteq B$). In this case, (C, D) is called a *superconcept* of (A, B). We write $(A, B) \leq (C, D)$ and define the relations $\geq$, $<$, and $>$ as usual. If $(A, B) < (C, D)$ and there is no (E, F) such that $(A, B) < (E, F) < (C, D)$, then (A, B) is a *lower neighbor* of (C, D) and (C, D) is an *upper neighbor* of (A, B); notation: $(A, B) \prec (C, D)$ and $(C, D) \succ (A, B)$.

The set of all concepts ordered by $\leq$ forms a lattice, which is denoted by $\underline{\mathfrak{B}}(\mathbb{K})$ and called the *concept lattice* of the context $\mathbb{K}$. The relation $\prec$ defines edges in the *covering graph* of $\underline{\mathfrak{B}}(\mathbb{K})$.

3 Stability

3.1 Rationale

An obvious solution to reducing the number of groups of individuals defined as concept extents by selecting “most interesting groups” is to compute only an upper part of the concept lattice: concepts with extents comprising at least $n\%$ of all objects. This approach produces an order filter of a concept lattice often called nowadays an “iceberg lattice”. There are several well-known top-down lattice construction algorithms (see a review in [9]) and algorithms for computing frequent itemsets [4,10] suitable for building such iceberg lattices. The reduction in the number of concepts, as compared to the number of concepts in the whole lattice, can be considerable. However, one should be careful not to overlook small but interesting groups, for example, “exotic” or “emergent” groups not yet represented by a large number of objects, or, groups that contain objects who are not members of any other group.

Undoubtedly, the size of the concept lattice is not only a computational problem. The lattice may contain nodes that are just too similar to each other because of noise in data or real minor differences yet irrelevant to a given purpose. In this case, taking an upper part of the lattice does not solve the problem, since this part may well contain such similar nodes.

To tackle the problem of selecting “meaningful” concept intents, the notion of concept stability was proposed in [7,8] and developed in [5]. The general idea of stability is as follows: A concept is stable if its intent does not depend much on each particular object of the extent.

3.2 Definition

In this section, we define the notion of stability of a formal concept introduced in [7] and [8] in a slightly different form than the one we use here. The definition given below is the one from [5].

Definition 1. *Let $\mathbb{K} = (G, M, I)$ be a formal context and (A, B) be a formal concept of $\mathbb{K}$. The* stability index, *σ, of (A, B) is defined as follows:*

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

In [5], it is shown that the following proposition holds:

Proposition 1. *Let $\mathbb{K} = (G, M, I)$ be a formal context and (A, B) be a formal concept of $\mathbb{K}$. For a set $H \subseteq G$, let $I_H = I \cap (H \times M)$ and $\mathbb{K}_H = (H, M, I_H)$. Then,*

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

Thus, the stability index of a concept is the probability of the intent B if all subcontexts of $\mathbb{K}$ over the attribute set M are equally probable. Stability indicates how much the concept intent depends on particular objects of the extent: a stable intent is less sensitive to noise in object descriptions. Besides, the extent of a stable concept is not “very close” to extents of its lower neighbors.

3.3 Properties

In [5], an algorithm for the computation of stability indices is given. In general, computing stability is a #P-complete problem [8]; hence, simple heuristics (easily computable sufficient conditions) to discard concepts with low stability would be useful. The following two propositions give conditions of this sort.

Proposition 2. *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 $\sigma(A, B) \leq 1 - 1/2^{|A \setminus A_1|}$.*

Proof. Since $A_1 \subset A$, $A' = B$, and $A_1' \neq B$, we have $B \subset A_1'$ and $B \subset A_2'$ for all $A_2 \subseteq A_1$. Therefore, $|\{A_2 \subseteq A \mid A_2' = B\}| \leq 2^{|A|} - |\{A_2 \mid A_2 \subseteq A_1\}| = 2^{|A|} - 2^{|A_1|} = 2^{|A \setminus A_1|}$ and $\sigma(A, B) \leq \frac{2^{|A|} - 2^{|A_1|}}{2^{|A|}} = 1 - 1/2^{|A \setminus A_1|}$. □

In particular, if $|A_1| = |A| - 1$, one has $\sigma(A, B) \leq 1/2$ and the concept (A, B) has fairly low stability: usually, one retains concepts with stability very close to 1. Testing if $A_1' \neq B$ for some A_1 with $|A_1| = |A| - 1$ takes $O(|A|^2 \cdot |M|)$ time.

Proposition 3. *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 $\sigma(A, B) \leq 1 - \frac{3}{2^{|A \setminus A_1| + 1}}$.*

Proof. By the condition, $A_1 \subset A$, $A_2 \subset A$, $A' = B$, and $A_1' \neq B$, $A_2' \neq B$. To obtain an upper bound of $\sigma(A, B)$ we need to consider the situation where the size of the set $\mathfrak{P}(A_1) \cup \mathfrak{P}(A_2)$ (by $\mathfrak{P}(X)$ we denote the powerset of X) is as small as possible. This is attained when A_1 and A_2 are as close as possible, i.e., $|A_1 \setminus A_2| = |A_2 \setminus A_1| = 1$. In this case (since $|A_1| = |A_2|$) we have $|\mathfrak{P}(A_1) \cup \mathfrak{P}(A_2)| = 2^{|A_1|} + 1/2 \cdot 2^{|A_1|} = 2^{|A_1|} \cdot 3/2$. Therefore, $\sigma(A, B) \leq \frac{2^{|A|} - 2^{|A_1|} \cdot 3/2}{2^{|A|}} = 1 - \frac{3}{2^{|A \setminus A_1| + 1}}$. □

In the same way one may obtain a condition for three, four, and so on extent subsets that do not give rise to intent B. However, it seems hard to get a useful general statement for an arbitrary number of such extent subsets, which is related to the #P-completeness of computing stability.

Now we describe how stability of concepts changes with the growth of the data sample. Consider the situation when a context (G, M, I) is updated with a new object g to form context $(G \cup \{g\}, M, J)$ such that $(G \times M) \cap J = I$. Then, according to [11], we distinguish three possible types of concepts of the context $(G \cup \{g\}, M, J)$: an *old* concept that is equal to a concept in the old context, a *modified* concept of the form $(A \cup \{g\}, B)$ such that (A, B) is the concept of (G, M, I), and a *new* concept of the form $((A \cup \{g\})'', B \cap \{g\}')$ such that (A, B) is a concept of (G, M, I) and $B \cap \{g\}'$ is not an intent of (G, M, I). We denote stabilities in contexts (G, M, I) and $(G \cup \{g\}, M, J)$ by σ_I and σ_J, respectively.

Proposition 4. *Given a concept (A, B) of a context (G, M, I), if a new object g is added to the set of objects to form context $(G \cup \{g\}, M, J)$ (such that $(G \times M) \cap J = I$), then for the stability of concepts of the new context $(G \cup \{g\}, M, J)$ there can be the following three possibilities:*

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*

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

Proof. (1) The extents of an old concept and all its subconcepts do not change, neither does the stability.

(2) First, we prove that $\sigma_I(A,B) \leq \sigma_J(A \cup \{g\}, B)$. Indeed, by definition of $\sigma_I(A,B)$, there are $\sigma_I(A,B) \cdot 2^{|A|}$ subsets $A_1 \subseteq A$ ("old" subsets) such that $A_1' = B$, and, since $g' \cap A_1' = B$ for every A_1 such that $A_1' = B$, we also have $\sigma_I(A,B) \cdot 2^{|A|}$ subsets A_2 such that $g \in A_2 \subseteq A \cup \{g\}$ and $A_2' = B$. Since all such A_1 and A_2 are different, we have

$$\sigma_J(A \cup \{g\}, B) \geq (2^{|A|} \cdot \sigma_I(A,B) + 2^{|A|} \cdot \sigma_I(A,B))/2^{|A|+1} = \sigma_I(A,B).$$

To prove $\sigma_J(A \cup \{g\}, B) \leq 1/2 + \sigma_I(A,B)/2$, note that the largest stability of a modified concept $(A \cup \{g\}, B)$ is attained when $A_2' \cap \{g\}' = B$ for each of the $2^{|A|}$ subsets $A_2 \subseteq A$. Taking into account $2^{|A|} \cdot \sigma_I(A,B)$ subsets $A_1 \subseteq A$ with $A_1' = B$, we have

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

(3) In the case of a new concept (A,B), we can have $A_1' = B$ for $A_1 \subseteq A$ only if $g \in A_1$. The largest stability will be attained if $A_1' = B$ for all such $A_1 \subseteq A$ with $g \in A_1$. Formally,

$$\sigma_J(A,B) = \frac{|\{A_1 \subseteq A \mid A_1' = B\}|}{|2^A|} \leq \frac{|\{A_1 \subseteq A \mid g \in A_1\}|}{|2^A|} = \frac{|2^{|A|-1}|}{2^{|A|}} = \frac{1}{2}.$$

It is easy to see that the equality $\sigma_J(A,B) = 1/2$ holds only for the object concept of g, and only this concept has stability $1/2$. All other new concepts are less stable. □

4 Applying Intensional Stability

As such, stability measures how much a community depends on some of its individual members. This may be useful in building attribute-based taxonomies representing intensional categories. In particular, this notion is likely to be relevant when investigating taxonomies of epistemic communities, i.e., groups of agents jointly interested in identical topics, sharing the same notions [12,6]. In this respect, contexts where scientists are objects and the topics on which they work are attributes are particularly adequate: here, formal concepts represent epistemic communities as groups of topics along with corresponding agents. Removing a few scientists from the context should not change the topics of an epistemic community—“real” epistemic communities ought to be stable in spite of noisy data. Apart from noise-resistance, a stable field does not collapse (e.g., merge with a different field or split into several independent subfields) when a few members stop being active or switch to another topic.

We illustrate this criterion using two case studies featuring scientists attending a particular conference and biologists working on a particular model animal. In both cases, while stability-based lattice reduction is significant—from thousands of concepts to less than 30—we are still able to tell a meaningful “story” with respect to what field experts may describe.

4.1 “European Complex Systems Conference”

Using the database of all papers submitted to the second European Conference on Complex Systems in 2006[1], we build a context made of authors and terms mentioned in article titles and abstracts. The resulting context contains 401 authors and 109 terms, which yields a lattice of 6011 concepts. The reduced substructure featuring the 25 most stable concepts is presented in Fig. 1. Note that the set of all stable concepts (for an arbitrary threshold) does not have to be a lattice, even if it is in the examples used in this paper.

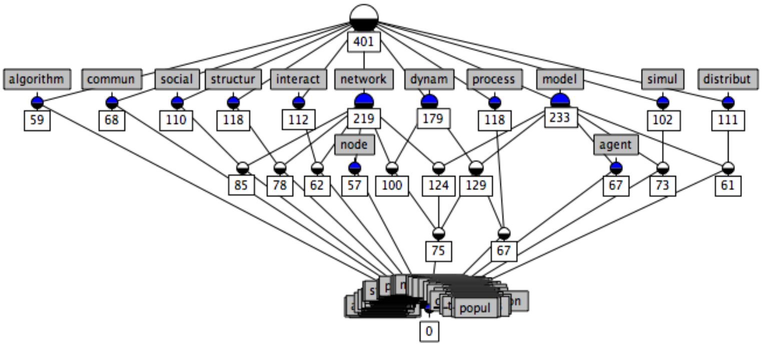

Fig. 1. The 25 most stable concepts in the ECCS dataset. Figures in squares show the sizes of concept extents.

From this lattice, it is possible to provide the following desription of the community attending the ECCS:

- The notion of “network” is obviously a central issue: in addition to being a large community, it is also a parent for several associated subtopics: “social network” (agent-based networks), “structure network” (topological issues),

[1] http://complexsystems.lri.fr/Portal/tiki-index.php?page=ECCS’06

"interact network" (networks as representation of interactions), "node network" (a node being a basic unit), "dynamics network" (evolution of networks) and "model network" (modeling of networks).

- "model" is an important topic too and is related to "agents" and "simulation", as well as "dynamics" (dynamical models in general), in addition to "networks"—it is also worth noting that there exists a sizeable community around "network dynamics model" which refers to scientists interested in the modeling of network dynamics (morphogenesis). The use of models to reconstruct distributions of any kind is represented by the "model distribut" community. Finally, the modeling of dynamical processes ("model dynam process"), although sensibly less significant, is an interesting field as well in this framework.
- Some topics are more isolated as they do not form any joint epistemic community in the stabilized lattice—such as "algorithm" and "community" (to the left). These concepts are likely to refer to minor fields focused on particular issues: community and cluster detection, or introduction and use of novel and general algorithms to achieve empirical measurements in a variety of cases.

In this lattice of 25 concepts it is not possible to see some fields which are actually representative of minor yet active subcommunities — such as "network distrib", "algorithm network" and "algorithm model", which are respectively the 40th, 50th and 75th most stable concepts. Nonetheless, on the whole and given a certain (high) level of epistemological description, the above story appears to be fairly consistent with what experts of the field would perceive as the main topics of complex systems science at that time.

4.2 Embryologists Working on the "zebrafish"

In [5], we have applied stability-based pruning to data obtained from the bibliographical database of `MedLine` abstracts coming from a well-bounded community of embryologists working on the zebrafish during the period 1998–2003, the goal of the application being to build a taxonomy of this research field.[2] Since the purpose of that paper was to illustrate the proposed technique, we used a small random sample context consisting of 25 authors and 18 words. The incidence relation of the context indicated which authors used which words in their papers on the subject. The lattice of this context consisted of 69 concepts, of which we selected the 17 most stable ones (with stability ≥ 0.52). They constitute a lattice shown in Fig. 2 (some of the 18 attributes are not contained in any stable intent except for the intent of the bottom concept; they are not shown on the diagram).

Taking a larger data sample, 250 authors using the same 18 words, we get 1146 concepts. Of course, in the larger structure, stability indices are also larger

[2] Data is obtained from a query on article abstracts containing the term "*zebrafish*" at http://www.pubmed.com.

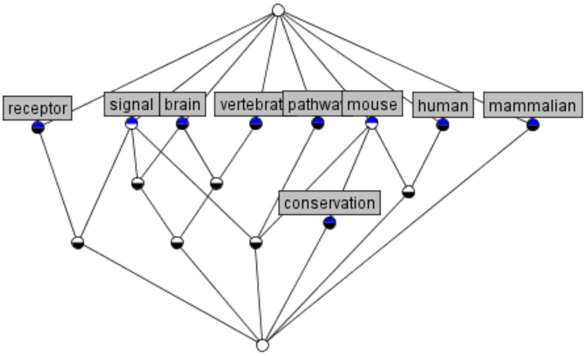

Fig. 2. The lattice of the 17 most stable concepts of a context built from 25 zebrafish researchers and 18 words they used in their papers (taken from [5])

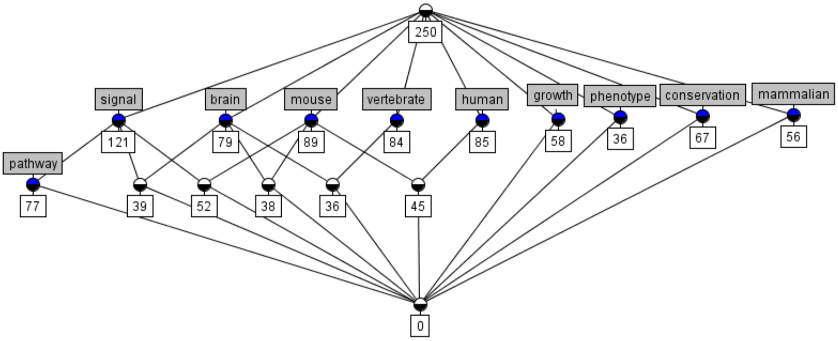

Fig. 3. The lattice of the 17 most stable concepts of a context built from 250 zebrafish researchers and 18 words they used in their papers

(see Section 3), which makes it impractical to use the same stability thresholds for pruning in both cases. Hence, we simply take the same number of the most stable concepts. Fig. 3 shows the lattice formed by the 17 most stable concepts of this context (figures in boxes indicate the extent size of corresponding concepts).

As should be expected, the lattices in Figs. 2 and 3 are not identical, but still share a lot of features in common. One interesting difference is that, in the structure based on the larger data sample, "pathway" occurs as a subconcept of "signal", which certainly makes sense from the domain point of view ("pathway" on its own is still a concept intent, but it is not sufficiently stable in the larger context). Some less important communities, like "mouse, conservation" or "signal, pathway, mouse" are missing from Fig. 3. Instead, the taxonomy resulting from a larger number of authors focuses on more solid associations ignoring some particularities, which can be reintroduced by increasing the number of stable concepts included in the taxonomy.

4.3 Improving the Quality of Taxonomies: Linguistic Processing

In our analysis of the ECCS and zebrafish data, we described authors in terms of words they used in their papers. We removed stop words (such as “and” or “was”), common words with no special meaning for the domain (such as ‘’size” or “function”), as well as “paradigmatic” words, i.e., those relevant to all members of the entire community even if they are not explicitly used by all members (the obvious examples are “complex” and “system”). The remaining words were stemmed using the Porter algorithm [13].

In the process, it has become clear that these techniques are certainly not sufficient: the resulting author–word tables contain a lot of noise coming, in particular, from homonymy and synonymy. We approached the latter by manually combining synonyms—or other semantically related words that could be considered equivalent for our purposes—into one attribute. Of course, this requires some expert knowledge of the domain and cannot be done simply using a general-purpose English thesaurus: words that are synonymous in everyday language can be used differently in the domain to be described or, on the contrary, there may be domain-specific associations between otherwise unrelated words. Homonymy is even more difficult to deal with: words used on their own, without taking the context into consideration, are not very informative; it seems more appropriate to use word phrases.

That we still get rather meaningful taxonomies from formal contexts obtained with such poor means suggests that the methods we use further on may, in general, be valid, but we believe that better linguistic preprocessing will have a significant effect on the quality of the resulting taxonomies.

5 Applying Extensional Stability

The stability index discussed so far relates actually to intensional stability. In a dual manner, it is possible to define an *extensional stability index*, which indicates how a concept extent depends on particular attributes: would the objects of a given concept still belong to the same category if they stop sharing some attributes? A stable extent is thus likely to indicate a group of objects which do not depend on particular attributes.

Definition 2. *The* extensional stability index σ^e *of a concept* (A, B) *is defined as follows:*

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

Like intensional stability, the relevance of this index depends on the domain and the aim of the lattice-based taxonomy. For instance, affiliation data, in social science, defines people related to some organizations or events; in this case, formal concept lattices represent taxonomies of agents who share identical affiliations. Extensional stability may be helpful in this situation in measuring how durable links between people within a community are. In this respect, it principally relates to the social aspect of the group: if some people are together because they have a given activity, one may wonder whether they will still be together if they stop doing this activity. People who are also doing something else together are more likely to belong to stable extents. Here, extensional stability tests how much the community as a group of people depends on particular activities. In other words, if one of the activities that unites them becomes less appropriate, will they still survive as a separate community?

To illustrate this, we focus on data stemming from a celebrated case study by Davis, Gardner and Gardner (DGG) [14] which features ladies attending particular events in a small Mississippi town in the 1930s. Using a context where objects are people and attributes are attendance to social events, it is possible to build a concept lattice representing groups of women attending jointly some sets of events [15]. However, even in this simple case the resulting lattice is already rather sizeable with 65 concepts; finding cohesive subgroups in such a structure could be uneasy. By contrast, the lattice corresponding to extensionally stable concepts (stability index strictly above .5) contains only three concepts, in addition to top and bottom nodes: their extents are {g14}, {g12, g13, g14} and {g1, g3}. The stabilized lattice is shown on Fig. 4.

The identification, in this data, of subcommunities together with core and peripheral members has already been the focus of several studies in social science. While interview-based identification in the original DGG study suggests that {g1, g2, g3, g4} and {g13, g14, g15} are respectively core members of two distinct groups, a comprehensive review given in [16] reveals a collection of remarkably diverse results, depending on whether subgroups were identified using, *inter alia*, principal component analysis, matrix algebra, information theory, as well as concept lattices—by means, in this latter case, of a relatively manual approach [15].

Most interestingly, a study by Doreian [17] agrees particularly well with our results: it yields the same core members as those found in our stabilized lattice, i.e., {g1, g3} and {g12, g13, g14}. His approach relies on Q-analysis [18], whose principles are unsurprisingly analogous to FCA: for a given context, each object

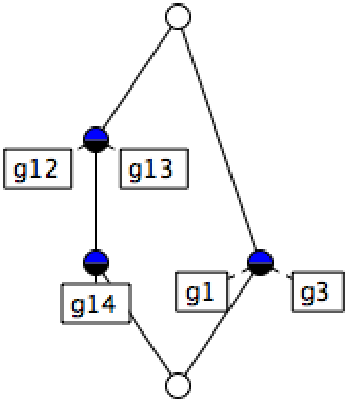

Fig. 4. The lattice of concepts with extensional stability above .5 for the DGG data

is defined as a “polyhedron” where attributes are edges. In this framework, a formal concept can thus be seen as an intersection of polyhedra—extents and intents are respectively defined by polyhedra-objects and edges-attributes participating in a given intersection [19]. Additionally, Q-analysis introduces the notion of connected paths between polyhedra, which are plausibly useful for dealing with connection patterns between objects, yet irrespective of the actual attributes underlying these connected paths. As Freeman underlines, “by considering subsets of women who were connected at higher levels, Doreian was able to specify degrees of co-attendance ranging from the core to the periphery according of each group” [16]. The idea of strong relationships between objects independently of particular attributes is not dissimilar to our notion of extensional stability and could perhaps account for our identical results.

More broadly, while extensional stability appears to yield a satisfying outcome in this small case study, it is nonetheless the matter of further research to check its adequacy on larger datasets and different domains of application.

6 Dynamic Mappings

Let $\mathbb{K}_1 = (G, M, I)$ and $\mathbb{K}_2 = (H, N, J)$ be two contexts describing the same domain in two different time points (or periods). How has the domain changed between these time points? In particular, if $(A, B) \in \underline{\mathfrak{B}}(\mathbb{K}_1)$ is a concept of $\mathbb{K}_1$, what has happened to it in $\mathbb{K}_2$?

Consider a concept $(C, D) \in \underline{\mathfrak{B}}(\mathbb{K}_2)$. If the closure of $B \cap D$ equals B in $\mathbb{K}_1$ and D in $\mathbb{K}_2$, we may say that (A, B) and (C, D) are *intensionally related.* In the case of the ECCS data, concepts intensionally related to (A, B) represent the evolution of the field B between the two periods.

Figure 5 shows two diagrams corresponding to the ECCS conferences in 2005 and 2006 (assuming that the words are the same, i.e., $M = N$). In both cases, we have selected the 15 most stable concepts. The differences are as follows: the diagram for 2005 contains concepts with intents {network, dynamics}, {dynamics, model, process}, {dynamics, process}, and {information}—all missing from the 2006 diagram, which contains its own unique intents: {interaction}, {network, social}, {model, agent}, and {simulation, model}. The only 2006 concept intensionally related to {network, dynamics} is the one with intent {network, dynamics, model}. This suggests that the 2005 topic described by {network, dynamics} has merged with the topic described by {network, dynamics, model}; at least, the difference between the two is no longer important at the given level of detail. The other three 2005-specific communities are intensionally related only to the bottom node of the 2006 diagram, which means that they have disappeared or become less important. On the other hand, the bottom node of the 2005 diagram is intensionally related to the four 2006-specific concepts, suggesting that they correspond to new subareas of research.[3] Even though {model, agent} has

[3] Again, as we deal with “stabilized” lattices, these new areas are such only at the chosen level of detail. It would be more accurate to say that their importance has increased.

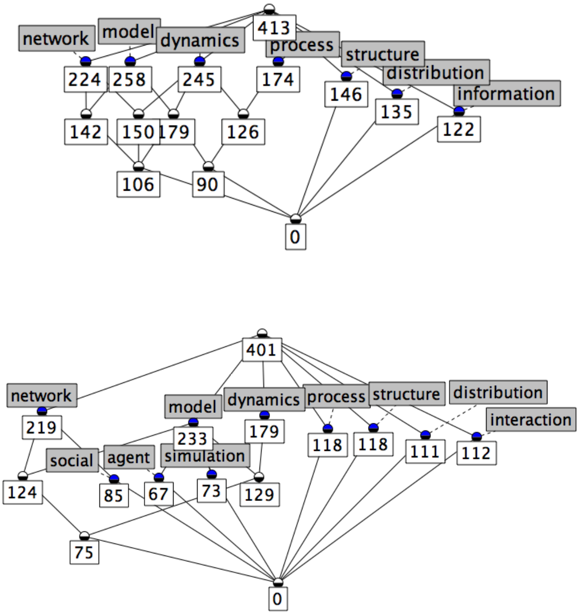

Fig. 5. Stabilized 15 concept lattice for ECCS 2005 and 2006

a parent topic, {model}, already present in the 2005 taxonomy, the "agent" aspect is new, thus, defining a new subfield, that has no corresponding nodes in 2005, which is indicated by the fact that it is intensionally related to the 2005 bottom node.

In the above discussion, the social aspect of the communities has been completely ignored. In some contexts, it is more appropriate to describe the history of a community in terms of what happens to its members. In this case, if all authors dealing with topic A in 2005 switch to topic B in 2006, B should be considered as the 2006 equivalent of the 2005 A-community, even if A is still an active topic in 2006 (supported by newcomers, for example). Such population moves can be captured by *extensional relations* between nodes defined dually to the intensional relations.

It is worth noting that extensional and intensional relations defined in this section originate from the mappings in nested line diagrams [3]. In the case of intensional relations, we assume that that $G \cap H = \emptyset$ (if this is not so, we can always time-tag the objects) and define a context $\mathbb{K}_3 = (G \cup H, M \cup N, I \cup J)$. If a nested line diagram of $\mathbb{K}_3$ is constructed so that G is used as the object set for the outer diagram and H is used as the object set for the inner diagram, then the nodes intensionally related to an outer node are the "realized" nodes of the inner diagram inside this outer node.

Another approach to dynamic mappings could be based on the theory of multicontexts [20], which however has to be adapted for our reduced lattice-based structures.

7 Conclusion

We extend a previous approach based on the notion of stability to build arbitrarily small concept lattices from sizeable contexts. After presenting theoretical properties of stability, introducing in particular several propositions useful for incremental computation of stability in evolving contexts, we distinguish intensional stability from extensional stability and illustrate them through selected case studies, where one or the other could be suitable. In particular, intensional stability appears to be useful for epistemic community taxonomy building, while extensional stability seems to be more effective for finding cohesive subgroups in communities of agents involved in common activities. These examples also demonstrate how different expectations regarding what makes a formal concept relevant for a given taxonomical description task may call for distinct usages of stability, extensional or intensional, which admittedly might not apply in all domains. We have also shown how it is possible to track taxonomy evolution using dynamic mappings between stability-reduced lattices.

Acknowledgements. The authors wish to thank Bertrand Chardon and Marc Schoenauer for providing us with the ECCS 2005 and 2006 data, as well as Nadine Peyriéras for her kind feedback regarding the zebrafish community. Diagrams have been produced with ConExp (http://sourceforge.net/projects/conexp), whose primary developer is Serhiy Yevtushenko. The first author was supported by the project COMO (Concepts and Models) of Deutsche Forschungsgemeinschaft (DFG) and Russian Foundation for Basic Research (RFBR).

References

1. Wille, R.: Concept lattices and conceptual knowledge systems. Computers & Mathematics with Applications 23, 493–515 (1992)
2. Ferré, S., Ridoux, O.: A file system based on concept analysis. In: Palamidessi, C., Moniz Pereira, L., Lloyd, J.W., Dahl, V., Furbach, U., Kerber, M., Lau, K.-K., Sagiv, Y., Stuckey, P.J. (eds.) CL 2000. LNCS (LNAI), vol. 1861, pp. 1033–1047. Springer, Heidelberg (2000)
3. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg (1999)
4. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC. Data & Knowledge Engineering 42, 189–222 (2002)
5. Roth, C., Obiedkov, S., Kourie, D.G.: Towards concise representation for taxonomies of epistemic communities. In: Ben Yahia, S., Mephu Nguifo, E. (eds.) CLA 4th International Conference on Concept Lattices and their Applications, Tunis, Faculté des Sciences de Tunis, pp. 205–218 (2006)
6. Roth, C., Bourgine, P.: Lattice-based dynamic and overlapping taxonomies: The case of epistemic communities. Scientometrics 69(2), 429–447 (2006)
7. Kuznetsov, S.O.: Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf. Ser.2 (Automat. Document. Math. Linguist.) 12, 21–29 (1990)

8. Kuznetsov, S.O.: On stability of a formal concept. In: SanJuan, E. (ed.) JIM, Metz, France (2003)
9. Kuznetsov, S.O., Obiedkov, S.: Comparing performance of algorithms for generating concept lattices. J. Expt. Theor. Artif. Intell. 14(2/3), 189–216 (2002)
10. Bayardo, Jr., R., Goethals, B., Zaki, M. (eds.) In: Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI 2004), CEUR-WS.org (2004)
11. Godin, R., Missaoui, R., Allaoui, H.: Incremental concept formation algorithms based on Galois lattices. Computational Intelligence 11(2), 246–267 (1995)
12. Haas, P.: Introduction: epistemic communities and international policy coordination. International Organization 46(1), 1–35 (1992)
13. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
14. Davis, A., Gardner, B.B., Gardner, M.R.: Deep South. University of Chicago Press, Chicago (1941)
15. Freeman, L.C., White, D.R.: Using Galois lattices to represent network data. Sociological Methodology 23, 127–146 (1993)
16. Freeman, L.: Finding social groups: A meta-analysis of the southern women data. In: Breiger, R., Carley, K., Pattison, P. (eds.) Dynamic Social Network Modeling and Analysis, pp. 39–97. National Academies Press, Washington, D.C. (2003)
17. Doreian, P.: On the delineation of small group structure. In: Hudson, H.C. (ed.) Classifying Social Data, pp. 215–230. Jossey-Bass, San Francisco (1979)
18. Atkin, R.: Mathematical Structure in Human Affairs. Heinemann Educational Books, London (1974)
19. Johnson, J.H.: Stars, maximal rectangles, lattices: A new perspective on Q-analysis. International Journal of Man-Machine Studies 24(3), 293–299 (1986)
20. Wille, R.: Conceptual structures of multicontexts. In: Eklund, P.W., Mann, G.A., Ellis, G. (eds.) ICCS 1996. LNCS, vol. 1115, pp. 23–29. Springer, Heidelberg (1996)