FCA in knowledge technologies: experiences and opportunities

Yannis Kalfoglou[1] Srinandan Dasmahapatra[1] Yun-Heh Chen-Burger[2]
y.kalfoglou@ecs.soton.ac.uk sd@ecs.soton.ac.uk jessica@inf.ed.ac.uk

[1] Advanced Knowledge Technologies (AKT), School of Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, UK.
[2] Advanced Knowledge Technologies (AKT), Artificial Intelligence Applications Institute,
School of Informatics, University of Edinburgh, Edinburgh EH8 9LE, UK.

Abstract. Managing knowledge is a difficult and slippery enterprise. A wide variety of technologies have to be invoked in providing support for knowledge requirements, ranging from the acquisition, modelling, maintenance, retrieval, reuse and publishing of knowledge. Any toolset capable of providing support for these would be valuable as its effects would percolate down to all the application domains structured around the domain representation. Given the generic structure of the lattice building algorithms in Formal Concept Analysis, we undertook a set of experiments to examine its potential utility in knowledge technologies. We elaborate on our experiences and speculate on the opportunities lying ahead for a larger uptake of Formal Concept Analysis approaches.

1 Introduction

A distinguishing feature of much of the knowledge technologies today is the attention paid to representations of appropriate domain knowledge as scaffolding around which automated and mixed-initiative processing systems are constructed to provide the required functionality. In particular, with the promise of a Semantic Web a great deal of effort is being directed at providing knowledge-level characterisations of both domains and functionality of processes to achieve inter-operability in an environment as open and distributed as the Web.

These characterisations are significant for representing and modelling domain knowledge in sound and machine-processable manners. Advanced Knowledge Technologies (AKT) [1] is a large interdisciplinary research collaboration between five UK universities working on developing technologies to address these issues. There are a number of technologies used for knowledge-level domain characterisations, namely ontologies, and tools to support their engineering. However, there is little support for the modeller to help in identifying appropriate conceptual structures to capture domain semantics. Formal Concept Analysis (FCA)[8] provides a fertile ground for exploitation with its generic structure of lattice building algorithms to visualize the consequences of partial order that the underlying mathematical theory builds on.

In this paper, we elaborate on our experiences with using FCA in conjunction with knowledge technologies in the context of the AKT project. We also identify joint points where prominent knowledge technologies, like Description Logics for the Semantic Web, could benefit from FCA. We do this in a speculative fashion in the next section, coupled with our example cases to show where possible collaboration could be achieved. We briefly report on similar work in section 3 and conclude the paper in section 4

[1] Accessible online from `www.aktors.org`

2 Experiences and opportunities for FCA

FCA has been applied at various stages of a system's life cycle: for example, in the early stages when analysing a domain for the purpose of building and using a knowledge-rich representation of that domain - like the work of Bain in [3] where FCA was used to assist building an ontology from scratch - or applied at later stages in order to enhance an existing system for the purpose of providing a specific service - like the *CEM* email management system described in [6].

It appears though that is being used selectively and opportunistically. The reason for this scattered area of FCA applications could be the fundamental ingredients of FCA and its underlying philosophy: FCA emerged in the 80s as a practical application of lattice theory. The core modelling ingredients underpinning FCA are *objects* and *attributes*[2] which stem from predicative interpretations of set theory. Thus, for a given object, one performs a "closure" operation to form a set of objects which is the intersection of the extension of the attributes that the object is characterised by. These are defined as the concepts in any particular formal context, with the order ideal (or down set) $\downarrow m$ of any attribute m.

In the AKT project, we experimented with identifying these concepts in a number of scenarios from the scientific knowledge management realm where we confronted with loosely defined objects and attributes. Our aim was to use FCA to help us identify the prominent concepts in the domain at question, and most importantly, provide us with a structured representation which we could use to perform certain knowledge management tasks, such as:

Analysing programme committee memberships: One could assume that programme committee membership for a conference or similar event requires that those on the programme committee (PC) are the current and prominent figures in the field at question. Using this as a working hypothesis, and the year in which they served at a specific PC as temporal marker of recognized prominence, we then applied FCA techniques like concept lattice exploration to visualize the distribution of PC members over a number of years. This could, arguably, give us an idea of how the specific event evolved over a period of time by virtue of the changes (or otherwise) in their PCs.

In our experiments (briefly described online in [11]), the objects were PC members and attributes were EKAW conferences in which these members served. A visual inspection of sort of lattice can reveal trends in how the event has evolved over the years. For example, we can identify people who where in PCs of early EKAWs but are not appearing in more recent EKAWs, whereas others have a sustainable presence in the PCs throughout the whole period of 1994 to 2002. If we correlate this information with information regarding the research interests of the PCs, we could end up with a strong indication of the evolution of research themes for the EKAW conferences. In what we present, we regard the extraction of research interests of these researchers as peripheral to our discussion but we point the interested reader to the work done on identifying communities of practice in [2].

Analysing the evolution of research themes: This analysis can be supported by another lattice which depicts the evolution of research themes in EKAW conferences, based on the designated conference session topics. We depict this lattice in figure 1. From the lattice drawing point of view, we should note that we used a publicly available FCA tool, *ConExp*[3] but we deliberately changed the position of the nodes in the line diagrams produced. We did that to enhance its readability and ease its illustration when

[2] Priss points out in [13] these can be *elements, individuals, tokens, instances, specimens* and *features, characteristics, characters, defining elements*, respectively.

[3] Presented in [17] and can be downloaded from http://sourceforge.net/projects/conexp

depicted on paper as we wanted to include all labels from objects and attributes. That compromised the grid projection property of the diagram without, however, affecting the representation of partial order between nodes.

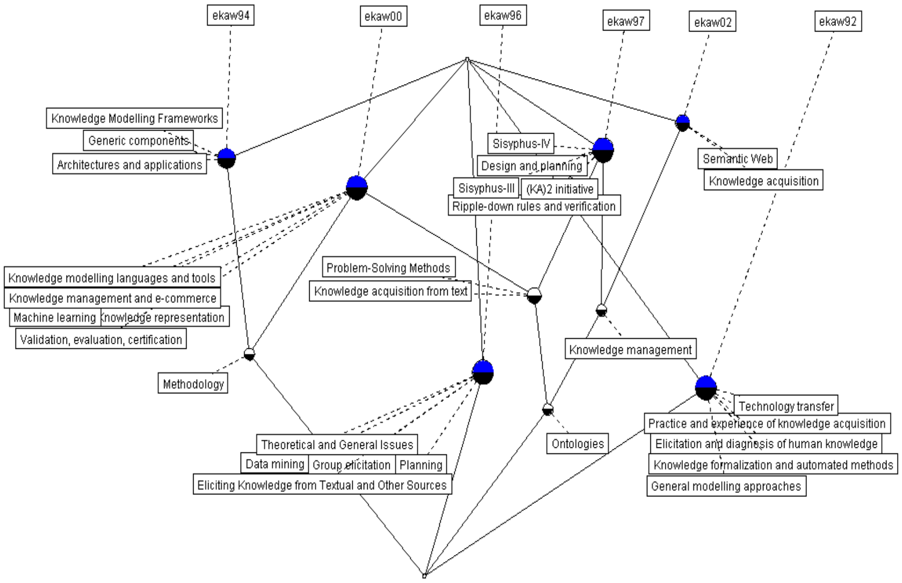

ekaw94
ekaw00
ekaw96
ekaw97
ekaw02
ekaw92
Knowledge Modelling Frameworks
Generic components
Architectures and applications
Sisyphus-IV
Design and planning
Semantic Web
Knowledge acquisition
Sisyphus-III
(KA)2 initiative
Ripple-down rules and verification
Problem-Solving Methods
Knowledge acquisition from text
Knowledge modelling languages and tools
Knowledge management and e-commerce
Machine learning
nowledge representation
Validation, evaluation, certification
Knowledge management
Methodology
Technology transfer
Practice and experience of knowledge acquisition
Elicitation and diagnosis of human knowledge
Knowledge formalization and automated methods
General modelling approaches
Theoretical and General Issues
Ontologies
Data mining
Group elicitation
Planning
Eliciting Knowledge from Textual and Other Sources

Fig. 1. Concept lattice depicting session topics of the EKAW conferences from 1994 to 2002.

Again, a close inspection shows some trends which are evident in today's research agendas in many organisations: *knowledge modelling frameworks* and *generic components* were popular in the early 90s whereas nowadays the research focus is on *semantic web* and *knowledge management*. The inherited taxonomic reasoning of concept lattices can also reveal interesting relationships between research topics, as for instance the subsumption of *ontologies* from *knowledge management*, *knowledge acquisition* and the *semantic web* topics.

Analysing research areas attributed to published papers: We also applied FCA techniques, in particular context reduction algorithms like those described in the FCA textbook (p.27 in [8]) to analyse the formal context of online academic journals. Our aim was to expose relationships between research areas used to classify published papers. The premise of our analysis is the very algorithm that Ganter and Wille describe in [8] for clarifying and reducing formal contexts: "[. . .] we merge objects with the same intents and attributes with the same extents. Then we delete all objects, the intent of which can be represented as the intersection of other object intents, and correspondingly all attributes, the extent of which is the intersection of other attributes extents.". This process, if captured in a step-wise fashion, will expose the objects and attributes that are about to be merged with others, hence allowing us to infer that they are related.

For our data sets, we used a small number of articles from the *ACM Digital Library* portal[4] focusing on the *ACM Intelligence* journal[5]. The formal context consists of 20 objects (articles) and 58 attributes (research areas). The research areas originate from a standard classification system, the *ACM Computing Classification System*[6]. We also used a second data set, the *Data and Knowledge Engineering (DKE)* journal from Elsevier[7]. In this context we had the same articles (objects) as in the ACM context, but this time we classified them against the *DKE*'s own classification system, Elsevier's classification of *DKE* fields[8], which used 27 research areas (attributes) for classifying the aforementioned articles.

For both data sets we chose as objects for their context papers that appeared in the journals. For instance, for the *ACM Intelligence* journal we chose papers that appeared over a period of three years, from 1999 to 2001 and were accessible from the *ACM Digital Library* portal. As there were already classified according to the *ACM Computing Classification System*, we used their classification categories as attributes. We then applied typical context reduction techniques in a step-wise fashion. While we were getting a reduced context, we captured the concepts that are deemed to be interrelated by virtue of having their extents (objects that represent articles in the journal) intersected. For instance, the *ACM classification category H.3.5* on *Web-based services* is the intersection of *H.5* on *Information Interfaces and Presentation* and *H.3* on *Information Storage and Retrieval* by virtue of classifying the same articles. This sort of analysis supports identification of related research areas using as supporting evidence the classification of articles against standardized categories as those originating from the *ACM Computing Classification System*, and the inherited taxonomic reasoning which FCA concept lattices provides to infer their relationship.

Although these experiments were based on loosely defined objects and attributes, it is evident that knowledge representation formalisms also deal with concepts in as much as they underpin systems that require expressive domain characterisation. Much of this effort involves a restricted predicative apparatus in order that the expressivity allowed by appropriate formalisms does not lead to well-known problems of decidability.

To illustrate the point, recall that the experiments we reported above were conducted within the umbrella of the AKT project which has been developing and applying knowledge technologies for the Semantic Web. In the Semantic Web enterprise ontologies have been identified as a key enabling construct, and Web Ontology Language (OWL) is currently going through the final stages of adoption as the W3C standard for the language of choice in which to express ontologies on the Web. OWL inherits characteristics of Description Logics (DLs) which are languages of restricted expressivity designed to facilitate tractable taxonomic reasoning. Some of the simpler and computationally less expensive logics are fragments of first-order predicate logic with two variables, small enough to express binary predicates[4]. A binary predicate can also be described graphically as an edge connecting two nodes each representing a variable, and the knowledge modelling experiments reported below use such a diagrammatic representation.

AKT Research Map: We obtained these kind of diagrams for the same domain as the the one described in the experiments reported above during a knowledge modelling exercise, the construction of the *AKT Research Map*[5], which would capture the expertise and areas of research conducted by AKT members. The specific case was to plan, build, check and publish the *AKT Research Map* and its ontology which is an

[4] Accessible online from http://portal.acm.org

[5] Accessible online from http://www.acm.org/sigart/int/

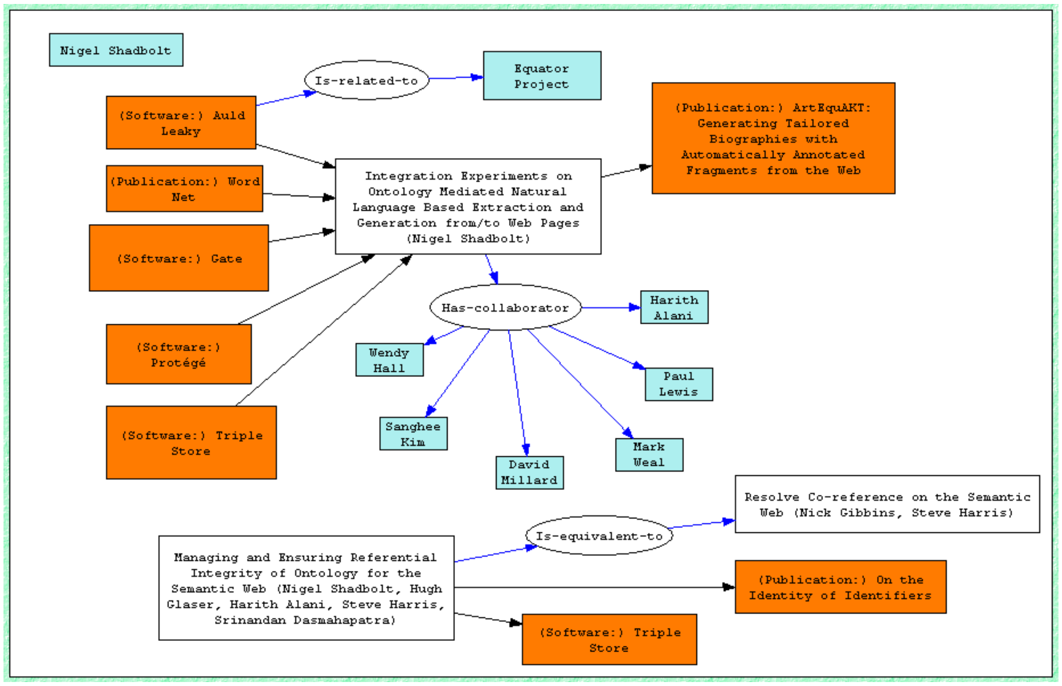
[6] Accessible online from http://www.acm.org/class/1998/

[7] Accessible online from http://www.elsevier.com/locate/issn/0169023X/

[8] Accessible online from http://www.elsevier.com/homepage/sac/datak/dke-classification-2002.pdf

extension of the *AKT Reference* ontology[1]. The motivation of producing such a map is to fill the knowledge gap that will not normally be filled by methods such as automatic information extraction techniques from Web pages. While the Web is very good in advertising the results of research work, such as publications and white papers, it is also poor in publishing the in-depth knowledge about how such research results were produced, what sort of resources were utilized, what contacts, best practices, etc. This information, however, is valuable for other researchers who have similar interests. To address this problem, an *AKT Research Map* has been built to capture such in-depth know-how as well as collaborators, software systems and literature that are contributors to forming the final result.

The *AKT Research Map* is written using a modelling method that is a specialization of Entity Relational Data Modelling method[15]. As a part of building the *AKT Research Map*, members of the AKT project participated in structured interviews to develop their own individual maps during knowledge acquisition (KA) sessions by extending concepts and relations expressed in the OWL-based *AKT Reference* ontology. As an example, figure 2 depicts such an individual map of the AKT member Nigel Shadbolt.



Nigel Shadbolt
(Software:) Auld Leaky
Is-related-to
Equator Project
(Publication:) ArtEquAKT: Generating Tailored Biographies with Automatically Annotated Fragments from the Web
(Publication:) Word Net
Integration Experiments on Ontology Mediated Natural Language Based Extraction and Generation from/to Web Pages (Nigel Shadbolt)
(Software:) Gate
(Software:) Protégé
(Software:) Triple Store
Has-collaborator
Harith Alani
Wendy Hall
Paul Lewis
Sanghee Kim
David Millard
Mark Weal
Resolve Co-reference on the Semantic Web (Nick Gibbins, Steve Harris)
Is-equivalent-to
Managing and Ensuring Referential Integrity of Ontology for the Semantic Web (Nigel Shadbolt, Hugh Glaser, Harith Alani, Steve Harris, Srinandan Dasmahapatra)
(Publication:) On the Identity of Identifiers
(Software:) Triple Store

Fig. 2. The *AKT Research Map* of Nigel Shadbolt.

Between the KA sessions and when a draft of the *AKT Research Map* was finished, iterative verification and validation was carried out to make sure that the map is consistent within itself and consistent with the underpinning ontology. This includes checking of inconsistency of the same information that has been described in different parts of the map, and that a same object has not been represented under different names. It also carries out a pair-wise model checking between the map and the ontology to make sure that the map ontology is complete and the research map is consistent with the ontology. This process could have been performed using FCA lattice drawing facilities had our tools been so integrated.

In making the comparison with FCA, every attribute column in the cross table of the formal context can be viewed predicatively in the sense of DLs or equivalently in terms of the diagrams in the *AKT Research Map* – a directed edge or a binary relation shows up at the corresponding cell marked by the row and column positions. To make the correspondence work, the formulae in the ontology need to be instantiated by *individuals*. Concept lattices built with FCA makes concept dependencies manifest, thus making it an appealing tool for any modeller. For instance, the fragment of the *AKT Research Map* in figure 2 depicts a link with the label "Has-collaborator" a relation that is not present in the underlying *AKT Reference Ontology*. By choosing the set of people as objects and a limited set of attributes – that of authorship of some individual papers, or being investigators of some individual projects – a concept can emerge (after suitably taking intersections of the appropriate sets) which includes Nigel Shadbolt and all of his collaborators. A modeller can choose to extract a relationship based on this derived concept, which can then be defined into the conceptual structure using the syntax of the logical formalism. This requires lifting the propositions represented into a first-order formalism, i.e., introducing a variable in place of the list of instances upon which the modeller makes her decision.

In DLs, concept labels are assigned a formal semantics by their extension, as sets of objects. These can be primitive or defined by conjunction with other concepts. Binary relationships are used with existential or universal quantifiers to define properties these concepts might possess. So, for example, in a domain Δ, concept C would have as the domain of its interpretation function $\mathcal{I} : \cdot \rightarrow \Delta$ the set $C^{\mathcal{I}} \subset \Delta$. Furthermore, predicative properties are defined by role restriction; hence for binary relation R and concept C, we can create $\forall R.C$ or $\exists R.C$ whose extensions are, respectively,

$$(\forall R.C)^{\mathcal{I}} = \{x \in \Delta | \forall y (x,y) \in R^{\mathcal{I}} \Rightarrow y \in C^{\mathcal{I}}\} \quad \text{and}$$
$$(\exists R.C)^{\mathcal{I}} = \{x \in \Delta | \exists y (x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}.$$

Classes are also defined by conjunction and disjunction which facilitates subsumption reasoning of the form $A^{\mathcal{I}} \cap B^{\mathcal{I}} \subseteq A^{\mathcal{I}}$ etc. These enable the modeller to check for consistency of the ontology under development in the cases when concepts are constructed in terms of others.

However, since every attribute is defined *solely* by its extension, concepts generated by join and meet operations on the lattice may not have obvious labels that a domain expert would be comfortable with, a problem associated with any bottom-up approach. Bain[3], notes that "a drawback in FCA is that it does not allow for the introduction of concept names as new intermediate-level terms [. . .] this is necessary for re-use of conceptual structures, so that they may be referred to by name, e.g. to be used in incremental learning or theory revision." while Priss records these as being characterised as "informationless" states[13].

In the DL world, the approach to modelling is much more top-down, and ontologies are often created (as terminologies or T-boxes) with no instance information, despite the semantics being defined extensionally (for example, in A-Boxes). This makes direct comparison of FCA lattices with taxonomic trees rendered in DLs approaches difficult. However, inheritance of properties in a taxonomic hierarchy is possible because the *is-a* relationship (set inclusion) is transitive, and one can identify the down set of an attribute in FCA with the range of the corresponding interpretation function in DLs. This points out the restricted nature of different types of relationships that are represented in FCA. It has been noted in the literature that FCA focusses on hypernym hierarchy of concepts but not arbitrary relations among them[14]. The trees drawn to express taxonomic relationships are, of course, the only ones that DLs formalisms typically generate, even though they were designed to formalise semantic nets and ER diagrams that were described in the *AKT Research Map* above. The work of Hahn *et al* [9] who introduced formal structures to express part-whole relationships as taxonomic ones is an interesting avenue to extend the expressivity of relationships captured in FCA. A detailed study relating these formalisms would be valuable to provide methodological support for knowledge representation.

3 Related work

FCA has been applied in a variety of applications along the whole spectrum of knowledge management with emphasis on the early stages of acquisition, analysis and modelling. For instance, in [12] the authors describe the *KANavigator*, a web-based browsing mechanism for domain-specific document retrieval. Their work is closely related to our experiments on analysing research interests using FCA. However, *KANavigator* is a more generic system aiming not only at navigational aid when browsing a domain but also providing support for incremental development and evolution of the concepts involved. This could be achieved by the open environment in which the *KANavigator* was deployed where users can update the concept lattice with new objects of interest as these emerge. On the other hand, our aim was to simply capture the dependencies between seemingly related research interests as our approach was to perform further analysis and modelling using richer knowledge modelling structures, like ontologies. Identifying related concepts then, was our priority as it was in [10] work where the authors applied FCA as a conceptual clustering tool by using the intentional descriptions of objects in a domain as indicators of cluster membership. As our KA experience script highlighted in section 2 the results of this sort of analysis need to combined with other technologies.

There are few examples of combining FCA with other popular knowledge technologies, one of them is the *Troika* approach described in [7]. The authors combined repertory grids, conceptual graphs and FCA in order to overcome the acknowledged KA bottleneck where there exist a plethora of methods, techniques and tools but none of them is adequate enough to carry out its task elegantly without needing the input of another. The authors argued that these three technologies could be combined in a streamlined fashion, where repertory grids are used for acquisition, FCA for analysis and conceptual graphs for representation. The *Troika* approach is a set of algorithmic steps which dictate the order and the way in which these three technologies can be combined.

There have been attempts in the literature to integrate different approaches in order to support knowledge modelling, as indicated above. Further work has been done by Tilley and colleagues [16] who proposed a solution to provide some more interaction with the modeller in order to question and verify the elements of the formal context by editing, removing or adding new ones. In the domain of software engineering they used FCA to model the class hierarchy in a given domain and compared it with a typical software engineering use-cases based modelling approach. Although identifying initial nouns from the textual specification as candidates for objects in the formal context was a laborious and time consuming exercise, the authors praised the value of approach as it makes clearer and more collaborative the design process.

4 Conclusions

FCA provides a set of tools that allows for formalising a set of informal descriptions of a domain thus providing the basis for ontology building. The model theory of DL formalisms allow for a direct comparison with the sets that form the extents of concepts in FCA. However, the relationship between subsumption hierarchies that are commonplace in ontologies to the partial orders in FCA lattices needs to be explored in detail if FCA techniques are to be extended to assist the knowledge engineer.

Acknowledgements. This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University.

References

1. AKT. AKT Reference Ontology, `http://www.aktors.org/ontology/` 2003.
2. H. Alani, S. Dasmahapatra, K. O'Hara, and N. Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, Mar. 2003.
3. M. Bain. Inductive Construction of Ontologies from Formal Concept Analysis. In *Proceedings of the 11th International Conference on Conceptual Structures (ICCS'03), Springer LNAI 2746, Dresden, Germany*, July 2003.
4. A. Borgida. On the Relative Expressiveness of Description Logics and Predicate Logics. *Artificial Intelligence*, 82(1-2):353–367, 1996.
5. Y.-H. Chen-Burger. AKT Research Map v.2.0, `http://www.aiai.ed.ac.uk/ jessicac/project/akt-map-html/top-level.html` 2003.
6. R. Cole and G. Stumme. CEM - a Conceptual email Manager. In *Proceedings of the 8th International Conference on Conceptual Structures (ICCS'00), Darmstadt, Germany*, Aug. 2000.
7. H. Delugach and B. Lampkin. Troika: Using Grids,Lattices and Graphs in Knowledge Acquisition. In *Proceedings of the 8th International Conference on Conceptual Structures (ICCS'00), Darmstadt, Germany*, Aug. 2000.
8. B. Ganter and R. Wille. *Formal Concept Analysis: mathematical foundations*. Springer, 1999. ISBN: 3-540-62771-5.
9. U. Hahn, S. Schulz, and M. Romacker. Part-Whole Reasoning: A Case Study in Medical Ontology Engineering. *IEEE Intelligent Systems*, 14(5):59–67, Oct. 1999.
10. A. Hotho and G. Stumme. Conceptual Clustering of Text Clusters. In *Proceedings of the Machine Learning Workshop (FGML'02), Hannover, Germany*, Oct. 2002.
11. Y. Kalfoglou. Applications of FCA in AKT, `http://www.aktors.org/technologies/fca` 2003.
12. M. Kim and P. Compton. Incremental Development of Domain-Specific Document Retrieval Systems. In *Proceedings of the 1st International Conference on Knowledge Capture (K-Cap'01), Victoria, BC, Canada*, Oct. 2001.
13. U. Priss. Formalizing Botanical Taxonomies. In *Proceedings of the 11th International Conference on Conceptual Structures (ICCS'03), Springer LNAI 2746, Dresden, Germany*, July 2003.
14. C. Schmitz, S. Staab, R. Studer, and J. Tane. Accessing Distributed Learning Repositories through a Courseware Watchdog. In *Proceedings of the E-Learn 2002 world conference on e-learning in corporate, government, healthcare 4 higher education, Montreal, Canada*, Oct. 2002.
15. B. Thalheim. *Entity Relationship Modeling*. Springer, Dec. 1999. ISBN: 3540-6547-04.
16. T. Tilley. A Software Modelling Exercise using FCA. In *Proceedings of the 11th International Conference on Conceptual Structures (ICCS'03), Springer LNAI 2746, Dresden, Germany*, July 2003.
17. S. Yevtushenko. System of Data Analysis Concept Explorer (in Russian). In *Proceedings of the 7th National Conference on Artificial Intelligence (KII-2000), Russia*, pages 127–134, 2000.