6 References

[Angele et al. 96] **J. Angele, D. Fensel, R. Studer**: Domain and Task Modelling in MIKE. In: [Sutcliffe et al. 96]

[Birkhoff 93] **G Birkhoff:** *Lattice Theory*. 3rd edition, American Mathematical Society, Providence, RI 1993.

[Breuker, van de Velde 94] **J. Breuker, W. van de Velde (eds.)**: *CommonKADS Library for expertise Modelling.* Frontiers of Artificial Intelligence and Applications, vol. 21, IOS Press, Amsterdam, 1994.

[Ganter, Wille] **B. Ganter, R. Wille**. *Applied lattice theory: Formal Concept Analysis*. URL: http://www.math.tu-dresden.de/~ganter/concept.ps

[Ganter, Wille 96] **B. Ganter, R. Wille**: *Formale Begriffsanalyse. Mathematische Grundlagen*. Springer, Berlin 1996. (in German)

[Jansen et al. 97] **M.G. Jansen, A.Th. Schreiber, B.J. Wielinga**: Rocky III --- Round 1. A Progress Report. SWI, University of Amsterdam, 1997.

[Khabaza, Shearer 95] **T. Khabaza, C. Shearer:** Data Mining with Clementine. *IEE Colloquium on Knowledge Discovery in Data Bases. IEE DIGEST*. no. 1995/021(B), London. 1995

[O'Shea 85] **T. O'Shea (ed.):** *Proceedings of the Sixth European Conference on Artificial Intelligence (ECAI-84).* Pisa, Italy 1984. Elsevier, North Holland 1985.

[Plaza, Benjamins 97] **E. Plaza, R. Benjamins (eds.)**: *Knowledge Acquisition, Modeling and Management.* Proceedings of EKAW'97, Sant Feliu de Guixols, Spain, October 1997. LNCS 1319. Springer, Berlin 1997.

[Richards, Compton 97] **P. Compton, D. Richards**: Knowledge Acquisition First, Modelling Later. in: [Plaza, Benjamins 97]

[Rumbaugh et al. 91] **J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, W. Lorensen**: *Object-Oriented Modeling and Design*. Prentice Hall, Englewood Cliffs 1991.

[Schreiber et al. 93] **G. Schreiber, B. Wielinga, J. Breuker (eds.)**: *KADS. A Principled Approach to Knowledge-Based System Development*, Knowledge-Based Systems, vol. 11, Academic Press, London, 1993.

[Schreiber et al. 94] **A. Th. Schreiber, B. Wielinga, R. de Hoog, H. Akkermans, W. van de Velde**: CommonKADS: A Comprehensive Methodology for KBS Development. in: *IEEE Expert*, December 1994, pp. 28-37.

[Shadbold et al. 96] **N. Shadbold, L. Crow, J. Tennison, J. Cupit**: *Sisyphus III Resource Page*. URL: http://www.psyc.nott.ac.uk/aigr/research/ka/SisIII/

[Shearer 96] **C. Shearer**: User Driven Data Mining. *Unicom Data Mining Conference*. London 1996.

[Sutcliffe et al. 96] **A.G. Sutcliffe, D. Benyon, F. van Assche (eds.)**: *Domain Knowledge for Interactive System Design. Proceedings of the TC8/WG8.2 Conference on Domain Knowledge in Interactive System Design*, Geneva, Switzerland, May 1996. Chapman & Hall, London, 1996.

[Wielinga, Breuker 84] **B.J. Wielinga, J.A. Breuker:** Interpretation of verbal data for knowledge acquisition. in: [O'Shea 85]

5 Conclusion and Future Work

In this paper we described a formal method to structure and visualize information in order to make it intelligible and interpretable --- Formal Concept Analysis. We claimed that methods like FCA could help knowledge engineers in the process of building a domain model. At least these methods can support him in acquiring a first impression of the concepts of the domain.

Another method to support automatical derivation of knowledge from a given set of information sources is Data Mining or Knowledge Discovery in Databases (KDD). Especially in the context of the Sisyphus-III project this approach seems interesting, since a large databases has been provided by the Sisyphus-III team. We already performed some tests using machine learning algorithms to learn correlations which are hidden in the 19,000 rows of the given data base with geological information. These tests have been performed using the data mining workbench CLEMENTINE from ISL ([Khabaza, Shearer 95], [Shearer 96]). First obtained results look promising and in the future there will be further evaluations of the strength of KDD in Sisyphus-III. In general the same as for FCA holds for KDD; it could be a means to acquire knowledge without having a prebuilt domain model and thus building the basis for a domain model.

A completely other way to tackle Sisyphus-III has been taken by the SWI in Amsterdam. In a talk at the Sisyphus-III session at EKAW 97 Machiel Jansen demonstrated how to built a KBS for Rocky-III in a conventional way [Jansen et al. 97]. Because the rules of the game forbid to consult a human expert, but allowed the consultation of textbooks, he became an expert of the domain himself and was able to model the needed knowledge as the knowledge engineer and the expert in one person. This approach is nearly impossible when confronted with a domain where no sufficient textbooks exist and the only sources of information are the experts. At least then, FCA or KDD can support the knowledge engineer in acquiring a first model of the domain.

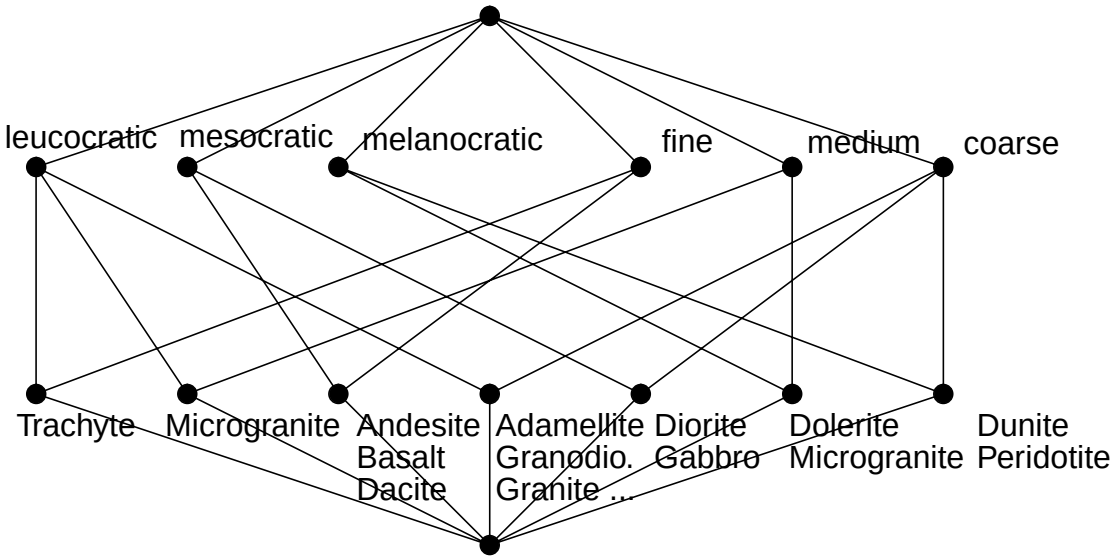
Integrating achievable results of FCA with those of conventional approaches for developing knowledge based systems such as the MIKE approach [Angele et al. 96] one can see obvious advantages in combining them. The MIKE approach is an incremental way to build a KBS and comprises such phases as *elicitation, interpretation, formalization*, and *implementation* of knowledge. The interfaces between these phases were built by a set of documents, e.g. the *elicitation model* which contains the set of expert protocols, or the *structure model* which is a semiformal representation of the model of expertise. The transition between these two models is achieved through interpretation and structuring of knowledge, performed by the knowledge engineer. This transition step can be supported by FCA because it is a means to structure and identify interesting concepts which should become part of the structure model.

Looking back at our work on the given resources, the Sisyphus-III project asks the right questions, we would summarize as follows:

> *"How do I acquire knowledge from ‘uncooperative‘ experts, from unstructured sources, and from contradicting and incomplete protocols?"*

We think these questions are important and have to be further investigated. Otherwise, building knowledge based systems will stay an art more than an engineering discipline.

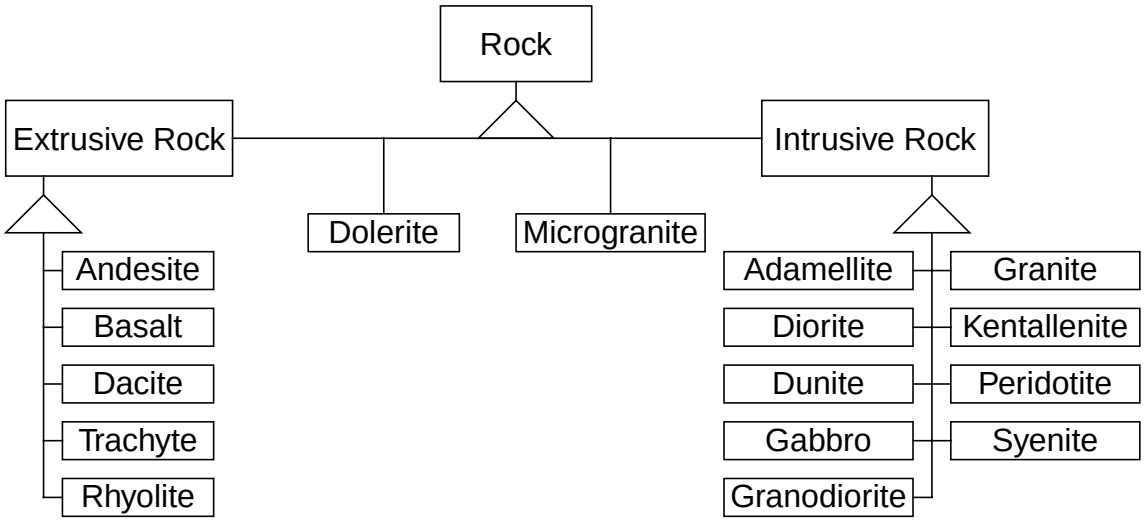
these two scales based on the information of card sort #1 (*Leucocratic, mesocratic,* and *melanocratic* thereby are values of the multi valued attribute *colour.*).



Grain size and *Colour* are two totally indepentent dimensions of rocks (based on card sort #1)

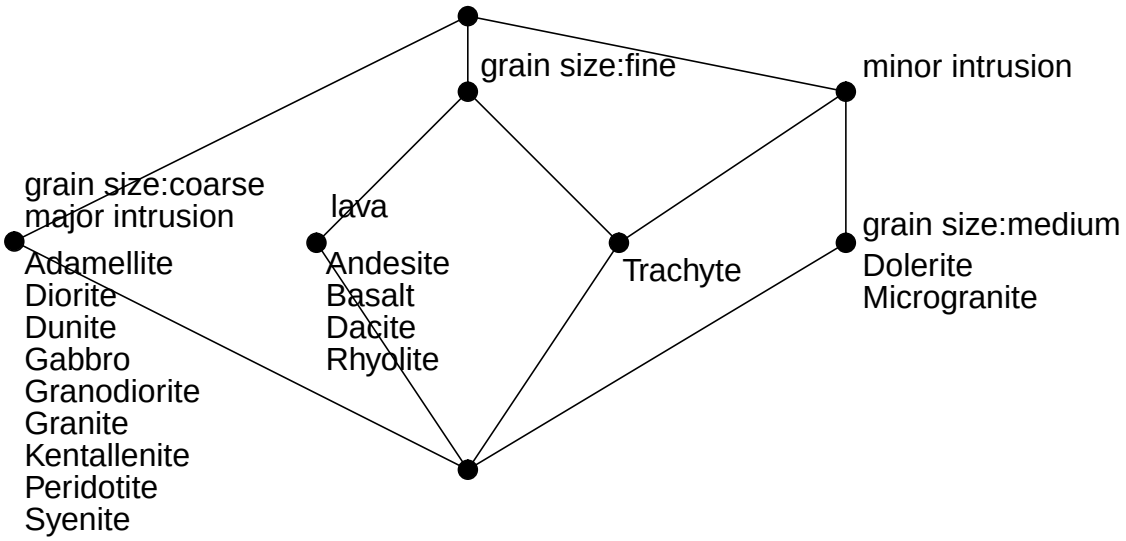
Other dimensions are very similar to each other, as e.g. *grain size* and *place of formation*. Knowledge gained in this way obviously helps the knowledge engineer in building a domain model. It helps in getting rid of ambiguities or providing missing information of expert interviews.

From the presented line diagrams which represent the place of formation of rocks a partial domain model (here represented in OMT's object model [Rumbaugh et al. 91]) can be achieved. Thus, the influence data analysis, esp. FCA, could have on domain modelling is visible.



OMT object model gained from the line diagrams

Very similar results can be found when analysing card sort #3 and looking at the scales for *grain size* and *formation environment*.



Line diagram for card sort #3 with scales for *grain size* and *formation environment*

Here we see that rocks are coarse grained exactly iff their formation environment is *massive plutonic intrusion*. All rocks stemming from lava are fine grained. And all medium grained rocks are formed in minor intrusions. Here again one rock type, namely Trachyte is special: although it is fine grained it is formed in minor intrusions.

These two examples may show how easy the visualization of conceptual relationships becomes when using line diagrams and formal concept analysis. The analysis of both line diagrams yields a general rule in the geological domain:

> *"Fine grained rocks were formed on or near the surface, and coarse grained rocks were formed in great depths beneath the surface."*

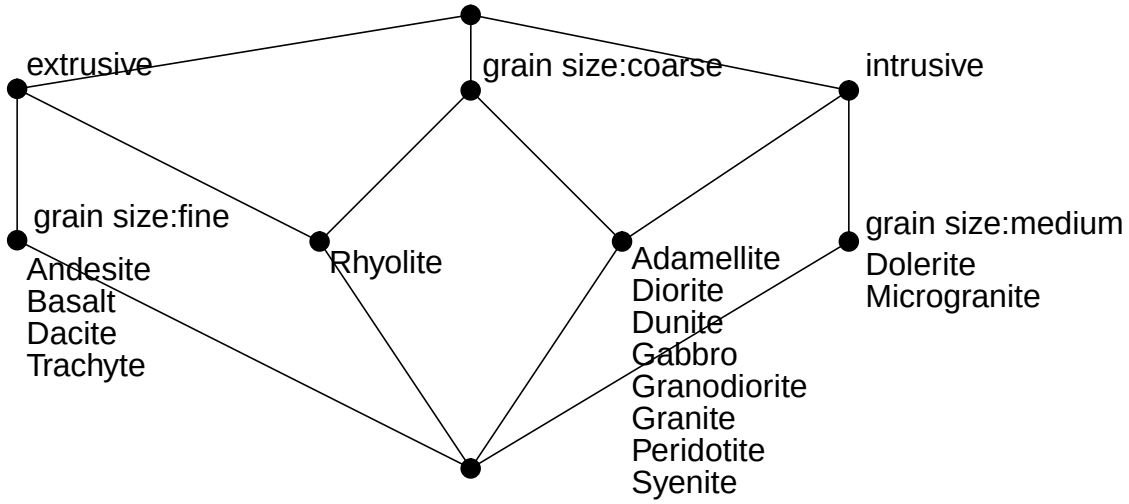
The two special cases Rhyolite and Trachyte are possibly only errors or noise in the card sorts. Looking at both card sorts, they only contradict in these two rocks, so that the general rule holds with high probability.

Not only the line diagrams provide valuable input for the knowledge engineer; the card sorts or formal contexts offer a list of interesting attributes associated with rocks as well. These attributes represent different dimensions along which rocks may be classified, e.g. grain size, colour, different minerals or place of formation. Some of these dimensions are orthogonal, e.g. grain size and colour. Their independence can be seen when computing the line diagram for

would contradict the idea of applying FCA to gain knowledge for building an initial domain model.

4.2 Application

Most of the card sorts contain too much information to present in a single line diagram, e.g. card sort #1 contains more 81 concepts. This means the line diagram contains 81 nodes and more than 200 subconcept/superconcept relations. Thus, it would be unreadable. To still be able to visualize the information from these card sorts the line diagrams were prepared representing only a subset of the given attributes. For example, concentrating on the scales *grain size* (*fine*, *medium* or *coarse*) and *emplacement* (*intrusive* or *extrusive*) of card sort #1 the resulting context contains five attributes that induce nine concepts. Therefore the resulting line diagram is easily intelligible.



Line diagram for card sort #1 with scales for *grain size* and *emplacement*

The knowledge engineer using this line diagram can easily read in it for example which rocks are coarse grained or extrusive, etc. The presentation allows the detection of implications between attributes, e.g. all coarse or medium grained rocks are intrusive, except Rhyolite, which is coarse grained but extrusive. Another rule found in this line diagram states that all extrusive rocks are fine grained, again except Rhyolite.

To illustrate the idea of this transformation procedure we will list a part of card sort #5 (taken from [Shadbold et al. 96]) together with its one valued transform.

Rock \ Attribute	**grain size**	**silica contents**	**...**
Adamellite	*coarse*	*very high (>70%)*	
Andesite	*fine*	*intermediate (60-70%)*	
Basalt	*fine*	*lowish (50-60%)*	
Dacite	*coarse*	*very high (>70%)*	
Diorite-Tonalite	*coarse*	*intermediate (60-70%)*	
Dolerite	*not coarse*	*ultrabasic (<50%)*	
...			

Excerpt of the original card sort #5

The original five columns (two shown) with multiple values were substituted by 14 columns with single values (seven shown), representing exactly the same information. For example the attribute *grain size* with its three possible values *coarse*, *not coarse*, and *fine* changes to three attributes with single values, namely *grain size:coarse*, *grain size:not coarse*, and *grain size:fine*. In a single row of the resulting table (i.e. for one object) at most one of these columns is checked. If none of these three is checked the grain size of the rock is not known.

Rock \ Attribute	**grain size: coarse**	**grain size: not coarse**	**grain size: fine**	**silica: >70%**	**silica: 60-70%**	**silica: 50-60%**	**silica: <50%**	**...**
Adamellite	*x*			*x*				
Andesite			*x*		*x*			
Basalt			*x*			*x*		
Dacite	*x*			*x*				
Diorite-Tonalite	*x*				*x*			
Dolerite		*x*					*x*	
...								

Transformation of card sort #5 into a one valued formal context

The straightforward way of converting multi valued contexts into one valued contexts (i.e. the use of elementary scales) does not require any interpretation or knowledge of the meaning of attributes and their values. Probably, the transformation could have been done more intelligently if additional domain knowledge had been used for formulating the scales, but this

diagrams also display object hierarchies and explicitly show why some concepts are specialisations of others. For example the line diagram shows, *cola* is a subconcept of *mineralwater.* Actually it says "*cola* is *mineralwater* with *caffeine*" which is very close to reality.

Another thing one can learn from line diagrams are **implications**, e.g. the example line diagram shows that any object that is hot also carries the attribute *non-alcoholic*. That is because all subconcepts of the concept annotated with *hot* are also subconcepts of the node annotated with *non-alcoholic*.

The easy perceivability of dependencies through line diagrams makes the method applicable to knowledge acquisition processes as we will demonstrate for the Sisyphus-III experiment in the next section.

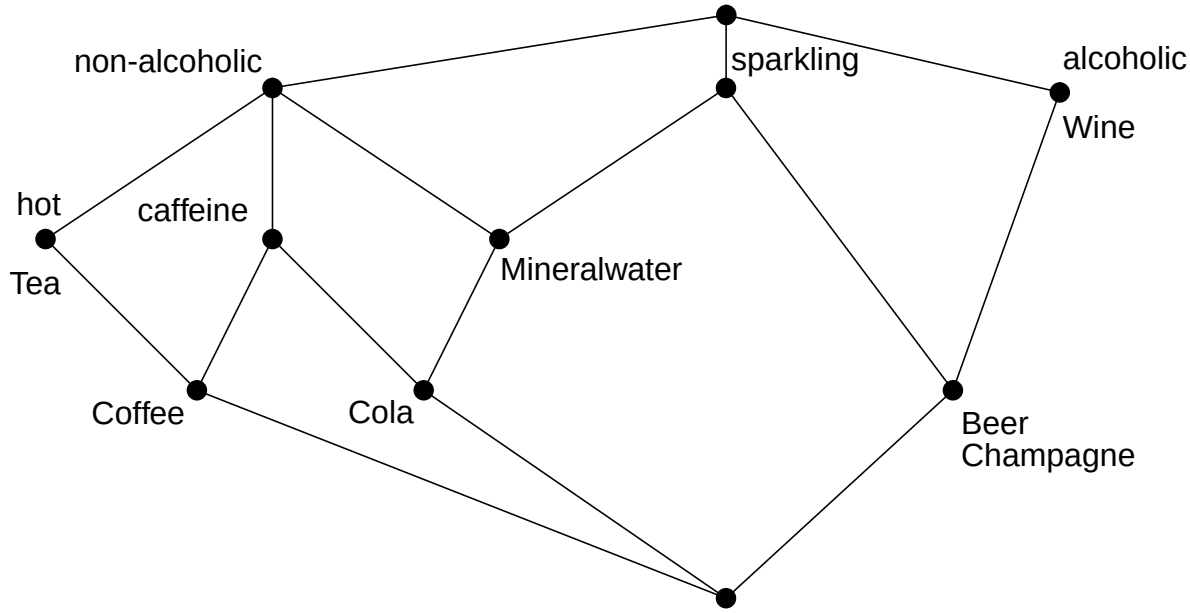
4 Application of FCA on Sisyphus' Card Sorts

In the context of the Sisyphus-III project FCA has been employed to analyse the card sorts provided by the Sisyphus-III team as KA resources. The aim of analysing the card sorts with FCA was to learn concepts and their conceptual dependencies in the domain (as far as contained in the card sorts) before trying to invent a domain model based on more or less vague and ambiguous natural language descriptions given by several experts of varying quality. A slogan characterizing this approach very well is the title of an EKAW paper by Debbie Richards and Paul Compton [Richards, Compton 97]: "Knowledge Acquisition first, Modelling later". We think doing this is necessary because otherwise --- without having an expert at hand to walk through and validate the created models --- the knowledge engineer has to become an expert himself (like in [Jansen et al. 97]). This expertise is hardly obtained solely by reading the protocols provided with Sisyphus-III. We propose to support this learning process by means of data analysis and visualisation, i.e. Formal Concept Analysis an line diagrams.

4.1 Preparing the Resources of Sisyphus-III for FCA

The input for our analyses are the provided card sorts. They represent discrete, formalized information, provided by experts. This level of detail is hardly present in the interviews or in the self reports of the experts. All card sorts contain information about rock types, thus it is quite obvious to associate the rocks of the card sorts with the objects of the formal contexts. During the whole section the set of objects will be *G = {Adamellite, Andesite, Basalt, Dacite, Diorite, Dolerite, Dunite Gabbro, Granodiorite, Granite, Kentallenite, Microgranite, Peridotite, Rhyolite, Syenite, Trachyte}*. The card sorts contain several columns which represent attributes with several ranges. Because the notion of a concept lattice only makes sense for one valued contexts we have to transform the card sorts (which actually are multi valued formal contexts) into one valued contexts [Ganter, Wille]. This can be achieved by so called transformational scaling, where the multiple values of an attribute were unfolded. This process of scaling opens several degrees of freedom and a number of possible decisions or interpretations of the attributes and values. We chose the simplest variant, i.e. plain scaling, where the objects remain unchanged and each column *x* of the original multi valued context is substituted by a set of columns $\{(x, i) \mid i$ is a value of attribute x in the original context$\}$. This set is called a scale of *x*.The resulting context has crosses at the cell *(g, (x, i))* iff object $g \in G$ carried the value *i* for attribute *x* in the original multi valued context, i.e. the card sort. We will represent the new columns *(x, i)* from now on as *x:i*.

The example shows the line diagram for the above presented context of beverages.



Example line diagram for the context of beverages.

The graph consists of nodes that represent the concepts and edges connecting these nodes. Two nodes C_1 and C_2 are connected iff $C_1 \leq C_2$ and there is no concept C_3 with $C_1 \leq C_3 \leq C_2$. Although the concept lattice is a directed acyclic graph (DAG) the edges are not provided with arrowheads, instead the convention holds that the superconcept always appears above of all its subconcepts. For example the line diagram shows that the nodes annotated with *coffee* and with *cola* are both subconcepts of the node annotated with *caffeine*. As a difference to usual lattice diagrams the labelling in line diagrams is reduced, i.e. each object and each attribute is only entered once. So the nodes are not annotated by their complete extents and intents. Rather, attributes and objects propagate along the edges, as a kind of inheritance. Attributes propagate along the edges to the bottom of the diagram and dually objects propagate to the top of the diagram. Thus the top element of a line diagram (the supremum of the context) is actually marked by $(G, \emptyset)$ if G is the set of objects. The bottom element (the context's infimum) is marked by $(\emptyset, M)$ if M is the set of attributes.

Attribute names are always displayed slightly above the node and object names are noted slightly below the respective node.

To read a line diagram you start at the object, attribute, or concept you are interested in, e.g. the node marked with *cola*. Following all paths from this node to the top element one visits all superconcepts of the selected concept. Collecting the attributes displayed along the paths one finds all attributes that the selected concept or object carries. Selecting a node and following all paths from this node to the infimum of the lattice one finds all sub- and subsubconcepts. If the selected node displays an attribute name all objects along these paths establish the set of objects carrying this attribute.

Thus, line diagrams display relationships between objects, attributes and concepts in an easily perceivable way. For example, the above given line diagram reveals that *beer* and *champagne* are equivalent objects. Of course, one has to pay attention to the context (in a colloquial as well as in a formal sense). Concerning the given formal context *beer* and *champagne* are equal because they carry exactly the same attributes, namely *sparkling* and *alcoholic*. Their equivalence can be seen in the line diagram by their appearance at the same concept node. Line

A' contains all attributes that are common to all objects in *A*. And *B'* is the set of all objects that carry all the attributes of *B*.

With that, the pair *(A, B)* is a formal concept iff

$$A' = B \text{ and } A = B'.$$

This property says that all objects of the concept carry all its attributes and that there is no other object in *G* carrying all attributes of the concept. When looking at the cross table this property can be seen if rectangles totally covered with crosses can be identified, e.g. the four cells associated with *tea*, *coffee*, *non-alcoholic*, and *hot* constitute such a rectangle. If we ignore the sequence of the rows and columns we can identify even more concepts, e.g. ignoring the row *cola* and the column *caffeine* (or moving them to another place) we achieve another rectangle/concept, namely the cells associated with the objects *beer* and *champagne* and the attributes *alcoholic* and *sparkling*.

Looking at the definition of a formal concept one can easily see that for all $A \subseteq G$ the pair *(A'', A')* is a formal concept. The dual holds for all $B \subseteq M$, i.e. *(B', B'')* is always a formal concept, too. Yet, the sets of concepts achieved in this way are equal and contain exactly the concepts existing in the given context.

For formal concepts a natural **subconcept/superconcept relationship** ≤ can then be defined as follows:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \; (\Leftrightarrow B_2 \subseteq B_1)$$

This relationship shows the dualism that exits between attributes and objects of concepts. A Concept $C_1 = (A_1, B_1)$ is a subconcept of concept $C_2 = (A_2, B_2)$ iff the set of its objects is a subset of the objects of C_2. Or an equivalent expression is iff the set of its attributes is a superset of the attributes of C_2. Actually, the set of all formal concepts of a context forms a so called **concept lattice**. The **infimum** of this lattice is formed by $(\emptyset, M)$ and its **supremum** is formed by $(G, \emptyset)$ if the context is given by *(G, M, I)*.

Because of the dualism between objects and attributes and the fact that data analysts or any other users of FCA are interested in investigating structures and relationships we need a representation of concepts that treats both objects and attributes alike. This representation is realised in a line diagram which will be presented in the next section.

3.2 Line Diagrams

A line diagram is a graphical visualisation of the concept lattice. It allows the investigation and interpretation of relationships between concepts, objects and attributes. This includes object hierarchies, if they exist in the given context. A line diagram contains the relationships between objects and attributes and thus is an equivalent representation of a context, i.e. it contains exactly the same information as the cross table. Thirdly dependencies and relationships between attributes can be easily detected in a line diagram.

FCA is widely used in numerous application domains, e.g. in psychology where repertory grids were analysed using FCA; in libraries where FCA and line diagrams were used to help readers retrieve desired literature; or in software reengineering where line diagrams were used to locate clusters of subroutines in 20 years old FORTRAN code. [Richards, Compton 97] apply FCA to reengineer, represent, and investigate knowledge bases created by ripple down rules in an iterative manner.

To illustrate the ideas behind Formal Concept Analysis a brief introduction of its mathematical foundations and an example is given (section 3.1). Because the main benefit of FCA in the context of knowledge acquisition for Sisyphus-III lies in displaying relationships between concepts with line diagrams we will stress this aspect of FCA (section 3.2).

3.1 Mathematical Foundation

To introduce the method FCA we first have to define the term context or **formal context**. A formal context is a triple *(G, M, I)* which consists of a set *G* of **objects**, a set *M* of **attributes** and a binary **(incidence) relation** $I \subseteq G \times M$ between objects and attributes. A context is typically represented in tabular form as a cross table, whose rows are represented by the objects, whose columns are represented by the attributes and whose cells are marked iff the incidence relation holds for the corresponding pair of object and attribute. As an example we will present a context of beverages. The different drinks form the set *G* of objects, and some possible features of drinks are collected in the set *M* of attributes. The incidence relation *I* is given by the cross table.

Objects \ Attributes	**non-alcoholic**	**hot**	**alcoholic**	**caffeine**	**sparkling**
Tea	*x*	*x*			
Coffee	*x*	*x*		*x*	
Mineralwater	*x*				*x*
Wine			*x*		
Beer			*x*		*x*
Cola	*x*			*x*	*x*
Champagne			*x*		*x*

An example of a formal context.

The table should be read in the following way: Each x marks a pair being an element of the incidence relation *I*, e.g. *(coffee, hot)* is marked because the object coffee carries the attribute hot, whereas *(mineralwater, hot)* is not marked because normally mineralwater is not hot. Thus $(g, m) \in I$ should be interpreted as "the object *g* carries the attribute *m*".

The central notion of FCA is the **formal concept**. A concept *(A, B)* is defined as a pair of objects $A \subseteq G$ and attributes $B \subseteq M$ which fulfil certain conditions. *A* is called **extent** and *B* is called **intent of the concept**. To define the necessary and sufficient conditions for a formal context we present two derivation operators. Given $A \subseteq G$ we define

$$A' := \{m \in M \mid \forall\, g \in A: (g, m) \in I\}$$

and dually for $B \subseteq M$

$$B' := \{g \in G \mid \forall\, m \in B: (g, m) \in I\}.$$

verbal data". These cites stem from [Wielinga, Breuker 84] and characterize typical KA material.

The huge set of more or less unstructured knowledge material as provided as verbal data contains statements of several experts concerning the classification of rocks. Due to the incorporation of knowledge from different sources the reliability of each statement is hardly predictable. There are contradictory statements, and imperfect, incomplete, or ambiguous informations. For example in self report #1 the subject classifies a hand specimen as Microdiorite although it indeed is Andesite.

> *"I'd go for a porphyritic microdiorite or something like that."*
>
> *"**It's an Andesite.**"*
>
> *"... I wouldn't have thought I'd have got it that wrong."*

Thus, without the information the interviewer gives a knowledge engineer would have no chance to assess the quality of the given classification.

The degree of expertise in the protocols varies widely. Some protocols contain information which could directly stem from textbooks, e.g. structured interview #2 contains a definition of the acidity of rocks in terms of the silica content:

> *"Silica percentage less than 42%, that's your ultrabasic, and from 42 to 52 is basic, and 52 to 66 is intermediate, and greater than 66% is acidic rock."*

Other protocols hardly contain any information, e.g. in laddered grid #2 the subject answers to a question to name coarse grained rocks:

> *"... Gabbro ... well, it might be Syenite or Trachyte, and it's not Andesite, I can't remember which is which."*

These examples are illustrations of the difficulties a knowledge engineer has to cope with during knowledge acquisition, i.e. while passing the knowledge acquisition bottle neck.

The aim of the Sisyphus-III project was the development of a decision support system to assist astronauts in the task of classifying igneous rocks and to develop a tutorial system for their (the astronauts' ;-) education. To build these systems a domain model is necessary. Due to the difficulties of acquiring domain knowledge from verbal protocols we instead exploited the other sources, esp. using Formal Concept Analysis, the basics of which will be presented in the next section.

3 Formal Concept Analysis

During the last years the amount of information being accessible is growing rapidly. To avoid getting lost in this information means to communicate, display and handle these masses are needed. One way of doing this is the Formal Concept Analysis (FCA) which was founded by Rudolf Wille in about 1980 at the TH Darmstadt, Germany [Ganter, Wille 96]. FCA is a mathematical approach to data analysis based on the lattice theory of Garret Birkhoff [Birkhoff 93]. It allows to take unstructured information, provide it with structure, and to display this well structured information to human users. The clear representation of the data allows investigation and interpretation and the acquisition of knowledge. Central are the so called concepts which in addition to their colloquial meaning have been defined in mathematical terms.

the card sorts offered by Sisyphus-III. At last, section 5 contains conclusions and an outlook of future work that could be done concerning the first phase of the KA process.

2 The Sisyphus-III Experiment

The Sisyphus-III experiment [Shadbold et al. 96] tries to explore a rather different area of the knowledge acquisition (KA) process than its ancestors. In Sisyphus-I ---the room assignment problem--- a very short description of the problem and the domain was given. The scope of this problem was quite narrow and the domain was easily understandable by any knowledge engineer. The second project in the Sisyphus series presented a broader and not easily perceivable domain ---the domain of elevators (VT). As a means to test and validate several approaches towards modelling of problem solving methods (PSM) the presented material was suitable. It contained information about all needed domain concepts and terms in a way non-experts can understand up to a convenient level. The presented document has been especially prepared as a source for knowledge engineers to be used in the process of developing a knowledge based system (KBS). The current experiment Sisyphus-III, aka. Rocky-III, is very different. It provides the participating researchers with real material, mostly transcripts of interviews or self reports of "experts" in geology. While the former Sisyphus experiments stressed the development of PSMs, settled in a more or less clear domain, work in the Sisyphus-III problem has to begin earlier in the phases of the Knowledge Engineering process. In Sisyphus-III the focus lies on the acquisition of knowledge from the set of given materials. The appearance of the material makes Sisyphus-III special. The documents contain information given by several experts (which show several degrees of actual expertise). Different experts use different terminologies. Transcripts of interviews are not always perfect. The interviews and self protocols have been already performed and are now canned, such that the participants of Sisyphus-III have to cope with which information is present and which is not, i.e. the participants cannot pose questions to the experts if they needed a special kind of information. Except the last point, these features of expert utterances are more realistic compared to actual KBS developing projects than Sisyphus-I and II.

The resources provided by the Sisyphus-III team were manifold and richer than in Sisyphus-I and II. They contain three different kinds of transcripts of experts' statements:

1. laddered grids,
2. structured interviews, and
3. self reports

The sources contain further repertory grids and card sorts which contain discreticized information about rocks and minerals.

At last a geological database (IGBA) has been provided which could be used for machine learning techniques for knowledge acquisition. This database has been originally designed to support human users (especially geologists, i.e. experts in the classification of rocks) in information retrieval. The IGBA database has not been designed to support machine learning algorithms or any other means of automatic processing.

Thus, there are two distinct subsets of resources: (i) natural language texts and (ii) discrete, structured information. The first subset seems to be "the most convenient source of information for knowledge acquisition ---due to their richness, expressiveness and the natural way in which they are used to communicate knowledge in general". Besides these obvious benefits there are "a number of serious problems with the elicitation, interpretation and quality assessment of

Formal Concept Analysis to Learn from the Sisyphus-III Material

Michael Erdmann
Institut für Angewandte Informatik und Formale Beschreibungsverfahren
University of Karlsruhe (TH)
D-76128 Karlsruhe (Germany)
e-mail: erdmann@aifb.uni-karlsruhe.de

Abstract: This paper presents a way to acquire the initial models of a domain when building a knowledge based system. This first step in the knowledge acquisition process has not been well investigated in the past. We propose to support the initial modelling by a technique called Formal Concept Analysis for data analysis. We will apply this technique to the knowledge acquisition material provided in the context of Sisyphus-III. This application yields an initial representation of the relevant concepts of the domain which then, can be incorporated in a domain model.

1 Introduction

In this paper we will present an approach to apply the method of Formal Concept Analysis (FCA) [Ganter, Wille 96] to knowledge acquisition tasks. FCA is a way to find, structure, and display relationships between concepts, which consist of attributes and objects. This method helps in understanding a given domain and in building a domain model for it. We will show how FCA has been adopted in the Sisyphus-III experiment.

We think this approach is especially fruitful in the setting of the Sisyphus-III project [Shadbold et al. 96]. There, a huge set of more or less unstructured knowledge material is provided which contains the knowledge of several experts concerning the classification of rocks and minerals. The aims of the project were building a decision support system to assist astronauts in this classification task and to develop a tutorial system for their education. To achieve these goals a model of the domain has to be built. Typically, this model is built by a knowledge engineer who familiarized himself with the domain terms, and thus has to gain a certain level of expertise. A large part of the KA research of the last decade has concentrated on this modelling approach, manifested in several approaches for development of knowledge based systems in the KADS tradition [Schreiber et al. 93], e.g. CommonKADS ([Schreiber et al. 94], [Breuker, van de Velde 94]), MIKE [Angele et al. 96], etc. Questions of how to achieve these models were only answered unsatisfactory; mostly, an iterative process is proposed in which an initial model is presented to an expert for validation. From this first model (and the experts' comments on it) in several steps a final domain model is achieved. None of the mentioned approaches supports knowledge engineers during the first step in the KA process, namely the acquisition of the initial model. At this point we propose to use techniques like FCA.

The FCA method will be briefly introduced in section 3 after a short overview of the Sisyphus-III experiment in section 2. After that, we will show some results achieved by applying FCA to