Uta Priss
School of Library and Information Science, Indiana University

A Graphical Interface for Conceptually Navigating Faceted Thesauri

Abstract: This paper describes a graphical interface for the navigation and construction of faceted thesauri that is based on formal concept analysis. Each facet of a thesaurus is represented as a mathematical lattice that is further subdivided into components. Users can graphically navigate through the Java implementation of the interface by clicking on terms that connect facets and components. Since there are many applications for thesauri in the knowledge representation field, such a graphical interface has the potential of being very useful.

1. Introduction

Thesauri are useful for many applications: they serve as subject access systems. They can be used to increase the recall in information retrieval. And they provide a means of storing semantic information (common sense as well as scientific knowledge) in a formal manner. They can even be used as ”knowledge bases” or as components in natural language computer interfaces. Considering the wide range of applications we do not restrict the term ”thesaurus” to its information science meaning. The artificial intelligence knowledge base CYC (Lenat, 1995), the natural language database WordNet (Miller, 1995) and, of course, Roget’s Thesaurus also represent types of thesauri because all of them group words into synonym sets and define semantic relations among the synonym sets. Although thesauri have many applications many existing thesauri are not anywhere near ”perfect” (compare for example Fischer (1993)). It seems that imperfection is due to four main factors: first, only faceted thesauri provide the flexibility of simultaneously presenting several possibly contradictory viewpoints. Only faceted thesauri are modular designed and therefore easier to maintain and to adapt to different applications - but most existing thesauri are not faceted. Second, manual thesaurus construction is expensive, but automated thesaurus construction is not yet sophisticated. Third, if faceted thesauri are to be used as knowledge bases they must be based on a philosophical and mathematical formalization of semantic relations. In a framework of a formalization, the semantic relations must be

formally defined and consistent. But most existing thesauri are not based on a formal theory. And, forth, without a graphical representation it is extremely difficult to maintain control over thesauri that have more than 100 terms. It is therefore not surprising that existing thesauri often contain irregularities.

In this paper we explore the possibilities of a graphical interface for a faceted thesaurus representation that is based on a mathematical formalization called formal concept analysis. While the mathematical formalization ensures the consistency of semantic relations, the graphical interface prevents users form ”getting lost” in a large set of terms and classes. Structurally related classes are connected by lines in a diagram. Users can navigate along those lines. Each facet is broken down into components. Several components can be pulled up on the screen simultaneously and therefore can be visually compared. Besides searching browsing is a further option. The thesaurus is stored as a relational (SQL) database. Although the main display in our approach is a graphical display it is easy to extract the data in any other format (such as SGML) directly from the database if a certain application requires a certain format. In our current implementation the code for the graphical interface is written in Java which is connected via JBDC to a relational database.

Previous research in the application of formal concept analysis to thesauri (Skorsky, 1997) uses TOSCANA (Vogt & Wille, 1995) a formal concept analysis tool that allows navigation through ”facets”. Since TOSCANA does not use information science terminology the ”facets” are called ”scales” or ”topics”. The connection between TOSCANA and facets in an information science sense has only recently been discovered (Viehmann (1996) and Kent & Neuss (1995). Our approach adds some further aspects, such as the decomposition of facets into components and a thesaurus construction design environment, to TOSCANA-like features. The decomposition of facets into components allows a domain specific design of facets instead of universal facets. For example, a facet ”mental states of human beings” tends to be more comprehensive than a similar facet for animals. In TOSCANA the same facet would have to be used for humans and animals. In our approach, humans and animals would share a generic facet for mental states that is enhanced in the case of humans. Stumme’s (1996) solution for this problem which is slightly different from ours is not yet included in TOSCANA. Other research in this area that is not based on formal concept analysis often does not result in fully graphical or fully structured representations, for example, Johnson’s (1995) hypertext thesaurus interface contains a traditional tree display and a random arrangement of related terms (”RTs”). Lin’s (1997) displays are based on term occurrences in documents, as many other displays in the information retrieval

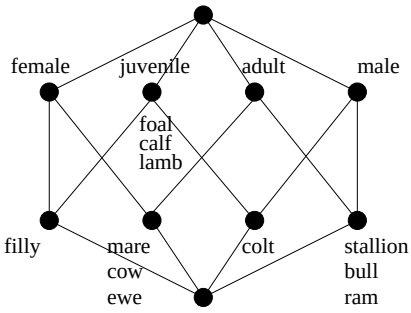
area, and represent a completely different field of research.

2. Formal concept analysis

Formal concept analysis (Ganter & Wille, 1996) starts with the definition of a *formal context* $\mathcal{K}$ as a triple (G, M, I) consisting of a set of *(formal) objects* (denoted by G), a set of *(formal) attributes* (denoted by M), and a relation I between G and M (i.e. $I \subseteq G \times M$). The relationship is written as gIm or $(g, m) \in I$ and is read as 'the formal object g has the formal attribute m'. A formal context can be represented by a cross table which has a row for each object g, a column for each attribute m and a cross in the row of g and the column of m if gIm. The upper half of Figure 1 shows two examples of formal contexts. They have 'filly', 'mare', and so on as formal objects, and 'female', 'juvenile' ('horse', 'cow', respectively) and so on as formal attributes. In a context (G, M, I) the set of all common attributes of a set $A \subseteq G$ of objects is denoted by $\iota A := \{m \in M \mid gIm \text{ for all } g \in A\}$ and, analogously, the set of all common objects of a set $B \subseteq M$ of attributes is $\varepsilon B := \{g \in G \mid gIm \text{ for all } m \in B\}$. For example, in the left formal context in Figure 1, $\iota\{\text{ram}\} = \{\text{adult, male}\}$ and $\varepsilon\{\text{female}\} = \{\text{filly, mare, cow, ewe}\}$ hold.

	female	juvenile	adult	male
filly	x	x		
mare	x		x	
colt		x		x
stallion			x	x
cow	x		x	
ram			x	x
bull			x	x
ewe	x		x	
foal		x		
calf		x		
lamb		x		

	horse	cow	sheep	animal
filly	x			x
mare	x			x
colt	x			x
stallion	x			x
cow		x		x
ram			x	x
bull		x		x
ewe			x	x
foal	x			x
calf		x		x
lamb			x	x



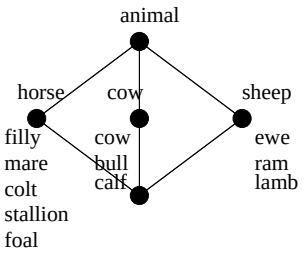


Figure 1: Formal contexts and line diagrams of their concept lattices

A pair (A, B) is said to be a *(formal) concept* of the formal context (G, M, I) if $A \subseteq G, B \subseteq M, A = \varepsilon B$, and $B = \iota A$. For a concept $c := (A, B)$, A is called the *extent* (denoted by $Ext(c)$) and B is called the *intent* (denoted by $Int(c)$) of the concept. In the right example of Figure 1, $(\{$cow, bull, calf$\}, \{$cow, animal$\})$ is a concept, because $\iota\{$cow, bull, calf$\} = \{$cow, animal$\}$ and $\varepsilon\{$cow, animal$\} = \{$cow, bull, calf$\}$. The set of all concepts of (G, M, I) is denoted by $\mathcal{B}(G, M, I)$. The most important structure on $\mathcal{B}(G, M, I)$ is given by the subconcept-superconcept relation that is defined as follows: the concept c_1 is a *subconcept* of the concept c_2 (denoted by $c_1 \leq c_2$) if $Ext(c_1) \subseteq Ext(c_2)$, which is equivalent to $Int(c_2) \subseteq Int(c_1)$; c_2 is then a *superconcept* of c_1. For example, $(\{$foal, calf, lamb, filly, colt$\}, \{$juvenile$\})$ as a superconcept of $(\{$filly$\}, \{$female, juvenile$\})$ has more objects but less attributes than $(\{$filly$\}, \{$female, juvenile$\})$. The relation '$\leq$' is a mathematical order relation called *conceptual ordering* on $\mathcal{B}(G, M, I)$ with which the set of all concepts forms a mathematical lattice denoted by $\underline{\mathcal{B}}(G, M, I)$.

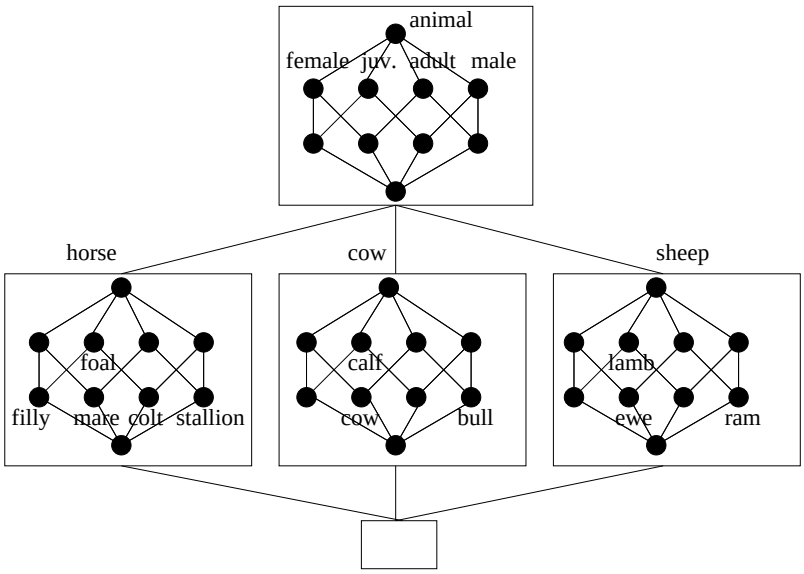


Figure 2: A nested line diagram

Graphically, mathematical lattices can be visualized by line diagrams which represent a concept by a small circle. For each object g the smallest concept to whose extent g belongs is denoted by γg. And for each attribute m the largest concept to whose intent m belongs is denoted by μm. The concepts γg and μm

are called *object concept* of g and *attribute concept* of m, respectively. In a line diagram it is not necessary to write the full extent and intent for each concept, instead the name of each object g is written slightly below the circle of γg and the name of each attribute m is written slightly above the circle of μm. The lower half of Figure 1 shows the line diagrams of the concept lattices of the examples. To read the line diagram, the extent of a concept consists of all objects which are retrieved by starting with the concept and then collecting all objects that are written at subconcepts of that concept. Analogously, the intent is retrieved by collecting all attributes that are written at superconcepts of the concept. Nested line diagrams facilitate the combination of several lattices that share the same set of objects, such as the two lattices in figure 1. One lattice is chosen as an *outer structure*, the other one is an *inner structure*. Parallel lines of the outer structure are omitted. In the example in Figure 2, the left lattice of Figure 1 serves as an inner structure, the right lattice of Figure 1 as an outer structure. The concepts in each box are subconcepts of the corresponding concepts in other boxes that are subconcepts in the outer structure. For example, the object concept of ‘foal’ is a subconcept of the attribute concept of ‘juvenile’. Since the bottom box does not contain any objects, it is not completely represented.

3. Faceted thesauri

In the case of a thesaurus the BT/NT hierarchy corresponds to the conceptual ordering in a concept lattice. Formal objects of a thesaurus lattice are objects to which the thesaurus terms are applied. They can be present during the design phase, but are omitted in the final thesaurus. Formal attributes are intensional descriptions of thesaurus terms. They are not identical to scope notes and usually also not represented in the final thesaurus version. The terms of a thesaurus correspond to names of concepts and the classes correspond to concepts. During the design of a thesaurus it can be useful to consider the formal object and attributes. Considering objects corresponds to a data driven, extensional, bottom-up approach whereas considering attributes corresponds to a top-down, intensional approach. Since extensions and intensions can both be used to define formal concepts, formal concept analysis makes it possible to switch between top-down and bottom-up views without losing information. A faceted thesaurus is a set of lattices that share a common set of formal objects. Each lattice represents a different viewpoint or aspect on how to classify the objects. It can best be displayed as a nested line diagram. A note on the difference between tree hierarchies and lattices may be useful: tree hierarchies can be modeled as lattices if a bottom concept is added. This bottom concept has all attributes of the lattice in its intent and since no object usually has all attributes (especially if the attributes are antonymous) the extent of the bottom concept is usually empty. Therefore it can be omitted in the graphical representation in which case the lattice representation of a tree simply is a tree. On the other hand, lattices do not need to be tree hierarchies. They are a special case of poly-hierarchies.

The terms for the thesaurus in the example in figures 3, 4 and 5 are taken from the ISO thesaurus of thesaurus terms (Henriksen & Lindh, 1996). But only the terms are taken from that thesaurus. The structure of our thesaurus is entirely different from that thesaurus. Figure 3 shows a component of the generic facet of the thesaurus. Users can click on the encircled terms to navigate to other components of the same facet or to combine this facet with other facets. The top term of this component is ”indexing language”. Three other facets can be combined with it: display methods, construction methods and applications of indexing languages. Components of a part-whole facet can be combined with the components that have ”classification scheme” and ”thesaurus” as top terms. By clicking on ”indexing language” users can navigate to the other components of the generic facet. The classes that cannot be expanded or combined with anything are represented as black boxes instead of circles. The four boxes on the bottom right side of the figure represent classes such as ”multilingual microthesaurus”. Instead of

representing them in this facet they could have been separated into a ”size” and a ”number of languages” facet. But since each of them contains only two classes we decided to precombine them with the generic facet and to add all possible combinations (the four unlabeled black boxes).

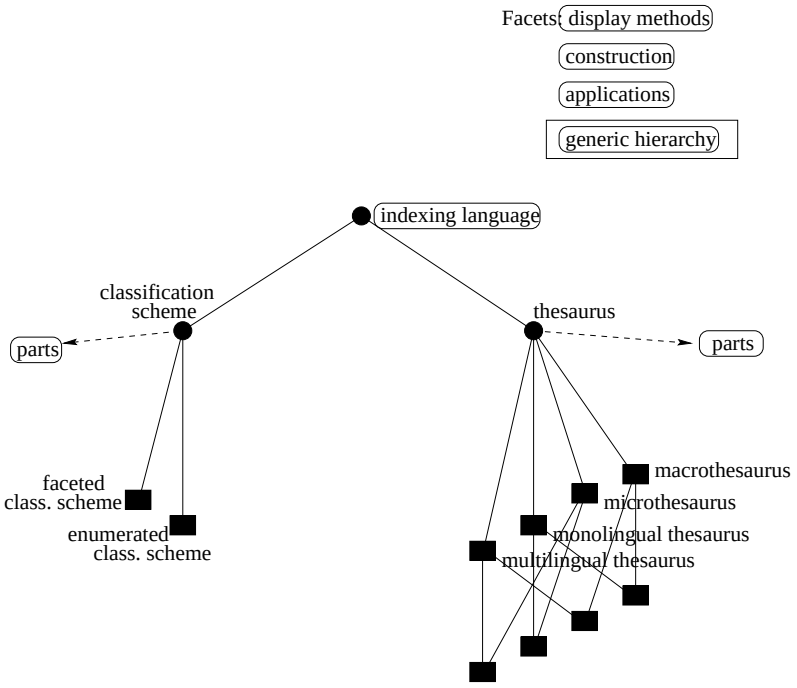


Figure 3: A generic facet

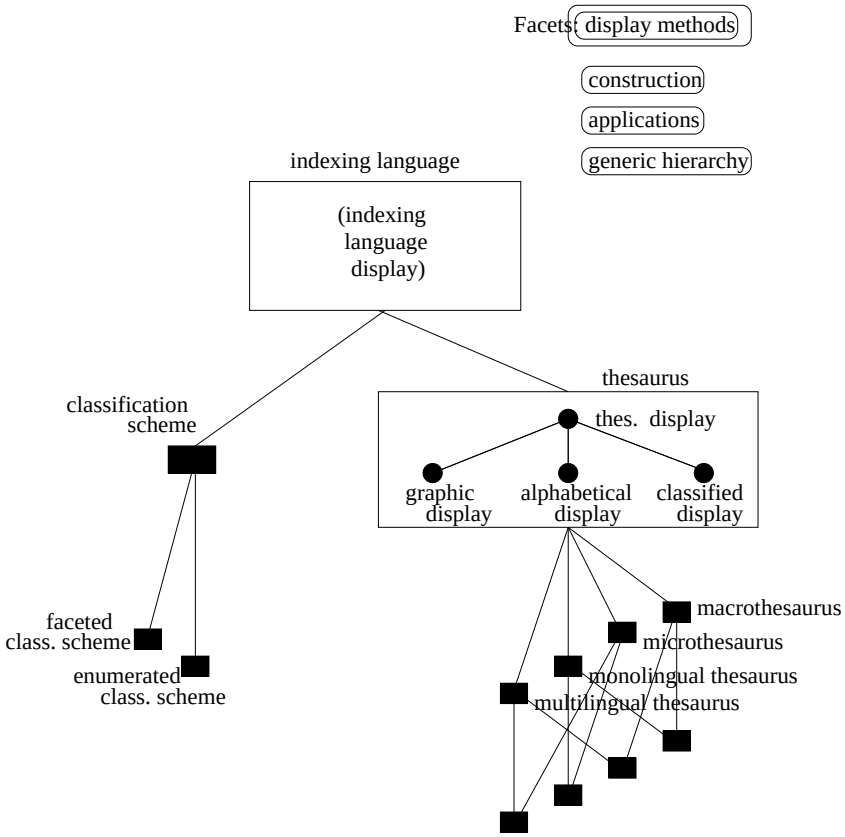


Figure 4: A generic facet combined with a display facet

Figure 4 demonstrates the combination of two facets. This is what users get to see if they click on ”display methods” in figure 1. The facets are expanded only where they add new terms to the thesaurus. There are no specific terms for the display of ”classification schemes” or ”indexing languages” in general. ”Thesauri” have three different display methods. Narrower terms always inherit all properties from their broader terms. Therefore ”multilingual thesauri”, ”multilingual macrothesauri”, and so on also have three different display methods. But since this information can be deduced from the diagram the box with the three thesaurus display methods is represented only once instead of nine times.

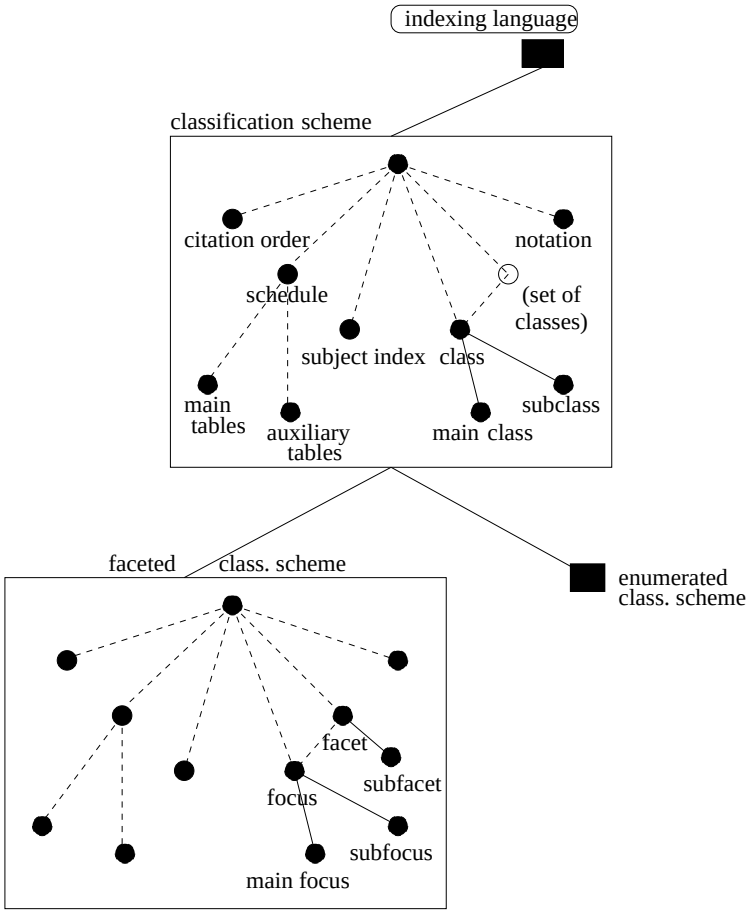


Figure 5: A part-whole facet

If users click on ”parts” of a classification scheme in figure 1, they get to see figure 5. The part-whole relation is represented as a dotted line. ”Enumerated schemes” use the same terminology as ”classification schemes”. ”Faceted schemes” change some of the terms and add a component (”facet”, ”subfacet”). Top terms of components that belong to narrower terms must be indicated in the component of the broader term. They are represented as unfilled circles which indicates that they can be expressed for that class but are not usually lexicalized.

4. Conclusion

This paper develops a formalization of faceted thesauri and shows screen shots of its implementation for a sample thesaurus. Users can navigate through the thesaurus horizontally from facet to facet and vertically among components of

the same facet by clicking on terms that serve as links between facets and components. Further options concerning search options and display features still need to be implemented. The interface is written in Java and therefore accessible through the World Wide Web. We are currently conducting extensive testing of the software and its usability. There is a further part of the software that provides an interface for faceted thesaurus design. Because of space limitations the other parts of the software and further details concerning the formalization and implementation cannot be described in this paper. They will be published elsewhere.

References

Fischer, Dietrich (1993). Consistency Rules and Triggers for Multilingual Terminology. TKE'93: Terminology and Knowledge Engineering. INDEKS-Verlag. p. 333 - 342.

Ganter, B.; Wille, R. (1996) Formale Begriffsanalyse: Mathematische Grundlagen. Springer-Verlag.

Henriksen, T.; Lindh, B. (1996). http://www.jbi.hioslo.no/publ/bibin/infoterm.htm

Johnson, Eric H. (1995) A Hypertext interface for a Searcher's Thesaurus. http://csdl.tamu.edu/DL95/

Kent, R.; Neuss, C. (1995) Conceptual Analysis of Resource Meta-Information. Computer Networks and ISDN Systems , vol. 27, pages 973–984.

Lenat, D. B. (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38, no. 11.

Lin, Xia (1997) Map Displays for Information Retrieval. JASIS, 48, 40 - 54.

Miller, George (1995) WordNet: a lexical database for English. Communications of the ACM 38 (11), pp. 39 - 41.

Skorsky, Martin (1997). Graphische Darstellung eines Thesaurus. Deutscher Dokumentartag, Regensburg.

Stumme, Gerd (1996) Local scaling in conceptual data systems. In: Eklund, Ellis & Mann (eds.). Conceptual structures: Knowledge representation as interlingua, Vol. 1115 of LNAI. Springer-Verlag, Berlin-Heidelberg, 121-131.

Viehmann, Viola (1996). Formale Begriffsanalyse in der bibliothekarischen Sacherschliessung. Master's thesis, TH-Darmstadt, 1996.

Vogt, F; Wille, R. (1995). TOSCANA – a graphical tool for analyzing and

exploring data. In: Tamassia; Tollis (eds.). Graph Drawing. Springer-Verlag, Heidelberg, 226-233.