Uta Priss
School of Library and Information Science, Indiana University Bloomington

Comparing Classification Systems using Facets

Abstract: This paper describes a qualitative methodology for comparing and analyzing classification schemes. Theoretical facets are modeled as concept lattices in the sense of formal concept analysis and are used as 'ground' on which the underlying conceptual facets of a classification scheme are visually represented as 'figures'.

1. Introduction

Classification schemes can be compared using quantitative criteria, such as number of classes, number of entry terms, number of cross-references and so on. They can also be compared using qualitative criteria, such as whether they are pre- or post-coordinated or whether or not they are faceted. But it is more difficult to identify the main principles of organization that underlie a scheme. For example, what precisely are the differences in organization between DDC, LC, Roget's Thesaurus, Yahoo, and WordNet? A model for representation of structural principles of classification schemes would facilitate a detailed qualitative comparison of classification schemes and would provide information about applicability and usefulness of schemes even prior to implementation and usability testing. Schemes that are similar in their organization should be similar in their performance. If a user's requests or a typical application demand are modeled in such a representational model then it would be possible to match schemes to users or applications. Identifying and visualizing principal characteristics of classification schemes can also provide valuable insight into the shared knowledge of a society or culture. Synchronic differences among cultures that coexist at the same time and diachronic changes within a single culture over time could be investigated.

A qualitative representation of a complete classification scheme would be very complex. To reduce the complexity of the representation, facets can be identified that apply to a scheme. For example, the top 1000 categories of the DDC can be represented as a combination of 9 facets, such as discipline, aspects of disciplines, type of document, time and space. It should be noted that a systematical application of these facets reduces the number of classes to about 500 or less instead of 1000. The hierarchy of a scheme implies a citation order. A classification scheme can thus be represented by a set of facets and their citation order - even if the system is not explicitly faceted. An obvious difference between the DDC and Roget's Thesaurus is, for example, that the primary facet in the DDC's citation order is 'discipline' whereas Roget's primary facet is a more philosophical one that is loosely based on Aristotle's categories. Yahoo also uses 'discipline' as a primary facet but other facets are mixed with it. Single facets of different classification schemes can then further be compared according to their details in arrangement and selection. For example, the arrangement of classes in the 'discipline' facet differs between DDC and Yahoo. The qualitative methodology proposed in this paper is subjective because underlying facets are assumed that are not explicitly there. But the transparency of the results visualizes underlying principles of classification schemes and opens them for critical analysis. A second benefit of the methodology is that a better understanding of classification schemes will improve the design of information retrieval systems.

2. Methodology: Formal Concept Analysis and Facets

Formal concept analysis (Ganter & Wille, 1999) is employed in this paper as the methodology for visualizing facet structures. In formal concept analysis every concept (or class) is uniquely defined via its extension, which is the set of objects to which it refers, or via its intension, which is the set of attributes, characteristics or features that describe the concept. Intensional and extensional definitions are equivalent which means that an extension corresponds to exactly one intension and vice versa. The concepts (or classes) are ordered according to their conceptual inclusion. That means that a concept A is a subconcept of concept B if the extension of A is contained in the extension of B or - and this is an equivalent condition - the intension of A contains the intension of B. For example, the extension of 'poodle' is contained in the extension of 'dog' and 'poodle' contains all features of 'dog' but may have additional ones. Formal concept analysis is thus a perfect model for traditional classification schemes: classes are ordered according to their inclusions, they are precisely defined and have fixed boundaries because objects either belong or do not belong to a class.

A major difference between formal concept analysis and traditional classification schemes is that concept hierarchies are mathematical lattices instead of being restricted to tree hierarchies. The mathematical formalization using extensions and intensions (or objects and attributes) and their duality to each other directly points to lattices as the most appropriate mathematical structure. Tree hierarchies are special types of lattices - if a shared lowest class that represents 'contradiction' and is empty is formally added - but not every lattice is a tree hierarchy. Polyhierarchies (or partially ordered sets) are not usually lattices but for every polyhierarchy there exists a unique smallest lattice in which the polyhierarchy can be embedded. Lattices can thus be thought of as polyhierarchies of concepts (or classes) that have some additional concepts. The additional concepts ensure that every set of concepts has a unique common superconcept and a unique common subconcept. The unique common superclass can be the top in the hierarchy, which means it represents 'universality or 'anything', and the common subclass can be the bottom in the hierarchy, which represents 'contradiction' or 'nothing'.

Lattices have more structure than polyhierarchies and are in many ways less 'disorganized'. Some computer scientists have recognized this and model their object-oriented class hierarchies as lattices. Information specialists on the other hand often do not know that there is a structure that imprints some order and control on polyhierarchies (or 'entangled hierarchies' as they are sometimes called). They prefer using tree hierarchies (such as Yahoo or DDC) which they 'entangle' with numerous cross-references instead of modeling them as lattices without cross-references. Figure 1 shows an example of a concept lattice. The formal attributes are 'contains articles', 'serial publication', 'contains entries' and 'contains bibliographic entries'. The formal objects are types of documents, such as 'encyclopedia' or 'journal'. The lattice demonstrates that a tree hierarchy would be fairly difficult to obtain for classifying types of documents without separating types that belong together. The lattice further demonstrates that any classification scheme is subjective and context dependent. Choosing different attributes would result in a completely different hierarchy.

Facets are viewpoints or aspects of classification schemes. Originally invented by Ranganathan (1962) they facilitate the modularization of a hierarchy into independent parts (Priss & Jacob, 1999). Facets are independent of each other because any facet can be combined with any other facet (although not every combination may be useful for every set of documents) and modifying a single facet does not have impact on other facets. Baseline facets describe a small aspect of a domain exhaustively and consistently. They can be combined to form larger facets so that a faceted classification scheme can be constructed as a hierarchy of facets. There are several methods of combining facets (Priss & Jacob, 1999) but they are not discussed in this paper. In this paper only single facets are considered. Using formal concept analysis, every facet is represented as a concept lattice. The example in figure 1 is a facet of 'types of documents'. Facets can be constructed using data-driven or theory-driven methods. In formal concept analysis, a data-driven concept lattice is generated by identifying a set of objects, a set of attributes and a relation among them. The resulting conceptual hierarchy should then be analyzed with regard to its relevance for a domain. That means that the conceptual relations that result from the data-driven approach must be valid in terms of the theoretical domain knowledge. For theory-driven facet construction, a polyhierarchy of concepts can be created according to domain knowledge and can be embedded into a concept lattice. A theory-driven lattice can then be verified by selecting typical objects and attributes from the domain and testing whether they can be appropriately integrated into the lattice.

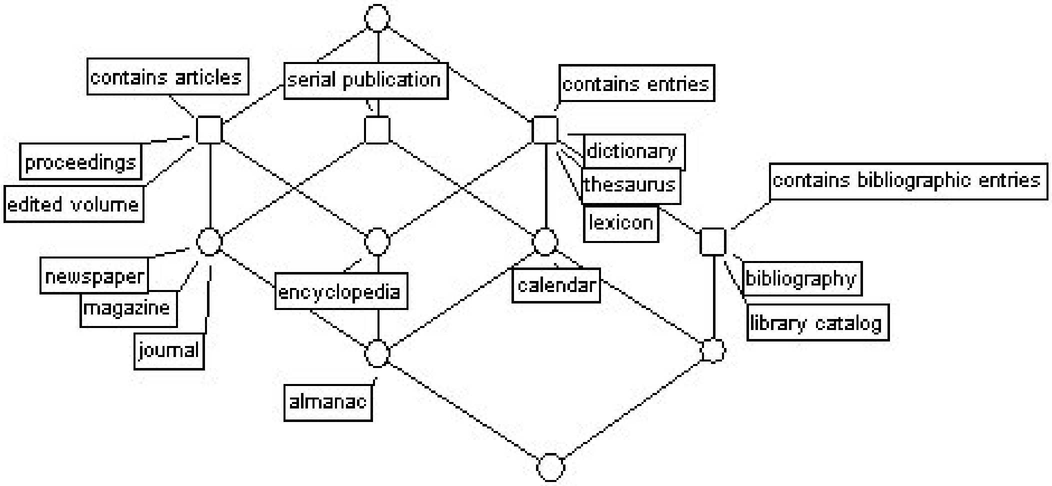

Figure 1: A concept lattice of document types

3. Application: A Facet for Disciplines in DDC and Yahoo

Data-driven lattices can be constructed for classification schemes by exploiting cross-references. In the DDC, for example, the relative index can be explored. The relative index shows close relationships between topics that may be far apart in the tree hierarchy. For example, ‘650 business’ is under ‘600 technology’ and thus far apart from ‘380 commerce’ and ‘330 economics’, which are under ‘300 the social sciences’. On the other hand, in the relative index under ‘business’ there are cross-references to 338, 322, and 368 and under ‘commerce’ there are cross-references to 350 and 658. It is thus obvious that there is a connection between the 650’s and the 320-380’s, which is not apparent in the hierarchy. These are the cases where patrons potentially need to walk long distances in the library building to retrieve closely related documents. Tinker et al (1999) provide further examples and describe a computer interface that facilitates simultaneous browsing through several facets of the DDC. Such a system could solve the problem and collocate related branches of the tree hierarchy according to the relative index.

A drawback of a lattice based on cross-references of the DDC is that the cross-references in the index represent different types. For example, ‘200 religion’ relates to other topics in different relationships: religion can influence scientific beliefs and be influenced by the observation of natural phenomena as evidenced in the link between ‘religion’ and ‘astronomy’. Religion can influence the ethical foundation of topics such as ‘politics’ and ‘education’. The artifacts of a society are often influenced by religious beliefs and often provide evidence of religious development, which is apparent in the links between ‘religion’ and ‘arts’ and ‘folklore’. Historical writings (‘history’) can shed light on the development of religious ideas. According to the relative index, ‘religion’ is thus strongly connected to other disciplines. But since the connections are of different types (‘influences’, ‘is influenced by’, ‘provides evidence for’) an automatically generated concept lattice from the cross-references would be entangled and may not pass the test mentioned above that a data-driven lattice should reflect the theoretical domain knowledge. To explore cross-references, sophisticated natural language processing software would have to be employed that distinguishes the different types of cross-reference links.

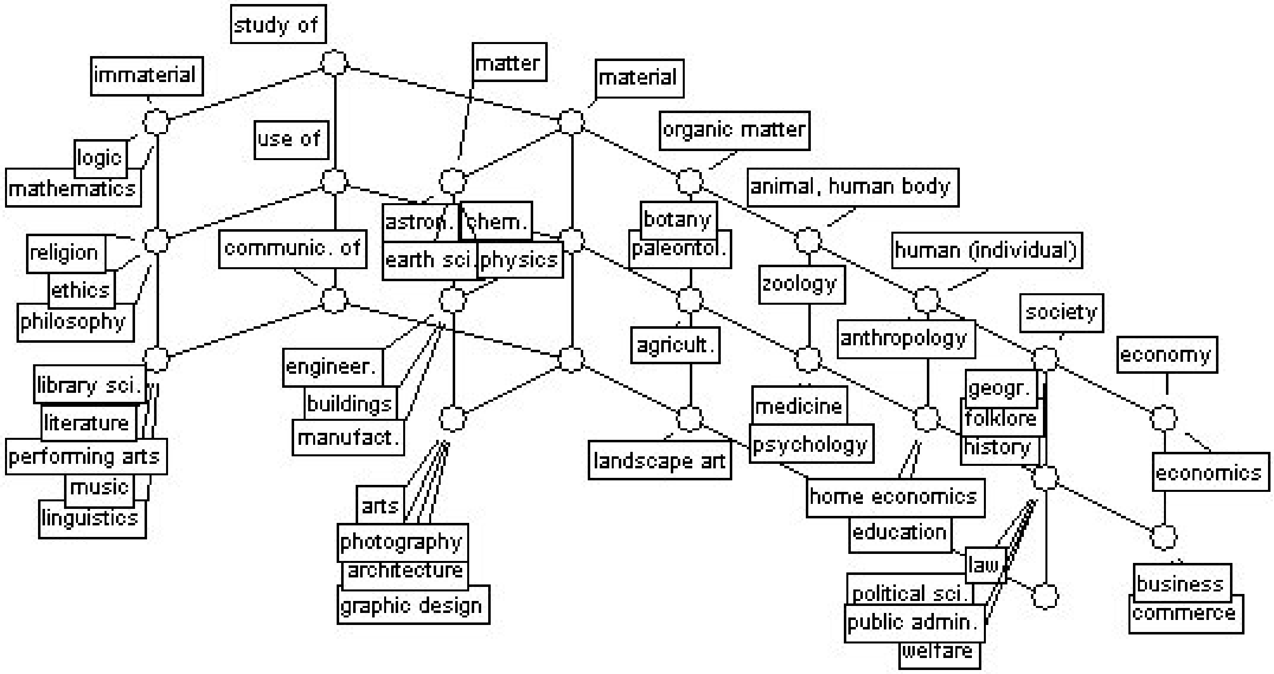

Figure 2: A facet for disciplines

A different approach is thus to construct a theory-driven scale as a 'ground' and to represent a classification scheme as 'figure' on that ground. The examples in figures 2 to 4 show a facet of disciplines as ground. The facet is probably best visible in figure 3. The facet is constructed as a combination of a three concept chain 'study of', 'use of' and 'communication of' together with a variation of the classical Tree of Porphyry that separates 'immaterial' from 'material', matter from organic matter, animal from plant (organic matter), and human from animal. Society and economy are modern additions. Figures 2 and 3 show the disciplines of the DDC in context of this ground facet. Some of them are difficult to place but mostly there is a strong congruence between the ground facet and the DDC classes because, as visualized in figure 3, the classes of the DDC correspond to the pattern of the lattice. The sciences are under 'study of' but above 'use of'. The 600's (technology) are under 'use of' but above 'communication of'. Arts and humanities are under 'communication of'. The social sciences are under 'society'. There are some obvious differences between DDC and the ground facet: 'psychology' is far apart from other 100 level classes, which means that the 'reasoning' aspect that combines logic and psychology is not represented. The distinction between 300 and 600 level classes is not always clear. 'Mathematics' is separate from the other sciences. The ground facet brings into proximity some of the classes that are cross-referenced in the relative index, such as business and commerce, architecture and buildings, logic and mathematics, engineering and physics. Some links that are obvious in the relative index are missing in the diagram, such as the connections between religion and other disciplines. This may be due to the different types of cross-references as mentioned above.

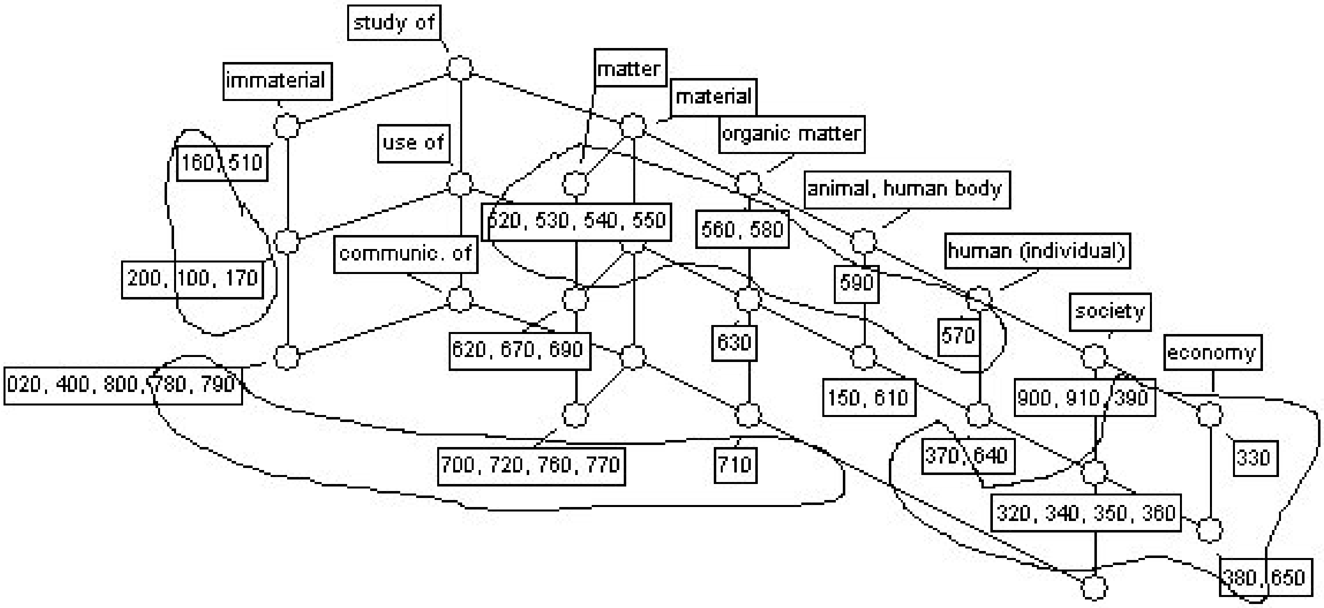

Figure 3: The DDC disciplines related to the ground facet

Figure 4 demonstrates that not all classification schemes are congruent to the ground facet in the example. Yahoo, whose 14 top-level categories are represented as a 'figure' on the ground facet in figure 4, does fit less well into the structure. First, there are several top-level categories that cannot be placed at all, such as 'reference', 'news & media' and 'regional' because they are not disciplines. 'Recreation & sports' cannot be placed because its combining feature is 'having fun', which is not contained in the ground facet. 'Entertainment' and 'computer & Internet' are grouped together because they both refer to immaterial communication - if 0's and 1's are considered in their conceptual instead of their physical existence. But that grouping is awkward. The equivalent to the DDC 100's is completely missing among the Yahoo top-level categories. That part of the ground facets thus has no objects in the lattice.

A conclusion is that Yahoo uses several other facets besides and instead of a discipline facet as underlying top-structures in its hierarchy. Furthermore, its disciplines are arranged differently from the DDC, which follows mostly a traditional classification of disciplines. Although this result may have been intuitively obvious from the start the methodology introduced in this paper provides the result in a formal framework that makes the underlying organizational schemes explicit. The schemes and the suggested ground facet can now be critically analyzed and discussed by identifying the precise attributes that generate the conceptual hierarchies. As explained in Priss (1997), other more practical applications are that documents could be attached to each class and document databases could thus be browsed visually.

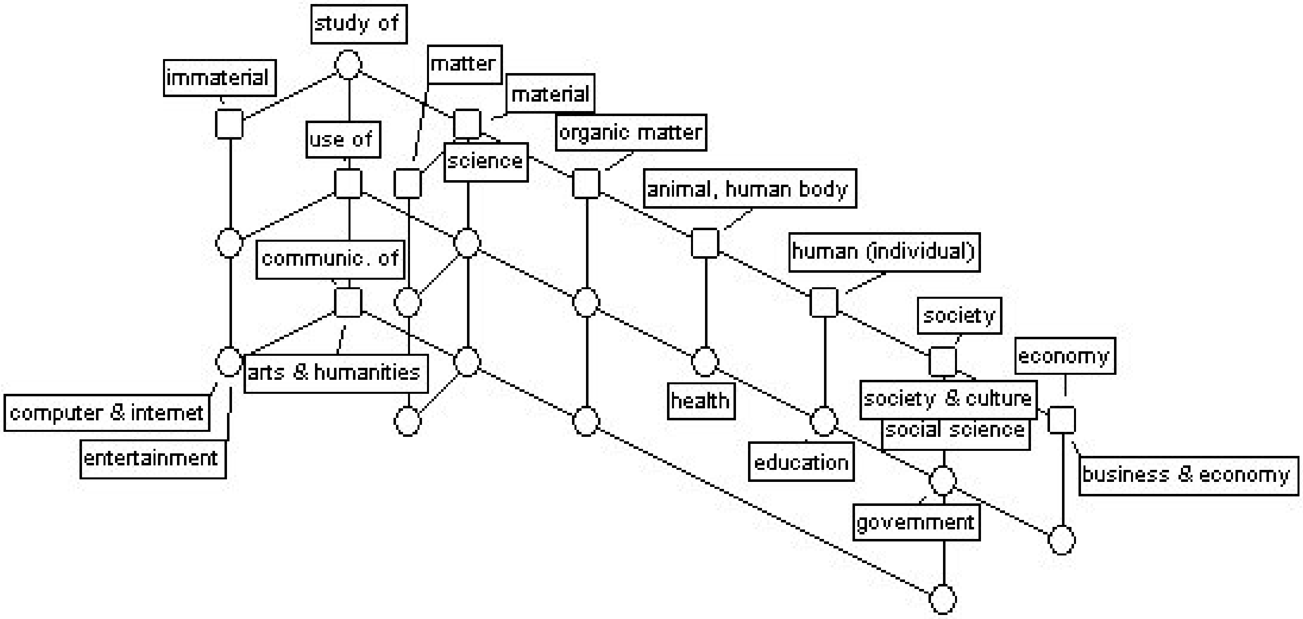

Figure 4: The Yahoo disciplines

4. References

Ganter, Bernhard & Wille, Rudolf (1999). Formal Concept Analysis. Mathematical Foundations. Berlin-Heidelberg-New York: Springer.

Priss, Uta & Jacob, Elin (1999). Utilizing Faceted Structures for Information Systems Design. Proceedings of ASIS'99 Annual Meeting, p. 203-212.

Priss, Uta (1997). A Graphical Interface for Document Retrieval Based on Formal Concept Analysis." In: Santos, Eugene (ed.), Proceedings of MAICS'97. AAAI Technical Report CF-97-01, 1997, p. 66-70.

Ranganathan, S. R. (1962). Elements of library classification.} Asia Publishing House, Bombay.

Tinker, A. J.; Pollitt, A. S.; O'Brien, A.; Braekevelt, P. A. (1999). The Dewey Decimal Classification and the Transition from Physical to Electronic Knowledge Organization. Knowledge Organization, 26(2): 80-96. mineau@ift.ulaval.ca