Concept Neighbourhoods in Knowledge Organisation Systems

Uta Priss, L. John Old

Edinburgh Napier University, School of Computing,
u.priss@napier.ac.uk, j.old@napier.ac.uk

Abstract. This paper discusses the application of concept neighbourhoods (in the sense of Formal Concept Analysis) to Knowledge Organisation Systems. Examples are provided using Roget's Thesaurus, WordNet and Wikipedia categories.

1 Introduction

Concept Neighbourhoods have been used as a means for exploring Knowledge Organisation Systems (KOSs) since at least the 1950s (although using different terminologies at different times). In the 1950s, Margaret Masterman a pioneer of Natural Language Processing and Artificial Intelligence suggested to use Roget's Thesaurus as a tool for representing what she called "semantic transformations" (and what would be called "conceptual structures" in modern terminology) for the purposes of machine translation (Masterman, 1956). Priss & Old (2009) show that algorithms similar to Masterman's have been independently rediscovered over the years, for example, by Dyvik (2004) for constructing dictionaries from a bilingual corpus, and by modern linguistic approaches to lexical databases using Formal Concept Analysis (FCA, a mathematical theory for representing conceptual hierarchies and classification systems, developed by Ganter & Wille (1999)).

Masterman saw a strong connection between the fields of machine translation and information retrieval because she argued that both fields require underlying semantic transformations (or conceptual structures) which could in both cases be represented by thesauri. Thus, it is of interest to apply her ideas, which in modern FCA-based terminology are called "concept neighbourhoods" and "neighbourhood lattices" (Priss & Old, 2004) and which so far have mostly been applied to Roget's Thesaurus, to a variety of other KOSs.

This paper provides a brief overview of the FCA technologies for concept neighbourhoods and neighbourhood lattices and their previous applications to an electronic version of Roget's Thesaurus. Previous research has shown that the formation of concept neighbourhoods depends on the structure of the underlying resources (Priss & Old, 2005). For example, Dyvik's (2004) method appears to be more successful when applied to a bilingual corpus than when applied to a bilingual dictionary because a corpus records more shades of meanings than a dictionary (Priss & Old, 2005). It is of interest to apply similar techniques to other KOS such as WordNet (Fellbaum, 1998). Even though Roget's Thesaurus and WordNet both contain a hypernymic hierarchy and

approximately the same number of words, these techniques require some modifications in order to be applicable to WordNet. Library classifications again pose a different challenge because they tend to be predominantly tree-like. On the other hand, a library classification scheme, together with a large set of documents that has been classified with the scheme, does represent structures similar to Roget's Thesaurus. Yet other techniques apply to ontologies, word association data and semantic networks.

The claim of this paper is that the extraction of concept neighbourhoods presents a useful tool for many types of KOSs, but different types of KOSs require subtle modifications to the algorithm that is used for deriving the neighbourhoods. This paper starts with an introduction to the theoretical background of concept neighbourhoods and an overview of previous applications. It then presents applications of these algorithms to WordNet and Wikipedia categories as examples of KOSs with slightly different structures. It is intended to extend this research to other KOSs in the future.

2 Concept Neighbourhoods and Neighbourhood Lattices

Formal Concept Analysis (FCA) is a mathematical theory for representing conceptual hierarchies and classification systems developed by Ganter & Wille (1999). FCA has many applications in library and information science as documented in the review by Priss (2006). Using FCA for KOSs is attractive because FCA provides a natural formalisation of hierarchical structures, called "concept lattices". Such concept lattices are derived from binary data tables called "formal contexts", where one set is called "formal objects" and the other set "formal attributes". Examples of formal objects and attributes in KOS applications are: documents and index terms or classes; terms and documents; words and their translations into a different language; lower and higher classes in a classification.

Concept lattices can be represented as browsable and expandable hierarchies (similar to how files and directories are often displayed in computers) or as graphs, which can have interactive, web-based interfaces. An example of a graphical representation of a concept lattice is shown in Figure 1. Each box represents a "formal concept". The formal attributes are written in the top half of the boxes and the formal objects in the bottom half. The extension of a concept is calculated by retrieving the objects that belong to a concept and all concepts below it. For example in Figure 1, the concept that has the word "slumber" as formal object has "sleep, slumber" as its extension. The intension of a concept consists of the attributes of a concept and all concepts above it. Concept lattices can represent tree hierarchies; but, as in Figure 1, they can also represent poly-hierarchies where each concept can have multiple parent nodes.

A concept neighbourhood is extracted by starting with an object and then retrieving all attributes that are related to the object, and so on. Or, alternatively, it starts with an attribute, retrieving all related objects, and so on. The data of a concept neighbourhood is stored in a formal context and can be displayed as a concept lattice. In some cases the process of building concept neighbourhoods can continue until it includes a complete KOS; in other cases it stops sooner. Once the process stops, a "neighbourhood closure" is formed. Usually the neighbourhood closures follow a power law distribution where one large neighbourhood closure includes most of the items; a large number of items

are isolated (i.e. form neighbourhoods of exactly one concept) and the remaining items are distributed linearly, logarithmically in between. This has been shown for Roget's Thesaurus by Priss & Old (2006).

Because neighbourhood closures are frequently too large to be useful, the process of building a neighbourhood is usually stopped after a predetermined number of steps. Alternatively, it is also possible to reduce a concept neighbourhood, after it has been calculated, by applying certain restrictions that limit a neighbourhood to its core connectivity (Priss & Old, 2004). A variety of software tools exist that facilitate the automatic extraction of neighbourhoods from a database. But in order to use these tools with KOSs that have not been previously modelled in this manner, it needs to be manually determined which combinations of database fields and which restriction algorithms yield the most interesting neighbourhoods. This changes with each type of KOS.

Concept neighbourhoods in Roget's Thesaurus have been studied in great detail (Priss & Old (2004), Priss & Old (2006)). Figure 1 shows an example of a concept neighbourhood for the word "sleep" in Roget's Thesaurus. The formal attributes in this example are senses from Roget's Thesaurus (e.g. "710:2:1" represents category 710, paragraph 2, synset 1). The formal objects are words that share at least one sense with "sleep" in the thesaurus. The lattice was generated using the interface available at www.roget.org.

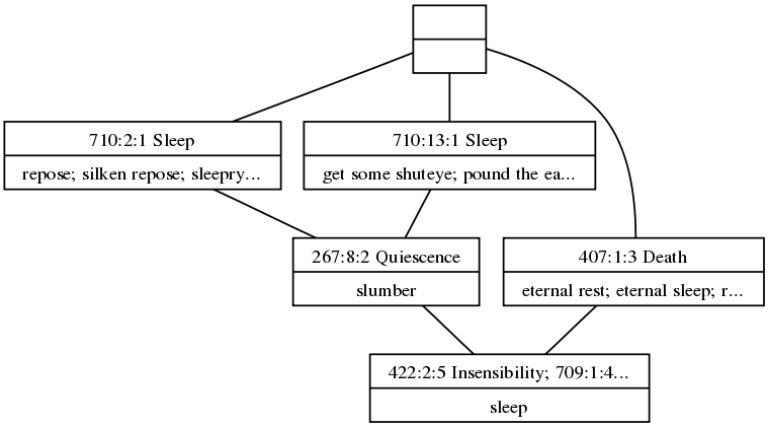

Fig. 1. A neighbourhood lattice for "sleep" in Roget's Thesaurus

Similar methods have also been used with other KOS. As mentioned above, Dyvik (2004) uses a method which is similar to neighbourhood lattices but not formulated in terms of FCA (Priss & Old, 2005). His "semantic mirrors" method automatically derives thesaurus entries from a bilingual corpus. His method finds the translations of a word, and other words which have the same translations. The resulting sets are then structured into lattices. Basili et al (1997) and Cimiano et al (2003) derive ontological or taxonomical structures from a corpus by building what are essentially concept neighbourhoods around words based on the words' syntactic frames in the corpus.

3 Concept Neighbourhoods in WordNet and Wikipedia

The methods used for building concept neighbourhoods in Roget's Thesaurus should be of interest for other lexical databases, such as WordNet. Although WordNet and Roget's Thesaurus both contain a hypernymic hierarchy, WordNet differentiates many semantic relations and its grouping of words into sets of synonyms (called "synsets") is more finely grained. Figure 2 shows the hypernymy relation of the word "sleep" in WordNet. It can be seen that "sleep" has six senses (the lower level in the figure), which all have different hypernymic synsets except for the first two which share their hypernymic synset. If one follows this hierarchy further up or down it forms mostly a tree. There seem to be implicit relationships which are not presented. For example, the word "rest" occurs both in a hypernymic and a hyponymic synset. The second synset from the left ("sleep", "slumber") is a subset of the fourth synset which also indicates an implicit relationship.

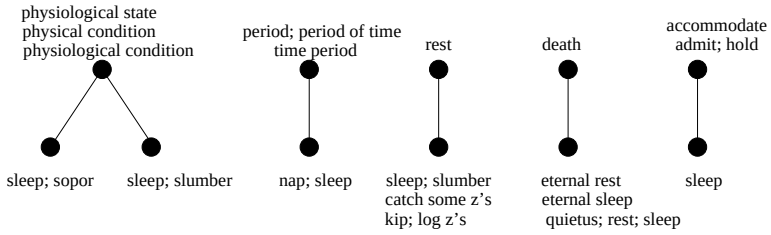

Fig. 2. WordNet's hypernymy relation

A neighbourhood lattice of the relation in Figure 2 can be formed as follows: the words of the hyponymic synsets are considered formal objects; the words of the hypernymic synsets are formal attributes. The hypernymy relation is then treated as a relation between words instead of synsets. The resulting neighbourhood lattice is shown in Figure 3. Most of the original synsets are maintained as extensions and intensions of the concepts. The only real difference between the synsets in Figures 2 and 3 is the synset "sopor, sleep" that is represented by "sopor, sleep, slumber" and the intension of the bottom concept in Figure 3 which contains all formal attributes. But that is not a problem because the bottom concept in a lattice represents a sort of "null" concept. Thus overall the synsets from Figure 2 are maintained in Figure 3 but Figure 3 shows additional structure and connectivity. The word "rest" still occurs in two different unconnected concepts in Figure 3, but that could be changed by adding an identity relation to the formal context (which is not further discussed in this paper).

A comparison between the neighbourhoods for sleep in WordNet (Figure 3) and Roget's Thesaurus (Figure 1) shows some remarkable similarities. Some of the synsets and concepts contain the same words ("Death: eternal rest, eternal sleep" and "sleep, slumber"). This might be partly influenced by the fact that Roget's Thesaurus was used as a source by the creators of WordNet.

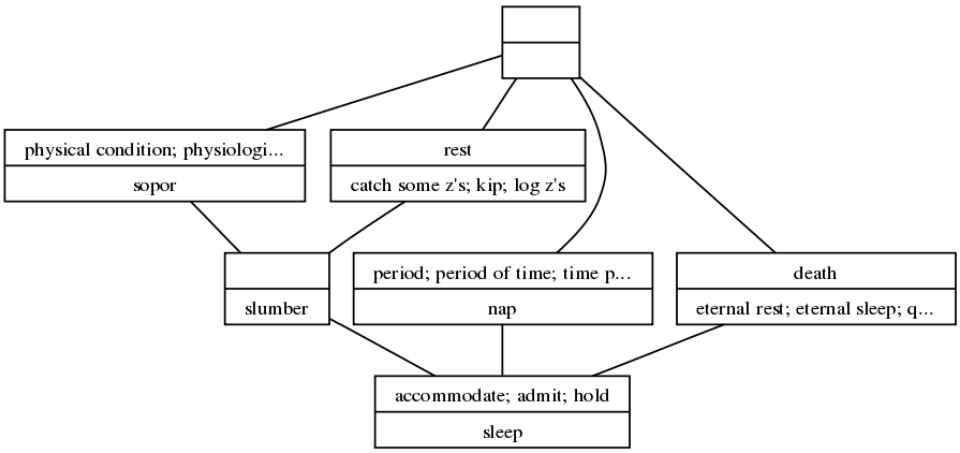

Fig. 3. A neighbourhood lattice for “sleep” in WordNet

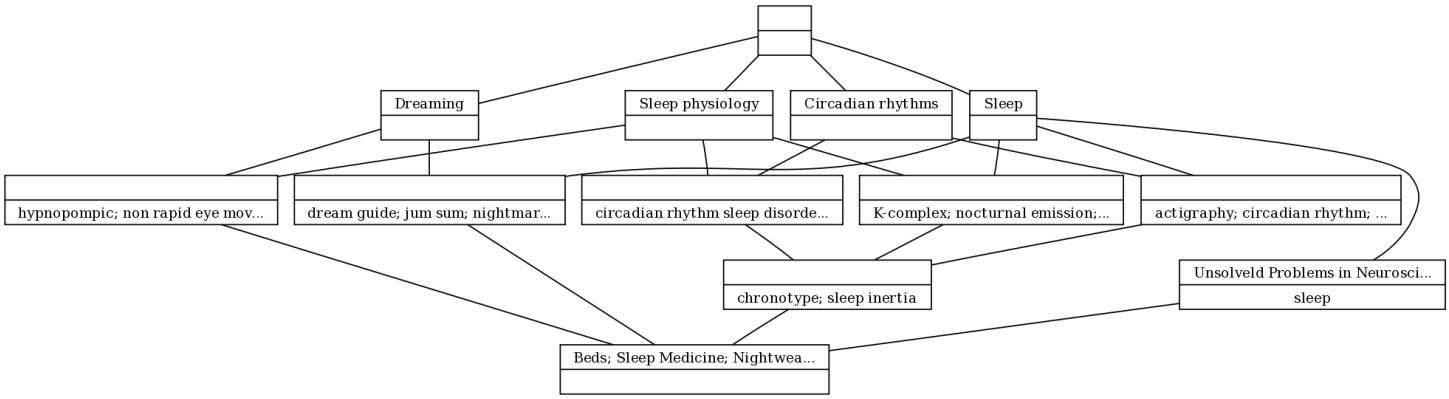

Fig. 4. A neighbourhood lattice for “Sleep” categories in Wikipedia

A similar approach can also be used for other KOS. Figure 4 shows a neighbourhood lattice for "sleep" in Wikipedia. This neighbourhood was started by selecting the category "sleep" and its direct subcategories as formal attributes. The formal objects are all pages that occur in at least two of the selected categories. Because the categories in Wikipedia are designed in an ad-hoc manner (any user can create categories and add pages to categories), it is not surprising that the resulting lattices can be complex and "messy". The lattice shows implicit relationships that could be used to improve the category hierarchy.

4 Conclusion

This paper demonstrates that concept neighbourhoods which the authors previously only applied to Roget's Thesaurus can also be applied to other KOSs such as WordNet and Wikipedia. An on-line interface for exploring Roget's Thesaurus in this manner is available at www.roget.org. The authors intend to add a similar WordNet interface to that same site in the near future. It will be more difficult to add an interface for Wikipedia because of the size and complexity of that resource. More research is needed to generate concept neighbourhoods for other KOSs.

References

1. Basili, R.; Pazienza, M.; Vindigni, M. (1997). Corpus-driven unsupervised learning of verb subcategorization frames. AI*IA-97.
2. Cimiano, P.; Staab, S.; Tane, J. (2003). Automatic Acquisition of Taxonomies from Text: FCA meets NLP. Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, p. 10-17.
3. Dyvik, H. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. Language and Computers, Vol. 49, iss. 1, Rodopi, p. 311-326.
4. Fellbaum, Christiane (ed.) (1998). WordNet - An Electronic Lexical Database. MIT Press.
5. Ganter, Bernhard, Wille, Rudolf (1999). Formal Concept Analysis. Mathematical Foundations. Berlin-Heidelberg-New York: Springer.
6. Masterman, Margaret (1956). Potentialities of a Mechanical Thesaurus. MIT Conference on Mechanical Translation, CLRU Typescript. In: Masterman, Margaret (2005). Language, Cohesion and Form. Edited by Yorick Wilks. Cambridge University Press.
7. Priss, Uta; Old, L. John (2004). Modelling Lexical Databases with Formal Concept Analysis. Journal of Universal Computer Science, Vol 10, 8, 2004, p. 967-984.
8. Priss, Uta; Old, L. John (2005). Conceptual Exploration of Semantic Mirrors. In: Ganter; Godin (eds.), Formal Concept Analysis: Third International Conference, ICFCA 2005, Springer Verlag, LNCS 3403, p. 21-32.
9. Priss, Uta; Old, L. John (2006). An application of relation algebra to lexical databases. In: Schaerfe, Hitzler, Ohrstrom (eds.), Conceptual Structures: Inspiration and Application, Proceedings of the 14th International Conference on Conceptual Structures, ICCS'06, Springer Verlag, LNAI 4068, p. 388-400
10. Priss, Uta (2006). Formal Concept Analysis in Information Science. In: Cronin, Blaise (ed.), Annual Review of Information Science and Technology. Vol 40, p. 521-543.
11. Priss, Uta; Old, L. John (2009). Revisiting the Potentialities of a Mechanical Thesaurus. In: Ferre; Rudolph (eds.), Proceedings of the 7th International Conference on Formal Concept Analysis, ICFCA'09, Springer Verlag.