Unlocking the Semantics of Roget's Thesaurus Using Formal Concept Analysis

L. John Old

School of Computing, Napier University, j.old(at)napier.ac.uk

Abstract. Roget's Thesaurus is a semantic dictionary that is organized by concepts rather than words. It has an elaborate implicit structure that has not, in the 150 years since its inception, been made explicit. Formal Concept Analysis (FCA) is a tool that can be used by researchers for the organization, analysis and visualization of complex hidden structures. In this paper we illustrate two ways in which FCA is being used to explicate the implicit structures in Roget's Thesaurus: implications and Type-10 chain components.

1 Introduction

Like The Bible and Shakespeare, Roget's Thesaurus, for English speakers, is a cultural artefact. School children are taught how to use it at school and it is found on educated English speaker's bookshelves. In real life it is mainly used for crossword puzzles, for finding synonyms to avoid repetition in written work, or to find out what a word means by viewing the company it keeps in the Thesaurus. Whatever its use, it is acknowledged to be a rich source of ”meaning.”

Roget's Thesaurus has been studied or used for the automatic classification of text, automatic indexing, natural language processing, word sense disambiguation, semantic classification, computer-based reasoning, content analysis, discourse analysis, automatic translation, and a range of other applications. This research has involved mainly the American edition, or Roget's International Thesaurus (RIT), and usually the 3rd Edition (Berrey, 1962). Roget's Thesaurus was also used as the basis for WordNet (Miller, G., Beckwith, Fellbaum, Gross, Miller, K., & Tengi, 1993), the electronic model of the mental lexicon.

For researchers the dream of actually capturing and utilizing the semantics, or meaning, of the word associations in Roget's has been elusive. In part this has been because of a lack of a visualization method that would allow the analysis and insights that would unlock the ”inner structure” (Sedelow, W. A. Jr., 1990). Formal Concept Analysis (Wille, 1982) has the ability to automatically classify and arrange information while retaining the complete data, to produce graphics of lattices (Hasse diagrams), and to make relational structures explicit. This gives researchers, we believe, the tool to unlock the inner structure of words and senses in Roget's Thesaurus.

2 The Structure of Roget's Thesaurus

Roget's Thesaurus's organizational structure is a classification tree, or conceptual hierarchy. At the lowest level are what is commonly known as synonyms. The explicit structure of the book consists of three main parts. Following the front matter is the top level of the hierarchy represented by the tabular Synopsis of Categories. This is followed by the body, or Sense Index of the book, which continues the hierarchy down to the lowest level. The Sense Index lists the 1,000 or so Categories (also called headwords, or lemmas, by some researchers) representing the notions found at the most detailed level of the ”Synopsis.” Categories generally occur in pairs as opposed notions, or antonyms. Each Category contains the entries[1]—instances of words ordered by part-of-speech and grouped by sense, or synset[2] (Miller et al., 1993). Synsets exist in groups of senses within a Category, so an index such as 227:1:1 is used, where 227 is the Category; the second number indexes the sense group or *Paragraph* within the Category; and the third number represents the sequence number of the synset, or sense, within that group. At the back of the book is the Word Index, listing the words in alphabetic order, along with their senses ordered by part- of-speech. The senses are represented in the Word Index as references to locations in the Sense Index. On any particular page of the Sense Index the relations of synonymy and antonymy can be seen. By tracing a word out from its synonyms, cross-references, or antonyms to its distant neighbours, all facets of the meaning or semantics of a word can be derived.

The explicit structure of Roget’s thesaurus is evident to any reader. The implicit, hidden, or inner structure is not. The relations between instances of a word located in separate parts of the Sense Index, or senses located at separate places in the Word Index, can be made explicit only by automated means. Formal Concept Analysis (FCA) is a natural fit for both analysis and visualization of Roget’s Thesaurus. Section 3 describes a formalization of Roget’s Thesaurus. Sections 4 and 5 illustrate examples of the application of FCA to words and senses from Roget’s International Thesaurus, 3rd Edition (1963).

3 Formalizing Roget’s Thesaurus with FCA

Several researchers have used so-called neighbourhood lattices to visualize parts of Roget’s thesaurus. The original formalization was suggested by Wille in an unpublished manuscript. Priss (1996) defines neighbourhood lattices as follows:

Instead of using the prime operator ($'$), the plus operator ($+$) retrieves for a set of objects all attributes that belong to at least one of the objects. Formally, for a set G_1 of objects in a context (G, M, I), $\iota^+(G_1) := \{m \in M \mid \exists_{g \in G_1} : gIm\}$. Similarly $\varepsilon^+(M_1) := \{g \in G \mid \exists_{m \in M_1} : gIm\}$ for a set M_1 of attributes. If two plus mappings are applied to a set G_1 it results in a set $\varepsilon^+\iota^+(G_1)$ (with $\varepsilon^+\iota^+(G_1) \supseteq G_1$) which is called the *neighbourhood* of G_1 under I. A neighbourhood of attributes is defined analogously. A neighbourhood context is a context whose sets of objects and attributes

[1] An entry is a particular sense of a particular word. In this way Synset 227:1:1-Covering contains one of the twenty-two senses of the word *over*. Synset 36:13:1-Superiority contains another instance of *over*. The two occurrences of *over* are two separate entries in Roget’s Thesaurus; but only one *word*.

[2] When referring to the sets of synonyms, the term *synset* will be used. When the *meaning* represented by the words in the Synset is referred to, the term *sense* will be used.

are neighbourhoods, such as $(\varepsilon^{+}\iota^{+}(G_1), \iota^{+}\varepsilon^{+}\iota^{+}(G_1), I)$. The resulting lattice is called a neighbourhood lattice.

4 Semantic Containment between Words

A word is called semantically contained in another word if the set of Synsets of the first word is a subset of the set of Synsets of the second word. In this case, the semantically contained word is more specific and *implies* the second. This Section shows an example of semantic containments among words from RIT. A sample of words from RIT, between which semantic containment exists, is given in Table 1. The words on the right are semantically contained in the words on the left. A containment relation is always a true subset, not equal, in order to exclude perfect synonyms (words that share every sense, and only those senses).

Table 1. Semantic containment among some pairs of words

SuperSet	SubSet
3-D	stereoscopic
abandoned	deserted
about	circa
allow	deem
allowance	stipend
bloody	gory
blunt	take the edge off
blush	turn red
brief	short and sweet
calm	tranquil
caustic	escharotic
cave	grotto

The examples in Table 1 are pairs (just twelve of about 10,000). A graph of all semantic containment relations shows an elaborate semantic topology. The semantic containments between words are, likewise, much more elaborate. For example *twaddle*, which has four senses occurring in two Categories, 545: Meaninglessness and 594: Talkativeness, shares these senses with *babble* and *jabber*. *Babble* and *jabber* have other senses in Categories 472: Insanity, Mania, and 578 Language, respectively (amongst others). Furthermore, there are words, such as *gibble-gabble* and *twattle* that are perfect synonyms of twaddle because they have exactly the same senses as twaddle. There are a number of words that share an intermediate number of senses between *twaddle*, and *babble* and *jabber*, such as *prate* and *gibber*. And there are words that share senses with various combinations of the given words, although this example will omit those from the discussion in order to reduce the complexity of the relationships, and the discussion. All of this is visible in the lattice in Figure 1, which was automatically derived from RIT.

The lattice in Figure 1 is not a complete neighbourhood lattice of any of the words. There are a further 51 words involved in the full lattice of Figure 1 if the semantic neighbourhood of *twaddle*, alone, is taken into consideration. Those words, such as *blether, chatter, palaver, gush, spout*, and *gab* form more-complex relationships where, for example, pairs of words can form semantic containments. If all of the words from the eighteen senses in Figure 1 were included, 93 words in all would be pulled in as neighbours. Only an automated method, as is facilitated by FCA, can deal with this type of semantic complexity.

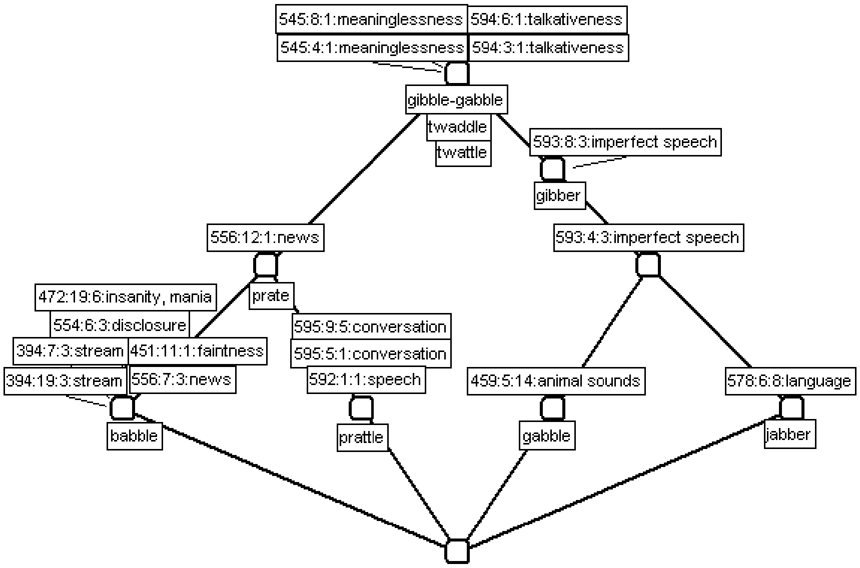

Fig. 1. A lattice showing semantic containment in RIT

5 Semantic Components in Roget's Thesaurus

This Section examines the automatically derived *and implicit* Type-10 chains and Components of the mathematical model of abstract thesauri, of which Roget's Thesaurus is one instantiation, developed by Bryan (1991). The elements in this model are word-strings – which may be single words, compound words, or phrases – and senses (sense definitions, or Synsets), and a relation between them. Bryan defined a series of chains linking the entries by word associations, sense associations, or both. The most restrictive, the Type-10 chain, is a double chain, which requires at least two words to share a sense or two senses to share a word (dubbed a *Quartet*), in order to participate in the chain. This ensures that links are not arbitrary, as happens when two senses are linked by homographs (identical spellings but with disjoint meanings) such as lead (the metal) and lead (to command).

Talburt and Mooney (1990) derived all possible Type-10 chain Components from the 200,000 entries in RIT and found that the majority of semantically strong connectivity formed one large component network of sense-sense and word-word Quartets. This was partitioned from 5,960 lesser component networks; and was dubbed TMC-69 (Talburt-Mooney Component number 69). TMC-69 included 22,242 inter-linked entries. Jacuzzi (1991) reduced this network by applying a further constraint—that a Quartet could not participate in a component if it shared only one entry with that component. TMC-69 was consequently reduced to 2,507 smaller Jacuzzi components, the largest, of 1,490 entries composed from 324 senses and 312 words, was dubbed VJC-184 (V.Jacuzzi Component number 184). The second largest Component was VJC-

1501, with 705 entries. While TMC-69 was a massive inter-connected network of word and sense associations, the resulting derived VJC-184 is a small, but extremely densely bound network—a core of the core connectivities of the semantics of Roget's Thesaurus, and of the English language.

5.1 The Semantics of VJC-184

The most prominent semantic features of VJC-184 emerge not from the numbers but from the Synsets (Synset Category labels) and words. The top most frequent Categories are 855: Excitement, 394: Stream, and 323: Agitation. On first sight, ”stream” appears to be semantically incongruent with *excitement* and *agitation* because it lies between Categories 393: Rain and 395: Channel in the RIT Synopsis of Categories and is certainly a classifier of ”water” words. But the semantic relationship becomes clearer on closer inspection. Some of the 29 words occurring in VJC-184 that are derived from senses in Category 394: Stream, in descending order of frequency, are *flood, gush, run, surge, flow, deluge, rush, race,* and *course*. These words, used in their metaphoric, non-liquid senses are in fact congruent with *excitement* and *agitation*.

Table 2. Top 20 words in RIT by polysemy.

Word	Polysemy	in VJC-184	in VJC-1501
cut	64		x
run	54	x	
set	51	x	
turn	45	x	
head	43	x	
pass	41		
charge	41		
close	39		
line	38	x	
beat	37	x	x
sound	37		
break	36		x
check	36		
discharge	36		x
drop	35		
cast	34		
go	34	x	
lead	34	x	
light	34		
form	34	x	

Table 3. Subcontext showing the connections (via Synset 707:1:1 Haste) between the motion (left) and commotion (right) groups of VJC-184

	Motion	Velocity	Stream	**Haste**	Agitation	Agitation	Activity
	266:2:1	268:10:1	394:17:1	**707:1:1**	323:1:1	323:10:1	705:4:1
bustle				**x**	x		x
flutter				**x**			x
flurry				**x**	x	x	x
rush	x	x	x	**x**			
scamper		x		**x**			
scramble		x		**x**			
scurry		x		**x**			
scuttle		x		**x**			
dash		x		**x**			

The top two senses—those contributing the most words to entries (and therefore the Quartets, and connectivity) in VJC-184—are: 323:1:1 {*fume, bluster, bustle, churn, commotion...* } (of 30 synonyms), and 62:4:1 {*row*[argue], *bluster, bother, brawl...* } (of 28). The overall most frequent words (out of the total of 312 words) begin, in order of frequency: *turn* (30 entries), *course* (20), *run* (18), *clash* (15), *bother* (15)... These words correspond closely to the top twenty words, by polysemy (number of senses the word has), in RIT.

The top (most polysemous, and therefore most frequent) twenty words in RIT (of 113,000 total) are listed in Table 2. The table indicates whether the words are found in VJC-184 or in VJC-1501 (the second largest sub-component of TMC-69). Some of the words are found in other Components. The word ”beat” occurs in both VJC-184 and VJC-1501[3].

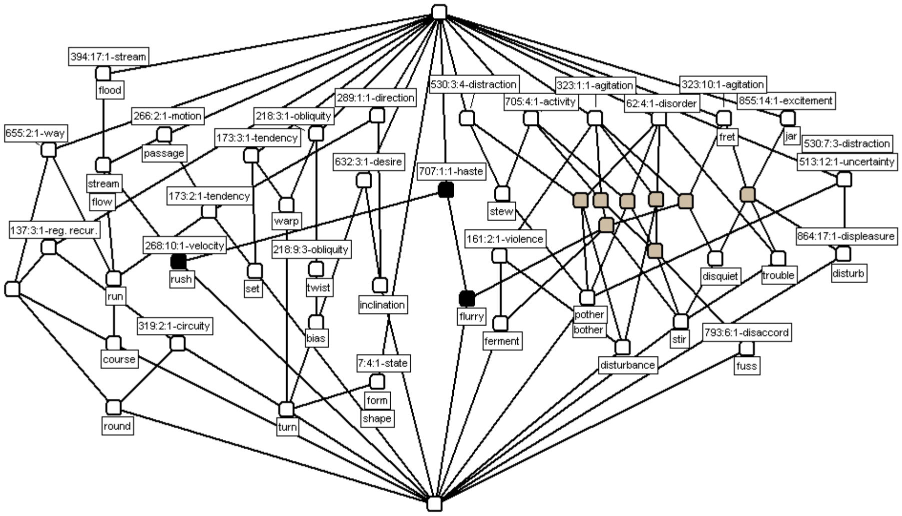

Fig. 2. Lattice of VJC-184 restricted to senses and words with ten or more instances (contributing to at least ten entries in RIT). Also visible are the ”motion” cluster (left) and the ”commotion” cluster (right).

An intuition about the semantics may be gained from listing these most frequent elements of VJC-184, but a different method is necessary to gain insights into the relationships among the elements. Figure 2 is a lattice of VJC-184. To reduce the complexity only words and senses that occur in at least ten entries of Component VJC-184 are included.

There is a clear left-right division within VJ-184, connected in the middle by Synset 707:1:1 Haste (the label above the top black-filled Formal Concept), through the sharing of *rush* (left black Formal Concept) and *flurry* (right black Formal Concept). The left

[3] VJC-1501 is characterized by words such as *cut, crack, hit, bust, gash, split, break*.

collection has semantics characterized by *turn, run, course*/Stream, Motion, Direction. The right hand collection has semantics characterized by *fuss, bother, trouble*/Agitation, Excitement, Disorder. For brevity they will be referred to respectively as the ”motion” and ”commotion” groups, or clusters. The motion and commotion groups of VJC-184 would have been extremely difficult to detect without the aid of a lattice diagram arrangement of the words and senses. Lindig and Snelting’s (1997) work on horizontal decomposition of lattices offers an algorithm solution to identifying such divisions. Table 3 shows how Synset 707:1:1 ties together the motion and commotion groups.

5.2 Unlabeled Concepts

The commotion group (right side of the lattice) displays a feature hidden by any other form of representation. When at least two objects share two attributes in a Formal Concept lattice, but in addition each of the four elements is differentiated by further objects and attributes that are not shared among the other three elements, an ”unlabeled Concept” emerges. These are the Concepts coloured in grey, in Figure 2. While unlabeled Concepts are always rare in lattice representations of semantic data, an entire cluster of unlabeled Concepts is has not been observed elsewhere. The cluster of unlabeled concepts in the commotion group suggests a large number of words with overlapping hues and tones of meaning, discriminated at the edges but not in the center.

Four of the words in the commotion group that contribute to the emergent unlabeled Concepts are *flurry*, *ferment*, *fuss*, and *stir*. These all classify under 705:4:1 Activity, and 3:1:1 Agitation Those words are also classified under other senses, but in different combinations (subsets); or independently of each other, alongside other words. *Flurry* is also found in 707:1:1 Haste alongside *rush*; and *ferment* is found in 161:2:1 Violence, alongside *disturbance*, for example.

Synset 323:1:1 Agitation holds sway over the majority of unlabeled Concepts. Synset 3:1:1 contributes the most entries to VJC-184 (all nouns). 323:1:1 is, however, not the key to the nest of unlabeled Concepts which is further linked to many, many other words and senses omitted from Figure 2, and which also hold this mesh together. If Synset 323:1:1 (or any individual Synset) is removed, other senses such as 161:2:1 Violence, 705:4:1 Activity, 62:4:1 Disorder, and 323:10:1 Agitation (the verb contribution from Category 323) would continue to hold the structure in place. Like a single strand plucked from a spider’s web, the web distorts but mostly holds in place—similarly if words are removed. Further discussion on unlabeled concepts, related to multiple inheritance in class hierarchies, may be found in Godin & Mili (1993).

Some of the dense connections seen in the VJC-184 (and other Components) are comprised of apparently etymologically unrelated words that in fact share common Indo-European roots. Examples from VJC-184 are: *flood*, *fluster*, *flutter*, *flight*, and *flow*, which all derive from the root, PLEU-, meaning ”flow.” *Warp*, *pervert*, *wring*, and *wrench* all derive from the root, WER-3, meaning ”turn, bend”-and there are others. Such ancient etymological threads may explain why some Synsets are so large, and why they interconnect so readily. Etymology alone can’t explain the cluster of unlabeled Concepts, however, as no single Indo-European root pervades that group. The underlying concept perhaps can be explained by proposing that it is an ancient concept at the root of human conceptual organization-not the central source, although it can be traced out to connect to more than 70,000 entries, but one of several primitive concepts possibly more felt than intellectualised, and a facet of consciousness connected to many other areas of thought.

6 Conclusion

We have shown that Formal Concept Analysis is a tool that can make explicit the implicit relationships in complex data. Roget's Thesaurus as an instantiation of what: ”might be accurately regarded as the skeleton for English-speaking society's collective associative memory” (S. Y. Sedelow, 1991, p.108). Insights into this semantic store can have implications for psychology and cognitive science, linguistics, and even anthropology. This will not be possible without the ability to automatically derive and visualize the implications, semantic neighbourhoods and implicit structures among the semantic elements in Roget's Thesaurus. Formal Concept Analysis is a flexible tool capable of facilitating this process.

References

1. Berrey, L. (Ed.). (1962). Roget's international thesaurus (3rd ed.). New York: Crowell.
2. Bryan, R. M. (1973). Abstract thesauri and graph theory applications to thesaurus research. In S. Y. Sedelow (Ed.), Automated language analysis, report on research 1972-73 (pp. 45-89). Lawrence, KS: University of Kansas.
3. Godin, R. & Mili, H. (1993). Building and Maintaining Analysis-Level Class Hierarchies Using Galois Lattices. In A. Paepcke (Ed.), Proceedings of the ACM Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA'93), (pp. 394-410). Washington, DC: ACM Press.
4. Jacuzzi, V. (1991, May). Modeling semantic association using the hierarchical structure of Roget's international thesaurus. Paper presented at the Dictionary Society of North America Conference, Columbus, Missouri.
5. Lindig, C., & Snelting, G. (1997). Assessing Modular Structure of Legacy Code Based on Mathematical Concept Analysis. Proceedings of the 19th International Conference on Software Engineering (ICSE 97), Boston, USA, pp. 349-359.
6. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., & Tengi, R. (1993). Five papers on WordNet. Technical Report. Princeton, N.J: Princeton University.
7. Priss, U. (1996). Relational Concept Analysis: Semantic structures in dictionaries and lexical databases. (Doctoral Dissertation, Technical University of Darmstadt, 1998). Aachen, Germany: Shaker Verlag.
8. Sedelow, S.Y. (1991). Exploring the terra incognita of whole-language thesauri. In R. Gamble & W. Ball (Eds.), Proceedings of the Third Midwest AI and Cognitive Science Conference (pp. 108-111). Carbondale, IL: Southern Illinois University.
9. Sedelow, W. A., Jr. (1990). Computer-based planning technology: an overview of inner structure analysis. In L. J. Old (Ed.), Getting at disciplinary interdependence, (pp. 7- 23). Little Rock, AR: Arkansas University Press.
10. Talburt, J. R., & Mooney, D. M. (1990). An evaluation of Type-10 homograph discrimination at the semi-colon level in Roget?s international thesaurus. Proceedings of the 1990 ACM SIGSMALL/PC Symposium, 156-159.
11. Wille, R., (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, (Ed.), Ordered sets (pp. 445-470). Dordrecht: Reidel.