A combined phrase and thesaurus browser for large document collections

Gordon W. Paynter and Ian H. Witten

Department of Computer Science, University of Waikato, New Zealand.
{paynter, ihw}@cs.waikato.ac.nz

Abstract. A browsing interface to a document collection can be constructed automatically by identifying the phrases that recur in the full text of the documents and structuring them into a hierarchy based on lexical inclusion. This provides a good way of allowing readers to browse comfortably through the phrases (all phrases) in a large document collection.

A subject-oriented thesaurus provides a different kind of hierarchical structure, based on deep knowledge of the subject area. If all documents, or parts of documents, are tagged with thesaurus terms, this provides a very convenient way of browsing through a collection. Unfortunately, manual classification is expensive and infeasible for many practical document collections.

This paper describes a browsing scheme that gives the best of both worlds by providing a phrase-oriented browser and a thesaurus browser within the same interface. Users can switch smoothly between the phrases in the collection, which give access to the actual documents, and the thesaurus entries, which suggest new relationships and new terms to seek.

1. Introduction

Browsing is an important activity in any large document collection (Chang and Rice, 1993). Previous work has shown how a browsing interface to a document collection can be constructed by extracting the phrases that occur more than once in the full text of the documents and structuring them into a hierarchy based on lexical inclusion—a phrase points to longer, and hence generally more specific, phrases that include it (Nevill-Manning *et al*, 1999; Paynter *et al.*, 2000a). The scheme is fully automatic and the phrase structure can be created without any manual intervention. Although it works on a purely lexical basis, it creates and presents a plausible, easily-understood, hierarchical structure for documents in the collection—a structure that conventional keyword queries could never reveal. This technique helps bridge the gap between standard term-based query methods and the more complex topics or concepts that readers employ.

Manually constructed subject thesauri also provide a very useful browsing structure. They provide a topic-oriented arrangement of documents, akin to a standard library subject heading scheme, that will generally be completely different from that described above—and far more soundly based. The thesaurus terms themselves constitute a carefully-constructed controlled vocabulary. Most thesauri identify, for each term, broader and narrower terms, and these permit users to navigate from broad

Figure 1 The FAO on the Internet (1998) collection

groups of items down to more manageable subsets in a well-defined topic-oriented hierarchy. Subject-oriented thesauri have been refined over decades to provide extremely useful browsing structures, and are universally used in all physical libraries—and many digital libraries—as the fundamental basis for the logical and physical organization of library holdings.

Clearly, high-quality subject headings that describe document content should be used wherever they are available, to assist users in their browsing activities. But manually classifying documents according to a thesaurus is expensive. In many digital library or Web-based document collections, subject heading information is unavailable, and infeasible to produce. Machine-readable subject thesauri provide invaluable searching and browsing tools for exploring document collections topically, but documents in digital libraries are rarely tagged with thesaurus metadata, and doing so manually is extremely time-consuming. Automated classification, an active research topic with great promise for the future (Giles, 1998), gives a handle on the problem, but it unlikely to solve it fully.

This paper describes an interface that combines a browsing hierarchy constructed from the full text of a document collection with a completely different hierarchy supplied by a standard subject thesaurus. Users can examine the phrases in the document collection, which give access to the actual documents that contain them. They can also examine the thesaurus terms, which are tagged with information about how often and in which documents they occur. Thesaurus entries suggest new relationships and new terms to seek. The user can switch smoothly between document

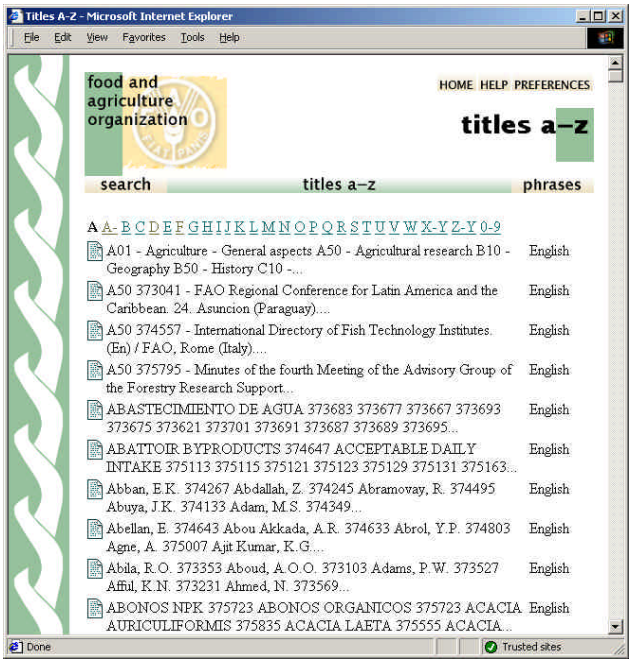

Figure 2 Browsing a list of *Title* metadata

phrases and thesaurus phrases. The result is a combined hierarchical browser based on both thesaurus phrases and all phrases that occur in the document collection.

The structure of the paper is as follows. The next section describes the document collection that we use as an example throughout the paper, and briefly discusses conventional non-hierarchical metadata-based browsing. Following that we describe the phrase interface, which is called Phind for "phrase index," and convey how it feels to browse a collection using it. We then discuss the process of identifying the phrase hierarchy in a document collection. Next we discuss a particular thesaurus that is used as an example, and show how thesaurus entries are presented, along with phrases, in the same interface.

2 Example Document Collection

Figure 1 shows the introductory page of a collection called *FAO on the Internet (1998)*, which forms the principal example used throughout this paper. It contains of the Web site of the Food and Agriculture Organization (FAO) of the United Nations, in a version that was distributed on CD-ROM in 1998. This is not an ordinary, informally-organized Web site. Because the mandate of the FAO is to distribute agricultural information internationally, the information included is carefully controlled, giving it more of the characteristics of a typical digital library collection. With 21,700 Web pages, as well as around 13,700 associated files (image files, PDFs, etc.), it corresponds to a medium-sized collection of approximately 140 million words

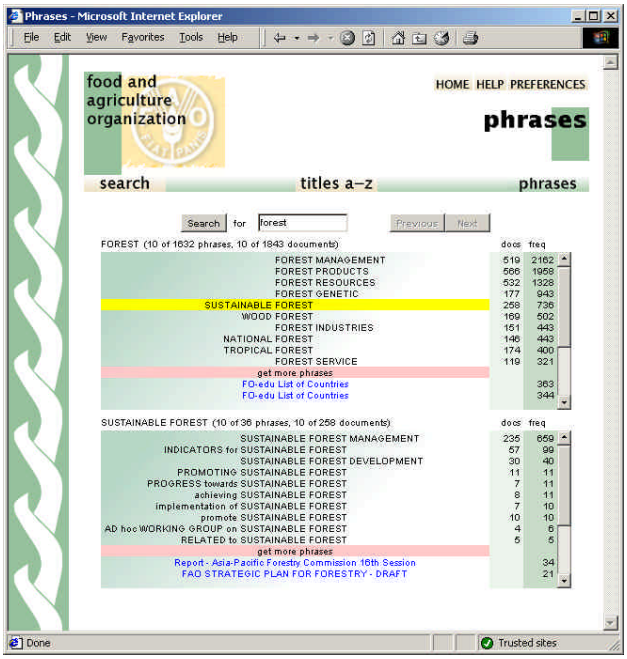

Figure 3 Browsing for information about *forest*

of text. The Web site (*www.fao.org*) has since grown to many times this size, but we use the 1998 version because it was selected by editors at the FAO, and contains no dynamic content.

Figure 2 shows a typical non-hierarchical browsing display, an ordered list of titles broken down by initial letter (*A* has been selected in the Figure) (Witten *et al*, 1999). This ordered list is selected by clicking the *titles a–z* button in Figure 1. However, it does not scale well (Paynter *et al*., 2000a). A user browsing the titles will find far too many to view at once—Figure 2, for example, goes only a very small distance through the *A*s. It is necessary to focus the browsing task, while retaining the simplicity and transparency of the interface presented to the user. Further refinement based on more initial letters is not a satisfactory solution.

3. Browsing the Phrase Interface

Clicking on the *phrases* button in Figure 2 takes users to an automatically-constructed phrase browser that lets them explore the collection according to a hierarchical structure built from all the phrases that occur in the full text of the documents. Unlike the title browsing discussed above, this does scale very well in practice and we have used it on some fairly large (around 0.5 Gb) document collections.

Figure 3 shows the interface in use. It is designed to resemble a paper-based back-of-the-book subject index. The user enters an initial term in the search box at the top.

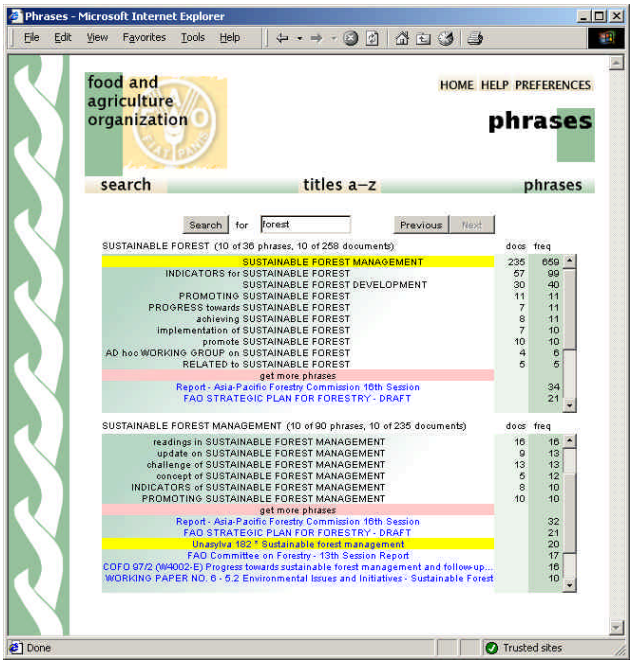

Figure 4 Expanding on *sustainable forest*

On pressing the *Search* button, the upper panel appears. This shows the phrases at the top level in the hierarchy that contain the search term—in this case the word *forest*. The list is sorted by phrase frequency; on the right is the number of times a phrase appears, and preceding that is the number of documents in which it appears.

Only the first ten phrases are shown, because it is impractical with a Web interface to download a large volume of text, and many of the phrase lists are very large. The total number of phrases appears above the list: in this case 10 phrases are displayed of an available 1632 top-level phrases that contain the word *forest*. At the end of the list is an item that reads *Get more phrases* (displayed in a distinctive color). Clicking it downloads a further ten phrases, which will be accumulated in the browser window so that the user can scroll through all phrases that have been downloaded so far.

The lower panel in Figure 3 appears when the user clicks one of the phrases in the upper list. In this case the user has clicked *sustainable forest* (which is why that line is highlighted in the upper panel), causing the lower panel to display phrases that contain the text *sustainable forest*. The text above the lower panel shows that the phrase *sustainable forest* appears in 36 larger phrases, and in 258 documents.

If one continues to descend through the phrase hierarchy, ever longer and more specific phrases will be found. The page holds only two panels, and if a phrase in the lower panel is clicked the contents of that panel move up to the top panel to make way for the phrase's expansion in the lower panel. In Figure 4, for example, the user has expanded *sustainable forest management*, and begun scrolling through its expansions.

The interface not only presents the expansions of the phrase, it also lists the documents in which the phrase occurs. Each panel shows a phrase list followed by a

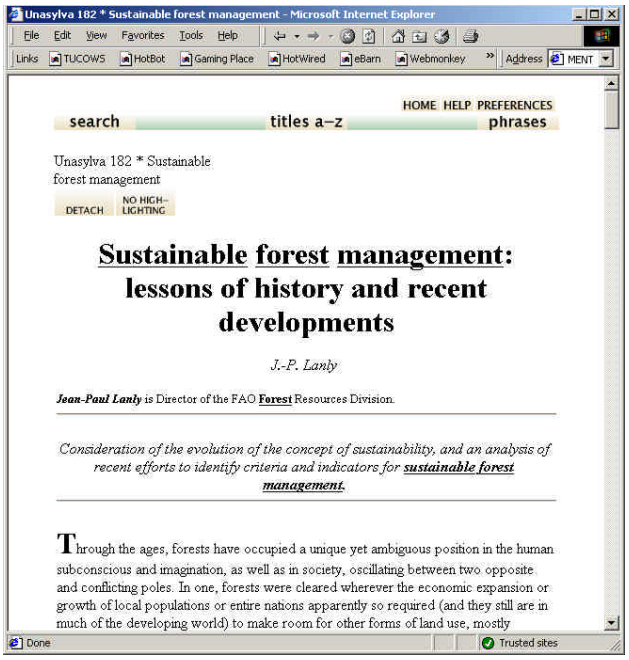

Figure 5 Example Web page

document list. The first ten document titles are loaded immediately, and become visible when the list is scrolled. In the lower panel of Figure 4, the user has scrolled down so that the first six document titles are visible. Document titles are easily distinguished on the screen because they appear in a different color from phrases. On the black-and-white rendition in Figure 4 they are distinguished by the absence of a "document" count, because this, by definition, is equal to 1 for the phrase in question (otherwise the document would appear not under the phrase itself but under an expansion of it.) Only the first ten document titles are downloaded, and (as with phrases) users can *Get more documents* by clicking on a special entry at the end of the list (which would become visible if the panel were scrolled down a little more).

Clicking on a document title opens that document in a new window. In fact, in Figure 4 the user has clicked on *Unasylva 182 * Sustainable forest management*, which highlights the phrase and brings up the page shown in Figure 5. As Figure 4 indicates, that document contains 20 occurrences of the phrase *sustainable forest management*. In Figure 5, each occurrence in the document text has been underlined (this can be turned off by clicking the *no highlighting* button).

4 Identifying Phrases

Underlying the Phind user interface is a hierarchy of phrases that appear in the document collection. For the purposes of this work, a "phrase" is a sequence of words

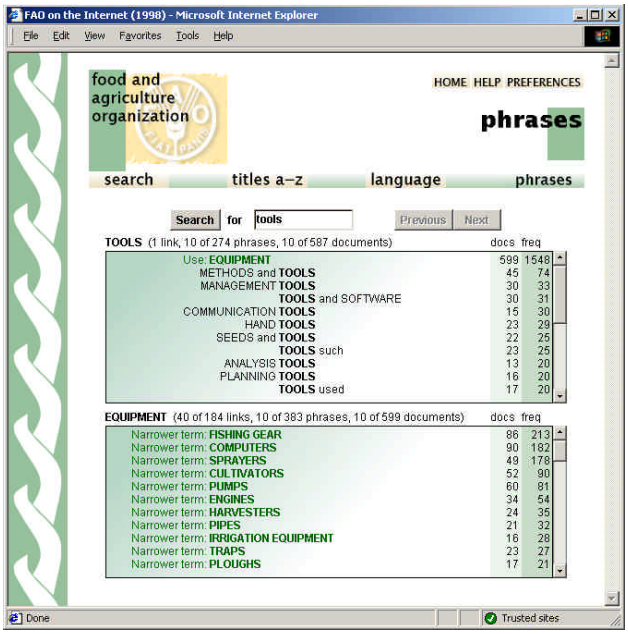

Figure 6 The Phind user interface again

that occurs more than once in the text—that is, we are talking about any phrases that repeat. To include *every* such phrase would clutter the interface with trivial phrases, so we add three further conditions to the definition. Phrases must be *maximal-length*, must not contain *phrase delimiters*, and must begin and end with a *content word*.

Phrases are *maximal-length* sequences if they occur in more than one context, where by "context" we mean the words that flank the phrase where it appears in the text. Phrases that are not maximal-length—ones that occur in a single unique context, in other words ones that are flanked by the same two words wherever they appear—are expanded to encompass that context. In the FAO collection, for example, the phrase *forest industries strategy* occurs only in the longer phrase *forestry industries strategy study*, so the latter term is displayed at the top level of the hierarchy in place of the former. On the other hand, the phrase *sustainable forest* occurs in many different contents—ten examples can be seen in the bottom pane of Figure 3.

If the text were treated as an undifferentiated stream of words, many of the phrases extracted from it would cross syntactic boundaries. To take an extreme example, the last word of one document and the first word of the next are unlikely to form a meaningful two-word phrase. For this reason, we impose the constraint that phrases may not include delimiters. Delimiters are defined as the end of documents, the end of sentences, and any punctuation characters. In practice, we tune the punctuation rule to account for common (and language-dependent) usage: in English, for example, neither the apostrophe in *don't* nor the hyphen in *language-dependent* are interpreted as phrase boundaries.

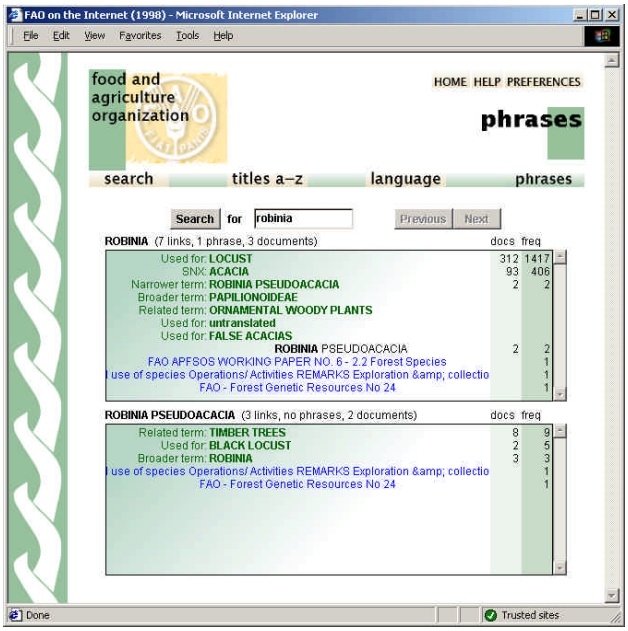

Figure 7 The Phind interface with thesaurus data

To suppress trivial phrases, we impose the condition that phrases must begin and end with "content words." Function words like *the*, *of*, and *and* occur very frequently (in English) but have no intrinsic semantic value. Without special treatment, the phrases that are extracted would include a myriad of trivial expansions like *the forest* and *of forest*—which would displace more useful terms by taking up space in the phrase list. For each language we further expand phrases whose first or last word appears in a predefined list of stopwords.

At the core of the phrase extraction process is a program that extracts the phrase hierarchy from a sequence of input symbols. It must identify the set of phrases that occur more than once, are maximal length, do not contain delimiters, and begin and end with content words. We will not describe the details of this lengthy and rather intricate process here, but refer the reader to other papers on the subject (Nevill-Manning *et al*, 1999; Paynter *et al.*, 2000a).

The result is a data structure that supports the online browsing interface. As well as the text of the phrases, the interface needs to know the structure defined by the subphrase relation, and what documents each phrase appears in. For every word and phrase there is a list of phrases in which that word or phrase occurs, and with each word or phrase is stored a list of the documents in which it occurs.

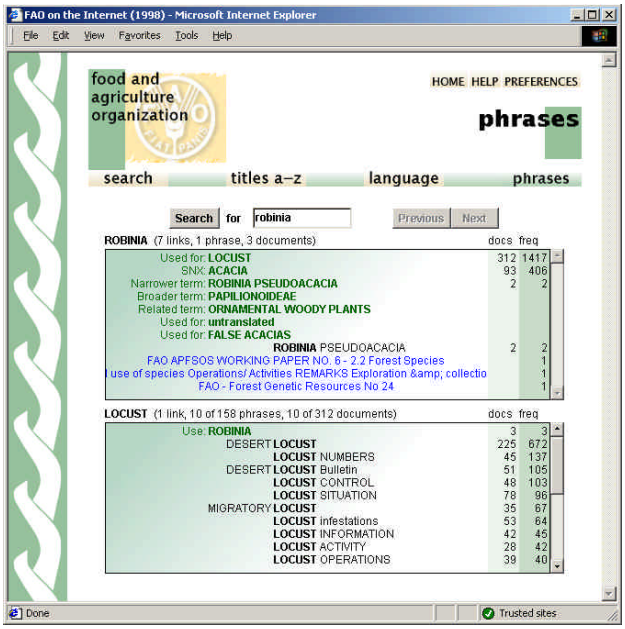

Figure 8 The Phind interface with thesaurus data.

5 Thesauri

AGROVOC is a multilingual thesaurus for agricultural information systems, developed by the FAO to provide subject control for the AGRIS agricultural bibliographic database and the CARIS database of agricultural research projects (FAO, 1995). The thesaurus supports the three working languages of the FAO—English, French, and Spanish—and versions in Arabic, German, Italian, and Portuguese are under construction. AGROVOC is actively supported by the FAO and its international community of users, and is periodically updated to reflect changing terminology or shifts in the boundaries of the research field.

Thesaurus terms are nouns or noun phrases, and are divided into “descriptors,” which are the preferred terms that the thesaurus compilers think should be used to describe documents, and “non-descriptors,” which are synonyms that are linked to a preferred descriptor that should be used in their place. Despite their name, non-descriptors are extremely useful when searching: they are meaningful domain terms that searchers might use in a query or that document authors might include in their writing. We have in fact studied the relationship between thesaurus phrases and the phrases extracted from a document collection as described above, and conclude that there is a significant degree of overlap (Paynter *et al.*, 2000b). However, many descriptors are rarely used in an actual document collection. For example, descriptors are often scientific names that do not form part of normal discourse, and equivalent non-descriptors may be far more prevalent in actual documents.

Each language version of AGROVOC includes approximately 15,700 descriptors and 10,000 non-descriptors. The terms were designed to be brief (three or fewer words if possible) and compact (at most 35 characters) due to limitations imposed by the original thesaurus software (FAO, 1995). The strict upper limit on characters requires that lengthy terms (such as the names of organizations, enzymes, chemical compounds, etc.) have had to be abbreviated, sometimes in arbitrary or non-standard ways. This impacts the overlap between extracted and AGROVOC phrases, though only slightly.

Manually constructed thesauri include a wealth of potentially useful terms and detailed information about their interrelations. Carefully compiled and edited by hand, they are an authoritative source of subject information and structure. Given a specialist thesaurus that relates to the area covered by a particular document collection, such as the AGROVOC thesaurus for the FAO collection, it seems likely that a phrase browsing scheme would be greatly improved by enriching the phrases with information from the thesaurus—such as descriptors, synonyms, and their relationships.

6 Thesaurus Integration

We have enhanced the Phind interface by including relevant information from the AGROVOC thesaurus alongside phrases extracted from the document collection itself. Given a search term, entries from the thesaurus that include the term are displayed first, followed by the phrases from the document that contain the term, extracted as described above, followed, finally, by documents that contain the term, also described above. The first ten items on each of these three lists are displayed immediately; more can be obtained by clicking the *Get more* links.

In the example of Figure 6, the user has performed a search for *tools*. The first result is a *Use* entry from the thesaurus; the results below are phrases from the document collection. The *Use* entry recommends that the term *equipment* be substituted for *tools* in the query (and states that it occurs 1548 times, in 599 documents). The user clicks on *equipment* and a list of narrower terms are displayed in the bottom pane.

The lower pane in Figure 6 is entirely taken up by thesaurus links. The user has expanded the thesaurus list to 40 terms by clicking the *Get more thesaurus links* bar; in fact, there are a total of 184 thesaurus links for *equipment.* These terms describe a diverse range of equipment, the vast majority of which would not be found by the phrase browser on its own because they do not include either of the words *tools* or *equipment*. Under the thesaurus links appear phrases containing the term; however none of these are currently visible in the Figure because too many thesaurus terms appear.

In Figure 7 the user is searching for information on locusts, using the Latin name *robinia*, which is the preferred descriptor in the AGROVOC thesaurus. On its own, the phrase browser would not lead the user to much useful information, because the term hardly appears at all in the documents in the collection. As the text above the first panel indicates, it occurs in only three documents, and one phrase. However, it also appears in seven thesaurus links, which are shown as the first seven lines in the

upper panel. The eighth entry gives the only longer phrase extracted, *robinia pseudoacacia*, which occurs twice, and in two documents (as noted in the columns on the right).

The thesaurus links given in the first seven entries in the top panel, however, provide several alternatives. The first entry indicates that *robinia* is used for the term *locust*, which occurs far more frequently in the text (1417 occurrences, 312 documents). Merely clicking on this term would bring up the display in Figure 8. Subsequent entries state that *acacia* is a synonym (and appears 406 times in 93 documents), that *robinia pseudoacacia* is a narrower term (which appears in two documents), and that several other terms in the thesaurus may be of interest but do not appear in the text of the collection. Clicking on the narrower term *robinia pseudoacacia*, the user discovers from its thesaurus entry that it is used for *black locust*, a subspecies (Figure 7, bottom pane).

The user is not interested in subspecies, so instead clicks on *locust*, the most frequently used term. The result, shown in the lower panel of Figure 8, gives only one thesaurus entry, which directs the user back to the preferred term *robinia*. However, there are 158 expansions, including *desert locust*, *locust numbers*, and *locust control*. The user can explore any of these topics, which may or may not have further expansions and thesaurus entries. Alternatively, they can explore another link from *robinia*, including those that occur in the thesaurus but not in the document text. For example, they may click on the broader term *papilionoideae*, which does not occur in the collection, but has 76 thesaurus entries for broader terms, narrower terms, and synonyms, many of which do occur.

These examples show how manually constructed thesauri can be used to significantly enrich the phrase browsing experience.

7 Conclusion

Both phrase browsing, based on phrases automatically extracted from the documents, and thesaurus browsing, based on manually tagging documents with thesaurus terms, are useful ways of examining the content of a document collection in a way that is more informal, but in many cases far more useful, than full-text searching. Manual tagging is the more powerful technique if the appropriate information is available and accurate, partly because it uses a controlled vocabulary and partly because it is based more directly on the content of the documents. Automatically extracted phrases are lexically based, which could prove limiting—even occasionally misleading—in the normal free-vocabulary situation. But manual tagging needs to be accurate, and is expensive.

We have demonstrated a novel interface that combines automatic phrase extraction, which is very cheap because it requires no manual processing, with the semantic benefits of a manually-constructed thesaurus, which normally presupposes expensive manual tagging of documents. Rather than tagging documents with thesaurus terms manually, the documents in which these terms occur are identified. The thesaurus hierarchy is presented in tandem with the automatically-extracted phrase hierarchy, and smooth transitions from one to the other are provided. For example, the number of occurrences of thesaurus terms and number of documents in

which they occur are noted in the display, just as they are for phrases. Thesaurus links are presented explicitly in the display. Clicking on a link to a non-descriptor, or a narrower term, or a broader term, or a related term, brings up all phrases that contain that term—just as clicking on an automatically-extracted phrase brings up all phrases that contain it.

A combined thesaurus/phrase browser gives the best of both worlds: an accurate, cheap, automatically-constructed phrase hierarchy, alongside a carefully-constructed and well-thought-out thesaurus.

References

1. Chang, S.J. and Rice, R.E. (1993) "Browsing: a multidimensional framework." *Annual Review of Information Science and Technology*, Vol. 28, pp. 231–276.
2. FAO (1995) *AGROVOC: multilingual agricultural thesaurus*. Food and Agriculture Organization of the United Nations, Rome, Italy.
3. Giles, C.L., Bollacker, K. and Lawrence, S. (1998) "CiteSeer: An automatic citation indexing system." *Proc ACM Digital Libraries*, Pittsburgh, PA, pp, 89–98.
4. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. (1999) "Lexically-generated subject hierarchies for browsing large collections." *Int J Digital Libraries*, Vol. 2, No. 2/3, pp. 111–123; September.
5. Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. (2000a) "Scalable browsing for large collections." *Proc ACM Digital Libraries*, San Antonio, TX, pp. 215–223.
6. Paynter, G.W., Cunningham, S.J. and Witten, I.H. (2000b) "Evaluating extracted phrases and extending thesauri." Proc Asian Digital Libraries Conference, Seoul, Korea, pp. 131–138.
7. Witten, I.H., McNab, R.J., Boddie, S., Bainbridge, D. (1999) "Greenstone: a comprehensive open-source digital library software system." *Proc ACM Digital Libraries*, San Antonio, TX, pp. 113–121.