Scalable browsing for large collections: a case study

Gordon W. Paynter, Ian H. Witten, Sally Jo Cunningham
Dept of Computer Science
University of Waikato
Hamilton, New Zealand
{gwp, ihw, sallyjo}@cs.waikato.ac.nz

George Buchanan
Dept of Computer Science
Middlesex University
London, UK
g.buchanan@mdx.ac.uk

ABSTRACT
Phrase browsing techniques use phrases extracted automatically from a large information collection as a basis for browsing and accessing it. This paper describes a case study that uses an automatically constructed phrase hierarchy to facilitate browsing of an ordinary large Web site. Phrases are extracted from the full text using a novel combination of rudimentary syntactic processing and sequential grammar induction techniques. The interface is simple, robust and easy to use.

To convey a feeling for the quality of the phrases that are generated automatically, a thesaurus used by the organization responsible for the Web site is studied and its degree of overlap with the phrases in the hierarchy is analyzed. Our ultimate goal is to amalgamate hierarchical phrase browsing and hierarchical thesaurus browsing: the latter provides an authoritative domain vocabulary and the former augments coverage in areas the thesaurus does not reach.

INTRODUCTION
Suppose you are browsing a large collection of information such as a digital library—or a large Web site. Searching is easy, if you know what you are looking for—and can express it as a query at the lexical level. But current search mechanisms are not much use if you are not looking for a specific piece of information, but are generally exploring the collection. Studies of browsing have shown that it is a rich and fundamental human information behavior, a multifaceted and multidimensional human activity [3]. But it is not well supported for large digital collections.

Web sites link together information in a way that is designed to help the browser. But as the scale of collections increases, links become very difficult to create and maintain. Inserting links manually is labor-intensive, and this kind of information rapidly goes stale as the collection grows. For large collections, the complexity of manually organizing the information is daunting.

Metadata provides information that can be used for browsing—given the appropriate metadata, it is possible to provide the human browser with indexes of authors and titles, classification hierarchies, and so on [19]. But as the scale of the information increases, the value of such lists decays—they become too large to be of much use. With large indexes one is reduced to searching rather than browsing.

We have been experimenting with different ways of automatically abstracting hierarchical structures of phrases from large collections of information and using them to facilitate browsing [10, 12]. This paper reports an application of these techniques to a large Web site. We discuss extensions to our earlier keyphrase extraction algorithm, and describe an improved browsing interface based on these keyphrases.

Our case study is based on the site of the United Nations Food and Agriculture Organization (FAO, www.fao.org), an international organization founded in 1945 whose mandate is to raise levels of nutrition and standards of living, to improve agricultural productivity, and to better the condition of rural populations. Web presence is seen as an important part of the FAO's information dissemination activities, and the site is organized and maintained by the World Agricultural Information Center (WAICENT), a subunit of the FAO. The version that we use in this study is dated 1998 and contains 21,700 Web pages, as well as around 13,700 associated files (image files, PDFs, etc). This corresponds to a medium-sized collection of approximately 140 million words of text. Figures 1 and 2 show typical pages from the site.

This site exhibits many problems common to large, public Web sites. It has existed for some time, is large and continues to grow rapidly. Despite strenuous efforts to organize it, it is becoming increasingly hard to find information. A search mechanism is in place, but while this allows some specific questions to be answered it does not really address the needs of the user who wishes to browse in a less directed manner.

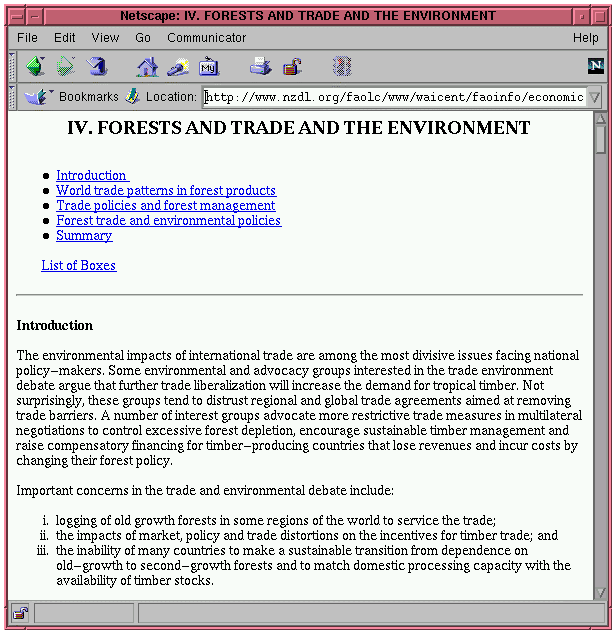

Figure 1: Example Web page (English)

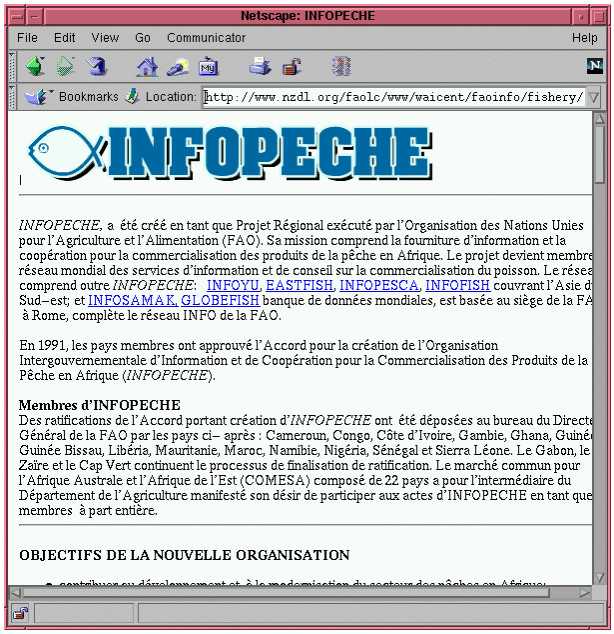

Figure 2: Example Web page (French)

We support browsing of the FAO site with an interactive interface to the phrases present in the documents; this interface is discussed in the next section, and the succeeding section describes the techniques used to create the underlying index of phrases. We then examine the potential usefulness of the phrases by comparing them with terms and phrases contained in AGROVOC [4], a manually constructed thesaurus for the field of agriculture. The extent of the overlap between the vocabulary of the thesaurus and that of the documents in a collection provides an indication of the applicability of the thesaurus to the collection—literally, examining whether the two describe the discipline in the same words.

PHRASE-BASED SUBJECT INDEX INTERFACE

The phrase-based browser that we have developed is an interactive interface to a phrase hierarchy that has been extracted automatically from the full text of the Web site. It is designed to resemble a paper-based subject index or thesaurus. Figure 3 shows the interface in use. The user enters an initial word in the search box at the top. On pressing the *Search* button the upper panel appears. This shows the phrases at the top level in the hierarchy that contain the search word—in this case the word *forest*. The list is sorted by phrase frequency; on the right is the number of times the phrase appears, and to the left of that is the number of documents in which the phrase appears.

Only the first ten phrases are shown, because it is impractical with a Web interface to download a large number of phrases, and many of these phrase lists are very large. At the end of the list is an item that reads *Get more phrases* (displayed in a distinctive color); clicking this will download another ten phrases, and so on. A scroll bar appears to the right for use when more than ten phrases are displayed. The number of phrases appears above the list: in this case there are 493 top-level phrases that contain the term *forest*.

So far we have only described the upper of the two panels in Figure 3. The lower one appears as soon as the user clicks one of the phrases in the upper list. In this case the user has clicked *forest products* (that is why that line is highlighted in the upper panel) and the lower panel, which shows phrases containing the text *forest products*, has appeared.

If one continues to descend through the phrase hierarchy, eventually the leaves will be reached. A leaf corresponds to a phrase that occurs in only one document of the collection (though the phrase may appear several times in that document). In this case, the text above the lower panel shows that the phrase *forest products* appears in 72 phrases (the first ten are shown), in 382 documents. The first ten of these are available too, though the list must be scrolled down to make them appear in the visible part of the panel. Figure 4 shows this. In effect, the panel shows a phrase list followed by a document list. Either of these lists may be null (in fact the document list is null in the upper panel, because we are not interested in single words, like *forest*, because they are equivalent to a full text search for *forest*). The document list displays the titles of the documents.

It is possible, in both panels of Figures 3 and 4, to click *Get more phrases* to increase the number of phrases that are shown in the list of phrases. It is also possible, in the lower panels, to *Get more documents* to increase the number of documents that are shown in the list of documents. Again, this option is displayed at the end of the list in a distinctive color, but to see that entry it is necessary to scroll the panel down a little more.

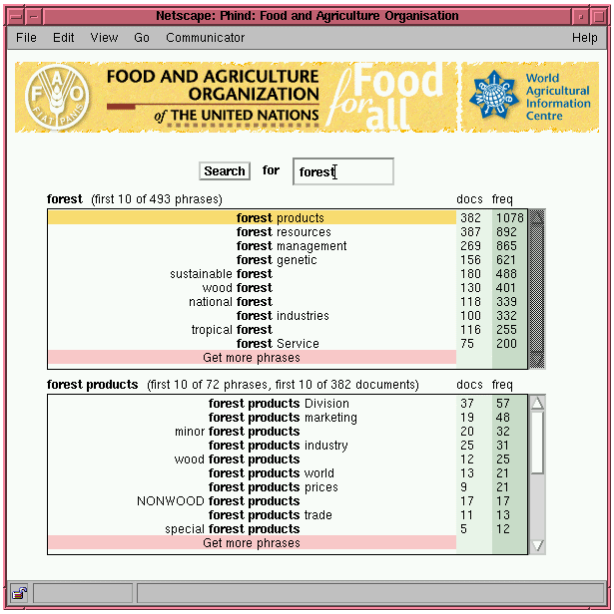

Figure 3: Browsing for information about *forest*

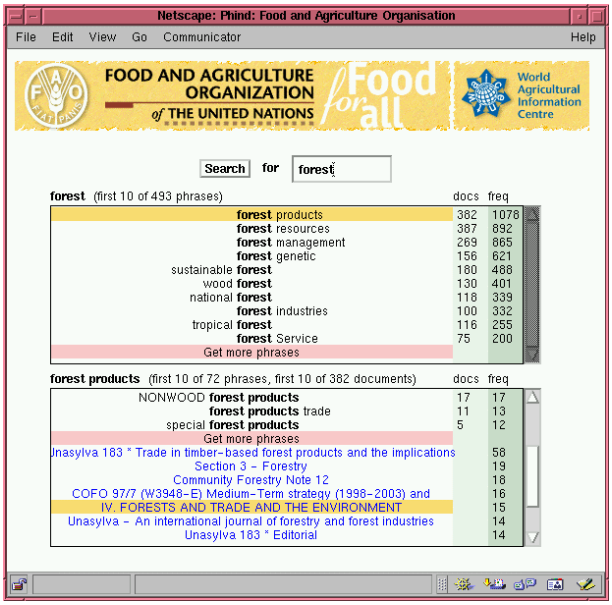

Figure 4: Expanding on *forest products*

Clicking on a phrase will expand that phrase. The page holds only two panels, and if a phrase in the lower panel is clicked the contents of that panel will move up into the top one to make space for the phrase's expansion. Alternatively, clicking on a document will open that document in a new window. In fact, the user in Figure 4 has clicked on *IV FORESTS AND TRADE AND THE ENVIRONMENT*, and this brings up the page shown in Figure 1. As Figure 4 indicates, that document contains 15 occurrences of the phrase *forest products*.

Figures 5 and 6 show some more examples of the interface in use. In Figure 5 the user has entered the word *dairy* and expanded on *New Zealand dairy* (note that this collection is from the FAO in Rome, Italy; it is impressive to be able to home in on information about the local dairy industry in New Zealand so rapidly). Figure 6 shows a French user typing the word *poisson*. The FAO site contains documents in French, but our phrase extraction system is tailored for English as described below. The French phrases that are displayed are of much lower quality than the English ones in Figures 3, 4 and 5; the list of ten phrases in the upper panel of Figure 6 contains only four useful ones. Phrases like *du poisson* (usually meaning *of fish*) are not useful, and can even obscure more interesting material. However, the system is still usable. Here, the user has expanded *commercialisation du poisson* and, in the lower panel, has clicked *INFOPECHE*, which brings up the page in Figure 2.

DERIVING THE PHRASES

We have experimented with several different ways of creating a phrase hierarchy from a document collection. Nevill-Manning *et al.* [10] describe an algorithm called SEQUITUR that builds a hierarchical structure containing every single phrase that occurs more than once in the document collection. We have also worked on a scheme called KEA which extracts keyphrases from scientific papers. This produces a far smaller, controllable, number of phrases per document [5].

In this section we first describe earlier work with SEQUITUR, which builds hierarchies from phrases, and KEA, which extracts keyphrases from documents. The scheme that we use for the interface described in this paper is an amalgam of the two techniques, designed to overcome difficulties encountered with SEQUITUR and KEA; that scheme is described in the final portion of this section.

Constructing phrase hierarchies using SEQUITUR

The basic insight of SEQUITUR is that any phrase that appears more than once can be replaced by a grammatical rule that generates that phrase, and this process can be continued recursively. The result is a hierarchical representation of the original sequence. It is not a grammar, for the rules are not generalized and are capable of generating only one string.

There exists a remarkably efficient algorithm to derive these phrases from an input sequence, and the time it takes is linear in the length of the input [11]. This has allowed us to investigate hierarchies formed from sequences of words containing up to 60 million tokens.

Nevill-Manning *et al.* [10] reported character-based hierarchies, formed by using characters as tokens, and word-based hierarchies, formed using words. Interesting effects occur in both cases, although word mode is most suitable for interactive browsing of large information collections.

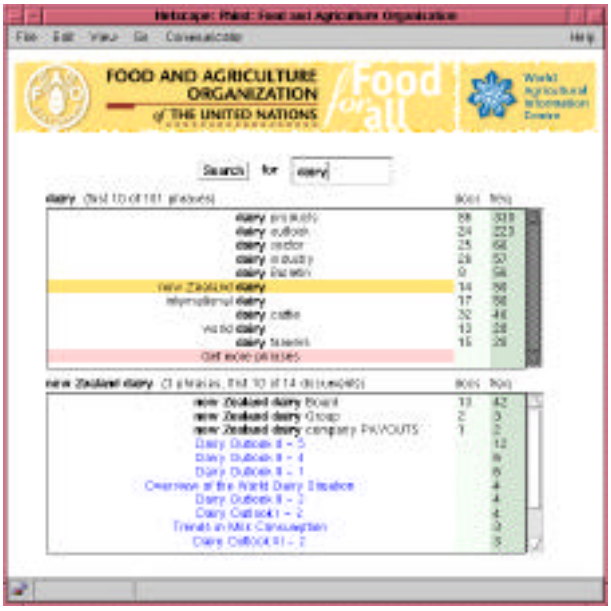

Figure 5: Browsing for information on *dairy*

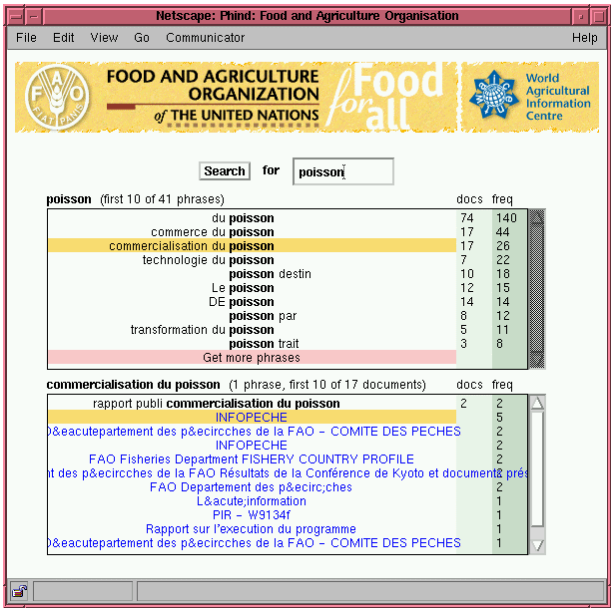

Figure 6: Browsing for information on *poisson*

The SEQUITUR algorithm only forms rules that represent phrases that occur more than once and in a unique context. Consequently subsets of lengthy phrases are only displayed if they occur independently; for example, in Figure 5 the two word phrase "Zealand dairy" never occurs independently of the phrase "New Zealand dairy", and so SEQUITUR forms only the three word phrase .

In order to display the phrase hierarchy interactively, a number of additional facilities are incorporated into the browser. Words like *a* and *the* cause problems because they are often used to form rules, but as far as the user is concerned they add little meaning to the phrase. Nobody really wants to know that the most common use of the word *index* is in the phrase *the index*. Hence we label as *common words* (stopwords) the one hundred most frequently occurring words in the collection, and weed out phrase expansions that differ from the original phrase only by the addition of common words. In practice, this simple stopword construction technique posed difficulties, and was abandoned in the amalgam system described at the end of this section. For example, in a computer science technical reports collection the terms 'system' and 'time' were labelled common words—which meant that users could not look up ordinary computing phrases such as 'operating system' and 'system time'.

At the other extreme, phrases that occur rarely increase the number of potential phrases but contribute little to our understanding of the collection. This effect is mitigated by the SEQUITUR algorithm, which ignores singleton phrases. Additionally, the phrase browsing interface more weight to frequent phrases and discards phrases whose frequency falls below a low-frequency threshold.

The SEQUITUR grammar forms the basis for our first phrase browsing interface [12]. Overlooking common and rare words greatly increased the usability of the interface, but poorly formed phrases were still a problem.

Extracting keyphrases using KEA

In a separate project, we investigated algorithms for extracting keyphrases from technical documents [5]. Keyphrases provide a kind of semantic metadata that is useful for a wide variety of purposes. It turns out that keyphrases can be extracted automatically from the full text of documents with surprising accuracy.

To do this, candidate keyphrases are identified, features are computed for each candidate, and machine learning is used to generate a classifier that determines which candidates should be assigned as keyphrases. One feature, TF×IDF [14], requires a corpus of text from which document frequencies can be calculated; the machine learning phase requires a set of training documents with keyphrases assigned. The success of various stages of the procedure was evaluated on a large test corpus, in terms of how many author-assigned keyphrases are correctly identified (a measure that is subject to some caveats).

In the final procedure we developed for keyphrase extraction (KEA), stop words were used to determine whether or not a phrase is a candidate phrase. Our experiments on keyphrase extraction also used a syntactic method for identifying candidate phrases: we tried to identify noun phrases. The two approaches are equally accurate on the keyphrase extraction task, but we used stop words in the final system because it is significantly faster.

The syntactic analysis first tags the input by assigning syntactic classes to each word. We use the Brill tagger [1,2]. Then we experimented with two heuristics for noun

length in words	AGROVOC number	AGROVOC percentage	Extracted phrases number	Extracted phrases percentage
1	12342	44.9%	58954	21.2%
2	13046	47.5%	126950	45.7%
3	1692	6.2%	57844	20.8%
4	327	1.2%	19356	7.0%
5	51	0.2%	7194	2.6%
6	7	0.0%	3271	1.2%
7	1	0.0%	1724	0.6%
8			1050	0.4%
9			639	0.2%
10 - 42			1109	0.4%
average length	1.64 words		2.37 words	

Table 1: Length of phrases (words)

length in characters	AGROVOC number	AGROVOC percentage	Extracted phrases number	Extracted phrases percentage
1 – 5	1207	4.4%	11513	4.4%
6 – 10	8089	29.5%	47665	18.1%
11 – 15	8737	31.8%	68471	26.0%
16 – 20	6146	22.4%	60861	23.1%
21 – 25	2477	9.0%	35598	13.5%
26 – 30	599	2.2%	17608	6.7%
31 – 35	211	0.8%	8950	3.4%
36 – 40			4752	1.8%
41 – 45			2690	1.0%
46 – 50			1544	0.6%
51 – 55			1037	0.4%
> 55			2674	1.0%
average length	13.58 characters		17.62 characters	

Table 2: Length of phrases (characters)

phrase identification. The first was suggested by Turney [18] as matching almost all of the keyphrases in the corpuses he used. It specifies zero or more nouns or adjectives, followed by one final noun or gerund:

(noun | adjective)* (noun | verb-gerund)

where a "noun" is either a singular or plural noun or proper noun. ("*" means repetition, appearing zero or more times.)

Although this structure resembles a noun phrase, it turns out that the notion of "noun phrase" is only loosely defined in the first place. Also, in our work we have encountered many author-defined keyphrases that are not noun phrases according to this regular expression

Consequently, we experimented with a different regular expression to locate candidate phrases, which we describe as "augmented" noun phrases:

[(noun | adjective | verb)+ (conjunction | prep)]* (noun | verb-gerund)

where conjunctions and prepositions are members of a predefined list defined by the Brill tagger (and "+" means one or more repetitions). This allows sequences of nouns, adjectives, and verbs to be interspersed with connectives, before the terminating noun or gerund, and permits phrases such as *programming by demonstration.*

Several browsing interfaces are based on keyphrases. Jones and Paynter [7] automatically insert hyperlinks into digital library collections using keyphrases as link anchors and document clusters as destinations. Turney [17] uses keyphrases to construct searchable subject indexes. Gutwin *et al.* [6] search for clusters of documents that share keyphrases. Phrases in the result list can be reused as search terms, allowing the user to search increasingly specific variations on a phrase. Price *et al*. [13] allow readers to 'highlite' interesting portions of documents, extract the text from these annotations, and then present the user with links to related documents in the collection. All four interfaces treat phrases as indivisible units; they do not exploit their hierarchical nature for browsing.

Constructing hierarchies of noun phrases

For the interface described in the present paper, we have employed a combination of the two approaches. As noted above, SEQUITUR produces all phrases that occur more than once. However, users who are browsing are generally far more interested in *noun phrases* rather than in other types of phrase. SEQUITUR, when applied to the full input text, tends to produce many other phrases that are not so useful for browsing information collections (though they are useful for other purposes).

If SEQUITUR produces too many phrases, then keyphrase extraction produces too few. A typical document contains thousands of candidate phrases, which the extraction algorithm pares down to fewer than a dozen. Inevitably, hundreds of valuable phrases are discarded. Further, by compressing every occurrence of a phrase to a single summary occurrence, the phrase's context and frequency are sacrificed. Without context and frequency—the *de facto* measure of relative importance—we are unable to construct a browsable hierarchy.

As a compromise, we extract just the noun phrases that appear in the full text of the documents, and base a SEQUITUR hierarchy on those. To do this we convert the Web pages to plain ASCII text, using the Lynx browser to strip out all HTML tags, then process the resulting sequence with the Brill tagger. We extract every sequence of words whose tags have the syntactic structure given above for augmented noun phrases, and insert a special delimiter symbol between noun phrases and at clause breaks like commas and the ends of sentences. The result is a long sequence of delimited noun phrases. This process sensibly limits the maximum possible length of a noun phrase to that of the longest 'sentence' in the document. For the FAO documents the longest phrase constructed is 42 words (Table 1)—and these extraordinarily lengthy

	AGROVOC thesaurus	Extracted phrases
1	*forest canopy*	forest Academy
2	*forest decline*	forest access
3	*forest dieback*	forest Act
4	*forest ecology*	forest activities
5	forest establishment	forest administration
6	*forest fires*	forest agencies
7	forest floor vegetation	forest agenda
8	*forest grazing*	forest animals
9	*forest health*	forest area
10	*forest industry*	forest assessment
11	*forest inventories*	forest authorities
12	*forest land*	forest authority
13	forest litter	forest base
14	*forest management*	forest benefits
15	forest measurement	forest biodiversity
16	*forest mensuration*	forest biology
17	*forest meteorology*	forest biomass
18	*forest nurseries*	forest Botany
19	*forest pathology*	forest boundaries
20	forest pests	*forest canopy*
21	*forest plantations*	forest capital
22	*forest policies*	forest certification
23	*forest products*	forest characteristics
24	*forest product industry**	forest charges
25	*forest protection*	forest clearance
26	forest range	forest co management regime
27	*forest regulations***	forest codes
28	*forest rehabilitation*	forest college
29	forest replanting	forest commons
30	*forest reserves*	forest communities
31	*forest resources*	forest companies
32	forest returns	forest composition
33	*forest roads*	forest concession
34	*forest soils*	forest condition
35	forest stands	forest conflicts
35	forest steppe	forest conservation
37	*forest surveys***	forest control
38	forest thinning	forest conversion
39	forest tree nurseries	forest cover
40	*forest trees*	forest crisis
41	*forest workers*	forest crops
…		…
235		forest zones
236		forest zoology

Table 3: Phrases beginning with the word *forest*

phrases are headers from the web pages, not properly constructed English phrases. In practice, a maximum phrase length could be set to eliminate this sort of 'phrase'.

There are many problems with this procedure, and the result is only an approximation to the actual noun phrases that occur in the input. First, the Brill tagger is not perfect—for example, unrecognized words are assumed to be nouns. Second, it is not easy to define a regular expression on the tags that captures all and every noun phrase. But most importantly, some of these documents (e.g. Figure 2) are in other languages—mostly French and Spanish—and this naturally plays havoc with the tagger. Non-English words are assumed to be nouns and used to build nonsense phrases. Another issue is whether or not to apply stemming before building the noun phrase list. Without stemming, we will get different versions of the same basic noun phrase. In our work on keyphrase extraction, we stemmed words and conflated different versions in order to remove duplicate phrases and count phrase frequencies, but kept a record of the most frequent unstemmed version of each phrase in order to reexpand the stemmed version for display to the user. This is also an option for the present system, although the illustrations in this paper do not use any stemming.

The final phase is to build a hierarchy from the noun phrases by running SEQUITUR over the sequence of noun phrases, specifying the delimiter symbol as a delimiter for SEQUITUR. In fact, the SEQUITUR algorithm is really designed for long undelimited sequences—the problem of generating a hierarchy from a set of short phrases in reasonable time is much easier than treating a single long sequence. And SEQUITUR makes some sacrifices in accuracy to operate in reasonable time. Thus this step also adds a degree of approximation to the phrase hierarchy that results, which could be avoided by using a more suitable method.

COMPARING THE PHRASES TO A THESAURUS

The phrases extracted represent the topics present in the FAO site, as described in the terminology of the authors of the documents. But how well does this set of phrases match the standard terminology of the discipline? We investigate this by comparing the extracted phrases with phrases used by the AGROVOC agricultural thesaurus. The degree of overlap between the two sets of phrases provides a rough indication of the suitability of the extracted phrases as subject descriptors—or conversely, the applicability of the AGROVOC thesaurus to the FAO site can be assessed by examining the extent to which the AGROVOC phrases appear in the natural text of the documents.

The AGROVOC thesaurus

AGROVOC is a multilingual thesaurus for agricultural information systems, developed by the FAO to support subject control for the AGRIS agricultural bibliographic database and the CARIS database of agricultural research projects [4]. 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. A searchable version is accessible at www.fao.org/AGROVOC.

The thesaurus is of a significant size—each language version includes more than 15,700 descriptors, and approximately 10,000 non-descriptors (also colorfully referred to as "forbidden terms", non-descriptors are synonyms or related terms that are linked to a descriptor

	AGROVOC thesaurus	Extracted phrases
1	coppice forest	actual forest
2	duff (forest litter)	aggregate forest
3	*high forest*	Amazon forest
4	*minor forest products**	amenity forest
5	mixed forest stands	American forest
6	monsoon forest	artificial forest
7	nontimber forest products	available forest
8	*nonwood forest products**	Bangladesh forest
9	*secondary forest products**	bavarian forest
10	semliki forest virus	Black forest
11	slash (forest litter)	boreal forest
12	thorn forest	Chimanes forest
…		…
204		world forest
205		Wright forest Mgt
206		young forest

Table 4: Phrases containing the word *forest*

by a "use" reference). Thesaurus terms are nouns or noun phrases, and all—including non-descriptors—were selected for inclusion on the basis of their common usage in the agricultural research literature. The AGROVOC vocabulary forms a rich semantic network describing the agricultural domain, with links between terms describing hierarchical relationships (*broader term*, *narrower term)*, associative relations (*related term*s), and synonym links between descriptors and non-descriptors (*use*, *use for*).

Tables 1 and 2 summarize the structural characteristics of the AGROVOC phrases and the extracted phrases. The AGROVOC phrases are taken from the English version only, and include both descriptors and non-descriptors. The non-descriptors appear in this analysis because, despite their title, they are useful in thesaurus searching, since they have a meaning synonymous or related to that of their associated descriptors. The algorithm extracts phrases of two or more words. The phrases in the hierarchy are drawn from a vocabulary of single word terms, and this vocabulary is the source of the single-word phrases in Tables 1–6.

The extracted phrases tend to be longer than the AGROVOC ones, measured both by the number of words and the number of characters per phrase (Tables 1–2). This difference was expected, since AGROVOC phrases were deliberately designed to be brief (three or fewer words) and compact (maximum of 35 characters). These limitations were imposed by the original thesaurus software [4]. The strict upper limit on characters has proven problematic, in 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 practice can make querying more difficult for users, who have to guess when and how a phrase has been abbreviated. The potential overlap between the extracted and AGROVOC phrases is also reduced, though only slightly—we estimate that just over 200 of the 12,342 AGROVOC phrases are abbreviated. If a list of abbreviations and their expansions were available (as is often the case), then this difficulty could be eliminated.

Overlap with AGROVOC phrases

We begin with an example to illustrate the degree and type of overlap found between the two sets of phrases. Table 3 shows phrases beginning with the word *forest* in AGROVOC and at the top level of the phrase hierarchy. Italics indicates that the AGROVOC phrase occurs amongst the extracted phrases (and vice versa). All italicized phrases occur at the top level except the ones marked with a single asterisk—in Table 3, just *forest products industry*—which appears at a lower level of the hierarchy. This distinction is visible in Figure 3, where *forest products industry* appears as an expansion of the top-level phrase *forest products* (as do the three asterisked phrases in Table 4). The doubly-asterisked phrases, *forest regulations* and *forest surveys*, appear in the plural only coincide with extracted phrases if they are stemmed—to *forest regulation* and *forest survey* respectively.

The overlap between the AGROVOC thesaurus and the phrases extracted from the FAO site is quantified in Tables 5–6. For comparison's sake, we also include statistics for the raw text and the keyphrases extracted from it by KEA. The former represents an upper bound for matches, and was generated by extracting every sequence of one to four words present in the FAO site. The latter emphasizes precision rather than recall in a match, since there are fewer keyphrases associated with each document (a maximum of six). The keyphrases are also more likely to be true indicators of the focus of the document, and so are closer to the intent of AGROVOC thesaurus entries.

As illustrated in the *forest* example, stemming can affect the degree of match. We examine this effect by comparing the overlap between unstemmed phrases and phrases stemmed using the Lovins and Iterated Lovins algorithms [9]. The Lovins algorithm stems words to their root form; for example, *dictionary* is reduced to *diction.* The iterated algorithm repeatedly applies the Lovins stemmer until the stem no longer changes; *dictionary* is thus stemmed to *dict.* When phrases are stemmed more severely, the number of unique entries decreases because similar phrases are stemmed to equivalent root terms, as can be seen in the top row of Table 6.

Note that slightly over half of the words appearing in the AGROVOC thesaurus phrases are also present in the FAO documents (Table 5). This overlap is a strong indication that AGROVOC is a suitable thesaurus to use with those pages. The coverage of the raw text by the AGROVOC phrases forms a baseline for coverage by the extracted hierarchy and the keyphrases. The proportion of AGROVOC words contained in phrases in the extracted hierarchy is smaller, as is expected, but still represents a respectable one-third of the AGROVOC terms. Including vocabulary terms from the extracted hierarchy increases the coverage of the AGROVOC terms. As expected, the

	Unstemmed	Lovins stemmer	Iterated Lovins
Number of unique terms			
Agrovoc	20574	17293	15670
FAO Web pages	169209	123975	107870
Extracted phrases	44226	30441	25013
Keyphrases	7886	5913	5284
Number of Agrovoc terms covered by words in...			
FAO Web pages	9945	8685	8210
extracted phrases	6186	5599	5384
keyphrases	2483	2356	2294
Proportion of Agrovoc terms covered by words in...			
FAO Web pages	48.3%	50.2%	52.4%
extracted phrases	30.1%	32.4%	34.4%
keyphrases	12.1%	13.6%	14.6%

Table 5: Term overlap between AGROVOC, extracted phrases, and keyphrases

Kea keyphrases cover a smaller proportion of AGROVOC terms.

The proportion of full (stemmed) AGROVOC phrases that are included in the FAO site and the extracted hierarchy is high—40% and 26% respectively (Table 6). This is particularly encouraging, as it indicates that a significant number of links exist between AGROVOC terms, documents, and the extracted hierarchy. These inter-relations could form the basis for a rich tool to support collection browsing. For example, the user interaction depicted in Figures 3 and 4 begins as the search term *forest* is entered into the phrase-based browser. The phrase hierarchy is scanned and the phrase *forest products* is selected. But this term is also represented in the AGROVOC thesaurus; access to the thesaurus would also have brought to the user's attention 44 specific types of forest product (for example, Christmas trees, charcoal, and particle boards), and 10 related topics (such as logging wastes, cellulose products, and tanning agents). These AGROVOC terms could then be browsed in the interactive interface. Interestingly, in the AGROVOC entry for *forest product*, three of the 54 narrower/related phrase links contain the word *forest*, one contains *forestry*, and six contain *products*. The majority of the AGROVOC links bring in new search or browsing terms for the user to consider.

Stemming increases the number of AGROVOC words and full phrases that can be matched to the FAO site, the extracted hierarchy, and the keyphrases, but only marginally. Iterated Lovins provides a higher degree of matching than Lovins, but again, the advantage is small.

DISCUSSION

A free-text index is the most common access method for Web collections, mainly because the index can be constructed automatically. Searchers typically experience difficulty in constructing effective queries, since they must match their personal vocabulary to that of the collection.

	Unstemmed	Lovins stemmer	Iterated Lovins
Number of phrases			
Agrovoc phrases	27466	26701	25901
FAO Web site phrases	19071445	18098815	17764015
Extracted phrases	278091	245374	233095
Keyphrases	13855	12183	11655
Number of Agrovoc phrases covered...			
by FAO Web site	9835	10750	10855
by extracted phrases	6166	6913	7014
by keyphrases	1447	1793	1874
Proportion of Agrovoc phrases covered...			
by FAO Web site	35.8%	40.3%	41.9%
by extracted phrases	22.4%	25.9%	27.1%
by keyphrases	5.3%	6.7%	7.2%

Table 6: Phrase overlap between AGROVOC, extracted phrases, and keyphrases

The interface presented in this paper provides a tool for spanning the gap between the two vocabularies. The phrases extracted from the document collection are noun phrases, and noun phrases are by far the most common queries submitted to retrieval systems. Users, then, can explore the collection's terms and term relationships through a display that mirrors the query construction naturally favored by users.

A controlled vocabulary such as a subject thesaurus is useful as a complement to free-text indexing: it can provide a framework for understanding the domain and learning its technical terminology [16]; as a primary interface for searching/browsing a document collection [15]; and as a supporting tool for query construction (typically in automated or semi-automated query expansion; for example, see [7]). Usually the information resource explored through a thesaurus is a bibliographic database, or (less commonly) a highly structured database such as the CARIS descriptions of agricultural research projects. In principle, users of an unstructured but focused document collection such as the FAO site should also benefit from the availability of a subject-specific thesaurus. However, the potential benefits are difficult to realize; the problems remain of matching the natural terminology of the searcher to the vocabulary of the FAO site and the thesaurus, and matching the terminology of the thesaurus to the site.

One approach to addressing the latter problem is to require the creator of a document at the FAO site to supply cataloging information that includes a set of applicable AGROVOC terms—in fact, this procedure is currently in use. But relatively few authors provide suitable AGROVOC keywords; perhaps the authors are themselves unfamiliar with AGROVOC and, like many searchers, find it difficult to select quality AGROVOC descriptors.

Approaching this problem from a different angle, the phrase hierarchies and frequency lists may be useful to

thesaurus developers. A thesaurus is a not a static object; it must be updated and revised to mirror terminology and usage changes in its application field. A comparison of thesaurus vocabulary to the phrases used in practice can give thesaurus constructors insight into the terms used by authors in the discipline.

Our next step will be to amalgamate the phrase and thesaurus hierarchies, both for searching and for AGROVOC term assignment during cataloging. Our analysis of the overlap between the AGROVOC and Web site vocabularies indicates that the two are similar enough that a tool linking the two hierarchies is likely to be useful. We envisage an interface that will allow users to gracefully navigate between their personal vocabulary, terms extracted from the FAO site, and AGROVOC terms/phrases. We can exploit the overlap between the extracted and AGROVOC phrases to support cataloging by running the extraction process over a submitted Web page and using the resulting phrases to link to potentially relevant portions of the AGROVOC hierarchy.

ACKNOWLEDGMENTS

Craig Nevill-Manning, Carl Gutwin, Eibe Frank and Steve Jones, have worked with us on phrase extraction and phrase interfaces, and all members of the New Zealand Digital Library project for their enthusiasm and ideas. We gratefully acknowledge financial support from the FAO in Rome, Italy. George Buchanan was employed on UK Engineering and Physical Sciences Research Council Grant No. GR/K79376; his visit to New Zealand, during which this paper was written, was funded by Middlesex University. We would also like to thank the anonymous referees for their comments.

REFERENCES

1. Brill, E. (1992) “A simple rule-based part of speech tagger.” *Proc ACL Conference on Applied Natural Language Processing*, pp. 152–155, Trento, Italy.
2. Brill, E. (1994) “Some advances in rule-based part of speech tagging,” *Proc AAAI-94*, pp. 722–727, Seattle.
3. Chang, S.J. and Rice, R.E. (1993) “Browsing: a multidimensional framework.” *Annual Review of Information Science and Technology*, Vol. 28, pp. 231-276.
4. FAO (Food and Agriculture Organization of the United Nations) (1995) *AGROVOC: multilingual agricultural thesaurus*. FAO, Rome.
5. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. (1999) “Domain-specific keyphrase extraction.” *Proc Int Joint Conf on Artificial Intelligence*, pp. 668-673, Stockholm, Sweden.
6. Gutwin, C., Paynter, G., Witten, I.H., Nevill-Manning, C., and Frank, E. (in press) “Improving Browsing in Digital Libraries with Keyphrase Indexes.” *J. Decision Support Systems*, Vol. 27, No 1/2, pp. 81-104.
7. Jones, S., Gatford, M., Robertson, S., Hancock-Beauliu, M., Secker, J., and Walker, S. (1995) “Interactive Thesaurus navigation: intelligence rules OK?” *JASIS*, Vol. 46, No. 1, pp. 52-59.
8. Jones, S. and Paynter, G. W. (1999) “Topic-based browsing within a digital library using keyphrases.” *Proc ACM Digital Libraries 99*, pp. 114–121.
9. Lovins, J.B. (1968) “Development of a Stemming Algorithm.” *Mechanical Translation and Computational Linguistics*, Vol. 11, pp. 2231.
10. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. (1997) “Browsing in digital libraries.” *Proc ACM Digital Libraries 97*, pp. 230-236, July.
11. Nevill-Manning, C.G. and Witten, I.H. (1997) “Identifying hierarchical structure in sequences.” *J Artificial Intelligence Research*, Vol. 7, pp. 67-82.
12. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. (1999) “Lexically-generated subject hierarchies for browsing large collections.” *Int J on Digital Libraries*, Vol. 2, No. 2/3, pp. 111-123; September.
13. Price, Morgan, Golovchinsky, Gene and Schilit, Bill N. (1998) “Linking By Inking: Trailblazing in a Paper-like Hypertext.” *Proc Hypertext '98* (Pittsburgh, PA), ACM Press, pp. 30-39.
14. Salton, G. (1989) *Automatic text processing*. Addison-Wesley Publishing Company, Inc., Reading, MA, USA.
15. Smith, M.P., Pollitt, A.S., and Li, C.S. (1992) “Evaluation of concept translation through menu navigation in the MenUSE intermediary system.” *Proc BCS IRSG Research Colloquium on Information Retrieval*, pp. 38-54, University of Lancaster, UK.
16. Soergel, D. (1985) *Organizing Information: principles of data base and retrieval systems.* Orlando: Academic Press.
17. Turney, P.D. (1999) “Learning to Extract Keyphrases from Text.” NRC Technical Report ERB-1057, National Research Council, Canada.
18. Turney, P.D. (in press) “Learning algorithms for keyphrase extraction.” *Information Retrieval*.
19. Witten, I.H., McNab, R.J., Boddie, S. and Bainbridge, D. (1999) “Greenstone: a comprehensive open-source digital library software system.” Research Report, Dept. of Computer Science, University of Waikato.