Journal of the American Society for Information Science

JASIS

VOLUME 46 NUMBER 1 JANUARY 1995

CONTENTS

Interactive Thesaurus Navigation: Intelligence Rules OK?

Susan Jones,* Mike Gatford, Steve Robertson, Micheline Hancock-Beaulieu, Judith Secker, and Steve Walker

Centre for Interactive Systems Research, Department of Information Science, City University, Northampton Square, London EC1V 0HB, United Kingdom. E-mail: sj@is.city.ac.uk

We discuss whether it is feasible to build intelligent rule- or weight-based algorithms into general-purpose software for interactive thesaurus navigation. We survey some approaches to the problem reported in the literature, particularly those involving the assignment of "link weights" in a thesaurus network, and point out some problems of both principle and practice. We then describe investigations which entailed logging the behavior of thesaurus users and testing the effect of thesaurus-based query enhancement in an IR system using term weighting, in an attempt to identify successful strategies to incorporate into automatic procedures. The results cause us to question many of the assumptions made by previous researchers in this area.

Survey of Approaches

A standard thesaurus comprises a set of terms making up a controlled indexing language, and a set of relationships between them. From the AI perspective, it can be considered as a knowledge base—a *semantic net* of concepts covering a particular domain. Current IR systems, as reported in the literature, use this knowledge in a variety of ways. For instance, we can distinguish between systems where the thesaurus is seen as the primary mechanism for matching queries to documents, and those where it is an adjunct to other look-up processes used for query enhancement during document retrieval.

An example of the first kind is described in Smith, Pollitt, and Li (1992), where the hierarchical structure of the thesaurus is exploited to provide a menu-like interface. Searchers move down the hierarchy from general to specific concepts, and retrieve documents indexed by the terms which best describe their information need. A more automatic approach is given in Rada and Bicknell (1989), where both documents and user-generated queries are located at points within the thesaurus structure, and matching involves finding "paths" (concatenated links representing term relationships) between them. Those documents on the shortest path from the query are deemed to be the best match for it.

That idea is extended in Kim and Kim (1990), where relationships between terms are assigned numeric "weights" according to their assumed strength of association, and the "distance" from query to document is computed by summing the weights of links along the path connecting them. A similar principle applies to the hierarchical concept representations used to aid retrieval in the commercial Topic system (Chong, 1989). The assignment and use of link weights is a topic which recurs quite often in proposals for thesaurus knowledge representation, and will be discussed in more detail later.

Systems in which the thesaurus is a means of *query enhancement* also exhibit a variety of approaches. Traditional IR services without pretensions to artificial intelligence may allow users to pick a controlled language descriptor from a displayed document record as a way into the thesaurus, from which further terms can be selected to form new document sets. For instance, the search mechanism provided with the INSPEC CD-ROM supports quite convenient navigation between document database and thesaurus using a mouse. More generally, it is recognized that a thesaurus can be explored as a *hypertext* document (McMath, Tamaru, & Rada 1989).

Researchers seeking to build "intelligent" IR systems, however, will wish to go beyond simple user-driven thesaurus navigation, and incorporate software components which act on behalf of or in conjunction with users to find the best routes to follow and the best terms to select. The declared aim is often to model the expertise of a human *intermediary*—see, for example, Shoval (1985); Chiaramella and Defude (1987); Smith, Shute, and Galdes (1989); and Chen and Dhar (1991). Following the normal methods for creating an intelligent knowledge-based system (IKBS), the behavior of the relevant human experts is monitored and recorded as a basis for automatically executable procedures.

* To whom all correspondence should be addressed.

Received October 14, 1993; revised January 24, 1994; accepted April 13, 1994.

Using a thesaurus is one component of the whole information retrieval task, which demands its own particular set of skills. The expertise consists of the ability to apply *heuristics* during the search process, that is, to decide:

- when to use the thesaurus;
- where to begin;
- which relationships (equivalence, hierarchical, or associative) to follow;
- which new terms to select; and
- how to include them in a query.

Besides an understanding of thesaurus structures, the intermediary brings to this task knowledge about the problem domain and about natural language. At a simple level, a standard thesaurus also holds such knowledge, but the standard is considerably enhanced in some experimental prototypes. Smith et al. (1989) give an account of a system where both document and thesaurus content is encoded into "frames" designed to answer the type of questions normally posed. As new items are added to the database these structures are updated dynamically—a process which clearly goes well beyond standard keyword indexing. The system reported by Chiaramella and Defude (1987) carries out extensive syntactic analysis of users' queries to map them onto the most appropriate index terms—the thesaurus is designed to be a "superset" of the index language and is used during the mapping process to help identify the meanings of unknown words. Any new terms generated during this process, which prove successful for retrieval, are saved in the thesaurus for future use.

Currently, however, most large-scale IR systems in general use consist of an indexed document database and a static thesaurus of terms and simple relationships. There are many such thesauri already in existence, designed in the first instance as printed documents to be consulted by human searchers, and there is an international standard setting out detailed rules for their compilation. So it is an important question how far it is possible to use them as a basis for an IKBS, without any enhancements except those that can be generated automatically.

A useful set of *general* heuristics for the intermediary was assembled by Harter and Peters (1985). The thesaurus is seen as a means to identify variant forms of index terms, and as a source of new terms during the process of "concept formulation" (or *re*formulation when the original query is unsuccessful). Proposals most relevant to the current discussion are the "heuristics for increasing or decreasing recall and precision", implementing strategies for *broadening* or *narrowing* boolean queries. For example:

- "use generic terms and related terms to obtain a very broad treatment of the search topic;
- move up or down in the thesaurus hierarchy to modify specificity;
- for high specificity, use only controlled vocabulary . . ."

Such heuristics could of course be offered as suggestions by a simple "advice-giving" expert system, but as they stand they are too weak to govern the action of a system attempting to do semiautomatic thesaurus navigation, even when converted into a more formal *rule* format, for example:

```
If ⟨size-of-hitlist⟩ is less than ⟨lower-limit⟩
  THEN expand query with broader terms
    ELSE
  IF ⟨size-of-hitlist⟩ is greater than ⟨upper-
limit⟩
  THEN refine query with narrower terms
```

Given the above rule, it would still be necessary to consult the user as to which of the many subordinate items which might be found in this way are really applicable to his query. In practice, since in a computer-based system there is no great penalty for following all paths simultaneously, designers tend to implement procedures for *ordering* the results of thesaurus navigation according to their likely usefulness. For instance, the prototype described by Shoval (1985) implements a "breadth-first" search through the thesaurus network from one or more starting nodes matching the query, picking up terms to augment or replace their predecessors in the search. The highest priority is given to terms which are reached by multiple paths, assumed to be most relevant to the original query. Terms are ordered by this "metric of strength" and presented to users for acceptance or rejection—there is also a possibility of backtracking to an earlier set of terms if the new ones are rejected.

Shoval (1985) considered that his system could be improved by adding *weights* to the network links to provide better ordering criteria. As mentioned earlier, such weights were used in the system reported by Kim and Kim (1990), though the authors comment that the task of assigning them was very difficult and knowledge-intensive work. In large-scale networks it would be almost impossible to handle it consistently, so objective methods must be sought.

Possible criteria for assigning weights are: the *number of connections* to the relevant terms (assumed to be related to their specificity), the *type of relationship* (e.g., *USE* > *NT* > *RT* > *BT*), and the *number of co-occurrences* of the terms as descriptors in a document database. Any general algorithm which weights successive "generations" of terms from a breadth-first search must also take into account the *path-length* from the starting node which is assumed to be the best match with the query.

Some interesting experiments on quite large networks have been carried out by Chen. In one case (Chen & Dhar, 1991), the relative weights for different relation-

ship types were derived by logging the behavior of a set of users navigating a thesaurus. In a later version (Chen et al., 1993), they were made adjustable by users as part of their search strategy. Weights based on term co-occurrences were taken from one thesaurus, then propagated to three others covering roughly similar ground, but again made adjustable according to users' ratings of the importance of each information source.

Once the numbers are there, sophisticated procedures certainly become possible—both Kim and Kim (1990) and Chen et al. (1993) recommend treating the thesaurus as a *neural network* and applying a spreading activation algorithm. There are, however, two basic problems:

- *Justifying the assignment of weights in the manner described above*—Weighting the links between thesaurus terms on the basis of their co-occurrence as document descriptors, for example, seems to confuse two rather different linguistic issues. Studies in natural language collocation over the last 30 years (see Sinclair, 1991) have shown that words very close in meaning are as likely to *repel* as to attract one another in short stretches of continuous text. In the more artificial context of controlled indexing this tendency may be less evident (or query enhancement via relevance feedback would not work as well as it does), but indexers are often instructed to treat a broader term and its subordinate as *alternatives* and assign only one of the pair to any document, so they may not co-occur as frequently as their semantic proximity would suggest.
- *Evaluating the results*—The complete system described in Chen and Dhar (1991) was subjected to statistical evaluation and performed well in comparison with traditional systems, but the thesaurus component was not considered separately so it is impossible to tell how much it contributed to the overall success. A strict evaluation of any term ordering would need to show that users had a greater than random probability of choosing the terms presented at the top of the list, and that those terms genuinely enhanced the original query. Even a very accurate set of link weightings would reflect facts about the language/subject domain as a whole, and might not produce the best results for any individual user.

Practical Investigations

We now give an account of some small-scale investigations in thesaurus navigation and query enhancement, carried out as part of the CILKS[1] project. Their intentions were:

- to study the behavior of thesaurus users under controlled conditions, and log their navigation paths automatically to identify strong patterns as a pointer to possible rule- or weight-based algorithms for term expansion; and
- to compare the effectiveness of queries enhanced by thesaurus use with that of the originals to determine what factors influenced success or failure in the task. Once again, the aim was to build successful strategies into automatic procedures.

The investigation was carried out with the INSPEC thesaurus and its associated document database. The retrieval engine was Okapi (Hancock-Beaulieu & Walker, 1992) which uses a "best-match" algorithm based on *term weighting* rather than boolean searching. This method tends to generate long hitlists, although the weighting algorithm attempts to show the most relevant items first. However, because most queries are short and the database is large and homogeneous, the hitlist often holds large "weight-blocks" of 50 or more equally ranked items.

Under these circumstances the simple notion of broadening or narrowing does not apply, and the most useful function of a thesaurus is to help users clarify their intentions more precisely, allowing the term-weighting process to bring the most relevant documents to the top of the hitlist.

The INSPEC thesaurus is stored as a separate Oracle database, and the user interface was developed with Oracle Forms3 and Pro-C. Jones (1993) gives an account of the data model designed for this system, and of the methods used to match an arbitrary user's query with a set of matching, or partially matching, thesaurus terms. Although it was slightly cumbersome to run the investigation on two different databases, it was useful to divorce thesaurus navigation from document retrieval so as to study it independently—also to exploit the resources of a general-purpose RDBMS to log and analyze details of each stage in the process.

The users who were subjects of the investigation were staff and students in the School of Informatics at City University. The main requirement was that they should be genuine users with real information needs, who might be expected to respond realistically (in terms of their internal perception of those needs) to offered thesaurus terms. However, this requirement conflicted somewhat with the need for controlled conditions (supervised thesaurus navigation sessions), and the fact that document retrieval with the expanded query was carried out as a separate operation. The method of obtaining subjects was to select (not sample) from the active regular network users of Okapi, and to invite them to conduct a search on a topic of current interest to them via the thesaurus navigation system. Clearly, no strong statistical validity can be claimed for the results of this investigation; it is intended only to indicate likely patterns of behavior.

The investigation was conducted in two phases, involving 18 and 21 users, respectively. (There was no overlap between the two groups.) The basic thesaurus

[1] The City Interactive Linguistic Knowledge Structure project is funded by the Joint Information Systems Committee of the UFC under the Knowledge-Based Systems Research Initiative.

TABLE 1. A log of one user's thesaurus navigation.
Query: "least squares Chebyschev polynomial."

DIST	PATH	TERM	EXPANDED_FROM
0		least squares approximations	
1	<	approximation theory	least squares approximations
1	–	curve fitting	least squares approximations
2	<–	function approximation	approximation theory
2	<–	interpolation	approximation theory
2	––	data reduction	curve fitting
3	<–>	Chebyshev approximation	function approximation
3	<––	function evaluation	function approximation
3	–––	spectral analysis	data reduction

navigation procedure was the same for both phases, and the 39 users were treated as a single group to produce the figures given in the next subsection. With regard to the handling of query enhancement and relevance feedback however, preliminary results for the first phase were studied leading to substantial changes in the way the second phase was conducted (see Query Enhancement and Relevance Feedback).

Thesaurus Navigation and Logging

Users were given a short explanation/demonstration of the interface, then asked to enter a query. The system listed thesaurus terms completely or partly matching the query, ranked according to the number of matching words and their inverse frequencies. Users could choose any number of these terms; if any were lead-ins there was an additional choice amongst the corresponding preferred terms. At any point during the interaction it was possible to view all terms selected so far—when there were options to pick a term for expansion, deselect terms, or save the whole set and exit. When a term was expanded all its broader, narrower, and related terms were displayed, and once again any number of these could be chosen and added to the current set. At the end of the session the list of selected terms was written to a file for later submission to Okapi.

During the navigation, details of all terms seen, selected, and expanded were recorded in the database, so that it was possible to look retrospectively at users' behavior, both individually and in total. Table 1, for instance, shows an analysis of one user's session with the thesaurus, giving a history of all terms expanded and selected, and showing, for each one, the distance from the start, its predecessor, and a symbolic representation of the path of relationships followed to reach it. The symbols $\langle$, $\rangle$, and – stand for NT, BT, and RT respectively. So, for example, the term "spectral analysis" was derived from the original query via three RT relationships, by following the chain from "least squares approximation" through "curve fitting" to "data reduction." In Table 1, only terms actually selected by the user are shown, although all the terms seen are recorded in the database and form the basis of the summary statistics discussed later.

Total numbers of terms seen and selected, by distance and by type of link followed, were also accumulated. Table 2 shows the averages for all participants in *both* phases of the investigation. Note that the breakdown by distance includes details of terms matching the original query (distance zero) while the breakdown by relationship includes only terms seen and selected *after* following a link. Hence, the difference in the overall totals for the two tables.

The raw data for the first part of Table 2 (in the form of terms chosen/rejected against distance, with the last two rows combined because of the small numbers involved) were highly significant on the chi-square test ($p < .01$) against the null hypothesis that the proportion chosen was independent of distance.[2] Impressionisti-

TABLE 2. Thesaurus navigation statistics.

Breakdown by distance (averages)

DIS	TERMS_SEEN	TERMS_CHOSEN	ROW %AGE
0	80.66	3.28	4
1	20.18	2.74	14
2	25.28	1.64	6
3	12.90	1.41	11
4	6.77	.61	9
5	1.66	.05	3
6	.30	.00	0
Overall	147.77	9.74	<–6.5

Breakdown by relationship (averages)

REL	TERMS_SEEN	TERMS_CHOSEN	ROW %AGE
BT	6.23	.51	8
NT	9.54	1.02	11
RT	51.38	4.92	10
Overall	67.10	6.46	10

[2] However, this result should be taken with a pinch of salt, because the true "sample" is the 39 users involved rather than the 5763 terms, and even the former was not a genuine sample.

cally, the data indicates that some users were prepared to navigate and select terms up to five moves away from their starting point, although a later comparison (see Table 7 below) suggests that that was not necessarily a good strategy.

The second part of Table 2 (breakdown by relationship) showed no significance on a similar test. The majority of terms retrieved by thesaurus navigation came through the associative relationship, reflecting the fact that the INSPEC thesaurus at the time contained 10,200 such links as against 5995 hierarchical ones. A common assumption in thesaurus research is that the hierarchical relationship is the "strongest" and most useful—for example, experiments reported in Rada (1991) show that a "distance metric" between query and document derived from the hierarchical links of a thesaurus correlates more reliably with expert opinion than one based on associative or other links, unless very specific query facet information is taken into account. By contrast, our users showed no strong preferences for hierarchical links. However, the context of use was very different—not direct query-document matching but query expansion, where users were not consciously trying to broaden or narrow their search by moving up and down hierarchies, but to find other terms to clarify their requirements. In these circumstances, "lateral" links which throw up genuinely new words and phrases may well be perceived as equally useful.

Perhaps the most important important point to note from Table 2 is that, on average, users selected only about 6% of the terms which they saw (or 10% following term expansion), and that no one factor was highly influential in their choice. No automatic procedure involving wholesale unselective term expansion is likely to mirror real user behavior, neither is there much evidence here to justify the use of differential weighting during the navigation process.

Query Enhancement and Relevance Feedback

In the first phase of the investigation, the main comparison was between the original query and the set of selected thesaurus terms, used in a *controlled language* search on the descriptor field of document records. A free-text search, using the component words from the thesaurus terms, was also performed. Users were presented with three lists comprising short details (author, title, journal, etc.) of the top-ranked 50 documents from each search. Of the original 18 users, 16 made relevance judgments (yes, no, possibly) on these lists, with results as shown in Table 3.

Clearly the differences between search types would not show significance on any statistical test. The lack of improvement in the overall performance of extended queries was disappointing, particularly as they generated fewer large weight-blocks than the originals, and should in theory have been more discriminating. However, for

TABLE 3. Relevance judgment summary. Number of documents found relevant, by type of search.

	Original query	Controlled language	Free-text
Range	1–42	1–39	2–34
Mean	12.00	11.62	13.00
Standard deviation	10.84	10.76	9.12

12 users, either or both of the new queries produced a modest improvement, while the others showed a stronger tendency in the other direction. It seemed that these users failed to find good matches for their original query in the thesaurus and, under the artificial conditions of the experiment, were forced to pick terms only tenuously associated with it. Another problem was that users were asked to judge relevance on comparatively little evidence (e.g., no document abstracts), and some of them commented on the difficulty of making sensible decisions in these circumstances. For the second phase, then, the following changes were made:

- As well as the query types used in the first phase, a *hybrid* query was generated containing both the original query words and the additional controlled-language descriptors.
- Document details shown to users included *abstracts* to allow more accurate relevance judgments. Judgments were made on the top 20 documents from each of the four searches, but before being shown to users the results were *pooled* and sorted alphabetically to eliminate bias caused by order of presentation.
- Details of the top 500 documents generated by each search for each user were saved in the database. This made it possible not only to compare the number of relevant documents in each case, but to see how the enhanced queries caused changes to individual document weights and positions in the hitlist.

Twenty-one users took part in the second phase. One did not select any terms from the thesaurus so only his original query was run, yielding 11 relevant records. We can conclude that, for him at least, the thesaurus was not useful, and his results are omitted from the following analysis.

One very obvious effect of query expansion was to raise document weights and break up large weight-blocks. Table 4 shows min, max, and average for document weight and size of weight-blocks, for all four query types. No less than 16 of the 20 original Okapi queries had a weight-block >20 at the top of the list, so users saw only an arbitrary subset of the highest-weighted records. Comparative figures for the other searches were three, two, and zero, respectively.

Unfortunately, the expanded queries once again failed to show a marked improvement in overall perfor-

TABLE 4. Weights and weight-blocks by query-type.

QTYPE	MINWT	MAXWT	AVGWT	MINBLK	MAXBLK	AVGBLK
Original	47	328	143.49	1	500	122.09
Controlled	0[a]	498	220.61	1	430	14.74
Hybrid	106	911	423.22	1	369	5.39
Free text	95	637	292.29	1	259	6.68

[a] Search based on one term so no document weight assigned.

mance. Altogether, 421 records were judged relevant[3] by the 20 users. Table 5 shows the number appearing on each of the four output lists at various document cut-off levels between 5 and 500. (Bear in mind that only those documents which appeared in the top 20 of any list were seen by users.) At level 20, relevant/nonrelevant records by query type gave a marginally significant result on the chi-square test ($p < .05$), mainly because of the free-text result. However, the same comments apply here as to Table 2.

Table 6 displays a more detailed comparison, by individual user, between results from the original query and the best-performing hybrid. The differences between success and failure are less clear-cut than in the first experiment—10 cases produced a better result, 9 cases a worse result, and in 1 case the result was the same. The table also shows how many thesaurus terms were selected by each user, but there is no apparent correspondence between this figure and the query performance.

The most striking point about Table 6 is the small overlap between the relevant document sets at the top of the two lists. At this level, the composition of the lists has changed almost entirely, with only ten relevant records retrieved by both original and hybrid queries. Further down, however, we find that query expansion is having a drastic *reordering* effect. Documents found by the original query will never be lost altogether by the hybrid search, but because their weights are low in comparison with those retrieved by a multiterm query, they are pushed so far down as to be effectively out of sight. In this case, 108 of the 157 relevant documents dropped below the 500 mark, 32 dropped below 100, 2 below 50, and 5 below 20. Perhaps only the last group would have a realistic chance of being seen by a user in normal circumstances.

Conversely, we can ask to what extent hybrid queries *promote* relevant documents from the original list to visibility, or bring in completely new ones. Although our data cannot give a complete answer, the sample shows that promotions account for at least 50% of the 152 (top 20) relevant documents from hybrid searches. In detail, 71 were not seen at all in the original list (they may or may not have appeared below 500), 38 do appear below 100, 16 below 50, and 17 below 20. Again, it may be fairly assumed that only the last set would be seen in a normal retrieval session, so perhaps the distinction be-

TABLE 5. Relevant records by query-type.

Level	Original	Controlled	Hybrid	Free-text
5	46	35	51	32
10	85	68	91	66
20	157	146	152	119
40	175	167	184	152
75	201	185	228	180
150	215	207	267	213
300	249	219	288	245
500	260	235[a]	305	270

[a] Two searches produced <500 documents.

[3] In spite of the fact that users saw document abstracts in the second phase, there were still 126 "don't knows" in the relevance judgments. Treating these as "yeses" made no significant change in the overall pattern.

TABLE 6. Comparison between original and hybrid query results (top 20 records) by user.

User ID	Original	Hybrid	Difference	Overlaps	Terms
17	11	1	−10	0	5
20	15	5	−10	2	2
9	5	1	−4	0	10
18	5	1	−4	0	3
14	7	4	−3	0	15
16	6	4	−2	0	3
3	4	3	−1	0	19
7	10	9	−1	2	6
21	20	19	−1	0	6
13	0	0	0	0	10
11	0	1	1	0	9
6	13	14	1	0	15
12	18	19	1	3	3
10	0	2	2	0	9
4	4	6	2	0	10
1	5	7	2	0	3
5	16	20	4	0	12
15	9	14	5	2	2
2	1	7	6	1	1
8	8	15	7	0	10
Totals	157	152	−5	10	153

TABLE 7. Thesaurus navigation: breakdown by distance. Comparison between "failures" and "successes."

	Failures			Successes		
DIST	TERMS_SEEN	TERMS_CHOSEN	COLUMN %AGE	TERMS_SEEN	TERMS_CHOSEN	COLUMN %AGE
0	82	7	15	382	19	44
1	46	13	28	130	13	30
2	126	9	19	133	8	18
3	71	12	26	69	3	7
4	37	5	10	13	0	0
5	5	0	0	—	—	—
Total	367	46		727	43	

tween finding new documents and promoting existing ones is somewhat artificial anyway.

The conclusion seems to be that thesaurus-based query expansion does indeed increase recall and bring new relevant documents to the top of the list, but that these records often usurp the places of those from the original query, leaving fairly constant the overall proportion of useful material which users are likely to see.

Successes and Failures

The last part of this analysis concerns our attempts to identify any aspects of users' thesaurus navigation behavior which contributed to "successful" query enhancement. Accordingly, two groups of users were isolated: the "best" five and the "worst" five from the list shown in Table 6. (The position of these users on the list was little affected by whether "don't-know" judgments were or were not treated as relevant, whereas some of those in the middle of the list were more volatile.) The original navigation statistics for these groups were analyzed separately, with results as summarized in Table 7. The "column percentage" figure here indicates what proportion of *all* terms chosen were picked at the relevant distance.

Although the average number of terms chosen was very similar, the pattern of selection was not. Briefly, the successful users had twice as many terms to choose from, tended to navigate less far through the network, and picked a high proportion of their terms at distances zero or one. Once again, success seems to be correlated with finding good matches for one's query in the thesaurus to start with, and having a selective approach. Comparison of the breakdowns by relationship type showed no differences worth discussing, but the verbal feedback from users obtained following thesaurus navigation indicated that those who felt their new query was in some sense "more specific" than the original one were more likely to get good results. One of these made the interesting comment that: "the new query is both more narrow and more broad: it has more specific topics but a larger number of topics". Clearly, the numbers under examination here are too small to do anything but suggest possibilities for future investigation—but it does seem that users should be invited positively to review their list of terms before leaving the thesaurus, and to delete any which look as if they might be too general.

Implications

What do the investigation described above imply about the possibility of automatic or "intelligent" thesaurus navigation? We have certainly found little evidence of patterns which could be used to justify strong rules or weighting algorithms, indeed even the basic assumption that thesaurus-based query expansion will improve retrieval performance has come under question. Nor is the experiment described above unique in this respect—an interesting comparable experience is reported in Voorhees (1993). The work there involved the use of Wordnet—a semantic structure which makes finer subclassifications of the basic equivalence, hierarchical, and associative relationships than a standard thesaurus. A number of trials in automatic query expansion were carried out, varying, for example, the relationships used, length of paths followed, and weightings applied. Using a fairly conservative strategy the retrieval effectiveness of original and expanded query was very similar; more aggressive strategies produced *poorer* performance overall. The authors concluded that thesaurus expansion is primarily a "recall-enhancing" technique and so will not improve queries which are already quite fully specified.

Our own small-scale experiments involved selective *user-driven* thesaurus expansion on initially short queries so in principle the results should have been better; nevertheless, as we have seen, the case for using a thesaurus is by no means cut and dried. It seems from our small sample that users get good results when:

- they have plenty of terms to choose from (so they need

well-designed thesauri giving a good coverage of their particular field); and
- they are discriminating in their choice (so they need to control the process of thesaurus navigation closely, rather than rely on automatic procedures based on *general* facts about term relationships).

Verbal feedback from users taking part in our survey indicated that many of them found the preliminary thesaurus navigation useful and informative. Future developments will be directed toward giving *all* Okapi users access to the thesaurus during normal searches, while continuing to log usage details automatically and quantify the results. Our basic principle is still that a term-weighted IR system like Okapi should perform better with multiterm queries which capture users' intentions more precisely and break up large weight-blocks in hit-lists, and that the thesaurus has a useful role to play here. But attempts to demonstrate "intelligence" must adopt a more subtle approach than the ones discussed so far. Our current work involves trying to identify separate topics or *facets* within queries, and ensuring that they are treated as separate entities both during thesaurus navigation and in the structure of the expanded query.

In a broader context, researchers trying to exploit thesaurus knowledge to improve retrieval performance still have many possibilities to explore. A thesaurus can be viewed as a *bridge* between queries phrased in natural language and an abstract classification structure which constitutes a "map" of a particular domain. In some systems, as we have already seen, it is the primary search mechanism for a set of documents which have been placed at some point on that map by an indexer. Further developments in this direction will require thesauri which are more highly structured than those defined by the current standard, richer in information (e.g., incorporating numeric weights on interterm links), and designed for/closely integrated with a particular document database. They should then be a suitable case for treatment by current IKBS techniques.

Alternatively, we can view the thesaurus mainly as a source of natural language terms for query enhancement in a more general context, not necessarily tied to a particular database, but useable with the free-text, unindexed documents that form an increasingly large proportion of the available online sources. The accuracy, depth, and coverage of thesaurus information will be the most important issue here, and the quality of the user interface for exploring it. However, there is still scope for improving performance by the application of AI techniques, particularly in natural language processing. For instance, the initial match between queries and terms has already been identified as crucial to the success of a thesaurus navigation. A system able to identify underlying syntactic structures could perform this task better than one relying on "brute-force" word and substring matching—likewise, the recognition of synonymy and homography, which current thesauri handle only in a very limited and superficial way.

REFERENCES

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management, 27,* 405–432.

Chen, H., Lynch, K. J., Bashu, K., & Ng, T. D. (1993, April). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert,* 25–34.

Chiaramella, Y., & Defude, B. (1987). A prototype of an intelligent system for information retrieval: IOTA. *Information Processing and Management, 23,* 285–303.

Chong, A. (1989). Topic: A concept-based document retrieval system. *Library Software Review, 8,* 281–284.

Hancock-Beaulieu, M., & Walker, S. (1992). An evaluation of automatic query expansion in an online library catalogue. *Journal of Documentation, 48,* 406–421.

Harter, S. P., & Peters, A. R. (1985). Heuristics for on-line information retrieval: A typology and preliminary listing. *Online Review, 9,* 407–424.

Jones, S. (1993). A thesaurus data model for an intelligent retrieval system. *Journal of Information Science, 19,* 167–178.

Kim, Y. W., & Kim, J. H. (1990). A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation, 46,* 113–116.

McMath, C. F., Tamaru, R. S., & Rada, R. (1989). A graphical thesaurus-based information retrieval system. *International Journal of Man–Machine Studies, 31,* 121–147.

Rada, R., & Bicknell, E. (1989). Ranking documents with a thesaurus. *Journal of the American Society for Information Science, 40,* 304–310.

Rada, R., Barlow, J., Potharst, J., Zanstra, P., & Bijstra, D. (1991). Document ranking using an enriched thesaurus. *Journal of Documentation, 47,* 240–253.

Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing and Management, 21,* 475–487.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Smith, M. P., Pollitt, A. S., & Li, C. S. (1992, April). Evaluation of concept translation through menu navigation in the MenUSE intermediary system. In T. McEnery, C. Paice (Eds.), *Proceedings of the 14th BCS IRSG Research Colloquium on Information Retrieval,* University of Lancaster. (pp. 38-54). New York: Springer Verlog.

Smith, P. J., Shute, S. J., & Galdes, D. (1989). Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems, 7,* 246–270.

Voorhees, E. (1993, April). *On expanding query vectors with lexically related words.* Paper presented at the Second TREC Conference, Gaithersburg, MD. Proceedings to be published by the U.S. Department of Commerce, National Institute of Standards and Technology.