Pergamon

Information Processing & Management, Vol. 31, No. 1, pp. 15-27, 1995

0306-4573/95 $9.50 + .00

0306-4573(94)E0004-L

TERMINOLOGICAL KNOWLEDGE STRUCTURE FOR INTERMEDIARY EXPERT SYSTEMS

RAYA FIDEL
Graduate School of Library and Information Science, University of Washington, Seattle, WA 98195, U.S.A.

and

EFTHIMIS N. EFTHIMIADIS
Graduate School of Library and Information Science, University of California, Los Angeles, CA 90024, U.S.A.

(*Received* 4 *November* 1993; *accepted in final form* 23 *January* 1994)

Abstract—An intermediary expert system (IES) helps both end users and professional searchers to conduct their online database searching. To provide advice about term selection and query expansion, an IES should include a terminological knowledge structure. Terminological attributes as well as other properties could provide the starting point for building a knowledge base, and knowledge acquisition could rely on knowledge-base techniques coupled with statistical techniques. The searching behavior of expert online searchers would provide one source of knowledge. The knowledge structure would include three constructs for each term: frequency data, a hedge, and a position in a classification scheme. Switching vocabularies or languages could provide a meta-schema and facilitate the interoperability of databases in similar subject domains. To develop such knowledge structure, future research should focus on terminological attributes, word and phrase disambiguation, automated text processing, and the role of thesauri and classification schemes in indexing and retrieval. In particular, such research should develop techniques that combine knowledge-base and statistical methods and that consider user preferences.

1. INTRODUCTION

An intermediary expert system (IES) helps users, professional searchers, and end users to conduct their searches of online bibliographic databases. Currently, most online bibliographic databases provide for searching the titles, abstracts, and sources of the bibliographic items to be retrieved in addition to descriptors and identifiers which have been assigned by human indexers, if they are available. This article examines the knowledge base of an IES that provides advice about the selection of search terms, or search keys. It presents a proposal for an integrated approach that would include various methods and techniques that are available today. Recognizing that these methods and techniques could be integrated in a variety of combinations, the article presents one option that focuses on terminological attributes that is based on knowledge acquired from professional searchers. This option creates a scenario that illustrates how various approaches can be used simultaneously and the effect such a combination would have on research. It examines what knowledge and information could be included in the knowledge base and how they could be organized. The article then shows what research would be required to support the development of this option.

Most commercially available search systems require the use of Boolean operators. Thus, before searching a request, a user breaks it down into concepts, the representation of which would be linked with Boolean AND operators. For actual searching, each concept is represented by one or more search keys. A search key is a string of characters to

Correspondence should be addressed to Raya Fidel, Graduate School of Library and Information Science, FM-30, University of Washington, Seattle, WA 98195.

be searched in the database. A search key, which represents a concept of a request, may consist of one or more words. The selection of search keys is at times a straightforward process; however, at other times it requires knowledge and expertise.

Consider the request "attitude of students toward themselves during examination period." The request can be broken down into three concepts: "attitudes toward themselves," "students," and "examinations." A straightforward approach to searching would be to search on the keys as they appear in the request in all available fields and then to intersect the resulting sets (using the AND operator). A professional searcher, however, would likely see much more complexity in the request and would probably try a variety of other search keys that would result in better retrieval. A searcher would probably decide to express the concept "attitudes toward themselves" in a phrase such as "self-image" or "self-esteem." Also, the searcher is likely to prefer to search the key "examination" only in the descriptor field because it is a common term; as a textword, it appears frequently in the text, often referring to concepts other than educational tests. An IES of the kind considered here would advise users of the most promising search keys to be used.

It is well established by now that relying only on the words in a request is not sufficient for satisfactory retrieval (Svenonius, 1986). Indeed, research into query expansion—the process of supplementing the original query with additional terms—has been motivated by this observation (Efthimiadis, 1991). In addition, databases that use an indexing language require users to make another decision: whether to enter the search key as a textword key, which would retrieve all bibliographic database records that include the key in any field of the record, or as a descriptor, which would retrieve only the records whose descriptor field includes the key. An IES of the type considered here should be able to help users in this decision as well.

The interaction that takes place in information retrieval between users and the database searched can be described with the use of a simple two-stage model (Efthimiadis, 1991; Efthimiadis & Robertson, 1989). The model includes the end user, the intermediary mechanism, and the database. The intermediary mechanism may be a human being or some software, such as a front-end system or an expert system. Here we consider an intermediary mechanism that is a machine, as described in Fig. 1. For simplicity in the discussion, the IES is treated as part of the retrieval system. However, an IES could reside anywhere; it could reside between the retrieval system and the user-supporting initial query formulation; it could be a front end or client at the user end, a front end at the database end, or an integral part of the retrieval mechanism.

Depending on the characteristics of the particular request searched, the user's level of expertise, and the database searched, an IES could provide help in three modes:

1. The system decides about the search key with no consultation with the user.
2. The system decides about the search key after interrogating the user.
3. The system presents options from which the user is asked to make a selection.

The decision about which mode of advice to provide is situational.

In general, we can identify two main sources which can provide knowledge to be utilized or incorporated in an IES (Efthimiadis, 1990). The first source of knowledge is the search intermediaries. Here, the approach that has been taken so far is to try to encapsulate their skills in a system, such as in PLEXUS (Vickery *et al.*, 1987), IR-NLI

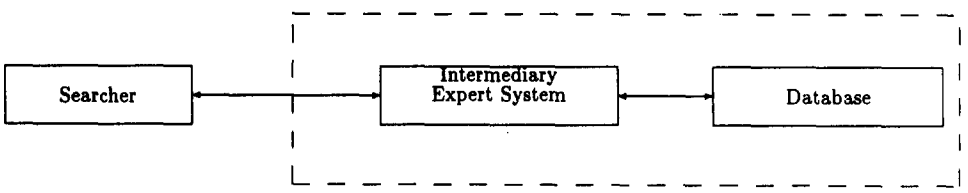

Fig. 1. The role of an IES in the two stage model of interaction in information retrieval.

(Brajnik *et al.*, 1986, 1988), and EP-X (Shute & Smith, 1992). The second source is the knowledge structures found in databases or embodied in search aids or indexing languages, such as thesauri or classification schemes like EP-X (Smith *et al.*, 1989a, 1989b), MENUSE (Pollitt, 1988), and UMLS (Humphreys & Lindberg, 1989).

It is promising, however, to integrate knowledge and information from both sources. For instance, terminological knowledge (i.e., knowledge about terms and their properties) acquired from professional intermediaries can point to terminological issues that are relevant to retrieval. At times, however, professional searchers may not be in a position to provide the best solution to a terminological problem because there is not enough information for them to make the most useful decisions. Given the existing techniques in information retrieval, help can come from additional sources. Associative retrieval techniques create one of these sources. While not yet widely available, various techniques have been developed and tested over the last two decades. These techniques are not incompatible. It is possible to devise methods based on more than one associative retrieval approach, and such mixed methods may be appropriate for certain retrieval situations. Furthermore, it is also possible to combine associative and Boolean techniques to enhance both through knowledge-based retrieval techniques.

Current prototype IES vary in the help they offer in terms of query negotiation aids, the selection of search keys, and query expansion. Because it is difficult to automate assistance offered at the query formulation stage, there have been various attempts to deal with it at the interface level (e.g., Pollitt, 1988; Thompson & Croft, 1989; Vickery, 1988; Vickery *et al.*, 1986). Some of the systems use thesauri and classification schemes to assist query formulation, for navigation and retrieval (e.g., Frei & Jauslin, 1983; Monarch & Carbonell, 1987; Pollitt, 1987, 1988; Shoval, 1981, 1985; Smith *et al.*, 1989a, 1989b; Vickery, 1988). A few experimental expert systems already incorporate in their knowledge bases knowledge that is pertinent to the selection of search keys. MedIndEx at the National Library of Medicine, for instance, incorporates the Medline indexing policy (Humphrey, 1989; Humphrey & Miller, 1987), a system at the American Petroleum Institute employs knowledge acquired from professional API indexers (Brenner *et al.*, 1984; Martinez *et al.*, 1987), and a system at BIOSIS incorporates knowledge on biological concepts in a semantic vocabulary (Vleduts-Stokolov, 1987). Although their knowledge bases could be incorporated into IESs, these expert systems were designed initially to assist indexing rather than searching. To date, terminological knowledge as employed by searchers has not been explored.

To focus the discussion on search-key selection, it is assumed that a request is already broken into its concepts, the databases are selected, and the user is prepared to enter search keys. Further, to provide advice in the selection of search keys and the field(s) to be searched, an IES must possess expert knowledge about the particular database that is being searched. Therefore, it is assumed here that such a system would provide advice for searching a defined set of databases covering a certain subject domain.

2. KNOWLEDGE USED BY EXPERIENCED SEARCHERS

The first type of knowledge to be incorporated into an IES is knowledge about the selection of search keys acquired from experienced online searchers (Fidel, 1991b). This type of knowledge was collected through observations of searchers performing their regular, job-related searches and through interviews with them. The study team analyzed search protocols, verbal protocols of thought processes while searching, and the transcripts of interviews, with 47 searchers performing a total of 281 searches in a variety of subject areas and library types. The analysis of the search and verbal protocols uncovered the intuitive rules that searchers used and resulted in a decision tree for the selection of search keys which is called the selection routine.

The selection routine embodied in the decision tree describes the conditions that searchers considered and the options that each condition generated. For example, the condition "a search key is mapped to a descriptor through an exact match" generated the following options: enter the descriptor, but if recall needs to be improved: add textword synonyms

to descriptors, or use generic descriptors in an inclusive mode ("explode," or "cascade"), or add the next broader descriptor in the hierarchy. Data were gathered on the frequency with which the different options were selected. Thus, of the 228 cases in which a descriptor was an exact match and searchers wanted to increase recall, 72% of the time they entered textwords as synonyms, 25% they did an inclusive search, and 3% of the time they selected a broader descriptor. Also, data were gathered on reasons associated with special conditions for each option. Thus, searchers selected textwords as synonyms because the user insisted on using the terms, because they needed to perform a multidatabase search, or because they did not trust the descriptors and/or the indexing of the database. They performed an inclusive search when the query formulation included a relatively large number of concepts, and they entered a broader descriptor when they thought the user would be interested in the broader descriptor as well.

Data of the type just described could be incorporated into the knowledge base and the inference engine of an IES for a specific set of databases. The frequencies with which options were selected could be used to handle uncertainties. The frequencies, together with other factors, could also determine the mode of advice to be given. For example, if a search key were matched to a descriptor through an exact match, the system might automatically enter the descriptor without consulting the user. This would be reasonable, given the searching behavior of the professional searchers: 100% of the time when there was an exact match they entered the descriptor; only if recall needed to be improved they selected additional keys. The next step, then, is for the system to inquire if recall is satisfactory. If it needs to be improved, other questions can be asked and advice given.

Selection routines by themselves are not sufficient for the IES to advise about the selection of search keys. Clearly, the system must include the database's thesaurus to be able to map search keys to descriptors. But it should include additional knowledge about the database as well. For example, it is important to include information about the indexing of the database—how often it is used by professional searchers, and the indexing policy that created it. This requirement is based on the finding that often searchers selected textwords when they did not trust the database's vocabulary and indexing, and they frequently referred to the indexing policy when giving reasons for their search decisions. For instance, if exhaustive indexing is mandated in an educational database, searchers may enter a frequently occurring descriptor such as "students" as a major descriptor, limiting the retrieval to documents in which "students" is a central topic.

In addition to the knowledge about the database, its controlled vocabulary, and indexing, the IES must incorporate terminological knowledge and information about individual search keys that is not available in thesauri and indexing manuals. For example, suppose a user with the request about students wishes to increase recall. The selection routine, as described earlier, shows that the most common option is to add the textword key "student" to the descriptor. A professional searcher, however, is not likely to select this option in an educational database because the textword occurs too frequently. How would an IES know to eliminate this option? To do so, the system would have to incorporate additional knowledge and information (for instance, information on the relative frequency of search keys in the database). Further, some databases do not have sources for terminological knowledge such as thesauri and indexing manuals because they are not indexed with controlled vocabulary. The system's knowledge base, however, should also be able to provide advice for the selection of search keys in such textword databases.

3. TERMINOLOGICAL ATTRIBUTES

Searchers considered a variety of attributes when they selected search keys (Fidel, 1991a). Some examples are the number of databases to be searched; the number of components in a request; and specific attributes of a term. The latter includes whether the term was added just to increase recall (ORing it) or used as a limiting factor (ANDing it); whether it occurred in the records of relevant citations and in which fields; whether it was mapped to a descriptor and through what kind of a match, exact or partial; and whether the user insisted it appears in the retrieved text. A close examination of these attributes

shows that most can be incorporated into an IES and used without additional knowledge. Terminological attributes are the one exception. Very few existing searching tools include terminological information perceived by searchers to be pertinent to the selection of search keys.

Each search key has a variety of terminological attributes which are determined by the terms it includes. Here we focus only on those attributes that require special considerations for searching. The study of professional online searchers revealed that searchers considered terminological attributes most often when they thought that a search key might not be a "good" textword; that is, if entered as a textword, it would produce unsatisfactory retrieval. Therefore, an IES would incorporate information about those attributes that determine the suitability of a search key to be entered as a textword.

The terminological attributes identified by the searchers as being relevant to the choice of search keys conform with those discussed in the literature; see Fugmann (1982b), Svenonius (1983), and Knapp (1988). Thus, according to the study's searchers, search keys with the following attributes are not suitable for textword searching:

- The key has many synonyms.
- The key is ambiguous; for example, archival "record" is different in meaning than database "record" or music "record."
- The key is vague; for example, "health promotion" has no agreed upon boundaries to its definition.
- The key occurs too frequently in the database's text.*

It is well accepted by now that controlled vocabularies and indexing are necessary to overcome the retrieval problems caused by these terminological attributes. Clearly, an index language controls for synonyms, ambiguity, vagueness, and context dependence, and it is also useful as an alternative to highly frequent textwords. For instance, a study of the ERIC database found that the search key "student" retrieved 115,061 citations when entered as a textword, but only 1,242 as a descriptor (Markey *et al.*, 1980). However, while incorporating a database's thesaurus into an IES knowledge base is necessary, it is not sufficient because two functions of an IES would require additional knowledge. First, to advise users about the selection of search keys an IES would recognize problems, or "diagnose" each key to determine whether it is good for textword searching. Second, to advise users searching textword databases, an IES would require terminological knowledge that is not included in current thesauri.

What then would be included in its knowledge base for an IES to diagnose and provide advice about the selection of search keys in all databases? The following are some exploratory ideas based on knowledge acquired from expert searchers.

4. TERMINOLOGICAL KNOWLEDGE STRUCTURE

The IES's knowledge base would include a variety of machine-readable sources of knowledge that are already available. It would include databases' thesauri that are pertinent to the subject domain, the relevant terminological databanks, machine-readable dictionaries, and Roget-like thesauri. In addition, the text stored in each database would be accessible to statistical manipulations.

Given these sources, a knowledge structure would be constructed to include knowledge generated from these sources and from external, intellectual sources, such as subject experts, indexers, and users. The knowledge structure may be an expanded thesaurus or a semantic net; parts of it may reside permanently in the knowledge base, and others may be generated ad hoc, triggered by specific situations. In addition, each content-bearing term that is relevant to the subject matter which appears in the databases' text would be included in the knowledge structure.

*Frequency data by themselves are not considered to be terminological attributes.

To diagnose problematic search keys and to advise in the selection of alternative or additional keys, the knowledge structure would include three constructs for each term, whether textword or descriptor: frequency data; a hedge; and a position in a classification scheme.

4.1 *Frequency data*

The number of times a search key occurs in the database can help determine whether the key occurs too frequently. Frequency data is commonly supplied by current search systems in the form of number of postings. These data by themselves, however, are not sufficient to determine whether a key is too frequent to be used for textword searching. For instance, databases vary in size so it is unlikely that a universal number could be found that would designate the threshold of too-high frequency.

One method to determine whether the frequency of a search key presents difficulties in searching is to compare the frequency with which the key occurs as a textword with that with which it occurs as a descriptor. An empirically based rule could then be established to determine which difference is significant. For instance, such a rule may state that if a textword occurs more than a certain number of times and if the difference is more than one order of magnitude, the search key is not suitable for textword searching. Note that even a highly frequent term might be used successfully (for instance, when it is entered in conjunction with other terms or when it is used as a limiting factor).

Because a terminological knowledge base is limited to a particular subject domain, it is plausible to assume that a search key that is too frequent in one database is likely also to be frequent in the other databases covering the same subject. While this assumption still requires empirical validation, it supports the expansion of the method described earlier to databases which can be searched only with textwords because of lack of a controlled vocabulary. For such textword databases, a list of the most frequently occurring keys would be compiled. These keys would then be checked for frequency in databases with controlled vocabularies; the difference between textword and descriptor occurrences would be measured to determine the suitability of these keys for textword searching. Using this method, then, all the search keys that are not suitable for textword searching because of high frequency would be designated as such. Finally, additional information drawn from the statistical approaches to information retrieval could be used to complement the approach discussed here.

4.2 *Hedges*

Frequency data, including frequency of co-occurrence, is also instrumental in the creation of hedges. A hedge is commonly understood to be an "OR string of terms that cover a topic for which no single term has sufficient extension" (Sievert & Boyce, 1983, p. 491). Hedges are essentially clusters of semantically related terms. Thus, they have been used by searchers as a list of synonyms to expand a concept of a request. As clusters, hedges are the equivalent of the term clustering approaches of automatic indexing (strings, stars, cliques, and clumps). However, the latter are based on statistical associations, usually term co-occurrence, rather than on semantic content.

The need for hedges was realized in the early 1970s, when online searching started to become common. Since then, hedges have been developed by individual libraries and searchers, as well as by database producers and search system vendors. Figures 2 and 3 show examples of predefined hedges.

Other terminological tools, however, can be viewed as hedges. Piternick (1984), for example, shows that as of 1984, a variety of enhanced thesauri, synonym listings, and early forms of switching languages were available to online searchers to enhance their searching vocabulary. Thus, an enriched lead-in (entry) vocabulary can provide a list of textword synonyms for a descriptor, and a list of equivalent descriptors in a number of databases can serve the same function. For instance, the now defunct BRS's TERM database incorporated the social science descriptors of five major thesauri (Knapp, 1992). In addition to these intellectually developed tools, more ambitious, computer-assisted projects, such as UMLS, the Unified Medical Language System (Humphreys & Lindberg, 1989), are

Topic: tests and measurements	
1	(test or tests or testing or subtests or pretesting).de.
2	(posttesting or inven or invent or inventory or inventories).de.
3	(surv or survey or surveys or scale or scales or scal or score).de.
4	(scores or measurement or measures or screening or exam).de.
5	(examination or examinations or questionnaire or questionnaires).de.
6	(rating or validity or psychometrics or sociograms).de.
7	(assessment or sociometry or piagetian-tasks).de.
8	semantic-differential or piagetian-tasks
9	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8

Fig. 2. "Hedges" (Source: Psychological Abstracts: database PSYC on BRS).

being developed to establish machine-generated tools to enhance the searching vocabulary in specific subject domains. Similarly, global thesauri, such as Roget's thesaurus or Wordnet,* are also being used in an attempt to test their applicability/suitability as vocabularies in searching.

An IES would integrate all such tools that are pertinent to its subject domain into its knowledge base. If a synonym list, or an enriched lead-in vocabulary, does not exist for a certain subject matter, it might be possible to develop it for the IES. While some automated methods have been applied, most are still experimental. Experience indicates, however, that it is realistic to develop such tools without automated techniques, or to verify the results intellectually of automated methods (e.g., Anderson & Rowley, 1992).

Hedges would be dynamic in nature. Although technical considerations would determine which parts of a hedge would be stored permanently in the knowledge structure and which parts would be developed ad hoc, hedges would evolve during the lifetime of an IES. Changes and developments in the vocabulary would be reflected in the text stored in the databases and in the revisions of thesauri and terminological databanks. Users and requests would also contribute to the refinement of hedges. Through machine-learning procedures, terms would be added to hedges or deleted from them. An example of such mechanism is provided by the experimental system TEGEN, which is designed to construct thesauri

*Wordnet is available from Princeton University.

Topic: Air pollution

Suggestions for terms to OR:

aerobic; AIR; air born; aerosol; asthma; aerial; bronch; dust; exhaust; EMPHYSEMA; fume; inhal; lung; nose; pleura; pulmon; respirat; smoke; SMOG; throat; trachea; centilat; vapor.

Fig. 3. A hedge from the Hedge Book of a medical library prepared for searching Medline. In this partial list, MeSH headings are printed in caps.

automatically. The system first acquires terms and relationships from users' requests and then verifies relationships with users' help (Guentzer *et al.*, 1989).

In the terminological knowledge structure, each node, or search key, would have a hedge that would include other keys, or hedge terms, which are associated with the node. For simplicity's sake we assume here only direct relationship between each member and the node. Such a hedge would have to support various searching decisions in addition to providing lists of synonyms for query expansion. As a result, it would be a special kind of hedge, an expanded hedge: It would not only cluster terms, but it would include information about the relationship between each term in the hedge and the search key (the node), such as frequency of co-occurrence and semantic relatedness, and it may also include terms that are not synonyms.

A hedge of a search key would include a list of all the terms, both descriptors and textwords, that co-occur with the key in the text stored in the databases. A designation of the relative level of frequency with which it co-occurs would be given for each term in the hedge, whether derived by co-occurrence analysis or otherwise. In addition, based on terminological tools such as machine-readable dictionaries, terminological databanks, and Roget-like thesauri, a semantic scale might be created that expresses the semantic relatedness between each term and the key would be designated; see Liddy *et al.*, 1991; Krovetz, 1991 and Nutter *et al.*, 1990.

With these relationships explicitly expressed, an IES would provide advice concerning most of the terminological properties. Although this advice would be generated by the inference engine supported by the knowledge base, it is useful to discuss a few examples to show how this knowledge could be used. Because much research is still required to find effective methods to deal with terminological attributes, we give these examples only to illustrate how the knowledge could be structured, not how it should be structured.

1. *A search key has many synonyms.* A diagnosis would be facilitated by checking the number of hedge terms that relate highly to the key on the semantic scale. If the number is relatively large, a ranked list of terms would be presented to the user. First on the list would be the descriptors (if available), followed by highly similar hedge terms arranged in descending order of co-occurrence frequency with the key. This would encourage the user to enter descriptors, which is the preferred action according to the selection routine for the case of a key having many synonyms. If the key has already been used as a textword and the user is looking for additional synonyms to increase recall, those textword synonyms that do not co-occur with the key, or those that co-occur the least frequently, would be suggested as most promising. The rule could be further refined based on frequency data for each hedge term. In addition, probability estimates could be used to determine which term to include and how to weight individual terms in the hedge.
2. *The search key occurs too frequently.* Such a search key is usually not suitable for textword searching and the IES would suggest alternative terms. It would first retrieve from the hedge descriptors that closely relate on the semantic scale to the key. Next, hedge terms that are semantically related would be listed in descending order of co-occurrence frequency, an order that could be refined by term-occurrence frequencies. The user would first view the descriptors and then the synonyms that co-occur frequently with the key but are themselves suitable for textword searching.
3. *The search key is ambiguous.* Various mechanisms have been suggested to disambiguate terms. Krovetz and Croft (1989) explain the early methods for word sense disambiguation using machine-readable dictionaries. Disambiguation tools used in the semantic vocabulary and the concept headings of the BIOSIS system are a combination of contextual restrictions, multiword entries, and an associated weighting technique, which is used when the corresponding meaning of the words cannot be derived from context (Vleduts-Stokolov, 1987). Ahlswede *et al.* (1988) and Nutter *et al.* (1990) describe the use of a machine-readable dictionary, the *Webster's Seventh New Collegiate Dictionary*, as a source for identifying semantic relationships between index terms and for linking phrases to index terms. For disambiguation

they make use of the lexical-semantic relationships, the selection restrictions, and the verb categories available in *Webster's* to create defining formulae. Wherever this information is not available, they rely on Sager's linguistic parser. Veronis *et al.* (1990) describe a method which combines different machine-readable dictionaries and connectionist models. The resulting neural network is used for word sense disambiguation.

The knowledge structure proposed here facilitates another approach. In addition to descriptors, the user would be presented with clusters, created according to semantic relatedness, of hedge terms that frequently co-occur with the ambiguous key. The ORed cluster that represents a desired point of view can then be ANDed with the ambiguous key to improve precision. For example, the ambiguous key "record" may have two clusters, one that includes terms typical of the database literature and another with terms pertinent to archives. A similar procedure can be used for search keys whose meaning depends on the context in which they appear.

With machine-learning techniques, hedges would increase their usefulness. For example, it might be possible to address the problem of nonlexical expressions. Fugmann (1982a, p. 141) defines a lexical search key as "one which consists of a linear sequence of alphanumerical symbols, which by general agreement is used to represent a certain meaning." He also explains why nonlexical search keys are not suitable for textword searching. For example, the phrase "attitudes of students towards themselves" is nonlexical. It could, however, be expressed by two lexical keys: "students" and "self-image" (or "self-esteem," etc.). Obviously, it is almost impossible to predict all the nonlexical expressions users would bring to an IES or that may occur in the text. With machine-learning techniques, however, it might be possible to analyze nonlexical expressions when they are presented to the IES. If a lexical presentation is identified, it could be stored for future requests which include this nonlexical phrase. For example, a proximity combination of the terms "attitudes" and "themselves" may reveal a high co-occurrence with "self-image." Once it is established and verified that "self-image" is representative enough of "attitudes toward self," the latter can be stored as a lead-in term.

If a new type of relationship is found to be useful for retrieval, such as the frequency with which terms co-occur in requests, the new relationship and the associated terms could be added to the hedge. Future advances in theoretical and computational linguistics, in terminological research, and in automated text processing and retrieval would enhance these expanded hedges and enrich their capabilities and effectiveness.

4.3 *Classification scheme*

A classification scheme would provide an overall semantic structure. It would establish hierarchical links between keys, or nodes, based on the keys' meaning and would designate the type of such link: member/class, part/whole, etc. As an overall structure, the classification scheme would also facilitate the combination of nodes into concepts. For example, the nodes "baby," "dog," and "puppy" would probably establish a variety of relationships: both hierarchical and nonhierarchical; semantic; or pure statistical relationships. It would be the function of the classification scheme to point to the semantic equivalence between the combination of the first two keys and the third one.

In a classification scheme, a concept may include more than one key, and a key may belong to more than one concept. Other relationships, which are neither hierarchical nor combinatorial but are relevant for retrieval, would already be expressed in the hedges.

The classification scheme would support browsing in hierarchy and would help users to navigate the particular terminological knowledge structure it represents. In addition to these overall functions, a classification scheme could help disambiguate search keys by pointing to their position in the classification structure or to resolve context-dependence issues.

5. DISCUSSION

The development of search interfaces for online bibliographic databases that are geared to end-user searching, and the growing body of research about expert systems, has made

the user the focus of research in information storage and retrieval. This trend motivated the proposal that terminological knowledge acquired from expert searchers could guide the construction of a terminological knowledge structure for IESs. Future investigations are necessary for the development of terminological knowledge structures for IESs. Three research areas that would contribute to this development are terminological research, automated text processing, and thesauri and indexing practice.

5.1 *Terminological research*

The terminological attributes discussed in this article are those that were defined by searchers as problematic for textword searching. Clearly, there are additional attributes of terms that are possibly important for retrieval. For instance, even well-defined terms vary in their level of specificity, in their stability (whether they are well established or just a fad), or in how likely they are to appear in a text about the concept they represent. Terminological research should identify such attributes and examine their effect on retrieval. As part of this task, methods for recognizing attributes should be developed. Such methods may make use of statistical or other automated text analyses.

5.2 *Automated text processing*

Research in automated text processing aims at developing methods to represent the content of documents for a number of purposes, including information retrieval. The purpose in information retrieval is to develop methods that would be effective in representing any text for all users and requests. Such a global approach has been assumed to be necessary.

After over four decades of research in automated text processing, there is a growing notion that such an all-encompassing goal is unattainable, and a more promising approach is to develop automated text analysis methods for limited subject domains and groups of users (e.g., Sparck Jones, 1991). Indeed, research in this direction has already begun. For example, Damerau (1993) developed computer-generated domain-oriented vocabularies with the aid of subject headings that were assigned intellectually to a text, and Ingwersen and Wormell (1988) suggested that different types of information needs may require different retrieval techniques (e.g., a search for a known item might be best searched with Boolean logic which is based on exact match rather than a partial match technique).

The proposed terminological knowledge structure illustrates that methods developed for automated text processing are relevant to the creation of a knowledge base that supports searching. It reinforces the somewhat neglected connection between indexing and searching. The idea that searching methods be considered when procedures for automated text processing are developed is not new. AID, CITE (Doszkocs, 1978, 1983), ZOOM on ESA/IRS (Martin, 1982), and OKAPI (Walker & de Vere, 1990) are among the attempts which use statistical text analysis techniques to suggest additional terms for query expansion. While some of these methods are limited to a subject domain, or even to a user group, they are all global on the terminological level; they do not consider terminological attributes. The importance of such attributes in determining search keys has been demonstrated empirically (Fidel, 1991a). Moreover, the description of the terminological knowledge structure shows that data generated from statistical or other automated analyses of text can increase retrieval effectiveness. An example of data about term co-occurrence can illustrate this point.

Co-occurrence data are used in automated text processing to indicate some similarity among terms. Thus, terms that co-occur with a search key can be used to expand a query. Recent research has demonstrated that the use of co-occurrence data is not effective in improving retrieval performance when used in searching to expand a query automatically, because terms that highly co-occur tend to also occur frequently in the database (Peat & Willett, 1991). However, research in interactive, semiautomatic query expansion demonstrated that the use of co-occurrence data to rank terms for query expansion does in fact result in improved retrieval performance (Efthimiadis, 1992).

It is possible that co-occurrence data could be used more effectively when terminological attributes of search keys are taken into consideration. If a search key has many synonyms, it is best to expand the query with terms that co-occur least frequently, but if

a search key itself occurs too frequently, it is best to substitute it with terms that co-occur most frequently but do not occur frequently in the database. Research in automated text processing that focuses on disambiguation, identification of semantic relatedness, and context dependency could provide a significant contribution to the creation of IESs. A close collaboration with terminological research would encourage automated text processing to go beyond its attempts to represent the text stored in a database and to concentrate on extracting terminological attributes from that text to help users' searching.

5.3 *Thesauri and indexes*

With IESs in place, the role of thesauri and indexing would change substantially. Both the process of indexing and the construction of controlled vocabularies would be limited to intellectual processes. The specific functions of thesauri and indexing would depend on the capabilities of the IES. For example, if an IES refers users to descriptors when a request includes terms that occur too frequently, indexers could be required to consult a list of terms that occur too frequently whenever they index a document to guarantee that they do not forget to assign the descriptors when relevant. Similarly, if the IES includes a semantic scale for each hedge term, a thesaurus may be developed that does not include the associative relationships (RT).

Indexing would not be eliminated because some of its current functions cannot be automated. It is useful to examine these functions more closely. First, indexing enhances retrieval effectiveness when it resolves terminological difficulties caused by vague, ambiguous, common, or synonymous terms and nonlexical expressions. The proposed IES would also resolve such difficulties. However, one of the options it would suggest in such situations would be the use of descriptors. In fact, using descriptors has proven to be an effective way to deal with terminological difficulties. Descriptors, of course, are assigned in indexing. But with IESs in place, indexing could be limited to assigning descriptors only for concepts whose terms represent terminological difficulties.

Second, human indexers assign weights to concepts, albeit in a subjective manner. When a term is assigned in indexing, whether it is a descriptor or a natural language term, it reflects the indexer's perception that the text is about the subject represented by that index term or that a user who is interested in this subject might want to retrieve the text. In addition, intellectual indexing uses some form of weighting when terms are assigned as major or minor terms. In automatic indexing a weight is assigned to each term in the text based on the statistical properties of the term. Term weighting is an active area of research in information retrieval.

Third, indexing provides explicit representation of information that is implicitly embedded in a text. That is, human indexers can infer perceived "aboutness" or potential relevance without textual clues. Given the present state of research in artificial intelligence (Sparck Jones, 1991), it seems that this important function would have to be performed intellectually. Thus, indexing would be limited to making explicit concepts that are implicit and to assigning descriptors for problematic concepts.

Vocabulary control would be exercised only for problematic search keys, and database thesauri would be limited to those keys which require the use of descriptors; all other search keys would be indexed and searched with natural language. In fact, with highly developed IESs in place, a database thesaurus would be derived from the terminological knowledge structure, tailored to the specific requirements of the database and its users.

In summary, a well-developed IES that advises users on the selection of search keys would not only help searchers, it would change the nature of indexing. It would transfer indexing, and possibly vocabulary control, from a primarily a priori process to a process that is determined by specific information needs and other situational factors. While intellectual indexing would be performed to resolve textual and terminological difficulties, indexing of other concepts and terms would be situation dependent and would be performed according to the requirements of each information request.

Acknowledgements—Elaine Svenonius played an active role in the creation of this article; we greatly appreciate her help and insights. We also thank Susanne Humphrey, Philip Smith, and Dagobert Soergel for their useful comments on an earlier version.

REFERENCES

Ahlswede, T., Anderson, J., Evens, M., Li, S.M., Neises, J., Pin-Ngern, S., & Markowitz, J. (1988). Automatic construction of a phrasal thesaurus for an information retrieval system from a machine readable dictionary. RIAO 88 Program. *Conference with presentation of prototypes and operational demonstrations: user-oriented content-based text and image handling* (vol. 1, pp. 597–608). March 21–24, 1988, Cambridge, MA. Paris: C.I.D.

Anderson, J.D., & Rowley, F.A. (1992). Building end-user thesauri from full text. In B.H. Kwasnik & R. Fidel (Eds.), *Advances in classification research: proceedings of the 2nd ASIS SIGCR Classification Research Workshop* (pp. 1–10). Medford, NJ: Learned Information.

Brajnik, G., Guida, G., & Tasso, C. (1986). An expert interface for effective man-machine interaction. In L. Bolc & M. Jarke (Eds.), *Cooperative interfaces to information systems* (pp. 259–308). Berlin: Springer-Verlag.

Brajnik, G., Guida, G., & Tasso, C. (1988). IR-NLI II: Applying man-machine interaction and Artificial Intelligence concepts to Information Retrieval. In Yves Chiaramella (Ed.), *ACM-SIGIR, 11th International Conference on Research and Development in Information Retrieval* (pp. 387–399). Grenoble, France: ACM Press.

Brenner, E.H., Lucey, J.H., Martinez, C.L., & Melekaa, A. (1984). American Petroleum Institute's machine-aided indexing and searching project. *Science and Technology Libraries, 5*(1), 49–62.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management, 29*(4), 433–447.

Doszkocs, T.E. (1978). AID—an associative interactive dictionary for on-line searching. *Online Review, 2*, 163–173.

Doszkocs, T.E. (1983). CITE NLM: natural-language searching in an online catalog. *Information Technology and Libraries, 2*(4), 364–380.

Efthimiadis, E.N. (1990). Online searching aids: a review of front-ends, gateways and other interfaces. *Journal of Documentation, 46*(3), 218–262.

Efthimiadis, E.N. (1991). *Approaches to search formulation and query expansion in information systems: DRS, DBMS, ES.* (Report No. RDD/G/102). Boston Spa: British Library.

Efthimiadis, E.N. (1992). *Interactive query expansion and relevance feedback for document retrieval systems.* Doctoral dissertation, City University, London.

Efthimiadis, E.N., & Robertson, S.E. (1989). Feedback and interaction in information retrieval. In C. Oppenheim (Ed.), *Perspectives in information management* (pp. 257–272). London: Butterworths.

Fidel, R. (1991a). Searcher's selection of search keys: I. The selection routine. *Journal of the American Society for Information Science, 42*(7), 490–500.

Fidel, R. (1991b). Searcher's selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science, 42*(7), 501–514.

Frei, H.P., & Jauslin, J.F. (1983). Graphical presentation of information and services: a user-oriented interface. *Information Technology: Research and Development, 2*, 23–42.

Fugmann, R. (1982a). The complexity of natural and indexing languages. *International Classification, 9*(3), 140–144.

Fugmann, R. (1982b). Natural versus indexing languages in chemical documentation. *Angewandte Chemie International Edition, 21*(8), 608–616.

Guentzer, U., Juettner, G., Seegmuueller, G., & Sarre, F. (1989). Automatic thesaurus construction by machine learning from retrieval sessions. *Information Processing & Management, 25*(3), 265–273.

Humphrey, S.M. (1989). MedIndEx system: medical indexing expert system. *Information Processing and Management, 25*(1), 73–88.

Humphrey, S.M., & Miller, N.E. (1987). Knowledge-based indexing of the medical literature: the Indexing Aid Project. *Journal of the American Society for Information Science, 38*(3), 184–196.

Humphreys, B.L., & Lindberg, D.A.B. (1989). Building the Unified Medical Language System. In *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care* (pp. 475–480). Washington, DC: IEEE Computing Society Press.

Ingwersen, P., & Wormell, I. (1988). Modern indexing and retrieval techniques matching different types of information needs. In S. Koskiala & R. Launo (Eds.), *Information, knowledge, evaluation: proceedings of the forty-fourth FID Congress* (pp. 79–90). Amsterdam: North-Holland.

Knapp, S.D. (1988). Free-text searching of online databases. *The Reference Librarian, 5*(6), 143–153.

Knapp, S.D. (1992). *The contemporary thesaurus of social science terms and synonyms: a guide for natural language computer searching.* Phoenix, AZ: Oryx Press.

Krovetz, R. (1991). Viewing the dictionary as a classification system. In S.M. Humphrey & B.H. Kwasnik (Eds.), *Advances in classification research: Proceedings of the 1st ASIS SIGCR Classification Research Workshop* (pp. 87–93). Medford, NJ: Learned Information.

Krovetz, R., & Croft, W.B. (1989). Word sensing disambiguation using machine-readable dictionaries. In N.J. Belkin & C.J. van Rijsbergen (Eds.), *SIGIR '89: Proceedings of the Twelfth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 127–136). New York: ACM Press.

Liddy, E.D., Hert, C.A., & Doty, P. (1991). Roget's International Thesaurus: conceptual issues and potential applications. In S.M. Humphrey & B.H. Kwasnik (Eds.), *Advances in classification research: Proceedings of the 1st ASIS SIGCR Classification Research Workshop* (pp. 95-100). Medford, NJ: Learned Information.

Makey, K., Atherton, P., & Newton, C. (1980). An analysis of controlled vocabulary and free text search statements in online searches. *Online Review, 4*(3), 225–236.

Martin, W.A. (1982). Helping the less experienced user. In *6th International Online Meeting* (pp. 67–76). Oxford: Learned Information (Europe) Ltd.

Martinez, C., Lucey, J., & Linder, E. (1987). An expert system for machine-aided indexing. *Journal of Chemical Information and Computer Sciences, 27*(4), 158–162.

Monarch, I., & Carbonell, J. (1987). CoalSORT: a knowledge-based interface. *IEEE Expert, 2*, 39–53.

Nutter, J.T., Fox, E.A., & Evens, M.W. (1990). Building a lexicon from machine-readable dictionaries for improved information retrieval. *Literary and Linguistic Computing, 5*(2), 129–137.

Peat, H.J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*, 378–383.

Piternick, A.B. (1984). Searching vocabularies: a developing category of online search tools. *Online Review, 8*(5), 441–449.

Pollitt, A.S. (1987). CANSEARCH: an expert systems approach to document retrieval. *Information Processing and Management, 23*(2), 119–138.

Pollitt, A.S. (1988). A common query interface using MenUSE—A menu-based user interface search engine. In *Proceedings of the 12th International Online Meeting* (vol. 2, pp. 445–457). Oxford: Learned Information.

Shoval, P. (1981). Expert/consultation system for a retrieval data-base with semantic network of concepts. *SIGIR Forum*, 16, 145–149.

Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing and Management, 21*(6), 475–487.

Shoval, P. (1986). Comparison of decision support strategies in Expert Consultation Systems. *International Journal of Man-Machine Studies, 24*, 125–139.

Shute, S.J., & Smith, P.J. (1992). Knowledge-based search tactics. *Information Processing & Management, 29*(1), 29–45.

Sievert, M., & Boyce, B.R. (1983). Hedge trimming and the resurrection of the controlled vocabulary in online searching. *Online Review, 7*(6), 489–494.

Smith, P.J., Shute, S.J., & Galdes, D. (1989a). In search of knowledge based search tactics. In N.J. Belkin & C.J. van Rijsbergen (Eds.), *SIGIR '89: Proceedings of the Twelfth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 3–10). New York: ACM Press.

Smith, P.J., Shute, S.J., Galdes, D., & Chignel, M.H. (1989b). Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems, 17*(3), 246–270.

Sparck Jones, K. (1991). Notes and references on early automatic classification work. *SIGIR Forum, 25*(1), 10–17.

Svenonius, E. (1983). Use of classification in online retrieval. *Library Resources and Technical Services, 27*(1), 76–80.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*(5), 331–340.

Thompson, R.H., & Croft, W.B. (1989). Support for browsing in an intelligent text retrieval system. *International Journal of Man-Machine Studies, 30*, 639–668.

Veronis, J., Ide, N.M., & Harie, S. (1990). Automatic building of large neural networks for natural language disambiguation. *Tenth International Workshop. Expert Systems and Their Applications. Specialized Conference: Natural Language Processing and Its Applications* (pp. 105–117). May 28–June 1, 1990, Avignon, France. Nanterre, France: EC2.

Vickery, A. (1988). The experience of building expert search systems. In David Raitt (Ed.), *Proceedings of the 12th International Online Meeting* (vol. 1, pp. 301–313). Oxford: Learned Information.

Vickery, A., Brooks, H.M., Robinson, B., & Vickery, B.C. (1986). *Expert system for referral (final report).* London: University of London, Central Information Service.

Vickery, A., Brooks, H.M., Robinson, B., & Vickery, B.C. (1987). A reference and referral system using expert system techniques. *Journal of Documentation, 4*, 198–203.

Vleduts-Stokolov, N. (1987). Concept recognition in an automatic text-processing system for life sciences. *Journal of the American Society for Information Science, 38*(4), 269–287.

Walker, S., & de Vere, R. (1990). *Improving subject retrieval in online catalogues: 2. Relevance feedback and query expansion.* British Library Research Paper 72. London: British Library.