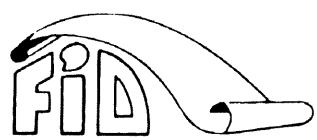
CLASSIFICATION RESEARCH FOR KNOWLEDGE REPRESENTATION AND ORGANIZATION

Proceedings of the 5th International Study Conference on Classification Research, Toronto, Canada, June 24-28, 1991



edited by
NANCY J. WILLIAMSON
FID/CR Secretariat
Faculty of Library and Information Science
University of Toronto
Toronto, Canada

MICHÈLE HUDON
Faculty of Library and Information Science
University of Toronto
Toronto, Canada

FID 698

1992
ELSEVIER
Amsterdam • London • New York • Tokyo

Classification Research for Knowledge Representation and Organization
N.J. Williamson and M. Hudon (Editors)
Elsevier Science Publishers B.V.

THESAURUS REQUIREMENTS FOR AN INTERMEDIARY EXPERT SYSTEM

Raya Fidel

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

Abstract

Direct observations and analysis of searching behavior of professional online searchers shed light on thesaurus requirements for an intermediary expert system--a system that mediates between online databases and end users. Examination of searchers' decisions about the selection of search keys, and of the knowledge about terminological and subject properties that they employed, illuminated the requirements for a thesaurus that will facilitate the selection of search keys. Expert knowledge is needed when: a term occurs very frequently in the database; it has many synonyms; it is ambiguous; it is vague; or its meaning is context dependent. To diagnose such terms and to give advice, a thesaurus would be used together with a variety of text sources such as databases' thesauri, machine-readable dictionaries and glossaries, and the databases' text. The thesaurus would be a knowledge structure that includes frequency data, hedges, and a classificatory structure; both intellectual and automated procedures would be used to create it. Such a knowledge structure in place would require a new approach to text analysis and to the construction of controlled vocabularies.

1. INTRODUCTION

An intermediary expert system (IES) helps users, both professional searchers and end-users, to conduct their searches of online bibliographic databases. Research about the requirements and design of IESs has just begun, and most experience has been gathered through experimental or prototype systems. This paper is an embryonic attempt to investigate the requirements on the knowledge base of an IES and it is limited to the selection of search terms, or search keys. It is brought before the participants of this conference to stimulate discussion, and not to present final and conclusive results.

Consider, for example, a user with a request: "The attitude of students towards themselves during examinations period." When searched with the help of an IES, the system is supposed to guide the user in all the stages of the search. To confine the focus of the paper, I assume that: the IES covers a given set of databases which are limited to a particular subject area, such as education; the databases to be searched for the request have been already selected; and the search system employs Boolean operators. The latter requires that the user had already defined the components of the request: "attitudes towards themselves," "student," and "examination."

Now, the IES is expected to help the user decide what search keys to enter to represent each concept. While a novice end-user may decide to enter the terms as they

appear in the request, an experienced searchers would probably know that these terms are not very useful. For example, the term "student" occurs too frequently in educational databases, and the term "examination" may not always mean "educational test," depending on the context in which it appears. Both search keys, therefore, should be entered as descriptors.

How can the IES "know" that problems exist? Presumably, it can consult the databases' thesauri. Unfortunately, common thesauri do not provide this information. Moreover, some databases have no indexing or controlled vocabulary and thus no sources for the IES to use. Clearly, to give advice about the selection of search keys, the IES requires more knowledge. It also requires a set of rules, such as: If a search key occurs too frequently, enter a descriptor. My research in the last years aimed at defining these knowledge and rules.

2. THE METHOD

My analysis is based on knowledge acquired from experienced online searchers. The data were 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.

This analysis uncovered the intuitive rules that searchers used and resulted in a decision tree for the selection of search keys, called the "selection routine." The routine 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 options: enter the descriptor, but if recall needs to be improved: add textword synonyms to descriptors, or use generic descriptor in an inclusive mode ("explode" or "cascade"), or add the next broader descriptor in the hierarchy. The data also provide the frequency in which each option was selected. Thus, of the 228 cases where 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.

These data are available and could be incorporated into an IES. They are not sufficient, however, to provide sound advice. Consider the request component "student." Suppose a searcher followed the recommendation of the IES and entered it as a descriptor, but after reviewing the results the searcher wants to improve recall. Following the frequencies in which options were used, the IES would advise the user to enter the textword "student." This might not be a useful advice because of the high frequency in which the key occurs in the database's text. A better approach might be the use of a generic search with descriptors, e.g., entering descriptors such as "college students," "commuting students," or "nonmajors."

This paper examines what knowledge is required for an IES to be able to give advice in such situations. While no clear-cut answer exists as yet, it is important to explore this issue to find out what is available now that could be used, and what issues should be investigated in future research.

3. TERMINOLOGICAL ATTRIBUTES

My study of online searchers showed that situations that require special consideration are terminological in nature. It revealed that professional searchers took into account the terminological attributes of a search key most often when they considered whether or not a term was suitable for free-text searching. According to the study's searchers, search keys with the following attributes are not suitable for free-text searching:

- o the key has many synonyms
- o the key is ambiguous; for example, archival *record* is different in meaning than database *record*
- o the key is vague; for example, *health promotion* has no socially-acceptable boundaries to its definition
- o the meaning of the key depends on the context in which it appears
- o the key occurs too frequently in the database's text.

This finding is no revelation. 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, vagueness and ambiguity, and it is also useful in balancing for highly frequent keys. It would seem, therefore, that an IES would need only the thesaurus of a database to resolve these issues, and would not require additional knowledge. This assumption is incorrect, however, because databases' thesauri cannot support two functions of an IES: diagnosing which keys are suitable for free-text searching; and providing advice on searching databases that do not have a thesaurus or indexing.

What then should be included in its knowledge base for an IES to diagnose and provide advice about the selection of search keys in *all* databases?

4. TERMINOLOGICAL KNOWLEDGE STRUCTURE

The knowledge base would include a variety of sources: databases' thesauri, existing terminological databanks, machine-readable dictionaries and glossaries, and a Roget-like thesaurus. Further, it would acquire knowledge from the text stored in the database. It would be based on these sources in addition to external knowledge which would be acquired from both system designers and users. The knowledge would be organized in a knowledge structure, or a semantic network, that is called here a terminological knowledge structure. This structure would consist of three constructs: frequency data, hedges, and a classification scheme.

4.1 Frequency Data

The frequency in which each term occurs would be given and compared with that of other keys. For example, the frequency in which a key occurs as a textword would be compared with its frequency as a descriptor to identify keys that occur too frequently. While such data about relative frequency support the diagnostic functions, other frequency data can be employed in the creation of hedges.

4.2 Hedges

A hedge is commonly a list of synonyms created by professional searchers, database producers or even search-system vendors, to expand a concept of a request. In the terminological structure, each node, or search key, would have a hedge that would include other keys, or hedge terms, which are associated with the hedge key.

A variety of methods could be used to create hedges. To start, many hedges already exist. Hedges developed by searchers, lead-in vocabulary in relevant thesauri, and a few switching languages that already exist (e.g., Unified Medical Language System) could be incorporated into the knowledge structure. Hedges, however, would include information about the relationships between a key and each of the hedge terms, and they would probably include terms that are not synonyms. Co-occurrence and relative semantic relatedness are examples of such relationships.

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 database. A designation of the relative level of frequency in which it co-occur would be given for each term in the hedge. The relative semantic relatedness between each term and the hedge key would be determined with the help of machine readable dictionaries, glossaries and Roget-like thesauri. A few example may show how this knowledge could be used.

(a) A search key has many synonyms. A diagnosis would be facilitated by checking the number of hedge terms that are highly related 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 any), followed by highly similar hedge terms arranged in ascending order of the frequency in which they co-occur with the key. This will encourage the user to enter descriptors, which is the preferred action according to the selection routine. If the key has been already 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 probably be the most fruitful.

(b) The search key occurs too frequently. Such a search key is not suitable for textword searching and an IES would suggest alternative terms. It would first retrieve from the hedge descriptors that are closely related to the key on the semantic scale. Next, other 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.

(c) The search key is ambiguous. Various mechanisms, which were tested primarily for automated indexing, have been suggested to disambiguate terms with the help of machine-readable dictionaries, glossaries and thesauri. The knowledge structure proposed here facilitates a further avenue. 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 key. The ORed cluster that represents a desired point of view can then be ANDed with the ambiguous key to improve precision. The same procedure could be used for search keys whose meaning depends on the context in which they appear.

Hedges would be dynamic in nature. 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 refine of hedges. Through machine-learning procedures, terms would be added to hedges or deleted from them.

4.3 Classification Scheme

A classification scheme would provide the overall structure. It would establish the links between keys, or nodes, and cluster them into concepts. In this scheme, a concept may include more than one key, and a key may belong to more than one concept. The classification scheme would support browsing in the hierarchy, and would help users to navigate in the terminological knowledge structure.

In addition to these overall functions, a classification scheme would help disambiguate search keys by pointing to their position in the classificatory structure, or to resolve context-dependence issues.

5. DISCUSSION

This view of the terminological knowledge structure has not been realized yet. Much research and development is needed before even an experimental prototype could be constructed. Exploring this idea, however, points to future trends that are necessary for the construction of such a knowledge structure.

The description above illustrates that methods developed for automated indexing are important for the creation of a knowledge base that supports searching. It reinforces the somewhat-forgotten connection between indexing and searching, and proposes that searching methods be of prime consideration when procedures for automated indexing are being developed. Moreover, this exploration shows that for research in automated text analysis to be useful for searching it should focus on resolving terminological issues that are important in searching. Thus, research about disambiguation, identification of semantic relatedness, and about context dependency could provide a significant contribution to the creation of IESs.

Finally, with IESs in place, the role of thesauri and indexing would change substantially. Indexing would be limited to the intellectual procedure that would provide explicit representation to information that is implicitly embedded in a text, and to concepts whose terms present terminological difficulties. 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 addition, automated indexing would abandon attempts to represent the text stored in a database, and would concentrate on extracting terminological attributes from that text to help users select search keys.