11-1-1999

Human or Automated, Indexing is Important

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Human or Automated, Indexing Is Important

CONTROLLED VOCABULARY indexing was developed well over a century ago to provide a consistent way for users to search for information. But human-assigned indexing has always been labor-intensive and thus costly.

One alternative devised by print index producers involves extracting words, known as keyword in context (KWIC) indexes. KWIC indexes were alphabetically arranged, with each major word in a document title extracted as an index term, e.g., "Indexing Is Still Important" could be found in a KWIC index under *indexing*, *still*, and *important*.

In the online environment, free-text searching of computer-produced extraction indexes serves the same function as KWIC indexing in print. A free-text search of the words in a title (and also abstract, full text, etc.) will retrieve all documents that contain those words, and many systems will let you view the words in context through a "KWIC" output option.

But KWIC and other automatic extraction indexes don't provide the intellectual advantages of assigned, controlled vocabulary indexing. Thus, the labor-intensive process of human-assigned indexing has persisted and coexists with free-text searching. In most bibliographic or full-text databases available today—on systems such as Dialog, Ovid, STN, FirstSearch, SilverPlatter, ProQuest Direct, and DataStar—users can choose to search free text or with controlled vocabulary descriptors.

Now two major information companies are trying automated approaches. The Institute for Scientific Information (ISI) adds computer-generated terms to its databases, and Lexis-Nexis has launched an automated indexing project.

Carol Tenopir (tenopir@utkux.utk.ed) is Professor at the School of Library and Information Science, University of Tennessee at Knoxville

Why add indexing?

Why do secondary publishers and online systems companies still spend the money and time to produce controlled vocabulary indexes when most people access indexes and abstracts online? Well, indexing improves search results. When authors express their ideas, they rarely consider the issue of consistent retrieval. Indexing imposes some linguistic consistency.

Because of the limitations of most search engines, controlled vocabulary indexing still must impose this consistency upfront. Search engines that rely on either Boolean logic or statistics (relevance ranking) are literal: they search character strings, not meanings. So, if a searcher enters the term "controlled vocabulary," knowing that it implies "assignment indexing" or "thesaurus," the major search engines will not turn up its broader meaning. Online searching instructors must teach students to think of all possible ways to express a concept, then link those terms with the Boolean OR operator or list them in a relevance ranking search engine.

Some better systems automatically search variant word form endings (such as plurals or -ing) or a limited number of automatic equivalents (such as British/American spelling variations or common abbreviations). A notable exception to this limitation is offered by linguistic search engines, which recognize parts of speech and discerns multiple meanings of words using natural language processing. So far, only one commercial online service offers this: DR-LINK (document retrieval using linguistic knowledge), developed by Elizabeth Liddy, professor at Syracuse University School of Information Management, NY, the search engine behind the Manning-Napier business online service (www.mnis.net).

Since most searchers, database producers, and online systems live with the limitations of the standard search engines, controlled vocabulary descriptors offer an important alternative to free-text searching. Descriptors allow a searcher to retrieve all documents on a topic, regardless of the words used by authors to describe that topic (e.g., motion pictures vs. cinema); clarify ambiguous terms (e.g., motion pictures vs. chemical films); and standardize spelling or word form variations (e.g., African Americans vs. Afro-Americans.) Cross references or narrower, broader, and related terms help further.

Traditional indexing

For years, indexes have relied heavily on human analysis. A database producer hires indexers who read the material, then locate the best terms in the database vocabulary list or thesaurus, and assign the terms to each document. Indexers go beyond surface meaning to assign terms that describe a document's real meaning. A complex system of checks, editing, and vocabulary updating keeps the vocabulary lists current and the indexing relatively consistent and error-free.

Some level of computer-assisted indexing has been commonplace for years. Online thesauri allow indexers to check for the correct term. Also, input errors are caught by spell-checkers, and seldom-used or overused terms are flagged by automatic matches against the existing database.

The H.W. Wilson Company has long been synonymous with the traditional form of controlled vocabulary indexing. Harold Reagan, president and CEO, said, "Quality indexing, with human intervention, has been and always will be important to librarians and other end users we serve."

Wilson approaches even computer-assisted indexing cautiously. VP Deborah Loeding noted, "Wilson has explored and will continue to aggressively investigate systems that provide machine-*assisted* indexing. However, until these systems are refined, we will not compromise the integrity of our content by integrating such services."

Wilson has been affected by electronic searching of indexes, however. Now that users can perform multifile searches on WilsonWeb, Wilson has

(Continued on p. 38)

launched a major editorial effort to standardize the vocabularies across its different indexes. The Wilson OmniFile Full Text project reconciles "selected subject headings" from six Wilson indexes. Internally, Wilson has increased automated processing with a new editorial system that combines terms across indexes. But staffers still analyze documents and assign terms.

ISI's project

While Wilson products have always relied on assigned indexing, the indexes created by ISI included no subject descriptors until 1995. Then ISI introduced KeyWords Plus, a form of fully automated indexing in which subject terms—augmenting or replacing author-designated keywords—are assigned automatically to records in ISI's citation and Current Contents databases.

ISI records do not have full texts, but each record provides a "cited reference" field, which includes the titles and source information for all items listed in the bibliography of the article described. KeyWords Plus uses the titles in this cited reference field.

The computer algorithm designed to assign the terms analyzes the titles of all major items included in each article's cited reference field. Candidate terms (phrases or single words) are selected by a process that takes into account frequency of occurrence, number of meaningful words, and lack of duplication with the original article's title. The terms are then added to the database record of the source article, providing additional subject-related terms to increase recall. Approximately two-thirds of ISI records include KeyWords Plus.

Nexis indexing initiative

Lexis-Nexis, long a bastion of free-text searching of full texts without subject-controlled vocabulary, has recently recognized the value of controlled indexing. Lexis-Nexis's approach differs greatly from the traditional indexing process typified by Wilson or the citation title analysis algorithm used by ISI. Its news indexing initiative provides computer-generated controlled vocabulary terms for topics to complement the full texts and the controlled fields of company, organization, people names, and geographic locations that have been available since 1992 in selected Nexis files.

Subject experts, computational linguists, engineers, and information professionals at Lexis-Nexis have been working since 1997 to develop the software, terms, and procedures for "SmartIndexing" of news documents. Lexis-Nexis put the first 400 subject terms (mainly for banking and pharmaceuticals topics) online in January 1998. Now there are over 800 subject terms; Lexis-Nexis expects that total to grow next year to 1000 in 30 general business-related subject areas. By the end of 1999, all "news" (newspapers, magazines, etc.) files will be reloaded so documents back to 1990 can be indexed.

> Some level of human analysis remains important for the best search results

Subject terms are selected controlled terms that are computer-assigned to documents after a term profile has been developed by indexers. The project team identifies words that collectively describe the controlled subject and assign weights to each term in the profile, depending on how precisely it defines the major subject. For example, Lexis-Nexis developers have identified synonyms for wineries such as "viticulture business" and "wine growing industry," strongly related terms such as "commercial winery," and weakly related terms such as "bulk wine prices," "barrel," and "cases per year." Some terms are assigned negative weights, such as "home wine-making," and some false matches are flagged as terms for the indexing algorithm to ignore (such as "Martha's Vineyard").

When any of these terms recur over a certain threshold in a Nexis document, the controlled term "wineries" will be added to that document, along with a numeric score that reflects how much the topic is discussed in the document. This numeric score may be used by a searcher to restrict a search to documents in which the term appears as a major, strong, or weak topic. Searchers may also restrict a search to the index term field to achieve higher precision, or, using the "thesaurus option," view words that are part of a controlled term's profile.

Contrasted with traditional indexing, Lexis-Nexis uses human intellectual intervention only as term profiles are developed. Also, it assigns broad terms, rather than the traditional practice of indexing documents at the level of specificity of the document. According to Lexis-Nexis indexing developers, this works for Nexis, because "about 1000 discrete, stand-alone subject terms filter general news very effectively. Some of these terms are quite broad such as Telecommunications, but others are very specific, such as Personal Communications Service or Paging. But many of these can be combined and postcoordinated, such as Telecommunications with Litigation (or, more narrowly, Class Actions) or Mergers & Acquisitions. When you start combining topics, you actually have tens of thousands of individual subjects."

The alphabetic term list is displayable and searchable on the web version of Lexis-Nexis, but staff members now aim to create a searchable thesaurus that will include broader and narrower terms.

Improving the search process

The vastly different approaches to assignment indexing followed by Wilson, ISI, and Lexis-Nexis have one thing in common: each aims to provide searchers with fields that go beyond mere word searching. The basis of each approach is human intelligence—throughout the process in Wilson's case; at the beginning in the term profile development for Lexis-Nexis; and by using the author's cited titles at ISI.

Traditional specific indexing can be used both to increase recall (bringing together like documents with disparate descriptions) and to increase precision (descriptors are only assigned to topics important to a document). The Lexis-Nexis approach, especially when a searcher designates a term as major, can increase precision in news full-text databases by limiting retrieval to documents covering a topic in depth. ISI's KeyWords Plus increases recall in the ISI databases (which include only bibliographic information, abstracts for some records, and citation titles). Although software aids the assignment of terms just as it aids searching, some level of human analysis remains important for the best search results.