

Automatic Indexing

FILE TRANSFER DOCUMENTS
ONLINE HIGH SPEED DSL MEGABIT
FTP ECOMERCE TELECOMMUNICATIO
TECHNOLOGY FUTURE MEMORY SPEE
APPLICATION

Comparing Rule-based and Statistics-Based Indexing Systems

By Marjorie M. K. Hlava

It's never been more fascinating—or more challenging—to be an information professional.

Surveys show that corporate librarians have more information to manage, and less staff to help manage it than ever before. Not only is the universe of data expanding by the minute, but also we are beset with confusion over the variety of tools and methods for managing and retrieving electronic information.

A striking example of this confusion is automatic indexing—the process of mechanically analyzing concepts and themes in a database's stored and newly added content to create links between keywords and phrases.

This is a critical area. The costs and benefits of how you categorize your information collection will cascade through your organization for years to come. In our increasingly knowledge-based economy, indexing amounts to basic infrastructure.

Why is automatic indexing important? It's the glue that holds your content together. It's the underlying layer of order that makes your database productive, robust, and responsive—and thus best able to serve the needs of your organization.

Without automatic indexing, you may find the precise bit of data that will ignite a new market, but at what cost, if you and your staff have spent hours wading through a river of irrelevant documents called up from an online search? Even more likely are instances of lost opportunities to deliver on requests for research, competitive intelligence, or industry awareness because you don't have the means to put missing or disparate pieces of information together.

So which system for automatic indexing of data is fastest? Easiest to implement

and update? Which provides the best return on investment? Which system will “understand” that when you're interested in the REM sleep phase you're not interested in the brainy rock band R.E.M.?

To better grasp the options for categorizing data, we'll be smart librarians and impose some categories here ourselves. We will divide the major systems for automatic indexing into two groups: rule-based and statistics-based.

We'll look at the return on investment, long and short term, for each, as well as how they compare for ease of implementation, user access, and accuracy. Most systems require a thesaurus to start, and we'll assume one here for each system. (The thesaurus is a controlled vocabu-

Marjorie M. K. Hlava is chairperson of Access Innovations in Albuquerque, New Mexico, provider of the Data Harmony line of software used for indexing and data structuring. You can reach her at mhlava@accessinn.com.



lary that lists the main components of the data collection, along with appropriate synonyms and antonyms. It helps the indexer and searcher to select the same terms to describe a particular subject.)

The critical notion for us is that one has to “teach” an automatic indexing system to identify relevant data and how to categorize it. How is this accomplished in each of the two automatic indexing systems?

Rule-Based Indexing

The newest type of automatic indexing system—rule-based—is a leap forward in the science of indexing. It offers greater precision while burning up fewer dollars and hours than previous systems. In a rule-based system, simple categorization rules are automatically generated, matching bits of text (“prompt words”) to the thesaurus or taxonomy terms, which tell the software how to categorize the document. Editors may further define the rules by telling the system what words must be present or absent in the text, or giving some other specific instructions that point the document to a particular category. For an organization conducting sleep research, you tell the software that a document with REM and “music” is a “miss,” and it doesn’t bring up any documents about the band. Think of a rule-based approach as driving a car using specific directions to get to a destination.

Statistics-Based Indexing

The second system—statistics-based—is “trained” by examining a set of 50 or so documents associated with each keyword in the thesaurus. This process creates scenarios from word occurrence and location in the training documents. In this system, the software would deduce that REM is part of the science of sleep and not a rock band, since none of the training documents mention music. Think of this as giving a driver a pile of maps and telling him to figure out the best way to get to the destination.

These are sophisticated systems, and the upfront investment for any automatic indexing system is substantial. For our comparison, let’s assume an existing thesaurus (sometimes called the controlled vocabulary) of 6,000 words, which is typical. We’ll assume hourly rates and units per hour using industry standards. And we’ll assume that 85 percent accuracy is the baseline required for implementation to save personnel time.

Now for a test drive—our experience with two different clients, using the two different systems.

The Rule-Based Approach

A simple rule-based system matches terms in the thesaurus to exact terms and synonyms in the documents to determine the appropriate indexing. With an existing thesaurus or authority file, this is a two-hour process. Rules for synonyms and preferred terms are generated automatically. So, for example, if the thesaurus category is “bush,” it might also recognize “shrub” as a synonym. Using the simple rule base alone usually provides 60 percent accuracy.

The editor can add more complex rules. For example, the index, in its search for shrubbery documents, might be trained to ignore a document with the word “bush” used within a few words of “president” or with a capital B. Complex rules such as these typically comprise about 10 percent of the terms in the vocabulary. An index editor can create four to six rules per hour, so it would take 2.5 weeks to create 600 complex rules for a 6,000-term thesaurus. Enhancing the simple rules through editorial rule building offers the potential to achieve 85 percent or higher accuracy.

The rule-based approach places no limit on the number of terms used in the taxonomy or the number of taxonomies held on a server. Our client was up and running with its rule-based index in a month.

So, let’s add it up:

- Software is about \$60,000, including training and support.
- Conversion of the thesaurus, about two hours at \$125 per hour in programming time (\$250).
- Loading the thesaurus and creating the rule base, two hours of editorial time at \$45 per hour (\$90).
- Complex rule building, 100 hours of editorial time at \$45 per hour (\$4,500). (Could be as much as 150 hours.)
- The total, based on those assumptions, would be \$64,840.

The client for whom we prepared the rule-based system reported 92 percent accuracy and a fourfold increase in productivity.

The Statistical Approach

We’ll start with the same preexisting 6,000-word thesaurus. The software for

this system starts at about \$75,000. Usually one week of training is required, at about \$10,000.

Now you must address the documents—news articles, for instance—that train the thesaurus. The documents can be collected using software programs, but document sets for each thesaurus term must be reviewed by editors to remove misleading records. If each thesaurus term requires 15 reviews, that’s a potential \$67,500 for editorial review of a training set.

Next, you run the training documents through the software, with programming time of 40 hours at \$125 per hour, or \$5,000. The index editor reviews the results, and collects new training sets for thesaurus terms that didn’t return good data sets. The second run is reviewed. Editorial time of 40 hours at \$45 per hour costs \$1,800.

The next step is to collect additional training data for bad sets. If 25 percent of the thesaurus term yields 1,500 terms, multiplied by one hour per term of editorial time is \$67,500. The training set is rerun, assuming 20 hours of programming time at \$125 per hour, or \$2,500. Reviewing the results requires 20 hours of editorial time at \$45 per hour, or \$900.

In our case study with a client, the accuracy at this point was 60 percent. To achieve reliable improvement in productivity requires 85 percent accuracy. At this point, an editor can write rules in a program language such as SQL. If you train editors to write these rules, you can avoid the higher programmer rates. Still, to write four SQL rules per hour for 1,500 terms (25 percent of the thesaurus terms), requires 375 more editorial hours at \$45 per hour, or \$16,875.

Again, adding up the costs:

- Total implementation time frame, 33 weeks.
- Total person hours, 6,488, plus 40 hours of editor training.
- Upfront cost, \$449,375.
- Maximum accuracy achieved, 72 percent; productivity doubled.

So, what is the return on investment? Assuming that six editors are involved in the process, a rule-based system recoups its value in one month, compared with almost five years under the statistics-based approach. It’s not hard to see why we prefer the rule-based index. ●