Building a Synonymous Search Index (Thesaurus)

December 10, 2009

Originally published October 30, 1998 in Web Review magazine.

Most of us are familiar with the traditional thesaurus or dictionary of synonyms.***Roget's Thesaurus*** was an invaluable tool in helping us spruce up our high school English papers with impressive, multi-syllabic words like **pulchritudinous**. Simply put, this traditional thesaurus helped us go from one known term to multiple synonymous terms.

In contrast, a thesaurus for your Web site works primarily in the opposite direction, mapping many known terms onto one acceptable term per concept. Its purpose is to help users find the documents they need within a large information system. Until recently, online thesauri were familiar only to librarians, expert searchers, and developers of high-end information systems such as **Dialog** and **MEDLINE**. However, as Web sites and intranets grow into large mission-critical information systems, we're seeing a rising need to employ online thesauri as tools to help users find what they're looking for quickly and effectively.

What is a thesaurus?

A thesaurus can be defined as "a controlled vocabulary that leverages synonymous, hierarchical, and associative relationships among terms to help users find the information they need." It sounds rather complex, but once you understand the challenges that a thesaurus is designed to address, things should become clearer.

The value of a thesaurus stems from the inherent problems of natural language indexing and searching. Different users define the same query using different terms. Document authors, indexers, and information architects describe the same concepts using different terms. Consider the following example:

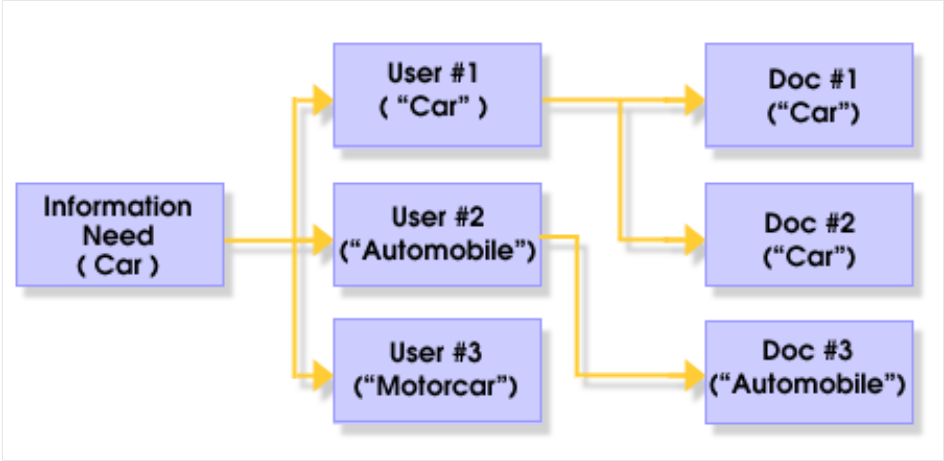

Figure 1: Many information needs go unanswered because a user's search terms don't map to the terms used by document authors and indexers.

Three users are looking for information about a car. However, they each use different terms to describe this same information need. Similarly, the people that indexed the documents selected different terms to describe the same concept. Each user has varying levels of success with no one finding all the relevant documents.

To address this problem, a thesaurus maps variant terms (synonyms, abbreviations, acronyms, and alternate spellings) to a single preferred term for each concept. For document indexers, the thesaurus tells them which index term must be used to describe each concept. This enforces indexing consistency. For users of the Web site, the thesaurus works in the background, mapping their keywords onto the single preferred term, so they find the complete set of relevant documents.

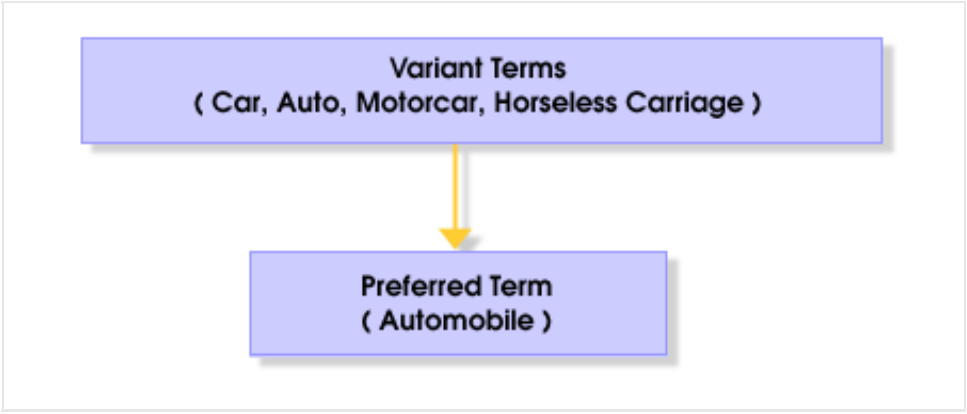

Figure 2: Variant terms serve as entry points into the information system, connecting the words that users have in mind with the preferred terms

applied by document indexers.

A thesaurus can also leverage the richness of hierarchical and associative relationships. Users may express their information need at a broader or narrower level of specificity than that used by the indexer to describe the documents. The mapping of hierarchical relationships addresses this problem.

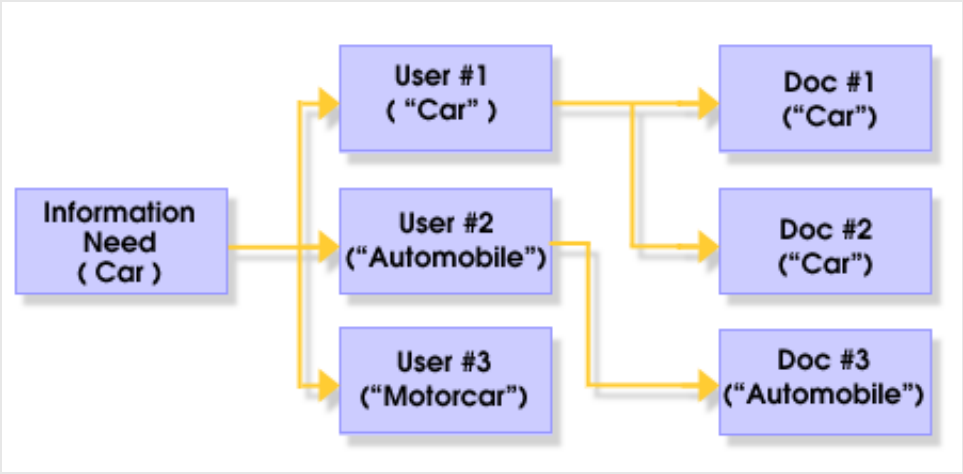

Figure 3: A thesaurus can be more than a dictionary of synonyms. You can also specify and leverage hierarchical and associative relationships.

Additionally, there may be value in mapping associations to related terms. In this example, the decision is made that users interested in automobiles may also be interested in the related terms, such as *mechanic* and *accident*. Identification of these subjective relationships increases the chances of success and promotes associative learning. In a commercial setting, the explicit suggestion that if you're interested in a particular product you may also be interested in other related products can be valuable to both buyers and sellers.

How Do You Build a Thesaurus?

The first step is to familiarize yourself with thesauri and their application in the Web environment. Fortunately, librarians and other information professionals have been building and using thesauri for quite some time, so there are many **examples** to review.

For an example of a very sophisticated implementation, take a look at **OVID's Medline** interface (select *Begin Demo*; you'll need to fill out a brief form to try the demo). This

interface is obviously designed for expert searchers who need power and flexibility to effectively search Medline's enormous database, but it can really get you thinking about the possibilities.

For example, you can perform a normal full-text search of the database, or you can select "Map Term to Subject Heading" and run your search against the thesaurus. You can then browse the hierarchy of subject headings until you find one that matches your topic and level of specificity. In such a large database, the ability to leverage a thesaurus through this integrated searching and browsing capability really helps you to narrow your query, and find what you're looking for.

Once you're done learning from the successes and failures of others, you can begin building a thesaurus through the process of term generation and consolidation. The basic steps include:

1. **Gather terms from as many sources as possible** (e.g., users, subject experts, the content itself, **existing thesauri**). These "entry terms" should include synonyms and abbreviations, acronyms, and alternate spellings for all of the important concepts in your document collection.
2. **Define the preferred terms.** You'll need to create guidelines for selecting preferred terms. For example, in a collection of health-related documents that include terms such as *cancer, oncology, skin,* and *dermatology,* you'll need to decide whether to select **medical terminology** or **regular English** as the preferred terms, based on the type of language most appropriate to your primary audience. For an audience of medical professionals, you would probably select *oncology* and *dermatology* as preferred terms and *cancer* and *skin* as the respective variants. Whichever terminology you choose, it's important to be consistent in your approach to defining the preferred terms.
3. **Link synonyms and near-synonyms.** This is where you map the synonyms, abbreviations, acronyms, and alternate spellings as "variant terms" to the preferred terms. Within reason, the more entry terms you have, the easier it will be for indexers and users to find the preferred terms.
4. **Group preferred terms by subject.** This forms the foundation of your

thesaurus' hierarchy. Definition of the subject hierarchy should be informed by a balance of top-down considerations (e.g., mission, vision, intended audiences) and **bottom-up** content analysis.

5. **Identify broader and narrower terms.** You're defining where each term fits within the hierarchy. Existing thesauri that cover your subject area or industry can prove extremely useful in generating ideas for broader and narrower terms.
6. **Perform associative linking.** The definition of related terms is highly subjective. For each term ask the question: "Where will users want to go from here?" Choose only the most obvious and important relationships.

As with everything on the Web, a thesaurus is never finished. The content and the terms used to describe concepts within that content will continue to grow and evolve. New terms must be added, old terms deleted, and relationships between terms revisited. And you should always be on the lookout for new variant terms. Thanks to *Roget's Thesaurus* and high school English classes, there's always someone out there with a new way to say the same thing.