DATABASE SEARCHING TIPS: PART 3

By Marvin Hunn

This is Part 3 in a three part series. Part 1 introduced some practical techniques of searching a structured, controlled-vocabulary metadata-based database using explicit operators and search syntax. Part 2 briefly explained some fundamental problems (steaming from the nature of language) that limit search engine performance. Finally here in Part 3 we turn our attention to a different approach to search and retrieval, based on keyword searching of documents using natural language input and document ranking. So in Part 3 we are getting away from metadata and away from explicit operators.

Is Metadata Really Necessary?

Why bother with metadata? It takes time and money to create metadata records (or to embed metadata tagging in structured documents). Why not just search the complete documents? Some arguments for controlled vocabulary metadata are as follows.

- Structured metadata allows field-specific searches (e.g., search for Calvin as author verses Calvin as subject)
- Metadata can be standardized. Names, topics, etc. can be expressed in a standardized controlled vocabulary. Standardization improves both precision and completeness of searches.
- Standardized metadata supports cross references and browsing of sorted entries.
- Metadata in one consistent language (e.g., English) simplifies searching a multi-lingual database. The alternative is searching for terms in all the languages.
- Non-textual documents (audio, video, still images) benefit from metadata.
- A mix of documents with metadata and documents without metadata is possible. This can lessen cost of creating metadata.

Above are some potential advantages of metadata. They are significant. But potential advantages are not always real advantages.

- Controlled vocabulary systems always lack terms for some specific user needs. So keyword searching of both controlled fields (e.g. subjects) and uncontrolled fields (e.g. titles, abstracts, full documents) is necessary.
- In practice, controlled vocabulary is implemented inconsistently. Controlled vocabulary is intended to collocate “everything” on a given concept under a single appropriate term. That is the goal. But in practice controlled vocabulary databases often list works on the “same” topic under multiple different headings. Likewise, I have seen controlled headings mis-assigned in hilarious ways. The claimed accuracy in assignment of controlled vocabulary terms to records is exaggerated.
- The need for standardization is true but somewhat exaggerated. Billions of Google users think non-standardized searches yield “good enough” results most of the time. They are satisfied by precise but incomplete search results; they are not looking for “everything” on the topic. For subject searches, natural language can be even more precise than commonly used controlled vocabularies. In theory, software can suggest synonyms to make recall more complete.

- The need for standardization is true but not fully solved by present systems because there is no standard standard for diverse databases. One common need for standardization in theological studies is bible passages. Is the book name “Canticles” or “Song of Songs” or “Song of Solomon” or “Hohelied” (German)? Is chapter three verse one “3:1” or “3,1” or “iii, 1”? Another common need for standardization is standardized transliteration of Greek and Hebrew using lemmas, not inflected forms. Personal name as subject also benefits from standardization. Interestingly, different databases try to standardize these things in different ways; there is no standard standard. The student must be aware of differences between WorldCat and ATLA and RAMBI and the catalog of the École Biblique et Archéologique Française. This diversity of standards somewhat resembles searching in non-standardized sources.

Natural language searching of full text documents with no metadata is often good enough. Indeed, sometimes natural language is better than controlled vocabulary metadata. Some arguments for searching based on natural language are as follows.

- The documents themselves use natural language and the terminology of the discipline. Why not use such language for searching? It is rich, diverse, vastly more comprehensive than a limited controlled vocabulary system, and always current
- Full-text provides deep access. Every word of every document is a potential access point. This is an enormous advantage over searching brief metadata records.
- Software is becoming clever enough to search non-textual documents without metadata. For example, software can convert audio recordings to text, and face recognition software can match facial images.
- Training is required to learn to search with explicit operators and controlled vocabulary. No training is required to do Natural Language Searching. This will make NLS very popular.
- Creating controlled vocabulary metadata for documents is very expensive. It is impossibly expensive to do it for everything in the world. (Perhaps it can be done for all scholarly documents, however.) So we must depend on matching full-text natural language to retrieve some documents. Why not spend our money developing software that will do this for all documents and abandon metadata?

Often a system only supports searching controlled vocabulary metadata or only supports searching natural language text. Learn how to do both.

Using Operators to Search Full-Text Documents

So how does one use standard operators to search full-text documents with no metadata? Some of the tactics you learned in *Part 1* can be applied to keyword searching of full-text documents.

- For any searchable textual content, using any search engine, to meet any information need, use successive searches and trial and error feedback to identify terms. Search, review results, note helpful terms, then search again.
- If the search engine supports operators, use exact phrases and proximity operators. The usefulness of logical "and" declines as the length of the document increases. Good search engines will allow you to use old-style operators and new-style document ranking at the same time. This is a powerful combination.

- If the search engine supports fields of any kind, and your initial searches yield very low precision results, then limit searches to titles to increase precision. Most internet search engines let you search for the title of a web page, for example.
- Experiment with synonyms to retrieve more.
- If you are just searching the open Internet (as opposed to peer-reviewed scholarly documents behind a pay wall), consider using Google Scholar (which includes some of Google Books) to improve quality of documents retrieved.

That is the low hanging fruit, the easy adaptations of what you learned in *Part 1*. Now we need to address the more radical differences between old-style and new-style retrieval systems.

Searching without Operators and with Document Ranking

Traditional library retrieval systems use explicit operators (AND, NEAR, etc.) to find matching records, and then sort records according to metadata fields such as date, author name, etc. Nearly all modern search engines de-emphasize search operators, work with little or no metadata, with little or no record structure, and rely on some kind of ranking to bring the most helpful items to the top of the list. They focus on ranking.

Ranking is usually discussed primarily in terms of two factors: relevance and reputation. Relevance has to do with subject matter or "aboutness." Is this document about the right topic? But relevance is viewed as a matter of degree so it is possible to rank by degree of relevance. How many of the search terms are in the document? Do those terms appear frequently in the document (thus suggesting a central topic of the document)? Are the terms close to each other (good) or far apart (bad)? Do search terms appear in a document title or in the URL/address (both suggesting a central role in the document)? When other documents mention or link to this document, what terms do they use to characterize the document? Those may be central ideas.

Reputation has to do with quality or accuracy or trustworthiness. Documents from websites known to have high quality material (like JSTOR) might be considered reliable. Page popularity is used as a surrogate for reliability. If many web pages point to the same web page, they are in effect voting for that page. If many documents footnote the same document, they are in effect voting for that document. That may indicate high quality or usefulness. Maybe. But it is really just an indicator of popularity. Not all votes are equal. A vote from (i.e., link from) a highly regarded source is weighted more highly than a vote from less highly regarded sources.

Using such ranking mechanisms, Google and its peers often work very well. But . . .

- Most web documents lack controlled vocabulary. This limits search magic, no matter how clever the software is in trying to deal with fundamental linguistic problems: synonyms, homographs, ambiguity and polysemy in general. Sometimes you must help the system by specifying synonyms and by quoting terms to get uninflected forms. This in turn requires that you understand why a given search is yielding bad results and what needs to be done to get good results. In other words, you must still think, not just rely on search magic.

- Most web documents lack fields or record structure (although html5 introduces some structural elements). This also limits search magic. (But internet search engines can parse structured documents which follow microdata format rules. See https://developers.google.com/custom-search/docs/structured_data. So an author search, for example, would be theoretically possible if restricted to such structured documents.)
- For the web as a whole, low quality documents/pages outnumber reliable, high quality documents. So searches of the entire web often retrieve many low quality documents, and may rank some of them highly. It is often better to search just scholarly documents. This is why Google offers Google scholar.
- Ranking by popularity is helpful but unreliable. Documents with many footnotes pointing to them do tend to be important, for example. But many popular websites (popular meaning many other sites link to them) are untrustworthy as sources of academic or scholarly information. Page popularity and footnote ranking require many links or citations to work. That is a problem. New pages/documents receive few links; it takes time for people to link to new pages. Old pages continue to receive links after they become obsolete. Pages on very narrow topics receive few links. There is some tendency to link to home pages rather than pages deep in a website, so deep pages receive fewer links. Many pages on the internet have few other pages pointing to them, so ranking by popularity does not help rank these pages. The network of footnotes in scholarly papers is a better ranking mechanism than links on the web.

So ranking is good, but it does not magically solve search and retrieval problems.

One mark of a really good ranking system is trainability. Good ranking systems allow the searcher to review results and indicate which documents were really desired. The system then examines words in the selected documents, learns what is desired, searches again, and ranks according to new criteria. If you don't see this feature, then you are not working with a good ranking system.

Although logical and positional operators are disappearing, we expect some specialized search engines will continue to rely on them. They are actually **necessary** when the search is targeting exact linguistic constructions. For example, a scholar might look for certain constructions in the Greek New Testament or in the Latin text of Aquinas. These texts might be tagged in special ways (e.g., grammatically, chronologically, geographically). The researcher needs to match exact field, tagging and wording.

State of the Art Search Engines[1]

What do state-of-the-art retrieval systems look like? They use natural language processing to parse search statements, so the searcher need not use logical and positional operators in a search statement. The searcher writes or speaks a few sentences describing what is wanted. State-of-the-art retrieval systems use statistical techniques to create document profiles (based on all the words in each document) to represent what each document is about, so the searcher does not have to match controlled vocabulary metadata headings. While processing search statements and while analyzing documents, these systems identify word roots, plurals, and otherwise compensate for word inflection. They recognize some syntactical

[1] Two good sources are C. Manning, P. Raghavan, and H Schütze, *Introduction to Information Retrieval* (New York: Cambridge University Press, 2008), and S. Büttcher, C. Clarke, and G. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines* (Cambridge, MA: MIT Press, 2010).

features like phrases and modifiers. They attempt to supply synonyms (to be more thorough) and to disambiguate homographs (to be more precise). They rank documents based on terminology in the documents, using approaches like the vector space retrieval model with tf*idf weighting (some function of term frequency times some function of inverse document frequency), and supplement that with other factors such as popularity (measured in several ways like sales, library circulation, clicks to display/download, explicit user ratings, number of citations/links to a given work, etc.), date/currency, or immediate availability of document (online, on shelf vs checkedout, interlibrary loan request). They may also consider anything known about the preferences of the person searching, including possibly private/confidential information, although there is much controversy about that.[2] One fear is that software will hide information it thinks you are not interested in, based on viewpoint (e.g., the software may show you only conservative or only liberal sources without telling you that is what it is doing, thus making it very hard for you to find other viewpoints or be aware of the range of views.)[3] Many of these so-called futuristic features have been available in experimental software for decades, [4] but they have been slow to reach libraries.

Further in the future, AI (artificial intelligence) software will play a role in search technology. Already primitive AI systems are learning to associate certain words and phrases with other words and phrases. They are learning to use linked data and taxonomies to categorize things (e.g. "this is a city; this is a person") and to draw inferences. In the near future AI systems might constantly monitor a person's activity and try to anticipate what that person would want next. For example, let's say you activate intelligent software to assist you with research. You train it by listing books and articles and websites that match your general interests. The software searches for all these documents, downloads as many as possible, analyzes them, and builds an interest profile for you. It follows the footnotes in all the books and articles, and analyzes those documents. It then constantly checks for new resources on those topics, day after day. It recommends new works. You rate those recommendations, and the software refines its profile of what you are interested in. Over a period of years the AI agent learns to anticipate what you need before you ask for it. AI may help in other ways in the future. It may study corpora of related documents and identify patterns and relations which scholars will then study. AI may also become a helpful servant in translating between languages and performing other linguistic tasks related to searching.

---

[2] Some non-library systems attempt to compile information about the person doing the search. What kind of information? They might track the websites you visit, what kind of documents you download and what links you click, what you buy, and similar information about your "friends" on social websites like Facebook, Twitter, etc. Some library systems keep track of books you have borrowed in the past, and offer to alert you when similar new materials are purchased by the library.

[3] For a description of how this already happens in search engines like Google, see Eli Pariser, *The Filter Bubble : What the Internet is Hiding from You* (New York : Penguin Press, 2011).

[4] All of these features (except gathering personal information about the searcher) have been around since the late 1960s. See G. Salton, *The Smart Retrieval System: Experiments in Automatic Document Processing*. (Englewood Cliffs, N.J: Prentice-Hall, 1971), which summarizes his work mostly from the 1960s.

The Need to Think

Future researchers will still need to think critically. That means there will never be a time when researchers can responsibly delegate their thinking to others, whether the others are AI agents or human colleagues. At the same time, the solo researcher cannot check everything and must rely on the work of others. Find the proper balance. Let that guide your efforts now.