December 2006

An Introduction to Relevance Ranking Systems

Scott Childers
University of Nebraska - Lincoln, scott@emeraldfusion.net

An Introduction to Relevance Ranking Systems

Guest Article
by
Scott Childers, Associate Professor
University of Nebraska-Lincoln Libraries

To be "relevant" is to be fitting or suiting given requirements, pertinent and applicable to the given situation. In Internet search engine or database terms, a relevance ranking is an attempt to measure how closely a web page or entry fits possible search terms. Search tools that display results in a relevance ranking order place their "best match," an entry with the highest relevance ranking on the top of the list, instead of using an alphabetical, date modified, or other more concrete sorting method.

Relevance ranking is usually best for searches that are not "either/or" types of searches. For example, in most traditional title searches, the result is either the library has the book, or it does not. The relevancy program would either show the entry for the book, or an alphabetical list that has a statement in the appropriate place that says, "Your search would be here." This is a very good place for this concrete, well-known sorting method.

Keyword searches, on the other hand, introduce vagueness and much more subjectivity on what constitutes acceptable entries as opposed to more "either/or" types of searches. Presenting an alphabetical list with results could bury the best results deep in the list, preventing the searcher from finding something useful in an efficient manner. Sorting results on relevancy gives us some hope to come up with that proverbial needle in the haystack by doing some preliminary work for the user.

Discussions about relevance ranking systems are nothing new in the library field. Paul Metz , Principal Bibliographer, Virginia Tech University Libraries, mentioned relevance rankings in computerized search results as early as 1989, in his article "Subject Searching in Libraries: Present and Future, Part 2," *Journal of Academic Librarianship* 14 (1989): 2. However, more people are becoming more familiar with the Internet search style of finding items on the World Wide Web (keyword searches and relevance ranked results lists) rather than the "traditional" OPAC choices of exact title or subject headings with an alphabetically sorted result list. This is creating more interest in relevance ranking systems for searching library holdings.

To better understand relevance rankings, a very simplified description of the typical process for assigning a relevance ranking to a web page is in order. In this example, the program that is being used examines the contents of the page, going through the text, counting how many times a word appears on that page. The ranking program then takes that information and uses that word count as the relevance value. This value is an attempt to measure how relevant the page is if that word appears in a search query. It stores the word, the relevance value generated, and the URL of the page in the search engine's index. This process is repeated over and over as more pages are found. Again, this is a very simplified example of the process.

Now that it has an index created, a search engine can use these relevance values to present the pages in some sort of order in an attempt to have the best results come up to the top of the list.

Below, I continue the oversimplified example using three pieces of information, with two lists and a declarative statement. We are assuming that the more times a word is in a set, the more relevant that page is to searches containing that word.

Sample data:
Set 1: apple, apple. orange, pear, apple, grape, grape
Set 2: orange, apple, grape
Set 3: An apple is not an orange.

Our index would look something like this:

Word:	Value:	Set
an	2	Set 3

apple	3	Set 1
	1	Set 2
	1	Set 3
grape	2	Set 1
	1	Set 2
is	1	Set 3
not	1	Set 3
orange	1	Set 1
	1	Set 2
	1	Set 3
pear	1	Set 1

Now we can search for terms and get results in some sort of order by displaying the results in descending order based on the relevance value. If a search for “grape” was done, the result list would be Set 1 first, Set 2 second, and Set 3 would be filtered out.

This example also shows that there are some flaws in this simple method. A set that would have arguably the most definitive information about apples (Set 3) would show up under a list that merely repeated the word multiple times. (Set 1). Our simple method did not take into account any type of context for the term being indexed.

Relevance rankings could be created simply by using counts of how often a search term is used on the page in question, as shown in our example above; however, most systems use methods that are much more complex to avoid the problem mentioned above and to provide more meaningful results. Some algorithms give more weight to those terms if they are in the title section of the page or if they are emphasized in any way such as being used as a section heading, or in a different font style than the rest of the page. Even the position of the word on the page may come into play. Examining any embedded metadata also takes place in many systems.

Another trend in relevance ranking systems for web pages is the use of information about what other pages link to the page being examined. What terms are being used to point to the page in question can also be used to determine how other sources see how relevant a page is to a search term. Some sites, such as Google, will use a recursive formula based on how popular the linking page is to help determine the relevance of the linked page.

As the relevance ranking is created automatically from information in the page, this leads to the ability to artificially inflate the relevance ranking the page would get, by placing terms either in the metadata section of the web page’s code, or throughout of the page. Early on in the World Wide Web’s history, unscrupulous web masters would add many lines of unrelated text to the bottoms of their pages to increase the number of hits they would receive. Often times the terms were synonyms, misspellings, or other variants of common words. There would also be the case where terms that had nothing to do with the content on the page were used, such as the words “sex”, “mp3”, or celebrity names in the hopes that this would inflate their rank in Internet search engine results.

There are many different algorithms that are in place that attempt to prevent artificial inflation of a page’s relevance ranking. They are usually closely guarded “trade secrets” and are often tweaked or changed when it is discovered that some entity is trying to “game” the system by finding ways to make their pages rise in the rankings as mention above. The other side to that is that some Internet search engines have a policy of allowing companies to purchase the top positions in their listings. In both cases there is skewing of the results lists in favor of someone’s financial gain, instead of helping the information gatherer.

In our example above you may have noticed that a search term that could be used, “fruits”, would find nothing relevant in this list and would result in no results being presented because that term was not listed in the set itself. If no page has been indexed with a particular search term, then there will be no results for that term. Human indexing could have identified synonyms or other types of appropriate access points, but

the focus is often to process items quickly and cheaply, not necessarily with an a eye toward accuracy, and certainly with no concern for authority controls that are typical of traditional cataloging.

Another problem with any relevancy ranking is the determination on when to have it say that there are no closely relevant items available. Some systems now show sort of "confidence" measure next to the result in some sort of percentage (90%") certain, color coding ("Green means very certain, yellow means sort of certain, and red means no that certain at all), or some sort of iconic symbolization (5 out of 5 stars, or bars like a cell phones signal strength) in order to convey just how relevant the result is to the initial search.

Determining context and meaning of terms in a source is also problematic. Ambiguous search terms such as "football" can leave the engine to guess if the user means the European meaning of football (soccer to Americans) or the American meaning of football (gridiron or American Rules Football to many in the rest of the world) and often creating muddled result lists.

Search engines are attempting to personalize their relevancy rankings to individual users, collecting data on what they click and what they do not, adjusting the rankings accordingly and creating a system more closely resembling a recommendation system. Google's Personalized Search beta is at http://www.google.com/psearch and is an example of this idea.

As more searchers depend on the system to do this preliminary sorting and filtering, libraries must think about their own data creation methods to assist these searchers. Expanding a record with a table of contents, synonyms of the subject headings, and the like will help create a more accurate relevance value. Embedding some of the authority control information directly may help as well; as an example books with Twain, Mark, 1835-1910 would also include "Samuel Clemens" in a searchable field somewhere in the record as well.

In conclusion, relevancy rankings, when implemented sensibly, can benefit the researcher by doing some preliminary sorting of results into a potentially helpful order, but they are not a replacement for a human's ability to determine what might be useful and what is garbage. While the algorithms and programs are constantly being updated, tinkered with, and otherwise made "better," there is something that can be done to help the searcher who uses keyword/relevance ranking systems. That something is to have improved and expanded records.