

Editorial

Bayesian indexing: the next craze in search algorithms?

The author

Michael Seadle is Editor of *Library Hi Tech*. He is also Digital Services and Copyright Librarian at Michigan State University, East Lansing, USA. <seadle@mail.lib.msu.edu>

Keywords

Bayesian statistics, Heuristics, Indexing

Abstract

Bayes' theorem is about updating assumptions. It provides a mathematical basis for a heuristic search algorithm. A system using such an algorithm could make a kind of "best guess" about what the query is likely to mean. Some disadvantages include the need for more individualized information, a possible tendency to focus on a limited set of works, and the potential for encouraging sloppy searching. The literature on heuristic systems is already substantial, and is relevant to all of us in the library and information field.

The first three times the statistician said, "You really ought to consider Bayesian indexing", I smiled politely and turned the discussion in another direction. This took place at a "National Gallery of the Spoken Word" project meeting that included librarians, computer programmers, engineers, historians, and this one lone statistician[1]. I knew a little about the Reverend Thomas Bayes (1702-1761), and had a vague recollection that Bayesian statistics had to do with updating probabilities. But I had no clue what Bayesian indexing might be, or how it related to our discussion.

When he repeated his comment the fourth time, I asked.

The indexing we do today for our OPACs and databases is not probabilistic. It takes each occurrence of a word, records where it can be found, and points to it when someone asks for that word or some portion of it. The indexing assumes that people want what they ask for and know what they want, which anyone who has worked at a reference desk realizes is simply not true.

If a student sits down at a search screen and types in "president", she may well get a screen full of works about President Clinton, because Clinton is the sitting president and most systems list current works first. She may in fact have been looking for some other president, or a list of presidents, or a corporate president who is not usually referred to as President Suchandsuch in titles, subjects, or other indexed fields. Eventually she might ask for help at the reference desk, and confess in the reference interview that she is writing a paper about the 1930s for a history class. The librarian then turns to the search screen with the new assumption that she wants something about Herbert Hoover or Franklin Roosevelt.

Bayes' theorem is about updating assumptions, as we do in the reference interview process. In more mathematical language, it is about revising prior probabilities with new information:

The so-called Bayesian approach . . . addresses itself to the question of determining the probability of some event E_1 , given that another event A has been observed. The event A is usually thought of as new information, so that Bayes' rule is concerned with determining the probability of an event given certain new information . . . (Kenkel 1984, p. 140)

In effect, Bayes' theorem provides a mathematical basis for a heuristic search algorithm. This has obvious advantages. Instead of locking the search results for "president" into an unvarying, chronologically sorted list, the system makes a kind of "best guess" about what the student is likely to mean. In other words, it might learn that Franklin Roosevelt is more likely to be wanted than Clinton when someone enters "president".

The approach has potential problems that are as obvious as its advantages. At a large university (mine has 43,000 students), several dozen classes are likely to have students searching for information about presidents at any one time. If students from the history class train the system to give them "Roosevelt" hits, a class of political science students the next week will not be well served, and will have to retrain the system to emphasize contemporary presidents. At the same time, the lone business student who enters the same search in the hope of finding something about corporate presidents, may despair of finding any – as might have happened anyway under a purely chronological approach.

A heuristic system would work better if it could know more about the searcher, and if it could build separate probabilities for distinct groups. For example, a student who first identified herself/hisself as a member of a particular history class could profit from prior searches by other classmates, without the noise of searches from political science or business. A business student might benefit even more, since works on corporate presidents occur significantly less often than those about the head of the US Government.

Unfortunately this solution too has its liabilities. It forces searchers to reveal more personal

information, some of which they might feel should remain confidential or simply not want to bother to key in. It also could increase the likelihood for students in the same class to look at the same materials, since they would get hits based on prior choices. Such a heuristic system may even reduce students' already modest inclination to learn better search techniques. It may be that anyone who enters so imprecise a search should automatically get a search-tips screen instead of any result list at all – some systems do this already, though searchers (rightly) regard this as an unfriendly slap on the wrist.

The literature on heuristic systems, whether using Reverend Bayes' theorem or not, is substantial. This editorial does not presume either to summarize or extend it. But the issues involved in heuristic search algorithms should interest and involve everyone who works in the library and information field. This journal in particular would welcome new articles, research, and commentary on the topic.

Note

- 1 The National Gallery of the Spoken Word is a Digital Library Initiative (Phase 2) project sponsored by the National Science Foundation, National Endowment for the Arts, the Library of Congress, and other federal agencies. Available at: www.ngsw.org

Reference

- Kenkel, J. (1984), *Introductory Statistics for Management and Economics*, 2nd ed., Duxbury Press, Boston, MA.