Economics and Search

Hal R. Varian*
School of Information Management & Systems
University of California at Berkeley
Berkeley, CA 94270-4600
hal@sims.berkeley.edu

They say that economists are people who are good with numbers but don't have the personality to become accountants. I want to belie that mild-mannered image by doing something rash and even downright dangerous: I want to talk to you about some work in my subject, economics, that I think might be relevant to the work in your subject, information retrieval.

I take this step with considerable trepidation, since I know that it is unlikely to be successful. It is rare that an outsider can really contribute anything useful to another subject, as I know from listening to innumerable speeches by the physicists, biologists, and mathematicians about what their subject has to say about economics.

However, it is also true that such cross fertilization can be extremely stimulating. Economics has, in fact, learned a lot from physics, biology and mathematics. And even when such attempts at interdisciplinary communication fail, as they often do, it is often interesting to see how others approach the questions that are your main business.

Here are three areas in economics where I think there might be some fruitful cross-fertilization:

Economic value of information. How economists define the value of information.

Estimating probability of relevance. How nonparametric econometrics might help in determining functional relationships.

Optimal search behavior. How the theory of optimal search offers some surprising advice on ordering retrieved results.

Economic value of information

Economists define the (economic) value of information in the context of an optimal choice problem. A consumer is making a choice to maximize expected utility or minimize expected cost. The value of information is the *increment* in expected utility resulting from the improved choice made possible by better information. Often this can be translated into some monetary equivalent representing how much someone would pay to acquire a given piece of information. (See Laffont [1989], page 61.)

To take a very simple example in an IR context, suppose that a user is given two sealed envelopes, one containing $100 the other containing $0. She is allowed to choose one, open it, and keep whatever is inside. To make things simple, suppose that she is risk-neutral, in the sense that she only cares about expected value.

In the absence of any information, she would choose an envelope at random, receiving an expected payoff of $50. If she had accurate information about which envelope contained the prize, she would, of course, choose it and receive $100. Hence the value of information about which envelope contains the prize is $50, the increment in value she would get by making the better choice. In this very simple case, it is also the upper bound on how much she would pay to acquire that information.

Now, the interesting thing about the economist's notion of the value of information is that it is only *new* information that matters. If she reads a document that says "the prize is in envelope on your left" then another document with the same information has no incremental value. This is quite different from "relevance" as it is usually defined since duplicate documents may well be relevant to a choice problem, even though the second instance of the relevant document is certainly not valuable.

We see this effect at work in the stock market where it is only *surprises* that move markets. If everybody expects the Fed to raise interest rates by 1/4 point in their next meeting, the market doesn't budge when this event actually occurs. But if they raise the interest rate by 1/2 point, the market may respond significantly.

How is this relevant to IR? The standard practice of "present documents in order of estimated relevance" doesn't take account of the incremental value of the documents already viewed. To take a trivial example,

*This was the invited plenary address at ACM-SIGIR 1999 held in Berkeley, California.

almost everyone tries to remove duplicate documents from retrieved lists since it is obvious that they add no new information. However, the same logic suggests that documents that are *similar* to those earlier in the list probably add little value to those already examined.

This in turn argues for strategies like post retrieval clustering of results. Several researchers have advocated such an approach for a variety of reasons such as reduction of cognitive load, disambiguation, and so on.[1] The reasoning may add another twist to this discussion, namely that clustering is good when it maximizes the difference (i.e., the incremental information content) across clusters.[2]

Estimating probability of relevance

The next topic I want to discuss is the problem of estimating the probability of relevance. I approach this problem in the tradition of Cooper et al. [1992]: given a training set of documents, queries, and 0-1 relevance judgments, estimate the probability that a given document will be relevant to a given query.

Cooper used a number of explanatory variables such as terms-in-common, length-of-document, length-of-query, inverse document frequency, and so on. In general, given a vector of explanatory variables, X, one can assume that the probability of relevance conditional on X is given by the logistic distribution

$$p(X) = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{1}$$

and then use maximum likelihood techniques to estimate the vector of parameters β.

The logistic parametric form is, of course, only one functional form, and maximum likelihood is only one estimation technique. It is a particularly convenient choice in a data-poor environment since it doesn't require a large training set and the likelihood function is concave, meaning that there is normally a unique maximum and standard optimization techniques work very well. These features made logistic estimation very attractive when Cooper did his work.

However, maximum likelihood does require assuming a particular functional form. In a data-rich environment, this may not be necessary or even desirable. In particular, it is possible to use nonparametric methods to examine how well any particular functional form performs in fitting the data.[3]

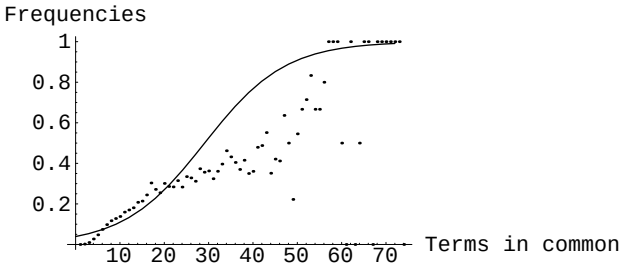

Figure 1: Frequencies of relevance and maximum likelihood estimate.

To illustrate this, I chose two TREC samples of document-query pairs from the *Wall Street Journal*. The data for fitting consisted of 100, 102 doc-query pairs; the data for extrapolation consisted of 173,330 doc-query pairs. The data were stemmed in the standard way.[4] To keep things simple I used only one explanatory variable, $x =$ terms in common.

I then calculate the *frequencies of relevance* by looking at all document-query pairs that had k terms in common, and calculated the fraction of these documents that were relevant. I then used these frequencies of relevance as inputs to a logistic regression and as inputs to a nonparametric regression, both of which will be explained in more detail below.

Figure 1 depicts the frequencies of relevance along with the maximum likelihood estimate of the probability of relevance, assuming a logistic functional form. Note that the frequencies-of-relevance and the maximum likelihood estimate have a somewhat different shape—in particular, the frequencies-of-relevance have a convex shape in the region involving 30-60 terms-in-common. This indicates that the logistic form isn't a particularly good fit in that region.

The logistic distribution in equation (1) implies that

$$\frac{p(x)}{1 - p(x)} = e^{x\beta}.$$

Taking logs of both sides of this expression gives us

$$\log \frac{p(x)}{1 - p(x)} = x\beta.$$

If we apply this so-called "logit transform" to the frequencies we have a simple linear regression:

$$\log \frac{f(x)}{1 - f(x)} = x\beta.$$

[1] See, for example, Hearst and Pedersen [1996] for a literature review and an argument about the value of clustering for disambiguation.

[2] An audience member told me that Carbonell and Goldstein [1998] have implemented a similar idea that they call "maximal marginal relevance."

[3] An audience member told me that Greiff [1998] also proposed using nonparametric methods for this end and conducted a more detailed analysis using related but different techniques.

[4] Thanks for Aito Chen and Fred Gey for providing me with these data.

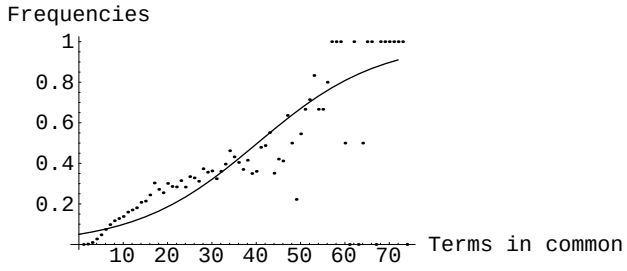

Figure 2: Frequencies of relevance and maximum likelihood estimate.

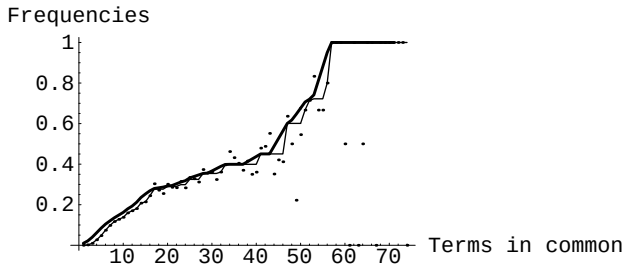

Figure 3: Frequencies of relevance, along with best-fit monotone function and a smoothed version.

Note that the logit transformation makes no sense when $f(x)$ equals 0 or 1, so we have to censor such observations. If we estimate this regression using ordinary least squares, we find the results in Figure 2.

Note that this provides a somewhat worse fit for low and high numbers of words-in-common, but fits much better over the middle range.[5] However, note that the curvature of the fitted function still seems to reverse of that implied by the data.

We can do somewhat better by using a nonparametric technique known as nonlinear regression. In this particular case, I used a technique known as PAV (pool adjacent violators) which finds the monotone function that minimizes the sum of squared residuals between the observed frequencies and the fitted function. (See Härdle [1989], page 218.) The results are depicted in Figure 3, along with a 4-term moving average of the fitted function.

We now have three candidates for the relationship between terms-in-common and probability of relevance: the maximum likelihood estimate, the logistic regression, and the nonparametric regression. These fits are compared to the second sample of *Wall Street Journal* doc-query pairs in Figure 4.

[5]Part of this is due to the fact that we have censored the observations with high number of terms-in-common.

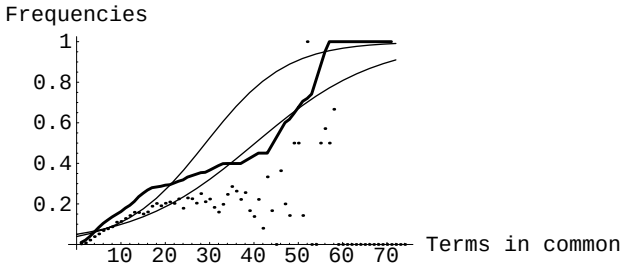

Figure 4: Performance on second sample.

The reader can judge for him or herself which method performs better. To my eye the nonparametric regression performs very well, picking up the same convexity in the 30-60 terms-in-common range as appeared in the earlier dataset.[6]

Of course this is only a toy model. In a real application it would be important to add some extra explanatory variables such as document and query length, document frequencies, and so on. Nevertheless, I think that this work (and the work by Greiff [1998] cited earlier) shows that EDA and nonparametric techniques may be very useful in a data-rich environment.

Optimal search behavior

Economists have been interested in search behavior ever since Stigler [1961]. (See Kohn and Shavell [1974] for a general model and Lippman and McCall [1982] for a survey.) This interest has been motivated by interest in modeling consumer behavior such as searching for the lowest price or the highest wage.

In the economic context, it is of interest to find the *best* price or wage, while in a document search one would presumably be interested in a *set* of good documents. Despite this difference, I think that the economics literature suggests a few interesting models and contains at least one surprising insight. To exhibit this insight, I will examine a simplified version of Weitzman [1979]'s "Pandora Problem."

Pandora has n boxes to open. The reward in box i is random with distribution function $F_i(x)$. It costs her c_i to open a box and she has a time discount factor of $d < 1$, which means that receiving a reward of R next period is worth dR to her today. Her payoff is the maximum value found up to the point where she stops opening the boxes.

[6]It is worth noting that the precision-recall measures for all three estimation techniques are the same since they are all monotone transformations of each other.

To fix ideas, consider the following practical interpretation of this abstract problem. You work in an airport bookstore selling travel books. People are in a hurry, they have a cost to examining the books, and they can only take one book with them. Due to your experience, you have a pretty good idea of the likely appeal of the books to potential customers. Your problem is to determine the order in which you show them the books.

One of my colleagues suggested "That's easy—just show them the most expensive book first!" Somewhat surprisingly, this was a computer science colleague, not an economist colleague. But to avoid this issue, we assume all books have the same price and that your goal is to satisfy the consumer.

Weitzman shows that this problem can be solved by dynamic programming. In particular, there is a way to assign a score to each box that depends only on the characteristics of that box and is relatively easy to compute. Once these scores are assigned, the optimal search can then be completely characterized by the following rules:

Selection rule: if you open a box, open the box with the highest score.

Stopping rule: stop searching when the maximum sampled reward exceeds the score of every closed box.

The interesting thing is that the score is *not* the expected value of the box. In fact, the optimal search strategy can easily involve opening a box with a low expected value *before* a box with a higher expected value!

The reason is that opening a box with a very spread out or risky distribution first allows for the possibility of terminating the costly search. As Weitzman puts it: "Other things being equal, it is optimal to sample first from distributions that are more spread out or riskier in hopes of striking it rich early and ending the search."

To see how this works, I have constructed a very simple example. Suppose that there are only two boxes. Box S gives a payoff of 6 for sure. Box R has an equal chance of 10 or 0. Note that S, the "safe" box, has a higher expected payoff than box R. If you were only going to choose one box, S would maximize expected payoff.

Let us consider the optimal search behavior. Suppose Pandora opens box S first; should she continue to open box R? Half of the time she will get $10d - c$, half of the time she will get $-c$. (The d is the discount factor indicating that the payoff comes next period; the c is the cost of opening the next box.) The expected payoff from continuing her search is therefore $5d - c$ which is less than 6. Therefore, if she opens box S first she will have a payoff of 6 and will not continue.

Suppose Pandora opens box R first. Half of the time she will get 10. Since she can't do any better than this, she will stop her search and go home happy. Half of the time she will get 0. In this case, she will continue if the net payoff from opening the safe box is positive:

$$6d - c \geq 0. \tag{2}$$

Her expected payoff from opening the risky box first and continuing is

$$\frac{1}{2}10 + \frac{1}{2}(6d - c) = 5 + 3d - c/2.$$

Hence opening R first is the best strategy if this expression is larger than 6, the payoff from opening the safe box first. After some simple algebra this condition becomes

$$6d - c \geq 2. \tag{3}$$

This inequality is stronger than (2), so if opening box R first is optimal, then it is always optimal to continue to the safe box if the payoff from R is zero.

In conclusion, the optimal strategy is to open the risky box R first if $6d - c \geq 2$. If you get the high payoff stop, otherwise continue. This is true even though the safe box has a higher expected value. The reason is that the risky box has what economists call "option value." Once you've seen the payoff, you have the option to continue, and this option is, itself, valuable.[7]

Option value plays a big role in search. Indeed, if you didn't have the option to truncate the search, it could hardly be called search. And option value is increasing in the riskiness of the choices—which means that it is better to look at risky choices early in the search process.

It follows that the standard practice of ordering retrieved documents by their expected relevance, or expected value, or any such expectation is not really right. The actual ordering should depend not only on the estimated first moment of the distribution, but on higher moments as well.

Let us return to our airport bookstore example. A patron rushes into your store and says "Quick, I need a guide to Borneo." You have two in stock, one by *Fodors* which most people find adequate, and one by *Lonely Planet*, which some people love and some hate. Which do you show first? I claim that it makes sense to show the *Lonely Planet* guide first. If the person loves it, they don't even have to look at *Fodors*. If they hate *Lonely Plant*, you can give them the *Fodors*. This example illustrates why a document that has a low expected payoff may still be presented earlier, if it has a chance of yielding a large payoff early in the search.

[7] For more applications of option value, see Dixit and Pindyck [1994].

Whether this principle is of practical import at this stage of development of IR is hard to tell. We certainly don't have very good estimates of any of the parameters—the discount factor, the cost of search, or the expected value of the alternatives. Until we can get better measurements of these key parameters, it will be difficult to apply the theory of optimal search to practical problems.

Conclusion

At the beginning of the talk, I said that I was doing something dangerous in talking about what economics might contribute to information retrieval. Such interdisciplinary trespassing is, by it's nature, a risky undertaking. But perhaps I have some justification with this last example—at least in some cases, an optimal search involves looking at the riskiest items first!

So thank you for your time and attention; it has been a privilege for me to talk with you this morning.

REFERENCES

1. Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, August 24-28, 1998 1998. SIGIR, ACM. `http://www.acm.org/pubs/contents/proceedings/ir/290941/index.html`.

2. W. S. Cooper, D. Dabney, and F. Gey. Probabilistic retrieval based on staged logistic regression. In Nicholas Belkin, Peer Ingwersen, and Annelise Mark Mejtersen, editors, *Proceedings of the 15th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 198–210, Copenhagen, 1992.

3. Avinash Dixit and Robert Pindyck. *Investment under Uncertainty*. Princeton University Press, Princeton, NJ, 1994.

4. Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, August 24-28, 1998 1998. SIGIR, ACM. `http://www.acm.org/pubs/contents/proceedings/ir/290941/index.html`.

5. Wolfgang Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1989.

6. Marti Hearst and Jan Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual ACM-SIGIR Conference*, Zurich, 1996.

7. M. Kohn and S. Shavell. The theory of search. *Journal of Economic Theory*, 9:93–123, 1974.

8. Jean-Jacques Laffont. *The Economics of Uncertainty and Information*. MIT Press, Cambridge, MA, 1989.

9. S. Lippman and J. McCall. The economics of uncertainty. selected topics and probabalistic methods. In Kenneth Arrow and Michael Intrilligator, editors, *Handbook of Mathematical Economics*, volume 2, chapter 6. North-Holland Press, 1982.

10. George Stigler. The economics of information. *Journal of Political Economy*, 69:213–225, 1961.

11. Martin Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, May 1979.