Reading Joachims et al.

Philip Levis
April 17, 2011

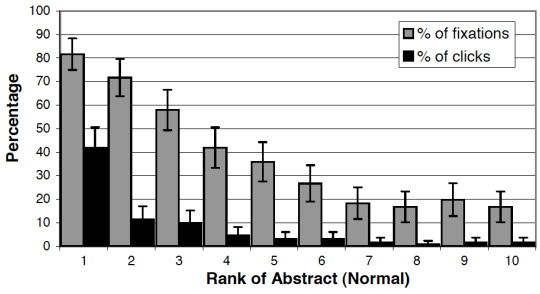
The reading for Friday's class is on web search rankings. The basic problem is that a search engine wants to use which links users click on as input to a learning algorithm that refines searches. For example, if searchers for "rhino club" always click on the second result, chances are it should be the first result.

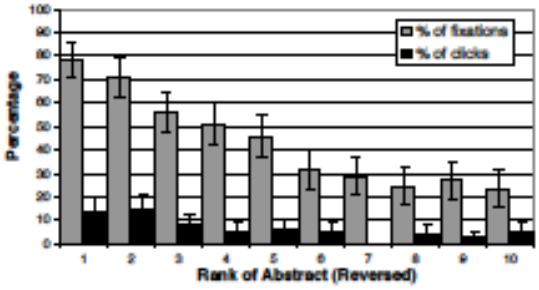
Challenge

While this idea seems simple in theory, in practice it's much more complicated. It turns out that the order in which a search engine returns results affects which results users click on. Therefore, simply saying that the most clicked-on result is the best one might not be correct. There are many possible reasons for this bias. For example, users often don't read all of the results and go with the first one that seems reasonable. Also, users may trust the engine enough to take its recommendation.

Experimental Setup

To try to quantify the effect order has on click-throughs, the paper uses a web proxy that modifies the returned results. They also use an eye tracker, a device that lets them see where and for how long users look at given results.





The two plots above show the distributions of the probability a given result has a fixation (user looks at it) or is clicked on. The plot on the left shows standard Google results, while the plot on the right is for when the results are reversed. The take-away from this plot is that the result order matters: while result 10 in the left plot is clicked on only a few percent of the time, on the right its corresponding position (1) is clicked on over 10% of the time.

One issue the paper faces is ground truth: which results are truly better than others? To establish ground truth, the paper uses a set of experts who rank results in terms of "relevance" to the question being answered. This ordering allows the paper to ask questions such as how

order of results relates to order of relevance (Table 3). E.g., how often will a user click on a later, more relevant result before an earlier, less relevant one?

What We Can Learn

This raises the question: how can one interpret click-through data as a signal to indicate which results are better or worse? To do this, the paper goes back to the expert results. It presents a variety of algorithms that one could use to infer that link A is better than link B, based on click-throughs, then measures how often these judgements agree with the expert ones.

What To Read

There are several places where the paper mentions the statistical test it uses without explaining why, such as the Wilcoxon test. Be sure to look these up to understand what they are; once you do, it should be very clear why that test is being used. For example, Wilcoxon is for ordinal, rather than numeric, values.

Table 3 can be a bit difficult to decipher on the first reading. The really critical columns are the two middle ones, where only one link is clicked on. The basic takeaway is that order matters: users select the first link more often than the second even when it is less relevant.