

Malcolm Slaney
*Yahoo! Research
and Stanford
Center for Computer
Research in Music
and Acoustics*

Precision-Recall Is Wrong for Multimedia

Precision-recall¹ is not the right metric. There, I've said it. I'm not the first to say this in regard to information retrieval, but my reasons are different because we are working with multimedia objects.

Search clickthrough ratio

Before we delve into this matter, I'd like to present an example that requires a little diversion to describe the foundations of search. Search engineers endeavor to find the right results, and they tune their algorithms on the basis of which items people click on the search results page (SERP). This is called the *clickthrough ratio* (CTR). In the ideal situation, a user clicks on the first item and never comes back, therefore hopefully satisfying their information need with one click. On the other hand, it's bad news when the user comes back to the SERP and clicks on the second result. This indicates that the first result did not have the right information. Perhaps the user comes back yet again, indicating that the second result wasn't right, and then clicks on the fourth result. This doesn't tell us anything about the third search result, but perhaps the fourth is correct (or he or she abandoned the search completely). Each click doesn't contain much information, but aggregated over billions of users, we have evolved the high-quality search engines of today.

Figure 1 shows an important metric my friends who do search use to summarize the performance of their systems. This line shows the expected clickthrough ratio of items on an SERP. On the whole, it's a monotonically decreasing function of position. On average, the first result gets the most hits, and the second gets less. This is a robust result; even when we randomize the search results, something our product managers don't like for us to do, users still click on the first result. This could be a sign of two different things: users trust search engines, or users are trained to click at the top of the page.

There are a couple of discontinuities in this curve. At the bottom of the screen, there is a big drop in the CTR because the user has to scroll the window to see the new results. Likewise, there is another discontinuity when the user has to click the "next page" button. (This drop in CTR at the end of the page is one reason that search engines have introduced infinite browse, where the search results fill in at the bottom of the page automatically.)

Now look at an image-search-result page for a popular celebrity (see Figure 2). At the top are some recent images of the celebrity, because users often want to see the most recent news about people. These images are there because people click on them. The left side of the page contains some related people and objects. These facets help people to explore different aspects of their search query.

But I want to draw your attention to the row of pictures along the bottom of the SERP (see Figure 3). These are related celebrities, and the name of the celebrity accompanies each picture. What is interesting is that these pictures get as many clicks as the pictures at the top of the page. Let's be clear here. Users have asked for *X*, we gave them clearly labeled pictures of *Y* and they clicked on them.

Editor's Note

I want to open a dialog about issues that are important to our field, and could change the direction of our work. I don't intend to write most of the columns, and I am soliciting ideas. Please send me your suggestions, either for issues you would like to discuss, or to nominate a colleague who has new and interesting ideas that deserve broader attention. Clear and perhaps controversial ideas are welcome.

—Malcolm Slaney

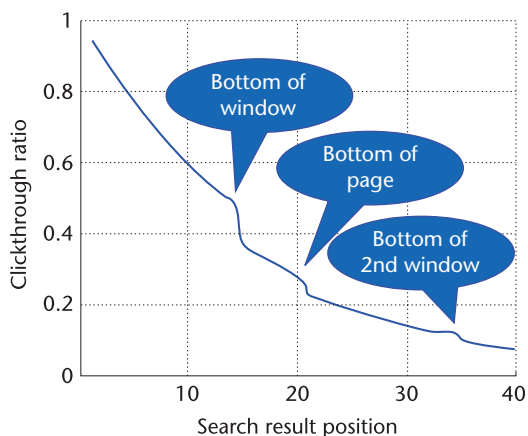


Figure 1. Relative clickthrough ratio (CTR) versus search result position. Search CTR declines with position on the page. (Adapted from Chapelle and Zhang.²)

We gave them the wrong answers, and they liked it. These aren't even query suggestions or enhancements. They are the wrong information! Why?

I propose that people's multimedia search activities are often driven by entertainment needs, not by information needs.

What would Shannon say?

In a Shannon sense (see http://en.wikipedia.org/wiki/Claude_Shannon), all queries are a request for information. The Xerox Palo Alto Research Center folks talk about information browsing or the scent of information.³ This is a brilliant formalism for analyzing people's informational quests. But perhaps people's multimedia motivations are different. You start reading a story about Nicolas Sarkozy. But then is a search for his wife's picture a genuine information request? And then what about a later search for the party they attended? And just how did you end up on a news article about Silvio Berlusconi's parties?

You could, postexperiment, formulate this Web-browsing session as a search for information. There is a scent. You can analyze the entire search path with an eye toward this eventual (semipurient) goal. But is information retrieval the right way to model this problem?

And what is the information content of a comedy film? I'm pretty sure that more people watch Web videos of silly pet tricks than they do of the world's best algebra lectures. If a user asks for early Laurel and Hardy films, would they be happy with any slapstick

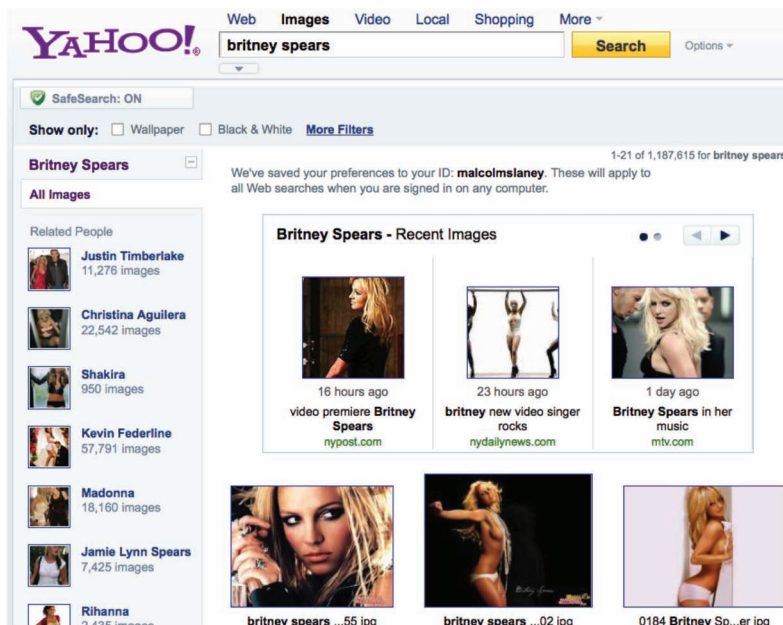


Figure 2. Image search results. The most important part of the search page shows current results on the top, and related people along the left.



Figure 3. The bottom of the image search results shows different answers—same type of person, but not the requested star. Users clicked on these images as often as they did the ones at the top. There is no subversion. The images are labeled correctly and are clearly not the ones requested.

silent film? I suspect in many cases the answer is yes.

Speech people like to separate the informational from the prosodic components of the acoustic signal. Speech-recognition systems endeavor to understand the information in the words, and they throw away the prosody. I don't know how well this example translates into other languages, but the way that most people say "yeah, right" is not a positive statement. Shannon would argue that this is all information, but we often drop the emotional messages when recognizing speech or parsing the Web.

Emotional messages

Thus, multimedia signals are complicated, often because they are used to convey emotion. It's certainly true that good authors can paint an emotional picture that can send chills up your back. But it's a lot easier to do this with a picture or sound. What is the information content of fingers screeching on a blackboard?

Arguably most of the information in music, especially those without vocals, is the emotional message. You can ask for recordings sung by John Lennon, and that is so well defined that precision-recall makes sense. But is that what the user really wants? How about songs written by John Lennon and sung in his style? And how do we ask or evaluate a query such as "happy songs?"

It's interesting to compare two different kinds of news outlets. Our local public radio station often rebroadcasts the audio of the Public Broadcasting System's "NewsHour." This television show prides itself on its high-quality information. And thus, most of the information is conveyed in the words. Even over radio, you don't miss much. In fact, it's hard to believe this signal started out as a video. Contrast this with the sensational approach of TV shows that cater to our emotional side. Pictures of large fires and action-packed videos set a tone that is hard to replicate with words alone. Often the sensational low-brow approach is more popular.

I suspect people searching for entertainment use our present-day search engines because that is the only tool they have. However, these are a poor substitute for what people really want to do. Video sites often show related videos to a user. They want to encourage more video consumption because it's good for their advertisers. But the only productive way they can do this is by organizing the content into channels, or by making it easy to find related videos. A search optimized for precision-recall doesn't help their users.

I'm not going to say much about so-called adult queries that we see in search logs. These queries aren't rare. Yes, some of them are genuine requests for information. But I was amazed to see the number of these queries when I looked at the image-search logs. It's not clear how many of these are coming from human searchers versus those coming from robots. I'm not going out on a limb by suggesting

that many of the adult queries are based on entertainment needs. How should precision-recall enter into this calculation?

Using precision-recall

The most common reason that we disparage precision-recall is because there are so many documents on the Web that to measure recall is nonsensical. This is definitely true of multimedia. And it's certainly true that precision is difficult to measure on a query that is essentially "show me interesting hair styles." Yet, these criticisms start with the assumption that information is our goal. There is a genuine informational need when you search for pictures of a 1974 Pinto to illustrate an article on automobile safety. Precision-recall might or might not be a good metric for these informational queries.

But the multimedia story is more complicated because people are often looking for entertainment, not information. If I had to venture a guess, I'd say that genuine informational queries are a small percentage of the requests seen by multimedia search engines. Precision-recall is an imperfect metric for text searches on the Web. Add in entertainment queries and precision-recall becomes all but meaningless. We know that we often use multimedia to entertain ourselves.

The collaborative-filtering folks have discovered the power of recommendations over informational requests. They don't measure their performance by asking which movies satisfy your entertainment needs (precision) or can we find them all (recall). Instead they ask whether you like this movie. Remember, Netflix didn't pay \$1 million to optimize their search engine.

Conclusions

As an applied researcher, I often joke that the only metric that matters is this one—and then I pull out my wallet. A related metric is whether you spend time looking at my webpage. In both cases, we are measuring whether people are willing to invest in our product by giving us something of value, either money or attention. But this is a hard metric to evaluate, often taking years for a new venture to get enough attention so the idea can be tested.

As scientists we need to find relevant metrics—something we can optimize or test

in the lab. The scientific metric is relatively easy if you are testing a new drug: does it kill the cancerous cells? However, what is the right metric for multimedia problems? You can always use precision-recall to evaluate your work, but does this really measure something that people care about?

I submit if you are using precision-recall to evaluate a multimedia system, the system is being optimized for a problem that is not of primary importance. When was the last time you have searched for a video that contains "Rock Hudson washing up on a beach" or put in the query "show me a video of a wing walker."

I think we can safely say that the Web is here to stay, and search will always be a part of the Web. I think it's also safe to conclude that precision-recall is not how the most popular multimedia sites will measure their performance. Here is something to consider when you think about your next research project: if precision-recall is the metric, are you asking the right question?

MM

Acknowledgments

The ideas presented in this column have evolved over the years, due to discussions with many people. I especially want to thank Dan Russell, Berthier Ribeiro-Neto, David Ayman Shamma, and Sara Anderson.

References

1. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, 2nd ed., Addison Wesley, 2011.
2. O. Chapelle and Y. Zhang, "A Dynamic Bayesian Network Click Model for Web Search Ranking," *Proc. World Wide Web Conf.*, ACM Press, 2009.
3. P. Pirolli, "An Elementary Social Information Foraging Model." *Proc. 27th Int'l Conf. Human Factors in Computing Systems*, ACM Press, 2009, pp. 605-614.

Contact editor and author Malcolm Slaney at malcolm@ieee.org.



Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

ONLINEPLUS™
publishing evolved

A new publication model that will provide subscribers with features and benefits that cannot be found in traditional print such as:

- More Rapid Publication of Research
- Online Access to the CSDL
- Interactive Disk and a Book of Abstracts
- Lower Price

Available Transactions Titles by 2012:

- TDSC
- TMC
- TPAMI
- TPDS
- TVCG

For more information about OnlinePlus™, please visit <http://www.computer.org/onlineplus>.

IEEE computer society