

A Bayesian Derivation of Personalized and Social Search

Sepandar D. Kamvar and Damon Horowitz

In this note, we derive the social ranking function described in equation 2 in [3], and put it into a context with the personalized ranking functions described in [5, 2, 6, 1, 7, 4].

Let's begin with traditional web search, where we wish to score how well document d answers query q . We can use Bayes' law to write:

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)} \quad (1)$$

where $p(d|q)$ is the probability that document d can answer query q .

This can be written as the composition of the query-dependent text IR score $p(q|d)/p(q)$, which gives how much more frequently the terms in query q appear in document d than they do in the overall corpus, and a document authority score $p(d)$, which can be given by a link analysis score like PageRank [8].

$$p(d|q) = \frac{p(q|d)}{p(q)}p(d) \quad (2)$$

In practice, text IR and document ranking scores are often more sophisticated, but looking at the simplified ranking function described in equation 2 is useful as most ranking functions build upon this basic model.

Let's use the same Bayesian framework to derive a similar ranking function for personalized search. We define $p(d|q, u)$ to be the probability that document d can answer the query q for a given user u . We can use Bayes' law to write

$$p(d|q, u) = \frac{p(q|d, u)p(d|u)}{p(q|u)} \quad (3)$$

If we make the simplifying assumption that $p(q|d, u)/p(q|u) = p(q|d)/p(q)$ and rearrange terms, we get:

$$p(d|q, u) = \frac{p(q|d)}{p(q)}p(d|u) \quad (4)$$

which is a composition of an unpersonalized query-dependent text IR score $p(q|d)/p(q)$ as in equation 2, and a personalized query-independent document score $p(d|u)$. Indeed, this is the framework that underlies the body of research on personalized PageRank [6, 7, 2, 1, 4] that resulted in the Kaltix personalized search engine.

Now, we can use a similar framework for social search. In this case, instead of $p(d|q, u)$, we would like to determine $p(u_i|q, u_j)$, the probability that user u_i can answer question q from user u_j . We use Bayes' law once again to write

$$p(u_i|q, u_j) = \frac{p(q|u_i, u_j)p(u_i|u_j)}{p(q|u_j)} \quad (5)$$

We again make the simplifying assumption that $p(q|u_i, u_j)/p(q|u_j) = p(q|u_i)/p(q)$ and rearrange terms to get:

$$p(u_i|q, u_j) = \frac{p(q|u_i)}{p(q)} p(u_i|u_j) \quad (6)$$

Note that

$$p(u_i|q) = \frac{p(q|u_i)p(u_i)}{p(q)} \quad (7)$$

If we take $p(u_i)$ to be uniform, then we have

$$p(u_i|q) \propto \frac{p(q|u_i)}{p(q)} \quad (8)$$

Plugging this into equation 6, we get:

$$p(u_i|q, u_j) \propto p(u_i|q)p(u_i|u_j) \quad (9)$$

which is exactly $s(u_i, u_j, q)$ from equation 2 of [3].

Note that in both personalized search and social search as described above, we begin by suggesting that the asker is important. But then we make simplifying assumptions that ignore the asker in the query-dependent text IR component of the ranking function. In personalized search, we make the simplifying assumption that $p(q|d, u)/p(q|u) = p(q|d)/p(q)$, and in social search, we make the simplifying assumption that $p(q|u_i, u_j)/p(q|u_i) = p(q|u_i)/p(q)$.

In both cases, the results are nevertheless personalized, because the document authority score (or in the case of social search, user intimacy score) is conditioned on the asker. However, this idiosyncrasy should be noted as it presents an interesting area of research around developing efficient personalized and social query-dependent text IR scores.

Mathematics aside, it is interesting to note that while there are many conceptual similarities between personalized search and social search as described above, in practice they work well for very different sets of queries. Personalized search as described above works best for short, ambiguous queries like “IR” [9], while social search as described above works best for long, highly contextualized queries like “I am looking for good books about IR, because I want to build a search engine. I am new to the field, but have some basic computer science background. Do you have any suggestions?”

In both paradigms, the results are personalized, but they are personalized in different ways, and thus lend themselves to different classes of queries.

References

1. T. H. Haveliwala. Topic-Sensitive PageRank. In *WWW*, 2002.
2. T. H. Haveliwala, S. D. Kamvar, and G. Jeh. An Analytical Comparison of Approaches to Personalizing PageRank. *Stanford University Technical Report*, 2003.
3. D. Horowitz and S. D. Kamvar. The Anatomy of a Large-Scale Social Search Engine. In *WWW*, 2010.

4. G. Jeh and J. Widom. Scaling Personalized Web Search. In *WWW*, 2003.
5. S. D. Kamvar. *Numerical Algorithms for Personalized Search in Large-Scale Self-Organizing Networks*. Princeton University Press, 2010.
6. S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Exploting the block structure of the Web for computing PageRank. *Stanford University Technical Report*, 2003.
7. S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation Methods for Accelerating PageRank Computations. In *WWW*, 2003.
8. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. *Stanford University Technical Report*, 1998.
9. J. Teevan, S. Dumais, and E. Horvitz. Characterizing the value of personalized search. In *SIGIR*, 2007.