

# A Generative Model of Tagging and Search

**Harry Halpin**

h.halpin@ed.ac.uk

## Introduction

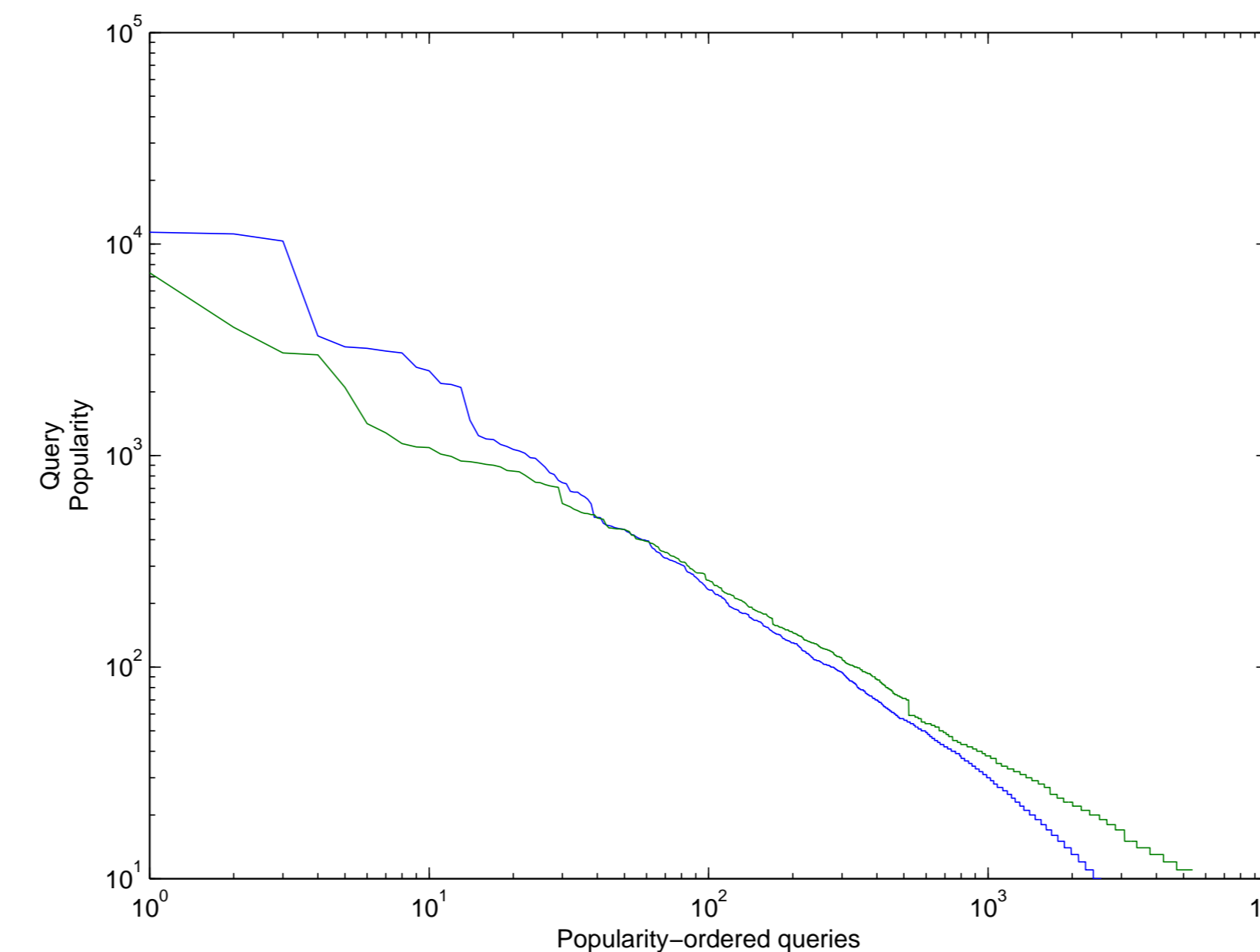
Is it possible that collaborative tagging and information retrieval can be united within a single theoretical framework? Although current models of information retrieval and tagging are separate, they can easily be incorporated into the principled *relevance model* framework, so that tags, search terms, and documents can all be thought of as samples from an underlying generative process.

## Web Science and Models

One of the major problems that Web Science needs to confront is the lack of theoretical models that combine the diverse areas that compose the Web. Take the two relatively well-defined and similar domains of collaborative tagging and Web search: both phenomenon involve users essentially labelling a resource with a set of natural language terms, with the former being done before the discovery of resources, and the other after. If these two phenomena are similar on the level of practice, on the level of theory their models seem wildly incompatible. Current generative models of tagging (Cattuto et al., 2006) propose that tagging can be modelled via a modified preferential attachment model with a feedback-based memory. In information retrieval, language models consider the probability that a query is a random sample from the list of relevant documents. Can these two divergent theories be unified?

## What are Relevance Models?

We believe that they can, in particular, be unified via the use of an extension of Lavrenko's *relevance models*, a theoretically elegant generative model of information retrieval (Lavrenko, 2009). Traditional language models had difficulty relating the query to the document in a simple mathematical model: relevance models cuts this Gordian knot by assuming that both the relevant document and query are samples from an unknown generative model, usually estimated via kernel-based density allocation. One can imagine that documents which do not even mention the same terms as the query being relevant to a query, as both have been generated by the same underlying process.



Power-law produced by tags (green) from del.icio.us and queries (blue) from Microsoft Live Search

## Sampling Relevance Models

The classical language-modeling approach to IR does not provide a natural mechanism to extend to tags and queries as *first-class citizens* like documents. However, a popular extension of the approach involves estimating a relevance-based model  $u_R$  in addition to the document-based model  $u_D$ , and comparing any resulting documents using information-theoretic measures that like cross-entropy to compare the relevance model to each document model retrieved. Since the  $u_r$  is an *underlying generative model*, we can imagine a sampling function that generates relevant samples, called  $s$ . So a tag  $t = s(u_R)$  in addition to  $q = s(u_R)$  and of course documents  $d = s(u_R)$  are all relevant samples produced by the *same* underlying process.

Let  $R = r_1 \dots r_k$  be the set of  $k$  relevant samples from the underlying process, so that  $r$  may be either tags  $t$ , documents  $d$ , and queries  $q$ . One way of constructing a language model of  $R$  is to average the language models of each sample in the set:

$$u_{R,avg}(w) = \frac{1}{k} \sum_{i=1}^k u_{r_i}(w) = \frac{1}{k} \sum_{i=1}^k \frac{n(w, r_i)}{|r_i|} \quad (1)$$

Here  $n(w, r_i)$  is the number of times the term  $w$  occurs in the  $i$ 'th relevant sample (tag, query, or document), and  $|r_i|$  is the length of that sample.

## Conclusions

On the surface, both information retrieval and tagging seem to be very similar, and we argue that this similarity can serve as the foundation for a generative model that unites both information retrieval and tagging systems theoretically. This violates a number of traditional assumptions about tagging. This violates the assumption that tags somehow are *special*. This is in-line with previous work that shows that "imitation-based" feedback is what causes the emergence of power-laws in collaborative tagging (Bollen and Halpin, 2009). Instead, it seems to be that the convergence of tags to power-law distributions is for the same reason as queries and documents, i.e. because both models are based on power-law distributions of natural language terms.

## Future Work

Current generative models of tagging propose that the model underlying tagging is a random walk over a labelled conceptual graph (Cattuto et al., 2009). In information retrieval, language models consider the probability that a query is a random sample from the list of relevant documents. Can these two divergent theories be unified? One can combine random walks with relevance models by characterizing the underlying generative relevance model not as a single generative process, but as a dynamic group of task-based processes that change over time. Walks (random or not) in this network are equivalent to mixtures of relevance models, allowing the underlying process to be updated over time.

## References

- C. Cattuto, L. Vittorio, and L. Pietronero (2006). Semiotic dynamics and collaborative tagging. PNAS, 104(5):1461–1464.
- C. Cattuto, D. Benz, A. Hotho, and G. Stumme (2009). Semantic Grounding of Tag Relatedness in Social Bookmarking Systems, International Semantic Web Conference
- D. Bollen and H. Halpin (2009). An Experimental Analysis of Suggestions in Collaborative Tagging. Web Intelligence Conference.
- V. Lavrenko (2009). A Generative Theory of Relevance. Springer.

