

Collaborative Biomedical Information Retrieval

Paul Thompson

Computational Genetics Laboratory
Department of Genetics
Geisel Medical School at Dartmouth
Lebanon, New Hampshire, U.S.A.
1 603 646-8747

Paul.Thompson@dartmouth.edu

Suzanne Thompson

Computational Genetics Laboratory
Department of Genetics
Geisel Medical School at Dartmouth
Lebanon, New Hampshire, U.S.A.
1 603 646-8739

Suzanne.Thompson@dartmouth.edu

Abstract

In the context of two related NIH projects supporting scientific collaboration we seek to implement an environment for collaborative information retrieval and analysis based on utility theory.

Introduction

Academic research on information retrieval has long followed a model of attempting to rank documents in response to an individual searcher's single query. Rankings have usually been determined based on a measure of similarity between a representation of the user's query with a representation of each of the retrieved documents, or on a calculation of the probability of relevance for each document with respect to the query. This model falls short of the needs of optimization for collaborative scientific research in two main ways. First, many scientists work in a collaborative environment. Second, it is the value of information which is to be optimized, not its relevance. Relevance, one of the fundamental notions in information retrieval theory, is generally considered to be a binary variable. A document is, or is not, relevant with respect to a query. In order to measure the value of information, it is important to consider the utility of information, rather than only the relevance of information. Utility-theoretic information retrieval has been relatively ignored as a field of research, but initial work in this area was done by Kraft (1973), Kochen (1974), Bookstein and Swanson (1975), and Cooper and Maron (1978), among others. We propose to build on Cooper and Maron's probabilistic and utility-theoretic approach to design a cost function to optimize a model for collaborative scientific research.

The Eagle-i and VIVO Projects

We will develop our approach to collaborative information retrieval in the context of the NIH Eagle-i and VIVO projects. There is much inefficiency and duplication in life sciences research because researchers are unaware of hidden resources available at other institutions or are unable to identify researchers working on similar projects. The Eagle-i project, led by Harvard Medical School, created a semantic repository and search engine that identifies hidden resources in participating institutions (eagle-i, 2012, Vasilevsky et al. 2012,). The VIVO project, led by the University of Florida developed a software infrastructure that enables national networking of scientists (VIVO 2012, Wolski et al. 2012). Although our focus in this position paper is on collaborative, utility-theoretic information retrieval, it is our intention to address the wider scope of the Eagle-i and VIVO projects' collaborative information seeking including user-centered design and Semantic Web technologies for collaborative querying and analysis in the life sciences (Chueng et al. 2008) in future work. The Eagle-i and VIVO projects are based on Semantic Web technologies and Linked Data (Bizer, Heath, and Berners-Lee, to appear)

Related Research

Related research has been conducted in information retrieval as well as in fields, such as machine learning and neural networks. Mateescu et al. (2002) compare work on relevance feedback in information retrieval to related techniques in these fields. In addition to the utility-theoretic approach of Maron and Cooper described below, our research will consider models from machine learning, such as reinforcement learning (Sutton and Barto 1998).

Although current research in probabilistic document retrieval, dominated by the perspective of the language modeling approach, is less able to incorporate relevance feedback, earlier Bayesian probabilistic models are better able to do so. The unified probabilistic model developed by Robertson, Maron, and Cooper (1982) did not include relevance feedback. However, this model was later extended to include relevance feedback (Thompson 1986). Laboratory studies of relevance feedback have generally shown it to result in improved retrieval, but it was not used in operational systems prior to the 1990s. When the Westlaw retrieval system introduced ranked retrieval to large online, commercial retrieval, relevance feedback was considered, but rejected. Lexis-Nexis, on the other hand, did include relevance feedback in its competing Freestyle ranked retrieval mode, which came out a year later. This relevance feedback capability, as was also later the case with several Web search engines, was based only on the searcher indicating that one of the retrieved documents was particularly relevant. Earlier laboratory research was generally based on all retrieved documents being judged for relevance. Relevance feedback was not adopted beyond this limited extent, because of the perceived difficulty in asking searchers to provide relevance feedback. The development of systems on the Web changed things dramatically. Implicit relevance judgments were taken based on clickstream mining (Cooley 2000). Recommender system technology, also referred to as collaborative filtering, has made use of relevance feedback, as well (Konstan et al. 1997). Additionally, some researchers have gone beyond the traditional focus of relevance feedback for a single user and have considered collaborative relevance feedback, e.g., (Wilbur 1998).

Cooper and Maron's Model of Probabilistic and Utility-Theoretic Information Retrieval

The Cooper and Maron model of probabilistic and utility-theoretic retrieval is based on classic decision theory as developed by Savage (1954) and others. Cooper and Maron describe a probability-utility space in which in addition to a probability measure there is a series of utility measures, one for each point in the space. For each future use of the system, there is a utility measure associated with whether or not the document was indexed with a given index term. As with probabilistic retrieval models, our concern is to define a cost function and a way to optimize it, given feedback from users on the value of retrieved information.

Collaborative Utility-Theoretic Information Retrieval: Metrics and Evaluation

Probabilistic information retrieval, though usually discussed in the context of a single searcher rather than collaborative searchers, is a well-developed area of research

going back to 1960s (Maron and Kuhns 1960). Nevertheless, most probabilistic information retrieval systems do not treat the probabilities that are generated to rank documents as serious probabilities. Rather these "probabilities", or scores, are seen as being useful in ranking documents with respect to each other. Utility-theoretic information retrieval has also been discussed, at least as far back as the 1970s (Kraft 1973, Kochen 1974, Bookstein and Swanson 1975, Cooper and Maron 1978), though much less extensively than probabilistic information retrieval. Utility-theoretic information retrieval has also primarily addressed the problem of retrieval of a document for an individual searcher, though Cooper (1978) discusses total utility, as the "... total utility experienced collectively as a result of the assignment [of an index term]." Cooper advocated that human indexers estimate this total utility in order to binarily assign an index term to a document, though in later work with Maron (Cooper and Maron 1978) utility-weighted index term assignment was proposed.

Since the 1970s the trend has increasingly been away from human indexing towards automatic indexing. While probabilistic information retrieval is arguably the dominant paradigm in academic information retrieval research, the current leading approach, the language modeling approach, does not treat relevance as central. Rather a system calculates the probability that the language model which generated a document to be retrieved could also have generated the user's query and uses this probability to rank documents (Ponte 1998). Even in settings where manual indexing, or classification, still takes place, e.g., in the biomedical or legal domains, indexing is binary. Often authors self-index their documents by selecting terms from a controlled vocabulary.

Applied to ranked retrieval, relevance feedback is a technique which has been shown in laboratory settings to lead to large improvements in retrieval effectiveness (Buckley and Robertson 2008). In the past using relevance feedback meant asking a user which of a set of retrieved documents were relevant and which non-relevant. Despite the laboratory success of the technique, relevance feedback was seen as impractical because real users would not be willing to provide relevance judgments. However, on the Web new techniques of implicit relevance feedback were developed with which to estimate relevance [17].

Projects such as VIVO, Eagle-i, and other collaborative environments for scientists provide opportunities to measure explicit and implicit feedback on relevance and on the utilities of indexed resources. As an example, consider a researcher who does not have access to a supercomputer at his or her university, nor to the software needed to perform some type of analysis on his or her research data. The Eagle-i project manually indexes resources in university labs, such as supercomputers and software. Depending on the nature of the computing resource or its associated software, a manual indexer might make an estimate of the utility of the resource to a researcher who found the resource using

the Eagle-i search engine. VIVO tracks the publications and grant awards of researchers at a given university (VIVO 2012). Using these two systems together, it will be possible to build a utility-theoretic retrieval system and to evaluate the accuracy of its performance, e.g., the utility of the retrieved resources could be inferred by considering citations in proposals which led to awards and citations in publications.

Acknowledgments

We acknowledge support of the NIH NCCR project *Networking Research Resources Across America*, 1U24RR029825-01.

References

- Bizer, C.; Heath, T.; and Berners-Lee, T. [to appear]. Linked Data – The Story So Far. In Heath, T.; Hepp, M.; and Bizer, C. eds. Special Issue on Linked Data, *International Journal on Semantic Web and Information Systems*
- Bookstein, A., and Swanson, D.R. 1975. A decision theoretic foundation for indexing *Journal of the American Society for Information Science* 26:45-50.
- Buckley, C. and Robertson, S. 2008. Relevance feedback track overview: TREC 2008 *Proceedings of the 2008 Text REtrieval Conference*
- Cheung, K.-H.; Frost, H. R.; Marshall, M. S.; Prud'hommeaux, E.; Samwald, M.; Zhao, J.; and Paschke, A. 2008. A journey to Semantic Web query federation in the life sciences *Semantic Web Applications and Tools for Life Sciences*.
- Cooley, R. W. 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. Ph.D. diss, Department of Computer Science, University of Minnesota.
- Cooper, W. S. 1978. Indexing documents by Gedanken experimentation *Journal of the American Society for Information Science*, 29(3):107-119.
- Cooper, W.S. and Maron, M. E. 1978. Foundations of Probabilistic and Utility-Theoretic Information Indexing *Journal of the Association for Computing Machinery* 25(1):67-80.
- eagle-i 2012. <https://www.eagle-i.net>, accessed 26 May 2012
- Kochen, M. 1974. *Principles of Information Retrieval* New York: Wiley.
- Konstan, J. A.; Miller, B. M.; Maltz, D.; Herlocker, J. L. ; Gordon, L. R. ; and Riedl, J. 1997. GroupLens: Applying collaborative filtering to Usenet news *Communications of the ACM* 40(3):77-87.
- Kraft, D.A. 1973. A decision theory view of the information retrieval situation: An operations research approach *Journal of the American Society for Information Science* 24:368-376.
- Maron, M. E. and Kuhns, J. L. 1960. On relevance, probabilistic indexing, and information retrieval *Journal of the ACM* 7(3):216-244.
- Mateescu, G.; Sosonkina, M.; and Thompson, P. 2002. A New Model for Probabilistic Information Retrieval on the Web *Second SIAM International Conference on Data Mining (SDM 2002) Workshop on Web Analytics*, 17-27.
- Ponte, J. 1998. A language modeling approach to information retrieval PhD. diss, Department of Computer Science University of Massachusetts, Amherst.
- Robertson, S. E.; Maron, M. E.; Cooper, W. S. 1982. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1(1):1-21.
- Savage, L. J. 1954. *The Foundation of Statistics* New York: Wiley.
- Sutton, R. S. and Barto, A. G. 1998. *Reinforcement Learning: An Introduction* Cambridge, Massachusetts: MIT Press.
- Thompson, P. Subjective probability, combination of expert opinion, and probabilistic approaches to information retrieval" Ph.D. diss, School of Library and information Studies, University of California, Berkeley, 1986.
- Vasilevsky, N.; Johnson, T.; Corday, K.; Torniai, C.; Brush, M.; Segerdell, E.; Wilson, M.; Shaffer, C.; Robinson, D.; and Haendel, M. 2012. Research resources: curating the new eagle-i discovery system. *Database* 2012.
- VIVO 2012. vivoweb.org accessed 26 May 2012
- White, R. W.; Ruthven, I.; Jose, J. M. 2005. A Study of Factors Affecting the Utility of Implicit Relevance Feedback *Proceedings of the 28th Annual ACM SIGIR Conference (SIGIR 2005)*.
- Wilbur, W. 1998. The knowledge in multiple human relevance judgments *ACM Transactions on Information Systems* 16, (2):101-126.
- Wolski, M.; Richardson, J.; Fallu, M.; Robollo, R.; Morris, J. 2012. Developing the Discovery Layer in the University Research e-Infrastructure *15th World Multi-Conference on Systemics, Cybernetics and Informatics, Proceedings. Vol. III*.

Acknowledgments

We acknowledge support of the NIH NCRR project *Networking Research Resources Across America*, 1U24RR029825-01.