

A Survey of Statistical Challenges in Web Search

David PURDY, *University of California, Berkeley, USA*, E-mail: dpurdy@stat.berkeley.edu

KEY WORDS: Dynamic Graphical Models, World Wide Web, Information Retrieval, Machine Learning

MATHEMATICAL SUBJECT CLASSIFICATION: 62-01

Abstract:

We will discuss several areas where statistical methods are important in improving web search, and where web search problems remain unsolved problems in statistics. In particular, sampling, dynamic graphical models, classification, unsupervised learning, cross-validation, time series analysis, and many other topics arise in the context of providing extensive search facilities. Due to the evolving nature of the web and growing understanding by those who build, use, or manipulate search engines, some of these statistical problems will exist for some time to come.

Special attention is given to the topic of useful dynamic graphical models and sampling methodologies. Dynamic graphs, in which vertices and edges are removed and added over time, challenge current techniques for representation and inference. Nonetheless, an arms race exists between those attempting inference on such graphs (e.g. search engines) and those who manipulate such graphs (e.g. spammers). We will discuss some of the types of inference that have been done and those that could become more prevalent. We believe this topic also has relevance to analyzing contours in videos, analyzing network stability (e.g. computer network intrusion, power network reliability), disease transmission networks, and other domains.

This presentation is accessible for a general audience and aims to appeal to graduate students in particular. The discussion draws upon the presenter's experience in research at a major search engine in the U.S. and as a consultant to others working in search.

References

- [1] Bar-Yossef, Z., Broder, A.Z., Kumar, R. and Tomkins, A. (2004). Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 328-337.
- [2] Cho, J. and Roy, S. (2004). Impact of Search Engines on Page Popularity, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 20-29.
- [3] Fetterly, D., Manasse, M., Najork, M. and Wiener, J.L. (2003). A Large-Scale Study of the Evolution of Web Pages, in *Twelfth International World Wide Web Conference, Budapest, Hungary, ACM Press*.
- [4] Gruhl, D., and Guha, R. (2004). Information Diffusion through Blogspace, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 491-501.
- [5] Guillaume, J.-L., Latapy, M. and Viennot, L. (2002). Efficient and Simple Encodings for the Web Graph, in *Eleventh International World Wide Web Conference, Honolulu, Hawaii, ACM Press*.
- [6] Henzinger, M.R., Motwani, R. and Silverstein, C. (2002). Challenges in Web Search Engines, in *ACM 18th International Joint Conference on Artificial Intelligence, ACM Press*, 1573-1579.
- [7] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. (2003). On the Bursty Evolution of Blogspace, in *Twelfth International World Wide Web Conference, Budapest, Hungary, ACM Press*.
- [8] McCurley, N.E.K.S. and Tomlin, J.A. (2004). Ranking the Web Frontier, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 309-318.
- [9] Ntoulas, A., Cho, J. and Olston, C. (2004). What's New on the Web? The Evolution of the Web from a Search Engine Perspective, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 1-12.

- [10] Pitkow, J. and Pirolli, P. (1997). Life, Death, and Lawfulness on the Electronic Frontier, in *ACM Special Interest Group on Computer-Human Interaction, Atlanta, GA, USA, ACM Press*, 383-390.
- [11] Yu, P.S., Li, X. and Liu, B. (2004). On the Temporal Dimension of Search, in *Thirteenth International World Wide Web Conference, New York, NY, ACM Press*, 448-449.