"Scalability and Relevance in an Internet-scale Persistent Search System"

Erich R. Schmidt

Abstract

Current persistent search systems come in two flavors: (1) single- or multi- source, email/RSS notifiers (e.g. New York Times, Pubsub.com, Google News Alerts), which provide fresh publications but cover a limited publication base; (2) persistent queries on traditional search engines (e.g. Google Web Alerts), which are limited by the search engines' low refresh rates. Besides coverage and freshness, a very important issue in persistent search systems is the quality and relevance of results; existing citation-based authority methods (e.g. PageRank) are strongly biased against new pages, and hence are not well suited to persistent search.

To address coverage and freshness, we propose Distributed Persistent Search, a new architecture based on a distributed publish-subscribe framework that achieves linear improvement in publication processing and notification routing, as a function of the number of servers used. To address citation-based authority for persistent search, we propose SiteRank, a new ranking mechanism that handles new publications well and also dramatically reduces the communication and computation costs.