Full-text federated search in peer-to-peer networks

Jie Lu
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
jielu@andrew.cmu.edu

Thesis available at: *http://www.cs.cmu.edu/~jielu/thesis_jielu.pdf*

Peer-to-peer (P2P) networks integrate autonomous computing resources without requiring a central coordinating authority, which makes them a potentially robust and scalable model for providing federated search capability to large-scale networks of text digital libraries. However, P2P networks have so far mostly used simple search techniques based on document names or controlled-vocabulary terms, and provided very limited support for full-text search of document contents.

This dissertation provides solutions to full-text federated search with relevance-based document ranking within an integrated framework of P2P network overlay, search, and evolution models. Previous notions of P2P network architectures are extended to define a network overlay model with desired content distribution and navigability. Existing approaches to federated search are adapted, and new methods are developed for resource representation, resource selection, and result merging in a network search model according to the unique characteristics of P2P networks. Furthermore, autonomous and decentralized algorithms to evolve the network topology into one with desired search-enhancing properties are proposed in a network evolution model to facilitate effective and efficient full-text federated search in dynamic environments.

To demonstrate that the proposed solutions are both effective and practical, two P2P testbeds consisting of thousands of real-content text digital libraries and hundreds of thousands of automatically generated queries are developed. Evaluation using these testbeds provides strong empirical evidence that the approaches proposed in this dissertation provide a better combination of accuracy, efficiency and robustness than more common alternatives.