Towards Peer-to-Peer Web Search (Extended Abstract)

Gerhard Weikum, Holger Bast, Geoffrey Canright, David Hales,
Christian Schindelhauer, Peter Triantafillou *

Contact author: Gerhard Weikum, Max-Planck Institute for Computer Science,
Stuhlsatzenhausweg 85, Saarbrücken, Germany, Email: weikum@mpi-inf.mpg.de

The peer-to-peer (P2P) computing paradigm is an intriguing alternative to Google-style search engines for querying and ranking Web content. In a network with many thousands or millions of peers the storage and access load requirements per peer are much lighter than for a centralized Google-like server farm; thus more powerful techniques from information retrieval, statistical learning, computational linguistics, and ontological reasoning can be employed on each peer's local search engine for boosting the quality of search results [1, 2, 10–12, 26]. In addition, peers can dynamically collaborate on advanced and particularly difficult queries. Moroever, a peer-to-peer setting is ideally suited to capture local user behavior, like query logs and click streams, and disseminate and aggregate this information in the network, at the discretion of the corresponding user, in order to incorporate richer cognitive models.

The DELIS project is aiming at a P2P system where each peer has a full-fledged Web search engine, including a crawler and an index manager. The crawler may be thematically focused or crawl results may be postprocessed so that the local index contents reflects the corresponding user's interest profile. With such a highly specialized and personalized "power search engine" most queries should be executed locally, but once in a while the user may not be satisfied with the local results and would then want to contact other peers. A "good" peer to which the user's query should be forwarded would have thematically relevant index contents, which could be measured by statistical notions of similarity between peers [3, 4]. Both query routing and the formation of "statistically semantic" overlay networks could greatly benefit from collective human inputs in addition to standard statistics about terms, links, etc.: knowing the bookmarks and query logs of thousands of users would be a great resource to build on [8, 14]. Note that this notion of Web search includes ranked retrieval and thus is fundamentally much more difficult than Gnutella-style file sharing or simple key lookups via distributed hash tables (DHTs) [24]. Further note that,

* The authors are with the Max-Planck Institute for Computer Science in Saarbrücken, Telenor in Oslo, the University of Bologna, the Heinz-Nixdorf Institute in Paderborn, and the University of Patras. The work presented in this paper is partially supported by the EU within the 6th Framework Programme under contract 001907 "Dynamically Evolving, Large Scale Information Systems" (DELIS).

although query routing in P2P Web search resembles earlier work on metasearch engines and distributed information retrieval [17, 18], it is much more challenging because of the large scale and the high dynamics of the envisioned P2P system with thousands or millions of computers and users. Finally, the P2P setting poses great challenges also for network-efficient top-k query processing [25, 16], decentralized and other advanced forms of link analyis [6, 7, 13, 14, 21], distributed gathering and dissemination of statistics about data, load, and user behavior [16, 19, 20], and the creation of self-organizing overlay networks [9, 15, 22, 23, 27].

A system architecture for the envisioned solution is currently prototyped, as an experimental platform within the DELIS project, under the name *Minerva* [5]. This system has all the characteristics and poses the challenges of a complex system. The autonomy of peers and the diversity of different behavioral patterns can be understood only by analyzing and controlling the system at different levels, ranging from the underlying physical network and the virtual overlay network layers to the level of intelligent search, query routing, and collaboration strategies of the individual peers. For cost-efficient solutions it is crucial to consider benefit and cost factors at all levels. Finally, a deep understanding mandates studying such complex systems at different scales in terms of time and space, for example, the short-term interactions of a peer with its immediate neighborhood, triggered by query routing and query execution, on one hand, and the long-term, long-range evolution of the entire system, to organize itself into effective and robust semantic overlay structures, on the other hand.

References

1. Holger Bast, Debapriyo Majumdar. Understanding Spectral Retrieval via the Synonymy Graph, ACM SIGIR Conf. on R&D in Information Retrieval, Salvador, Brazil, 2005.
2. Holger Bast, Ingmar Weber: A Framework for Ranked Retrieval based on Rank Aggregation, Int. Workshop on Web Information Retrieval and Integration, Tokyo, Japan, 2005.
3. Matthias Bender, Sebastian Michel, Gerhard Weikum, Christian Zimmer: Bookmark-driven Query Routing in Peer-to-Peer Web Search, ACM SIGIR Workshop on Peer-to-Peer Information Retrieval, Sheffield, UK, 2004
4. Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: Improving Collection Selection with Overlap Awareness in P2P Search Engines, ACM SIGIR Int. Conf. on R&D in Information Retrieval, Salvador, Brazil, 2005.
5. Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: Minerva: Collaborative P2P Search, Demo Paper, Int. Conf. on Very Large Data Bases (VLDB), Trondheim, Norway, 2005.
6. Klaus Berberich, Michalis Vazirgiannis, Gerhard Weikum: T-Rank: Time-aware Authority Ranking, Int. Workshop on Algorithms and Models for the Web Graph, Rome, Italy, 2004.
7. Geoffrey Canright and Kenth Engo-Monsen. Roles in Networks. Science of Computer Programming 53: 195–214, 2004.

8. Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma: Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering 15(4), 2003
9. David Hales: From Selfish Nodes to Cooperative Networks – Emergent Link-based Incentives in Peer-to-Peer Networks. IEEE Conf. on Peer-to-Peer Computing, Zurich, Switzerland, 2004
10. Thomas Hofmann: Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning 42(1/2): 177–196, 2001
11. Georgiana Ifrim, Martin Theobald, Gerhard Weikum: Learning Word-to-Concept Mappings for Automatic Text Classification, Int. Workshop on Learning in Web Search, Bonn, Germany, 2005.
12. Amy N. Langville, Carl D. Meyer: A Survey of Eigenvector Methods of Web Information Retrieval. The SIAM Review, 47(1):135-161, 2005.
13. Amy N. Langville, Carl D. Meyer: Deeper Inside PageRank. Internet Mathematics, 1(3):335-400, 2004.
14. Julia Luxenburger, Gerhard Weikum: Query-log based Authority Analysis for Web Information Search, Int. Conf. on Web Information System Engineering (WISE), Brisbane, Australia, 2004
15. Peter Mahlmann, Christian Schindelhauer: Peer-to-Peer Networks based on Random Transformations of Connected Regular Undirected Graphs, ACM Symp. on Parallelism in Algorithms and Architectures (SPAA), Las Vegas, 2005
16. Sebastian Michel, Peter Triantafillou, Gerhard Weikum: KLEE: Internet-scale Distributed Top-k Query Algorithms, Int. Conf. on Very Large Data Bases (VLDB), Trondheim, Norway, 2005.
17. Weiyi Meng, Clement T. Yu, King-Lup Liu: Building efficient and effective metasearch engines. ACM Computing Surveys 34(1): 48-89, 2002.
18. Henrik Nottelmann, Norbert Fuhr: Evaluating different methods of estimating retrieval quality for resource selection. ACM SIGIR Conf. on R&D in Information Retrieval, Toronto, Canada, 2003.
19. Nikos Ntarmos, Peter Triantafillou: AESOP: Altruism-Endowed Self Organizing Peers, 3rd Int. Workshop on Databases, Information Systems and Peer-to-Peer Computing, Toronto, Canada, 2004.
20. Nikos Ntarmos, Peter Triantafillou: SeAl: Managing Accesses and Data in Peer-to-Peer Data Sharing Networks, 4th IEEE Conf. in Peer-to-Peer Computing, Zurich, Switzerland, 2004.
21. Josiane Xavier Parreira, Gerhard Weikum: JXP: Global Authority Scores in a P2P Network, 8th Int. Workshop on Web and Databases (WebDB), Baltimore, USA, 2005.
22. Josiane Xavier Parreira, Sebastian Michel, Gerhard Weikum: p2pDating: Real Life Inspired Semantic Overlay Networks for Web Search, ACM SIGIR Workshop on Heterogeneous and Distributed Information Retrieval, Salvador, Brazil, 2005.
23. Christian Schindelhauer, Gunnar Schomaker: Weighted Consistent Hashing, ACM Symp. on Parallelism in Algorithms and Architectures (SPAA), Las Vegas, USA, 2005
24. Ion Stoica, Robert Morris, David Liben-Nowell, David R. Karger, M. Frans Kaashoek, Frank Dabek, Hari Balakrishnan: Chord: a scalable peer-to-peer lookup protocol for internet applications. IEEE/ACM Transactions on Networking 11(1): 17-32, 2003
25. Martin Theobald, Gerhard Weikum, Ralf Schenkel: Top-k Query Evaluation with Probabilistic Guarantees, Int. Conf. on Very Large Data Bases (VLDB), Toronto, Canada, 2004
26. Martin Theobald, Ralf Schenkel, Gerhard Weikum: Efficient and Self-Tuning Incremental Query Expansion for Top-k Query Processing, ACM SIGIR Conf. on R&D in Information Retrieval, Salvador, Brazil, 2005.
27. Peter Triantafillou, Chryssani Xiruhaki, Manolis Koubarakis, Nikos Ntarmos: Towards High Performance Peer-to-Peer Content and Resource Sharing Systems, 1st Int. Conf. on Innovative Data Systems Research (CIDR), Asilomar, USA, 2003.