P2P Web Search with MINERVA: How do you want to search tomorrow?

Sebastian Michel[1], Matthias Bender[1], Peter Triantafillou[2], Gerhard Weikum[1], Christian Zimmer[1]

[1]Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany
{smichel, mbender, weikum, czimmer}@mpi-inf.mpg.de

[2]University of Patras
Rio, 26500, Greece
peter@ceid.upatras.gr

Abstract. MINERVA[1] is a novel approach towards P2P Web search that connects an a-priori unlimited number of peers, each of which maintains a personal local database and a local search facility. Each peer posts a small amount of metadata to a physically distributed directory layered on top of a DHT-based overlay network that is used to efficiently select promising peers from across the peer population that can best locally execute a query. This paper proposes a live demonstration of MINERVA, showcasing the full information lifecycle: crawling web pages, disseminating metadata to a distributed directory, and executing queries online. We additionally invite all visitors to instantly join the network by executing a small piece of software.

1 Introduction

In recent years, research on the Peer-to-Peer (P2P) networks has been receiving increasing attention [9, 8] and is rapidly making its way into distributed data management and information retrieval (e.g., [5, 11]) due to its potential to handle huge amounts of data in a highly distributed, self-organizing way. These characteristics offer enormous potential benefits for search capabilities powerful in terms of scalability, efficiency, and resilience to failures and dynamics. Additionally, a P2P web search engine can also facilitate pluralism in informing users about internet content, which is crucial in order to preclude the formation of information-resource monopolies and the biased visibility of content from economically powerful sources. Last but not least, such a search engine can potentially benefit from the intellectual input (e.g., bookmarks, query logs, etc.) of a large user community in order to improve a user's search experience.

The crucial challenge in developing successful P2P Web search engines is based on reconciling the following conflicting goals: on the one hand, delivering high quality results with respect to precision / recall, and, on the other hand, providing unlimited scalability in the presence of a very huge peer population and the very large amounts of data that must be communicated in order to meet the first goal. We put forward MINERVA whose architecture, design, and implementation satisfies these conflicting goals.

As an application scenario, consider a community of scientists interested in diverse fields, such as computer science, physics, and bioinformatics, as well as various hobbies, such as cooking and climbing. Each scientist acts as a completely autonomous peer that has gathered high-quality information by performing focused crawls of thematically relevant portions of the web and built an according

local index. By sharing portions of their local indexes at their own discretion, the peers form a collaboration that can be used by our innovative architecture to satisfy highly specific information needs. Queries posed by a scientist are executed locally first and, if the results are not satisfactory, the query is forwarded to carefully chosen, thematically related peers.

The demonstration aims at illustrating the complete querying process in a live demo, as we invite all visitors to join our network instantly by starting a small piece of software from a local web server: we have prepared collections representing peers that have crawled thematically focused portions of the web. Alternatively, visitors can do a live crawl originating from arbitrary starting points using the BINGO! crawler [3]. Peers can join the DHT-based overlay simply by connecting to one already existing peer. After joining the P2P network, every peer disseminates its metadata to the physically distributed directory and can subsequently pose arbitrary keyword queries. The system identifies promising peers within the network based on the peers' metadata. Carefully selected peers execute the queries leveraging their local indexes and forward their local results to the query initiator, where these local results are combined to form one global result set.

2 System Design

MINERVA assumes an architecture of a P2P Web search federation as follows. Each peer is fully autonomous and has its own local search engine and has a local index that can be built from the peer's own crawls or imported from external sources and tailored to the user's thematic interest profile. Peers are willing to share metadata about their local indexes (or specific fragments of local indexes) by publishing it into a P2P network. This conceptually global but physically distributed directory, which is layered on top of a Chord[10]-style Distributed Hash Table (DHT), contains compact statistics and quality-of-service information. For failure resilience and availability, the responsibility for a term is shared and replicated across multiple peers. Notice that, unlike [6], we use the DHT to partition the term space, such that every peer is responsible for the *metadata* of a randomized subset of terms within the global directory. We do *not* distribute documents or index lists across the directory.

Query processing works as follows. In a preliminary step, every peer publishes statistical metadata (*Posts*) about a subset of terms in its local index to the directory. A hash function is applied to the term in order to determine the peer currently responsible for for this term. This peer stores all Posts for this term from across the directory in a *PeerList*. Posts contain contact information about the publishing peer together with statistics to calculate IR-style relevance measures for a term (e.g., the size of the inverted list for the term, the maximum average score among the term's inverted list entries, or some other statistical measure) and other information, e.g., regarding quality-of-service. The query initiator collects the PeerLists for all query terms from the distributed directory and combines this information to find the most promising peers for the current query. This step is refered to as *query routing*.

Query routing has been a research issue for many years [4, 7], but typically focuses on disjoint data sets. A number of these strategies have been evaluated in previous work [2] using MINERVA. However, naturally, the peers' data collections often highly overlap, as popular documents are highly crawled. We have de-

veloped strategies to combine overlap estimation with the available score/ranking information into an overall quality-novelty measure that can boost the effectiveness of query routing [1] in such an environment.

3 Demonstration

Our demonstration aims at illustrating the whole functionality of our system as well as its ease of use in a live demo using three notebook PCs. To make this a true P2P demo, we additionally invite all visitors to join our network instantly with their notebooks. After physically connecting to our ad-hoc LAN (cable and wireless; using network equipment we bring along) and automatically obtaining an IP address, visitors can browse a local web page, hosted on one of our notebooks, to instantly start the prototype using Java Web Start technology (Java 1.4+ required).

To ease live deployment, we have prepared thematically focused collections using the BINGO [3] crawler and stored them in a local database. Alternatively, provided an outbound network connection, visitors can perform a live crawl originating from arbitrary starting points using BINGO.

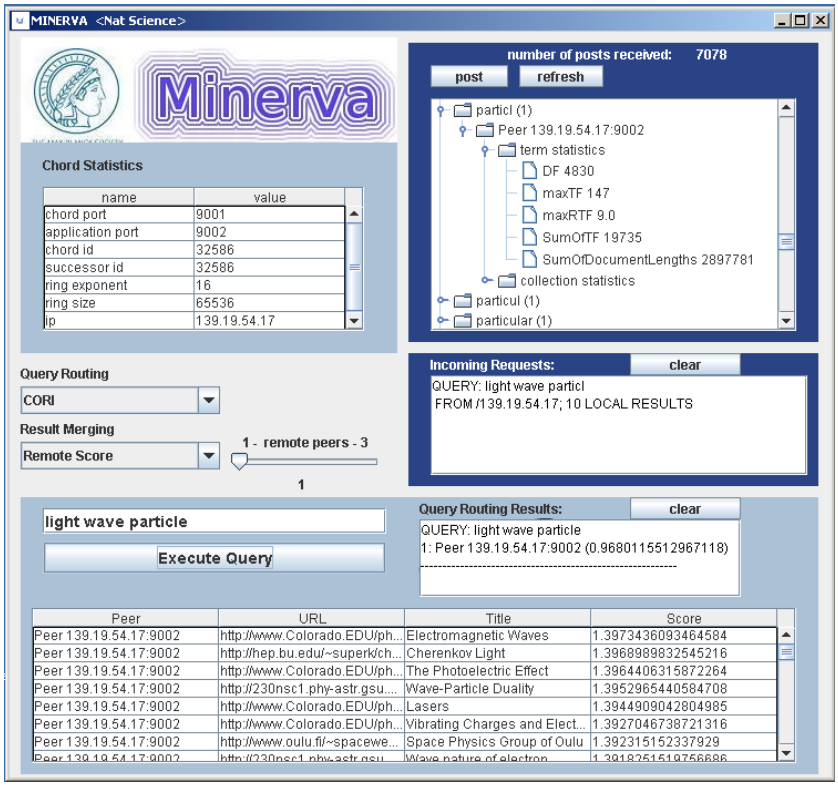

Fig. 1. MINERVA GUI

After joining the DHT-style directory, users can inspect the state of the DHT network (Fig 1, upper left part). Next, peers publish statistical metadata about their local indexes to the directory. This metadata is subsequently used to identify promising peers for particular queries. Peers can inspect the metadata they received from remote peers (Fig 1, upper right part), as every peer maintains a random subset of the directory.

Arbitrary keyword queries can be entered into a form field, just like in one of today's popular web search engines (Fig 1, lower part). The metadata is used to identify a tuneable number of promising remote peers for a query using query routing strategies such as CORI[4]. Users can instantly inspect the resulting peer ranking (Fig 1, Query Routing Results). The query is sent to these selected peers who indicate this fact in real-time (Fig 1, Incoming Requests). The user can also interactively validate this decision by inspecting the peers' metadata.

The results obtained from the remote peers are merged into one global result list by the peer initiating the query and is presented to the user in form of a result list indicating the URL, the page title, the origin peer, and the document score (Fig 1, lower part). Cached copies of the documents in the prepared collections have also been stored to the database, so that the relevance of a document to a query can instantly be validated by the user.

References

1. M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap awareness in p2p search engines. In *SIGIR*, 2005.
2. M. Bender, S. Michel, G. Weikum, and C. Zimmer. The minerva project: Database selection in the context of p2p search. In *BTW*, pages 125–144, 2005.
3. Bookmark-Induced Gathering of Information with Adaptive Classification into Personalized Ontologies. http://www.mpi-sb.mpg.de/units/ag5/software/bingo/.
4. J. Callan. Distributed information retrieval. *Advances in information retrieval, Kluwer Academic Publishers.*, pages 127–150, 2000.
5. R. Huebsch, J. M. Hellerstein, N. Lanham, B. T. Loo, S. Shenker, and I. Stoica. Querying the internet with pier. In *VLDB*, pages 321–332, 2003.
6. J. Li, B. Loo, J. Hellerstein, F. Kaashoek, D. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *In 2nd International Workshop on Peer-to-Peer Systems (IPTPS)*, 2003.
7. H. Nottelmann and N. Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. In *SIGIR2003*, pages 290–297. ACM Press, 2003.
8. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *SIGCOMM*, pages 161–172. ACM Press, 2001.
9. A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *IFIP/ACM Middleware*, pages 329–350, 2001.
10. I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *SIGCOMM*, pages 149–160. ACM Press, 2001.
11. T. Suel, C. Mathur, J. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long, and K. Shanmugasunderam. Odissea: A peer-to-peer architecture for scalable web search and information retrieval. Technical report, Polytechnic Univ., 2003.