Project Number 001907

DELIS
Dynamically Evolving, Large-scale Information Systems

Integrated Project
Member of the FET Proactive Initiative Complex Systems

Deliverable D6.5.5

Final Prototype System
Start date of the project: January 2004

Duration: 50 months

Project Coordinator: Prof. Dr. math. Friedhelm Meyer auf der Heide
Heinz Nixdorf Institute, University of Paderborn, Germany

Due date of deliverable: December 2007
Actual submission date: December 2007
Dissemination level: PU – public

Work Package 6.5: P2P System Architecture and Testbed

Participants: Max Planck Institute for Computer Science (MPII), Saarbrücken, Germany

Authors of deliverable: Gerhard Weikum (weikum@mpi-inf.mpg.de)
Matthias Bender
Tom Crecelius
Sebastian Michel
Josiane Xavier Parreira
1 System Overview

Minerva is a peer-to-peer (P2P) prototype system that comprises a focused crawler, a local (per-peer) search engine, and global search engine that connects peers and users query routing strategies for collaborative search.

The focused crawler Bingo!, which is included in the Minerva package, mimics a human user browsing the Web by only indexing documents that are thematically related to a predefined set of user interests. Bingo! is a multi-language parser, i.e., it can detect the language of documents and restrict the crawl to documents of a language of choice. Bingo! learns the user interest profile by running a feature analysis over the bookmarks that it can import from the user’s Web browser. Within the user’s interest, Bingo! can further classify the documents it indexes into predefined and automatically trained categories. Crawling is continuously performed in the background, without manual user interaction. Bingo! automatically parses and indexes all applicable content types (currently text, html, and pdf) to build a local search index from these documents. It utilizes stemming and stopword elimination. The search index (in form of inverted index lists) is stored in the embedded Cloudscape/Derby database. Different score values are computed without any user interaction, to support ranked retrieval queries. Bingo! can compute link-based authority scores (PageRank, HITS) on its local Web graph.

For collaborative P2P search, Minerva is equipped with strategies for query routing and result merging. Minerva computes compact statistical synopses based on the local search index created by Bingo! that describe the quality of the index w.r.t. particular terms. Minerva publishes this information into a fully distributed directory of Minerva instances. The directory implementation is based on FreePastry, a freely available implementation of a distributed hash table.

Minerva offers both APIs for program development on top of Minerva and a search GUI that allows a user to enter query terms, which (transparency to the user) starts the global query execution. Upon receiving a query, Minerva retrieves applicable synopses from the directory and selects a small subset of promising peers that are most likely to provide high-quality results for a particular query. Minerva sends the query to these selected peers, which evaluate the query using their local search functionalities on top of their local indexes and return their top-matching URLs to the query initiator. Minerva appropriately combines the URLs from these autonomous sources and displays the results to the user.

Additionally, Minerva offers the functionality to annotate (local and remote) content with attribute-value-style annotations, much in the spirit of a social tagging community. Based on these tags, Minerva offers the following search functionality: it can retrieve all documents that were annotated with a certain attribute-value-combination, and it can retrieve all annotations that were provided for a particular document.

2 Installation

We have been developing and improving Minerva since March 2004. Minerva is implemented entirely in Java. As of end of 2007, the core component of Minerva consists of approximately 25,000 lines of code, which are organized in almost 200 classes in 35 packages. The Minerva software, including Bingo! as the focused crawler, can be downloaded from the URL

This software archive is fully self-contained, i.e., does not impose any prior software requirements (this is the reason for its large size). However, it does require (free) licenses for the underlying external open-source software FreePastry, Apache Derby/Cloudscape, and Tomcat. The Minerva software has been successfully tested on various versions of Windows and Linux. The included readme file explains the installation itself.

To cite Minerva, please use the following reference:

References


