DELIS
Dynamically Evolving, Large-scale Information Systems

Integrated Project
Member of the FET Proactive Initiative Complex Systems

Deliverable D6.5.4

Functionally enhanced and stress-tested prototype system
Start date of the project: January 2004
Duration: 48 months
Project Coordinator: Prof. Dr. math. Friedhelm Meyer auf der Heide
Heinz Nixdorf Institute, University of Paderborn, Germany
Due date of deliverable: December 2006
Actual submission date: January 2007
Dissemination level: PU – public

Work Package 6.5: P2P System Architecture and Testbed
Participants: Max-Planck-Institut für Informatik (MPII), Saarbrücken, Germany
Authors of deliverable: Gerhard Weikum (weikum@mpi-inf.mpg.de)
Matthias Bender
Tom Crecelius
Sebastian Michel (smichel@mpi-inf.mpg.de)
Josiane Xavier Parreira
**System Overview**

BINGO! is a focused Web crawler that 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.

MINERVA is a research prototype P2P Web search engine that was originally designed and implemented as an internal tool to evaluate the strategies we develop for query routing and result merging in the course of our research on P2P Web search. 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 that 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 (DHT).

MINERVA offers a simple search interface that allows a user to enter query terms, which (transparently to the user) starts the global query execution: 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 attribut-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.

**Installation**

The software is available at

http://www.minerva-project.org/downloads.html

The software archive provided is fully self-contained, i.e., it does not impose any prior software requirements (this is the reason for its large size). The software has been successfully tested on various versions of Windows and Linux.

Unpack the provided archive file to a directory of your choice (Windows users: preferably on the C: drive). This will create a subdirectory named bingo3 at the location of your choice.

As there is no further installation necessary, the software can be uninstalled simply by deleting the directory again.