WebRACE: A Distributed WWW Retrieval, Annotation, and Caching Engine.

Marios D. Dikaiakos
Dept. of Computer Science
University of Cyprus, PO Box 537, 1678
Nicosia, CYPRUS
mdd@ucy.ac.cy

Demetrios Zeinalipour-Yazti
WinMob Technologies
Ag. Antoniou 5, #103
Nicosia, Cyprus
csyiazti@ucy.ac.cy

Abstract

In this paper, we present the architecture and implementation of, and early experimentation with WebRACE, a prototype HTTP Retrieval, Annotation and Caching Engine. WebRACE retrieves from the Web documents according to XML-encoded user profiles that determine the urgency and relevance of collected information. The system subsequently caches and processes retrieved documents. Processing is guided by pre-defined user queries and consists of keyword-searches, title-extraction, summarization, classification based on relevance with respect to user-queries, estimation of priority, urgency, etc. WebRACE processing results are encoded in a WebRACE-XML grammar and fed into a dissemination server, which selects dynamically among a suite of available choices for information dissemination, such as "push" vs. "pull," the formatting and transcoding of data (HTML, WML, XML), the connection modality (wireless vs. wire-based), the communication protocol employed (http, GSM/WAP, SMS), etc.

WebRACE is part of a more generic system, called eXtensible Retrieval Annotation Caching Engine (eRACE) [1], which collects, annotates and disseminates information from heterogeneous Internet sources and protocols (Web, email, newsgroups), according to pre-specified user profiles and interests. The eRACE system consists of protocol-specific proxies, like mailRACE, newsRACE and dbRACE, that gather information from POP3 email-accounts, USENET NNTP-news, and Web-database queries, respectively.

In our work we address the challenge of designing a modular, open, distributed, and scalable architecture for WebRACE. To this end, we employ XML to encode and standardize information maintained within our engine, and Java-based Mobile Agents to enhance the distribution of WebRACE processing, achieve load-balancing and the extensibility of the software architecture. The WebRACE architecture includes:

- A Proxy Server, which consists of a Distributed Crawler fetching Web resources and an Object Cache storing multiple versions of retrieved resources.
- An Annotation Engine that indexes collected resources, executes user-queries, and produces "user alerts," encoded in the form compliant to the "ACI" (Annotated Cached Information) XML-grammar [1].
- A garbage-collector implementing cache replacement policies.

Furthermore, it interfaces with:

- A dispatcher that translates user-profiles into "jobs" for our proxy server, and schedules the execution of these jobs in a distributed cluster.
- An Alerting Server that interprets user alerts and transforms them into messages sent to users via different channels.

WebRACE is developed with Java, XML and Concordia, the Java-based Mobile Agent platform by Mitsubishi Electric Information Technology Centre. In our work

we discuss some of the most serious Java pitfalls in the development effort of our indexing and our distributed crawlers and discuss how we worked around them.

In summary, WebRACE represents an effort to build “next-generation” Internet servers that provide scalable, user-centric, content-distribution services. Such an approach can be used to support personalized, Internet-based information provision, on top of different communication protocols (wire-based and wireless) and information encodings.

References

[1] D Zeinalipour-Yazti. “eRACE: eXtensible Retrieval Annotation Caching Engine”, Diploma Thesis, University of Cyprus, June 2000.

[2] M. Dikaiakos, D. Gounopoulos, “FIGI: The Architecture of an Internet-based Financial Information Gathering Infrastructure,” in *Proceedings of the International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems*, pp. 91-94, IEEE-Computer Society, April 1999, Santa Clara, California.