AQR-Toolkit: An Adaptive Query Routing Middleware for Distributed Data Intensive Systems

Ling Liu, Calton Pu, David Buttler, Wei Han, Henrique Paques, Wei Tang
Georgia Institute of Technology, College of Computing
{lingliu,calton,buttler,weihan,paques,wtang}@cc.gatech.edu

Overview. Query routing is an intelligent service that can direct query requests to appropriate servers that are capable of answering the queries. The goal of a query routing system is to provide efficient associative access to a large, heterogeneous, distributed collection of information providers by routing a user query to the most relevant information sources that can provide the best answer. Effective query routing not only minimizes the query response time and the overall processing cost, but also eliminates a lot of unnecessary communication overhead over the global networks and over the individual information sources.

The AQR-Toolkit divides the query routing task into two cooperating processes: query refinement and source selection. It is well known that a broadly defined query inevitably produces many false positives. Query refinement provides mechanisms to help the user formulate queries that will return more useful results and that can be processed efficiently. As a complimentary process, source selection reduces false negatives by identifying and locating a set of relevant information providers from a large collection of available sources. By pruning irrelevant information sources, source selection also reduces the overhead of contacting the information servers that do not contribute to the answer of the query.

The system architecture of AQR-Toolkit consists of a hierarchical network (a directed acyclic graph) with external information providers at the leaves and query routers as mediating nodes. The end-point information providers support query-based access to their documents. At a query router node, a user may browse and query the meta information about information providers registered at that query router or make use of the router's facilities for query refinement and source selection. The meta information about an information provider includes content summary, query capability description, and information quality of that provider, and is called the provider's source capability profile. A source capability profile contains a query that is true of all documents available from the information provider. The precise nature of source capability profiles is not specified by the query routing architecture, but left to the information providers and individual query routers themselves. A source capability profile may therefore be a simple predicate consisting only of the host name of the computer acting as the information server or a disjunction of all the terms in all the documents at the information server. It may also be a subset of the attributes derived from the indexed documents plus value-added attributes that describe the collection of documents but may not occur in the documents themselves.

Description of Demo. The AQR-Toolkit is written in Java, running on both Windows NT and Solaris machines, and requires a Java-compliant web browser such as Netscape 4.0 or above. We demonstrate the latest version of our AQR-Toolkit, including the query refinement component, the source selection component, and the multi-level pruning strategies as described in the previous sections. Specifically we show the effectiveness and scalability of our AQR-Toolkit using a comparative shopping application, where multiple web-based information servers are used. The toolkit also includes facilities for performance monitoring, which allows a system developer to examine the impact of using different query routing mechanisms. In addition, we will use the execution monitoring capability of the AQR-Toolkit to show how the relevant information sources for a given query are dynamically discovered through the step-wise utilization of the user query profile and the source capability descriptions in the multi-level relevance pruning process, and how to ensure that each step considerably reduces the number of candidate rewritings considered in the next step.