WSQ/DSQ: A Practical Approach for Combined Querying of Databases and the Web*

Roy Goldman, Jennifer Widom

Stanford University
{royg,widom}@cs.stanford.edu
www-db.stanford.edu

Abstract

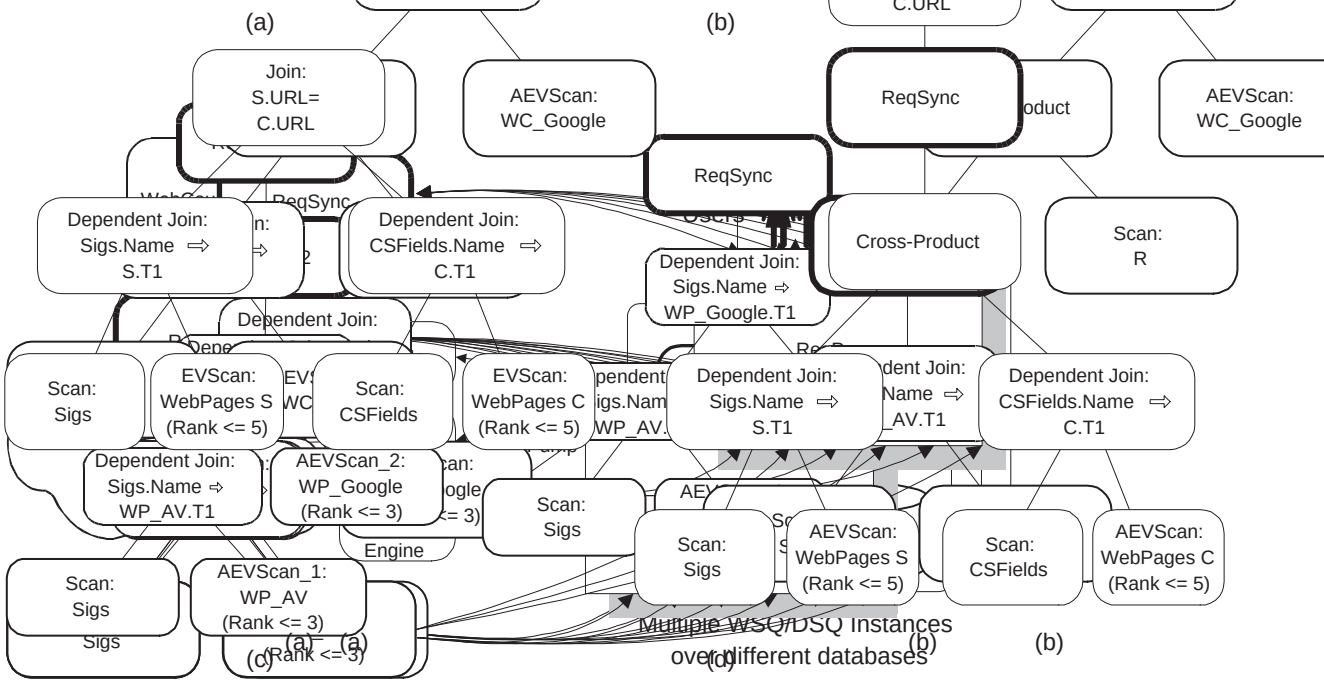
We present WSQ/DSQ (pronounced "wisk-disk"), a new approach for combining the query facilities of traditional databases with existing search engines on the Web. WSQ, for *Web-Supported (Database) Queries*, leverages results from Web searches to enhance SQL queries over a relational database. DSQ, for *Database-Supported (Web) Queries*, uses information stored in the database to enhance and explain Web searches. This paper focuses primarily on WSQ, describing a simple, low-overhead way to support WSQ in a relational DBMS, and demonstrating the utility of WSQ with a number of interesting queries and results. The queries supported by WSQ are enabled by two *virtual tables*, whose tuples represent Web search results generated dynamically during query execution. WSQ query execution may involve many high-latency calls to one or more search engines, during which the query processor is idle. We present a lightweight technique called *asynchronous iteration* that can be integrated easily into a standard sequential query processor to enable concurrency between query processing and multiple Web search requests. Asynchronous iteration has broader applications than WSQ alone, and it opens up many interesting query optimization issues. We have developed a prototype implementation of WSQ by extending a DBMS with virtual tables and asynchronous iteration; performance results are reported.

1 Introduction

Information today is decidedly split between structured data stored in traditional databases and the huge amount of unstructured information available over the World-Wide Web. Traditional relational, object-oriented, and object-relational databases operate over well-structured, typed data, and languages such as SQL and OQL enable expressive ad-hoc queries. On the Web, millions of hand-written and automatically-generated HTML pages form a vast but unstructured amalgamation of information. Much of the Web data is indexed by search engines, but search engines support only fairly simple keyword-based queries.

In this paper we propose a new approach that combines the existing strengths of traditional databases and Web searches into a single query system. *WSQ/DSQ* (pronounced "wisk-disk") stands for *Web-Supported (Database) Queries/Database-Supported (Web) Queries*. WSQ/DSQ is not a new query language. Rather, it is a practical way to exploit existing search engines to augment SQL queries over a relational database (WSQ), and for using a database to enhance and explain Web searches (DSQ). The basic architecture is shown in Figure 1. Each WSQ/DSQ instance queries one or more traditional databases via SQL, and keyword-based Web searches are routed to existing search engines. Users interacting with WSQ/DSQ can pose queries that seamlessly combine Web searches with traditional database queries.

*This work was supported by the National Science Foundation under grant IIS-9811947 and by NASA Ames under grant NCC2-5278.


(a)
(b)
Join:
S.URL=
C.URL
AEVScan:
WC_Google
ReqSync
AEVScan:
WC_Google
Dependent Join:
Sigs.Name ⇨
S.T1
Dependent Join:
CSFields.Name ⇨
C.T1
Dependent Join:
Sigs.Name ⇨
WP_Google.T1
Cross-Product
Scan:
R
Scan:
Sigs
EVScan:
WebPages S
(Rank <= 5)
Scan:
CSFields
EVScan:
WebPages C
(Rank <= 5)
Dependent Join:
Sigs.Name ⇨
S.T1
Dependent Join:
CSFields.Name ⇨
C.T1
Dependent Join:
Sigs.Name ⇨
WP_AV.T1
AEVScan_2:
WP_Google
(Rank <= 3)
Scan:
Sigs
Scan:
Sigs
AEVScan:
WebPages S
(Rank <= 5)
Scan:
CSFields
AEVScan:
WebPages C
(Rank <= 5)
Scan:
Sigs
AEVScan_1:
WP_AV
(Rank <= 3)
(a)
(b)
(c)
(d)
Multiple WSQ/DSQ instances
over different databases