Harvest: A Scalable, Customizable Discovery and Access System

C. Mic Bowman
Peter B. Danzig
Darren R. Hardy
Udi Manber
Michael F. Schwartz

CU-CS-732-94

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

Harvest: A Scalable, Customizable Discovery and Access System

C. Mic Bowman
Peter B. Danzig
Darren R. Hardy
Udi Manber
Michael F. Schwartz

CU-CS-732-94 July 1994

University of Colorado at Boulder

Technical Report CU-CS-732-94
Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309

Harvest: A Scalable, Customizable Discovery and Access System

C. Mic Bowman
Transarc Corp.
mic@transarc.com

Peter B. Danzig
University of Southern California
danzig@usc.edu

Darren R. Hardy
University of Colorado - Boulder
hardy@cs.colorado.edu

Udi Manber
University of Arizona
udi@cs.arizona.edu

Michael F. Schwartz
University of Colorado - Boulder
schwartz@cs.colorado.edu

July 1994

Abstract

Rapid growth in data volume, user base, and data diversity render Internet-accessible information increasingly difficult to use effectively. In this paper we introduce *Harvest*, a system that provides a set of customizable tools for gathering information from diverse repositories, building topic-specific content indexes, flexibly searching the indexes, widely replicating them, and caching objects as they are retrieved across the Internet. The system interoperates with Mosaic and with HTTP, FTP, and Gopher information resources. We discuss the design and implementation of each subsystem, and provide measurements indicating that Harvest can reduce server load, network traffic, and index space requirements significantly compared with previous indexing systems. We also discuss a half dozen indexes we have built using Harvest, underscoring both the customizability and scalability of the system.

1 Introduction

Over the past few years a progression of Internet publishing tools have appeared. Until 1992, FTP [43] and NetNews [39] were the principal publishing tools. Around 1992, Gopher [38] and WAIS [31] gained popularity because they simplified network interactions and provided better ways to navigate through information. With the introduction of Mosaic [2] in 1993, publishing information on the World Wide Web [3] gained widespread use, because of Mosaic's attractive graphical interface and ease of use for accessing multimedia data reachable via WWW links.

While Internet publishing has become easy and popular, making *effective use* of Internet-accessible information has become more difficult. As the volume of Internet accessible information grows, it is increasingly difficult to locate relevant information. Moreover, current information systems experience serious server and network bottlenecks as a rapidly growing user populace attempts to access networked information. Finally, current systems primarily support text and graphics intended for end user viewing; they provide little support for more structured, complex data. For a more detailed discussion of these problems, the reader is referred to [10].

In this paper we introduce a system that addresses these problems using a variety of techniques. We call the system *Harvest*, to connote its focus on reaping the growing crop of Internet information. Harvest supports resource discovery through topic-specific content indexing made possible by a very efficient distributed information gathering architecture. It resolves bottlenecks through topology-adaptive index replication and object caching. Finally, Harvest supports structured data through a combination of structure-preserving indexes, flexible search engines, and data type-specific manipulation and integration mechanisms. Because it is highly customizable, Harvest can be used in many different situations.

The remainder of the paper is organized as follows. In Section 2 we discuss related work. In Section 3 we present the Harvest system, including a variety of performance measurements. In Section 4 we offer several demonstrations of Harvest, including WWW pointers where readers can try these demonstrations. In Section 5 we discuss work in progress, and in Section 6 we summarize Harvest's contributions to the state of resource discovery.

2 Related Work

While impossible to discuss all related work, we touch on some of the better-known efforts here.

Resource Discovery

Because of the difficulty of keeping a large information space organized, the labor intensity of traversing large information systems, and the subjective nature of organization, many resource discovery systems create indexes of network-accessible information.

Many of the early indexing tools fall into one of two categories: file/menu name indexes of widely distributed information (such as Archie [18], Veronica [19], or WWWW [37]); and full content indexes of individual sites (such as Gifford's Semantic File System [21], WAIS [31], and local Gopher [38] indexes). Name-only indexes are very space efficient, but support limited queries. For example, it is only possible to query Archie for "graphics packages" whose file names happen to reflect their contents. Moreover, global flat indexes are less useful as the information space grows (causing queries to match too much information). In contrast to full content indexes that support

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.