Integrating Search an
with

By Edward A.
an
Dep
Virginia F

Introduction

While hypertext and hypermedia have numerous applicati
our problems relating to information access. Such access
or automatically created indices, search algorithms, feedb
techniques whose value has been proven on collections m
can be managed solely by browsing and following links.
model for what we believe is an urgent need, that of i
trieval with hypertext and hypermedia. Further, we des
designed around that model which would facilitate acc
types of information.

Current Status

In this chapter, we focus on the problem of how to effici
wide range of disparate information (including data stored
media archives, knowledge bases, and other organized col

lections were managed almost exclusively by mainframes, which ensured centralization and often enforced vendor, and sometimes industry-wide, standards. With the shift to personal computers and workstations, there has been a rapid proliferation of specialized packages, which make interchange of information between software and hardware platforms difficult.

Accumulations of information in forms that are hard to integrate are being created and expanded at a rapid rate. Large online database vendors such as Dialog, Maxwell Online, STN, and others provide access to enormous collections of bibliographic records (usually titles, abstracts, and other citation data) or full text (reference works, newspapers, legal documents). In house, corporate employees are generating electronic mail messages, database files (separately organized for geographic information, CAD/CAM, chemical registry data, and image databanks), and word processing documents. The serious problems of providing democratic access to all this information and of supporting manipulation techniques as versatile and natural as speech or pencil-and-paper methods have yet to be solved.

Prospects for Improvement

The emergence of networked computer systems and optical storage devices have exacerbated the problem of incompatible data formats and, as a result, forced vendors to seek solutions. For example, software designed to access data scattered across a heterogeneous network needs to follow rigorously enforced conventions.

Optical publishing systems increase the amount of storage capacity available at a reasonable cost. This allows in-house access to huge databases of text, graphics, images, and even interactive digital video [Fox, 1988d, 1989d, and 1990a]. The huge amount of information that can be stored optically can make particular facts difficult to find. Fortunately, locally stored information can be organized and more easily accessed with more modern software than is provided with on-line services such as Dialog. This allows more advanced information retrieval techniques to be implemeted [Fox, 1986a].

Our work is motivated by an analysis of current potentials and problems with information access. In the first part of this chapter, we describe a framework for integrating hypertext systems with techniques for discovering information quickly, leading readers to particular nodes efficiently even in very large, unfamiliar hyperbases. In the second section, we consider some of the capabilities that must be provided by information systems, pointing out which are provided by various types of existing systems. The third section discusses progress made in that direction and its sequel provides further background. The fourth section discusses the VPI&SU CODER project, which provides a means of exploring a broad class of information access approaches. The final section of the chapter outlines how we are extending CODER for hypermedia and interoperability, and includes recommendations for conceptual and practical integration of information retrieval with hypertext/hypermedia.

Necessary Capabilities

Computer systems which are intended to facilitate us
should:

- Include help systems that can explain how to perform
  as part of that search;
- Adapt to user requirements;
- Help users describe and elaborate their information ne
- Be built in a flexible and modular fashion so that th
  [Belkin, et al., 1983; Belkin, et al., 1987a]

A user-centered, mixed initiative interaction (where b
erate concurrently and can each seize control of input o
any time) is preferred, so that:

- the user can request an explanation or follow a new tr
- the computer can request clarification or display tent
  to further what it believes are the user's current goals.

Since these goals often relate to finding relevant i
repository, it is important to pay special attention to eff
ing, as is discussed in the next subsection.

Finding Barn Doors and Stringing Pearls

The current generation of information retrieval syste
techniques to find relevant items. These can be viewe
system-specific levels [Bates, 1981]. One approach is t
looking for a neighborhood where useful items can be
related or linked items from among the items at hand.

One can think of first finding the *barn door* and then,
of interest, i.e., a valuable *pearl*, following leads to clo
analogous to searching and then browsing, or using
niques and then applying hypertext methods.

Most information being accessed has been processed
make access easier. Outlines, tables of contents, and
names, or specially selected descriptors) are usually dev
ing to some theory or organizing principles. Readers
organizational systems can often find items of interest;
times applied to map these structures to available regio
lated items may be found if they occur nearby, if an i
items, or if explicit cross references are provided.

Unfortunately, the searcher's perspective and/or vocabulary does not always match that of the author or indexer. This is particularly important when a complex idea is involved and when a very large collection of items is being searched. Here precoordinated systems (where some of the most important combinations of ideas are precombined and added to indices, sometimes as phrases like "information retrieval") may fail and force readers to build post-coordinated queries (like "multimedia and retrieval"). Narrower queries can be formed using proximity operators. These include:

- searches for adjacent words (two words placed side by side);
- two or more words located in a single sentence;
- within field combinations (e.g., two words located within a title line);
- within a paragraph.

Boolean query construction (i.e., using ANDs, ORs and/or NOTs to specify matches) can be exceedingly difficult, and indeed ten-to-one variations have been observed in the ability of different individuals (and even of the same individual at different times) to form good queries. Frequently, repeated efforts are required to form queries that do not retrieve many nonrelevant items (i.e., have high precision) and yet do retrieve a significant fraction of the relevant items (i.e., have high recall).

Finding relevant items often involves one or more of the following steps:

- Translating an information need into the organization of concepts represented by a table of contents, index or thesaurus.
- Adopting a scheme to descend through some hierarchical organization of topics or descriptors to find specific section(s) of interest.
- Combining words or terms used within an index or thesaurus or forming the free text into a formal Boolean query.
- Refining the query by adding synonyms or related terms (broadening), by increasing specificity (narrowing) or by reorganizing the query structure.
- Visually examining all selected items to find those of value.

Sometimes the task at hand makes the choice of search method obvious. If you are looking for a single known item, a Boolean search will work best. Looking for multiple items on the same topic calls for browsing. Attempts to retrace a previously viewed presentation suggests the use of links. Sometimes a combination of methods works best.

Unified Metaphors

The steps involved in searching and browsing hyperdocuments should be hidden behind a task-oriented interface that lets readers search at a conceptual, descriptive level instead of at either a word-oriented or a procedural level. Boolean queries, for instance, fail to fulfill readers' information access needs because they focus

attention on the wrong objects and force users to think i
of search algorithms. Conversely, properly implemente
readers to explore a range of conceptual dimensions
graphical presentation coupling the spatial metaphor of n
metaphor of links supports such explorations using si
Most importantly, both searching and browsing should
consistent metaphors and abstractions that encourage th
cess techniques.

Using naturally occurring metaphors as the backbone
several advantages. Unfortunately, while paper sketches
may be easy for humans to work with, they are often b
the most expensive of today's computer systems.

Another possibility is to have a standardized unified
manipulation system. We all take for granted the fact tha
ple all around the world, and that they do this because o
on hardware and software standards. We can envision a
hardware, software, and information structures or data f
built around both existing and exciting new technologies
requires management of multimedia including: text,
audio, and video. It involves interconnection of multip
networks, and use of a variety of peripheral and input/
requires *intelligent* processing to improve efficiency and
sure smooth human-computer interaction.

Progress Toward Unified Access

Networking

A great deal of progress has been made towards integrat
[Tannenbaum, 1981]. Thousands of computers are int
change electronic mail. Many are connected to allow eve
tion: passing of files, remote log-ins, and even communi
In the area of library information retrieval, the Z39.50
oped so that a user of one library system can cause tha
processed on another system, and then indirectly receive
ous organizations, including universities such as VPI&SU
ing prototype systems for information interchange that
tions. In addition, the X.500 standard has enabled many p
bine into a partial global directory system—a computer
white pages.

Much of the success in internetworking can be credite
approach to software, whereby physical connection, data
agement, presentation handling, and application program
interfaces and specifications. The *International Organiz*

(ISO) *Open Systems Integration* (OSI) model uses physical, data link, network, transport, session, presentation, and application levels to encourage interoperability of software on heterogeneous computers [Zimmerman, 1980]. OSI has been the basis for Z39.50, X.500, and many other related standards.

This layering approach also applies to storage systems. On the one hand, there are levels of physical units, operating system-identified devices, and network wide file systems. On the other hand, there are software divisions, such as the proposed standard of the Air Traffic and Air Transport Associations, which separates the search engine component of CD-ROM software (which itself builds upon the volume and file description layer specified by the ISO 9660 standard) from the user interface layer.

Object Model

While computer users generally search for relevant concepts, computer scientists have tried to meet user's needs by studying and building specialized processing systems for particular classes of information. Thus, the term "data" is often associated with database management systems, "information retrieval" is often limited to text databases, and "knowledge bases" are viewed as the proper repositories for collections of complex list structures and rules.

To unify these approaches, object-oriented databases have been proposed that might contain and support processing of any class of information. In hypertext systems that build upon these object bases, completing a search or following a link would lead to presentation of any of various objects: a screen of text, a bitmap display, an audio segment, a spreadsheet entry, an individual plane of a complex image, an explanation of some expert rules, or a tabular report.

However, to date only limited progress has been made in integrating database management systems (which at present usually follow the relational model) with text retrieval systems (which frequently employ inverted files and require access by way of Boolean queries). One issue is the differences in items being manipulated; tables of numbers or short character strings have few data types as opposed to the variety encountered in multimedia systems, and database structures are much more regular than those found in the world of compound documents (e.g., books, magazines, letters, reports, and encyclopedias). A possible solution is to use abstract data types (e.g., sets, lists, vectors) as elements of a relational DBMS [Fox, 1981]. Some of these ideas were used in the redesign of the SMART system for the UNIX environment [Fox, 1983a], which eventually led to a version of that system that is widely used by document retrieval researchers. Another approach is to directly use the relational model, with performance tuning and limited extensions where needed, to handle bibliographic records [Lynch, 1987].

Artificial intelligence (AI) research extends these efforts to solving the problems of knowledge representation. Initially, languages like LISP and Prolog focused on

symbol and list manipulation. Real world objects were could be located by their names. Property lists (sets o were attached to these atoms.

But to better match human ability to deal with defa attributes with various types of objects, and taxonom classes of objects, the concept of a *frame* was develop useful for information retrieval [Fox, 1987b, Weaver, e

Essentially, a frame class (e.g., U.S. postal address) ant aspects of the object (e.g., state, zip code), and ma classes (e.g., postal address with slots for name and lo tiated for an individual (e.g., John Smith's U.S. post grouped together into extremely regular knowledge bas

A semantic network is an alternate knowledge repre tion to hypertext [Findler, 1979, Morgado, 1986]. A se visioned as a graph where nodes are used to represent c significant, meaningful relationships. One well-known SNePS [Shapiro, 1989], supports knowledge represe proving or drawing conclusions based on knowledge system). In SNePS a distinction is often made betw whereby the reader follows a succession or chain of identify some relationship, and *node-based inference*, w are represented in the network and are combined with regions of the network in accord with the rules of logi be very efficient, and corresponds to following links i works can also support *spreading activation,* where sev the network, all paths of length 1, 2, etc. from each are several concepts meet—a type of parallel search f Spreading activation has been used in the GRANT sys related to various topics, projects, and grant propos 1987].

Semantic networks are useful for handling the inter At VPI&SU, for example, we have taken machine-read and restructured the important data [Fox, 1986b], arch Prolog facts, and loaded them into a semantic network t ing and access [Fox, 1988a; France, et al., 1989a]. In s helpful to build an elaborate taxonomy of relations (i.e word senses in our lexicon. One goal is to extend ear cated improvement in retrieval effectiveness could result lexically related to query terms [Fox, 1988e]. Thus, t about the meaning of a query to search for "captain" i officers," or "general" when "army leader" is sought. O works as important support structures for probabilistic tems that prove when a document is relevant to a query Ultimately, semantic networks can be combined with hy lead to a uniform representation of objects, as discussed

Integration of Various Search Techniques

Research suggests that the best retrieval performance results when many different forms of information about documents are utilized by a variety of search methods [Belkin and Croft, 1987b]. Integration of various approaches has this same goal [Fox, 1987b]. Thus, one might merge the results from using Boolean, vector, and probabilistic searches for a given information statement. Any of these search methods can also be modified by the use of feedback, where a reader indicates what documents or sections thereof are relevant, and the computer uses all the information it has available about that sample to perform another, better search. One can even throw in use of AI techniques, as discussed in the section on our work with CODER. Reference models provide a clear framework for integration and thus have become popular in regard to networking, where different cabling, interconnection, specialized hardware, firmware, and layers of software allow users of similar applications on different computers to collaborate.

Reference Models for Hypertext

In order to develop interoperable information systems and a unified representation model, a reference model for information management systems must first be developed. Work on hypertext reference models has arisen in part as a result of standardization efforts for hypertext/hypermedia [see Chapter 26 by Devlin on standards]. We have found the Dexter model [Halasz and Schwartz, 1990] and the *r-model* [Furuta and Stotts, 1989; Furuta and Stotts, 1990; Stotts and Furuta, 1989] to be particularly insightful. However, interoperable information systems should properly include not only hypertext/hypermedia, but also searching, networking, and other applications and levels of manipulation.

A More Complete Reference Model

Our proposal for a reference model that integrates hypertext and other sources of information is shown in Figure 21.1, and explained in more detail in subsequent sections. Essentially it combines the seven OSI layers with seven other layers relating to hypertext/hypermedia and other types of information systems.

The bottom four layers are OSI layers that together support the secure transport of messages. Atop the transport layer is a layer that provides essential support for files and process communication (messages). Processes operate above this layer to support high-level communication between machines, including necessary translations between data representations. These six layers complete an extended groundwork for communication among machines, languages, and environments at a highly abstract level.

The node layer comes next, supporting atomic, structured, and multimedia objects. The anchor layer allows points or spans inside nodes to be addressed in ways appropriate for each node type. Links connect anchors, thus providing a

Layer	Typical Contents
Application	hypertext, hypermedia, inf image management, autho CAD/CAM, interactive di
Presentation	devices: windows, pointin media: text, audio, video; operations: generalization browsing, selection, pickin
Session	base selection, user identif versioning.
Base	base: knowledge, informat operations: search, inferen
View	graphs, lists, sets, vectors,
Link	link Ids, labels.
Anchor	anchor Ids, span descriptio
Node	textual strings, integers, ol
Communication	processes.
Physical machine	file systems, storage medi
Transport	ISO lower level equivalen
Network	ISO lower level equivalen
DataLink	ISO lower level equivalen
Physical	ISO lower level equivalen

Figure 21.1 Information Management System Refere

layer that is essentially a graph or network of informati
allows various aggregations and associations as well as c
tion and knowledge structures to be manipulated (e.g., h
view of vectors, a view of frames, a view as linked colle
layer allows coordinated access to data, information
through search, navigation, and related operations. Hig
upper level OSI layers, but are organized to support cor
hypertext/hypermedia, and presentation/application progr

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.