

Accessing Information from Globally Distributed Knowledge Repositories

(Extended Abstract)

Alfred V. Aho

Department of Computer Science
Columbia University
aho@cs.columbia.edu

Abstract

This paper discusses some of the major technical obstacles standing in the way of achieving cost-effective universal access to multimedia information stored in globally distributed knowledge repositories. Opportunities for contributions from the database research community are highlighted.

1 Introduction

The goal of developing an information infrastructure is to provide affordable universal access to multimedia information services on a global scale. Strong forces from business, technology, and users are influencing the evolution of this infrastructure.

A rapidly expanding market is developing for personalized communication and information services whereby people can get access to others and the information they need in the form and media they want, any time, any where. Businesses are integrating their operations on an international scale and people want to access their information environments and coworkers from their offices, homes, and remote locations with equal ease.

An increasingly diverse set of new services needs to be supported as a significant fraction of the populace gains access to the new infrastructure. Digital libraries, healthcare, email, entertainment, and electronic commerce are emerging as important application areas. Unresolved, however, is the question of who will pay for the new services and how.

Rapid improvements in computer hardware, software, and communications technology are accelerating the arrival of the information age. In the past decade we have continued to see order of magnitude improvements in the price/performance of processors and computer memories, as well as exponential growth in the number of people and the amount of traffic carried on the

Internet. Virtually every home in the U.S. has some form of information appliance such as a personal computer, telephone, or television.

2 The New Issues

The purpose of this paper is to highlight some of the key technical problems that must be solved if we are to achieve the goal of affordable universal information services. We shall focus on issues that the database research community is well positioned to attack. The problems that we discuss come from

- scalability
- integration of multiple media
- organization and integration of knowledge
- systems integration and evolution
- information quality
- searching and browsing
- universal access

The remainder of this paper will discuss these issues in more detail.

3 Scalability

The most challenging technical problem is how to design and implement a scalable information infrastructure architecture that supports the creation of effective information and communications services to a significant fraction of the people on the planet.

The current telecommunications infrastructure gets high marks for near-universal connectivity and high reliability, but it was not engineered to support a rich set of evolvable interactive multimedia information services on a global scale.

The most striking aspects of the current information infrastructure are how rapidly the number of people connected to it is growing and how much new on-line information is becoming available. Estimates indicate that at the beginning of 1996 more than 10 per cent of the population of the U.S. had access to the Internet.

The volume of on-line information is growing at a staggering rate. Tens of millions of Web pages and terabytes of information of various kinds are already

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.

PODS '96, Montreal Quebec Canada
© 1996 ACM 0-89791-781-2/96/06..\$3.50

available over the Internet. The Earth Observing System (EOS), a collection of satellites to be launched by NASA towards the end of this decade, is expected to generate about a third of a petabyte of information annually about the atmosphere, land, and oceans of the earth. This information will be combined with data from other sources and countries and stored in the EOS Data and Information System producing a database of unprecedented size. The genome centers at various universities throughout the world are expected to produce a comparable amount of information about the human genome.

The amount of traffic carried by the information infrastructure will grow dramatically with video-intensive applications such as teleconferencing, movies on demand, and video instruction. Delivering video and audio information puts more stringent quality-of-service requirements on the transmission of data. For example, the audio needs to be synchronized with the video in a movie, and the video has to be delivered within certain time constraints for it to appear natural. New network protocols are needed to meet the diverse quality-of-service requirements for multimedia information services. Guarantees of security and privacy will also be necessary for sensitive applications such as electronic commerce and healthcare.

At the physical level of the infrastructure we need storage devices capable of holding petabytes of data. In the immediate future, the most likely form these devices will take is tertiary-storage systems such as disk juke boxes or tape silos. However, such devices have data-access times measured in seconds, and thus are two or three orders of magnitude slower than secondary storage devices such as disks. The systems containing these storage devices will require new approaches to buffering and caching to ensure the smooth delivery of video and audio information to end users.

4 Integration of Multiple Media

The integration of multimedia data types is a significant challenge.

Digital documents are starting to combine new data types involving audio, images, and video with those dealing with text. These new types have widely varying needs and characteristics. Each type has its own preferred operators and access methods. The relational data model that the database research community has developed for handling tabular textual information does not extend to interactive multimedia types. We need to develop the appropriate operators and analyze the implementation tradeoffs for the new multimedia data types.

The problem becomes much more complex when we are forced to deal with composite data involving multiple types. We need a data model that can integrate

all forms of multimedia information and we would like associated languages with which we can describe the different types of information stored in our multimedia knowledge repositories. We need algorithms and data structures that can efficiently accommodate objects as small as a single byte of text or as large as a megabyte of video.

The presentation of new multimedia data types requires the design of information appliances with multimedia capabilities. The creation of appropriate authoring tools, user interfaces, browsers, and methods for displaying these data types in an understandable and esthetically pleasing manner are topics for future research.

5 Organization and Integration of Knowledge

The organization and integration of growing knowledge from disparate knowledge domains is an even more challenging problem.

Every field of human endeavor has its own jargon and way of organizing knowledge. For example, the fields of chemistry, dance, literature, medicine, and music have their own distinctive notations, taxonomies, and classification schemes for organizing their works. When we interconnect information systems from disparate knowledge domains, many difficult problems arise. At the organizational level, what concepts and terms should we use to describe the aggregated knowledge among the combined heterogeneous areas? This question is sometimes called the "ontology problem". In effect, we are asking what is an appropriate ontology for all of the Internet.

At the user level, we would like to have tools that would make information access from interconnected disparate systems as easy to perform as using a single system. We would like to work towards creating interoperable interfaces among different systems so application programs can gather and process information from different domains.

Several approaches are being explored for dealing with heterogeneous information sources. In one approach, an integrating model is defined and each source is fitted with a "wrapper" that translates the information from the source into the notation of the integrating model. Another approach is to use agents called mediators to collect and filter information from the disparate sources.

At the representational level, existing systems use a wide variety of formats and conventions for representing data. Translators are needed to convert data from one representation to another. Some standardization is taking place, but it is likely the need for translators and converters will continue for the foreseeable future.

The classification and organization of information facilitates its storage and access. In a rapidly changing world, classification schemes need to evolve to accommodate integrated approaches, changing business conditions, and new forms of knowledge. However, changing the classification scheme can cause existing applications and procedures to stop working. The problem of keeping old programs working when performing schema evolution has been widely recognized as an important problem in the database community. A number of approaches have been taken to address this problem including integrating schema evolution with view facilities, maintaining versions of the complete scheme hierarchy, and keeping different versions of individual types and classes. It is well worth trying to generalize these approaches to attack the problems arising from the integration of distributed heterogeneous information systems.

6 Systems Integration and Evolution

The global information infrastructure is the biggest systems integration challenge faced by engineers.

To facilitate the creation of new services and the introduction of new technology, we need interoperability at all levels in the infrastructure from the physical layer to the applications. Open interfaces between levels and components are critical for system interoperability and evolvability.

Many existing systems do not have well-defined interfaces and often do not expose interfaces that facilitate systems integration. Because the existing international information infrastructure contains many billions of lines of software, we cannot throw out the existing systems, but must learn how to interwork with them. A number of national and international organizations are addressing various aspects of interoperable open systems. Clear, precise, and unambiguous definitions of all systems interfaces are needed. From these specifications we can construct conformance tests for system components that would help assure system interoperability.

7 Information Quality

With the explosion of new data, the existing problem with data quality is likely to get worse.

Existing database systems are already plagued with the problem of “dirty data.” For a variety of reasons, legacy data can be incorrect, inconsistent, or incomplete. The corruption can spread when data derived from invalid sources is used to populate new information systems. When systems with inconsistent data are integrated, users are confronted with the problem of which source, if any, to believe.

An important research area is the assurance of information integrity. One approach is to trace and record the lineage of information – its origins, sources,

and annotations. This information can be used by users and query languages to determine the accuracy or reliability of particular information items.

8 Searching and Browsing

We need effective ways to find the information we need in the growing ocean of distributed multimedia data.

The database community has developed efficient techniques for answering precise questions such as “What is Jane Smith’s telephone number?” or “What is the cost of the textbook *Foundations of Computer Science*?” Query languages such as SQL facilitate access to conventional record-based data stores. The algorithms community has developed effective techniques for indexing and searching textual data for boolean combinations of keywords.

Many information retrieval tasks use much less precise concepts: “What are the causes of decline of public transportation in certain cities?” or “What are the factors influencing air quality in California?” Answering questions such as these require hybrid search strategies where the user may need to examine maps, satellite images, and other databases, and then correlate data among the various knowledge repositories.

There is a need for searching tools that can retrieve and process data from a variety of multimedia sources using imprecisely prescribed characteristics such as color, texture, shape, or approximate values. Effective methods for searching multimedia databases by content or concept are still open research areas. Some queries may well be best answered through a combination of browsing and interactive graphics. Designing a set of interoperable querying tools with these kinds of capabilities is a significant technical challenge.

9 Universal Access

Our final question is both a technical and a societal one: What facilities and services should be part of the basic information infrastructure that should be available to everyone at an affordable price?

At the technical level, this would involve asking what services could be provided at what cost. The database community has studied system and query optimization questions for years. However, the issues for the global information infrastructure are ones of vastly increased scope and scale.

On top of the basic infrastructure information service providers could offer customized services that would appeal to specialized communities with presumably an increased willingness to pay. The tradeoff is between what goes into the basic infrastructure and what is assigned to enhanced services.

If information is going to become essential for the educational, economic, and health well being of the average individual, then information services need to

be universal and affordable. Otherwise, we'll have a world of haves and have-nots differentiated by their access to information. The distinction between basic and enhanced services, however, is not well understood and is likely to be a subject of much debate in the near future.

10 Conclusions

We have taken a top-down view of some of the major technical challenges that must be met if people of all nations will have a universal and affordable global information infrastructure. While we have focused on technical issues, we should point out that there are substantial nontechnical questions that must be answered as well. These include safeguards for individual privacy, protection of intellectual property, and guarantees of freedom of expression.

At the technical level, all of the questions we have mentioned already being investigated to some degree in their respective technical communities. The issues that transcend individual communities are being shaped by sometimes conflicting forces from several communities, but there is relatively little critical examination of affordability/universality tradeoffs for various sets of basic services. Since the total new investment in the global information infrastructure is likely to run into the hundreds of billions of dollars over the next decade, this kind of research would have enormous economic impact. It is also the kind of work the database research community is well positioned to conduct.

References

- [A90] Alfred V. Aho. Algorithms for finding patterns in strings. In *Handbook of Theoretical Computer Science* J. van Leeuwen, Ed., Elsevier, 1990.
- [SSU95] Avi Silberschatz, Mike Stonebraker, and Jeffrey Ullman, eds. Database research: achievements and opportunities into the 21st century. *Report of an NSF Workshop on the Future of Database Systems Research*, May 26-27, 1995.
- [U82] Jeffrey D. Ullman. *Principles of Database Systems*, second edition. Computer Science Press, 1982.
- [WMB94] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes*, Van Nostrand Reinhold, New York, 1994.
- [CSTB94] *Realizing the Information Future* National Academy Press, 1994.
- [NRC94] *The Changing Nature of the Telecommunications/Information Infrastructure* National Research Council, 1994.