

RESEARCH STATEMENT

My principal research interests lie in the areas of database systems and mobile computing. I am especially interested in the application of database technologies to new domains such as the Internet and bioinformatics. My thesis presents an examination of techniques that we have discovered for representing data on the Web more efficiently. Another facet of my work is concerned with the discovery of data sources on the Web. While we anticipate that the standardization of rich Web service metadata descriptions and directory services will facilitate integrative online data processing in the next-generation Web, these solutions are still on the horizon. My service discovery work addresses current service discovery problems that will retain their relevance in the future Web Services paradigm.

Automatic Discovery and Classification of Bioinformatics Services The Large-scale Data Access Project addresses the interesting problem of providing integrated access to a large number of bioinformatics Web sources. The major challenges are to locate new Web sources, evaluate them to determine if they provide a BLAST interface, construct a wrapper for the source, and integrate the source into a mediator system that can provide a single point of access to all known sources conforming to the interface. Source autonomy complicates this problem: a cursory Web search yields hundreds of sources that provide a BLAST interface, many of which do not appear in bioinformatics directories. Manually maintaining a wrapper library will not scale to accommodate the growth of genomics data sources on the Web, challenging us to produce an automated system that can find, classify, and wrap new sources without constant human intervention. Our work to date focuses on BLAST interfaces to concretely demonstrate the approach, but these flexible techniques are generic and can be easily applied to other domains. Our approach to this problem combines service class descriptions with analysis techniques that map sources on the Web back to that description. We have shown how these concepts can be applied in an existing application scenario, Web-based BLAST genome sequence search. Finally, we have verified our claims experimentally by using a BLAST service class description to classify a group of Web sources.

Our approach to discovery and classification of Web sources groups them into *service classes* that share common functionality but not necessarily a common interface. Service classes are specified by a *service class description*, which uses an XML format to define the relevant aspects of a category of Web sources from an application's perspective. The service class description format supports the source discovery problem by providing a general description of the type of source that is considered interesting. It defines the data types that comprise the service arguments as well as any intermediate types that may appear in a source. It establishes a general description of the interface used by source class members and outlines intervening control points. Finally, it lists examples that are employed during source evaluation.

We are continuing development of new heuristics for site processing and recognition. In particular, significant progress has been made towards automatically identifying common types of indirection pages encountered during BLAST searches. The system will also be extended to support aggregation of data from hyperlinks—e.g. gene summaries commonly found in BLAST results. Longer term work will examine applying existing and novel information retrieval techniques to increase the number of recognized sources and further improve performance. For example, an

advanced classification system could compare new sources to those it has already classified: if the new source matches a previously discovered source, the information from the existing match can be used to guide analysis of the new source.

Web Change Monitoring. The World Wide Web offers a unique publishing medium that enables information broadcast with few of the traditional barriers to widespread communication. However, the rate of change of published data is highly variable: some pages—e.g. stock quote services—are updated frequently, requiring those who need the latest information to constantly check them. Other, more static data sources may only update a few times per year on an irregular basis. The user cost of manually monitoring even a small set of documents quickly overshadows any perceived benefit. Infrequently updated pages risk being forgotten, while highly variable documents quickly overwhelm the user with data.

Automatic Web change monitoring provides several compelling advantages even for simple scenarios. First, automatic systems remove the burden of monitoring from the user, allowing them to concentrate on other efforts while being assured of receiving timely notification when an interesting change occurs. Second, Web change monitors can track many different sources simultaneously: users can handle more data effectively, making them more productive and increasing the quality of their decisions.

We have developed an automatic Web change detection system that provides a mechanism for monitoring Web information sources. Rather than expending energy checking sites of interest for changes, the system allows users to concentrate on finding innovative applications for monitored information. Our design provides a framework for flexible and scalable Web change monitoring through the use of efficient data management, rich processing semantics, and grouping. Documents are stored and compared in an efficient format that allows change detection to be focused on areas of interest to the user. The system also groups compatible monitoring requests to actively reduce computation, network usage, and local I/O.

Web Document Encoding. One important consideration for large Web services is data storage and processing efficiency. Much of the data on the Internet is contained in HTML documents that are useful for human browsing but incur significant drawbacks from a data management perspective. HTML has no real type information aside from layout instructions, so any data contained in the document is mixed with formatting and layout constructs intended to help browser software render pages on screen. Automated data extraction or comparison of Web pages is expensive and slow.

We have developed a new document encoding scheme to address some of the problems associated with storage and processing of Web documents to enable Web service applications to operate efficiently on a large scale. The Page Digest encoding is designed to bring some of the advantages of traditional string digest algorithms to bear on Web documents. A Page Digest of a Web document is more compact than HTML or XML format but preserves the original structural and semantic information contained in the source document. The Page Digest uses a document tree model and explicitly separates the structural elements of the document from its content. This feature allows many useful operations on the document to be performed more efficiently than op-

erating on the plain text. Unlike schemes using general compression algorithms, a Page Digest is not “compressed” from the source nor does it need to be “decompressed” to be used, which minimizes the processing needed to convert a document from its native format. Page Digests can be efficiently converted back to the original document and are significantly smaller than the original document, thereby providing a scalable solution for large-scale Web services.

Other Interests. Mobile computing is an exciting research area, especially with the advent of low-cost global positioning and wireless Internet access. These tools create additional research problems since the constraints on a mobile device are radically different from those of a desktop machine. Data transformation for mobile devices, server-device data interactions such as broadcast scheduling, and human interfaces to mobile devices will all continue to be important topics as computers continue their migration away from the desktop.

I am keenly interested in applications of voice-recognition technology, both for mobile devices and for the desktop. Computer interaction via voice-recognition is fundamentally different than interacting with a keyboard and mouse. Both approaches have strengths and weaknesses: entering large blocks of prose text can be considerably faster with voice-recognition, but traditional input devices better facilitate tasks such as document navigation. This is not a failure of voice-recognition per se but is rather an opportunity to challenge the fundamental assumptions made in computer interface design. With computer usage continuing to rise, it will become increasingly important to provide machine interfaces that better match the needs and preferences of their varied users.