Future Directions and Research Problems in the World Wide Web

Udi Manber
University of Arizona
udi@cs.arizona.edu
http://glimpse.cs.arizona.edu/udi.html

(This is just a short abstract. The full paper will be posted on the web; see http://glimpse.cs.arizona.edu/PODS.html)

The web is changing the world of computing. This is no longer a prediction; it's a fact. The need for sophisticated techniques to handle data, be it text, images, numbers, video, sound, molecular sequences, or whatever, will grow more rapidly than ever before. Databases, and in particular database research, has not played a major role in the web so far, just as it didn't play a major role in the first 20 years of the history of computers. We are moving in an order of magnitude faster pace now, so this may be exactly the right time for a major database influence.

As an introduction to the kind of challenges we face, let's look at an obvious example: How does one find on the web current information about database research related to the web? It's possible, and even easy, to find information about specific topics that have unique names, like finding the manual for *agrep*. But it's much harder to find topics like current database research, current operating system advances, or new ideas in WWW search. It's even harder to find "quality" information related to such topics, of the kind that research journals or conferences like this one contain.

This tutorial will look at database research directions opened by the web. I will discuss the current place of databases on the web and the many possibilities and potential for the future. Since my work centers around the web (I am not a database expert), I will look at databases from the point of view of the web, rather than at the web from the point of view of databases. I will try to survey the areas where the web can benefit from database research and database expertise. The following issues will be discussed. (This is not a complete list. The web is moving too fast for anyone to be able to predict anything three months in advance.)

Integration with Information Retrieval There is a lot in common between IR and databases, but historically these two disciplines have gone their own ways. One way to look at the difference is that databases have mostly dealt with very well defined precise queries on structured information, and IR mostly dealt with unstructured textual information and imprecise queries. (There have been, of course, many exceptions in both areas.) Text databases have become more popular recently, but they have not yet found their way into mainstream databases. Since textual information will continue to dominate the web for a while, adding more power to text processing will continue to be important.

Handling Diversity The web is indeed a *world*-wide web, and its diversity is therefore inherent. Not only are there numerous types of information, numerous methods of presenting it, and numerous tools for accessing it, there are different interpretations, different meanings, and different needs for the same data. The tradeoff between ease of publishing (writing) and ease of querying (reading) on the web has tilted dangerously in favor of the former. Database technology could help balance this tradeoff, but only if we find more effective and general ways to handle diversity. We need to develop new publishing techniques that are general enough, easy enough to use, and yet have some useful organization. And we need to develop new querying mechanisms to handle diversity better.

The traditional problems of dealing with different processors, different operating systems, and different languages pale in comparison to the data diversity problem on the web. It's like the difference between the mechanic who has to deal with different cars vs. the teacher who has to deal with different students. (And not just because students are more complex than cars, but because technology has always been more attentive to the mechanic than to the teacher.)

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

Handling Scale Scale is not a new issue in database research. However, the scale of the web – and its growth – stretches everything to the limit. Problems with scale manifest themselves in many forms including size of databases (e.g., a list of all web sites), number of users (a million queries a day is not out of the question), number of changes (e.g., data that is updated every 10 seconds), diversity (mentioned above), and security issues. But the most difficult issue is that we need to address all of these scale problems together.

Integrating Query Mechanisms A particular subproblem of the data diversity problem is how to integrate different query mechanisms. This may be, in the short term, the most pressing database problem on the web. Consider the following three queries:

1. Find a "nice" public-domain color drawing of a tree.
2. Where's the "best" place to buy memory?
3. What's "good" on TV tonight?

The information required to answer these queries is already available on the web. Moreover, there are, or soon will be, search facilities attached to this information. But finding the starting points for each one of these queries, learning how to use the search, and adapting the search to one's personal preferences (so that "nice," "best," and "good" are defined meaningfully) are still hard problems.

Extending Query Mechanisms The problems of extending the usual query mechanisms to handle more types of data are already being addressed in mainstream database research. Besides textual data, examples include temporal data, spatial data, video data, and biological data. The challenge is to provide query mechanisms that can handle several types of data, for example, numeric, textual, and spatial, *at the same time*. These problems existed before the web, and they are not inherently dependent on the web, but the web makes them much more important.

Collecting Data The web is not only a means to find answers to questions, read the news, see images, and chat; it can also be a means to collect diverse information from many sources. This part has not been sufficiently addressed yet. Several new systems that deal with collection of information from the web will come out very soon, but there is a lot of room for more research.

Indexing Mechanisms The diversity and scale of the web call for many types of indexes. In some cases speed is the only factor, in other cases space is the major concern, and in most cases price and availability are the most important features. Better indexing mechanisms will no doubt appear. The issue of compatibility of different indexes has not yet received the attention it deserves. Several *meta*-search facilities, which use more than one search engine, have been developed for the web. But there are no facilities to combine several indexes. This is particularly important in the context of the next issue.

Client/Server and Hybrid Paradigms Databases started as centralized services and have moved to the client/server model in recent years. The web may open the door to other architectural models, although the client/server model will probably dominate for a while. In particular, the distinction between data and programs may become weaker. Clients may have unique query mechanisms which they will use on data in large servers, shipping the programs rather than the queries. The ability to contact a variety of indexes, understand them, and manipulate them will be essential. In addition to queries and query mechanisms looking around for data, we may also have data in search of good analysis.

Customization vs. Indoctrinization Shall we let each user dictate exactly how she wants to access any particular data? Shall we invest in complex customization schemes (e.g., agents) to allow users to create their own unique working place (like the UNIX shell, for example)? Or shall we try to unify as much as possible and ask users to learn one or very few methods for access, user interfaces, visualizations, etc.? I believe the answers to these 3 questions are yes, yes, and yes. Customization is still one of the most underutilized powers of computers. We can do a whole lot better. But, contrary to popular belief, I believe that steep learning curves are not inherently bad. No one is designing new pianos with only white keys or new guitars with fewer strings. You could learn to play such instruments much faster, but you wouldn't be able to produce much music. It sometimes pays to invest in learning to get more powerful computing techniques. We need to be careful to get the right balance of power, standards, and customization.

Security Security concerns have usually been an add-on; something to worry about after the system has been built or after something has happened. This will have to change. We cannot hope to let security experts put a magic spell on systems *after* they are designed, just as we cannot hope to let efficiency experts make programs fast after they are written. Security will have to be an integral part of every system from day one. With the web, local information can be made available everywhere, but at the same time any small weakness in a local system can quickly become "available" everywhere. Powerful global search facilities help legitimate and illegitimate users equally. A prime example of this is searching for sensitive (e.g., password) files using the information collected accidently by one of the web robots.

User Interfaces Again, these are not new problems, but the web provides a new scale for them. We will have to support users who will come to a database, expect to ask a couple of queries, get their results, integrate them with their own information, and leave within a matter of minutes, not to return for a while. They cannot afford a significant learning curve. More friendly user interfaces are needed and not just for comfort and aesthetics. Supporting feedback, hand holding, customizing on the fly, adapting to user preferences and history, and visualizing progress are some of the issues that will need to be addressed.

Browsing One of the main eye-openers from the success of the web is the importance of browsing. People like to be active! There is a benefit in "travelling" to the destination; often it gives unexpected information. I have developed over the years scripts to obtain ftp information efficiently, but I don't use them anymore, even though I can often save time. I prefer browsing. (I actually prefer, and have developed, more sophisticated web techniques, but that's another story.) Finding ways to include browsing in even relational databases would be a great step. In a sense, we have that already – it's called refinement of queries. But this is still far from being as user friendly as browsing.

Visualization Another eye-opener in the success of the web is the value of visualization. A picture is worth a thousand words (in terms of bandwidth too!) and people are willing to pay a thousand times more. Again, this is not a new issue, but a new scale. The "democratization of data" will push for more visualization than ever before.

Consistency The web, for the most part, has found a trivial solution to the consistency problem – it managed to completely ignore it. The fact that almost everyone got away with it should make us think. Strict consistency requirements are probably unachievable for the web as a whole. Some of the research questions that should be addressed are how to define more relaxed consistency requirements, how to detect consistency problems in the web, and how to integrate systems of different levels of consistency requirements.

Pricing Schemes Charging for access is not yet common on the web, but it may become so. Phone companies have incredible systems able to charge transactions of a few cents. We may need an order of magnitude lower charges and an order of magnitude increase in scale. It's way too early to predict how commerce, and in particular pricing, will be established on the web. But it will be naive to ignore this issue, which could become a major issue in the design of future databases. Copyright issues will also be of major concern. Already database companies try to change their licensing to allow companies to allow queries on the web. But things may change completely. Database techniques can also be *used* for supporting and enforcing copyright.

Doing Research on the Web The web presents unique opportunities for research. Not only can we disseminate results instantly, conveniently, and cheaply, we can discover – by searching the web – the impact of these results much better. Along with doing research about the web, it's time to rethink the way we do research with the web.

Acknowledgements Rick Snodgrass contributed many ideas and comments to this article. Partial funding was provided by the National Science Foundation under grant CCR-9301129, and by the Advanced Research Projects Agency under contract number DABT63-93-C-0052. The information contained in this paper does not necessarily reflect the position or the policy of the U.S. Government or other sponsors of this research. No official endorsement should be inferred.