**PIONEERING TECHNOLOGIES
FOR A BETTER INTERNET**

From the
World Wide Web (WWW)
to a
World Wide Data Base (WWDB)

Dr. Donald Cohen
Dr. K. Narayanaswamy
(Contact Email: swamy@cs3-inc.com)

Cs3 would like to explore the possibility of building new information delivery services that seem to have tremendous technology and business potential. The basic idea is to push the boundary of what is possible for Internet search.

THE OPPORTUNITY

The WWW makes a lot of data available to users but provides no mechanisms to help them process that data automatically. As a result, there is no way to combine the data from different web pages. Even in the case where the user has a conceptually simple question in mind and knows how to locate all the data, he may need to expend an inordinate amount of manual effort to retrieve and combine many pages to derive the ultimate answer to his question. We think this kind of work can, and should, be done automatically by software in response to a single user query. We view the services we describe as extending Google and Yahoo "advanced" search facilities.

As a concrete example of this problem, consider a user viewing a list of Neil Young concert events in one page. Suppose he wishes to find the cheapest way to attend a concert from his hometown. The current WWW infrastructure requires him to copy the data from that page (dates and cities) into other web pages where he can find airfares to those cities on those dates. In fact, he may have to enter the same data into many different pages (for different airlines), along with additional constant data, such as his hometown. He then has to read the result of each query and keep track of the best combination of city, date, flight and price.

The user benefits of a solution to the "*web page data fusion*" problem, as we term the above problem, are quite evident even in a small example as above. The benefits will be even more pronounced with more data (e.g., more concerts or airlines) or more complex queries. A solution to the problem also appears to have good market potential. As a start, such a solution might most easily be provided in the context of "advanced search" over corporate Intranets. Gradually, we would hope to extend such search facilities to work for external web pages to which we have legal access. We also believe that the solution could result in a product for users to search different data repositories to which they currently have access, for example their own file systems.

TECHNICAL APPROACH

We defer major technical details for later discussion, but we do wish to provide a feel for the overall technical approach. Our solution to the problem is to build software that allows end users to view web pages, files, etc. as sources of data. We call the resulting system "*World Wide Data Base*" or *WWDB*. Much as relational database users can formulate queries that span different relations, users of *WWDB* can formulate queries that combine the data from different web pages.

The *WWDB* would require a small amount of human effort per web page that contains machine-readable data to identify and describe the data to be extracted. This data is then interpreted as sets of tuples of "virtual" relations[1]. Certain data sources might offer complete database functionality, either as a web page or other interface, and should be easy to integrate into *WWDB*. Therefore, we focus on the harder case, where the available data offers only a small part of full database functionality, e.g., web pages with search URL's typically offer the ability to generate tuples related to particular user inputs.

The other important aspect of *WWDB* is the ability to search a space of possible algorithms in order to find the cheapest one to answer a query. The algorithms must be composed using just the interfaces that are actually available for the virtual relations involved, making this problem quite different from traditional query optimization for relational databases.
Cs3 personnel have been involved in developing this technology since the 1980s. Additional technical background can be found in:

❖ http://www.ap5.com

❖ http://www.triggerware.com

We also have additional technical papers that have appeared in the literature, which we will be happy to forward upon request.

TECHNICAL FEASIBILITY

Many of the key building blocks described above (e.g., support for non-standard implementations of relations, query optimization over such relations, triggering support for such relations) have been developed and commercialized in other contexts by the Cs3 team. This technology and Cs3's expertise will be available to the proposed collaboration, thereby reducing the "time to market" for a capability like *WWDB*.

While we believe that we have a good understanding of many of the technical issues involved with implementing the *WWDB* facility, we are certain that additional technical issues will emerge as we endeavor to make this into a commercial service. A service such as *WWDB* also raises a variety of business issues and legal issues (e.g., agreements with information providers about the terms of use of data), which we think we understand to a limited extent. It will take a company with sufficient resources to surmount these significant business and legal challenges.

If there is any more information we need to provide, please do not hesitate to ask us. We look forward to a response from you soon.

---

[1] There is no requirement of a 1-1 correspondence between web pages and virtual relations).