Displaying Dynamic Clusters in Haystack

Jaime B. Teevan

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: When information changes while being used, it is difficult to understand how the modified information should be presented to the user. One could imagine imposing the requirement that the data a user works with not be modified while it is being used. But this means that a user might miss valuable new information. Or one could imagine a system where data is changed immediately without concern for the context the user may already have developed. The challenge is to find a system which allows a user to have access to new information while not confusing him or her.

Motivation: It is important to be able to display information that is changing beneath a user because people are no longer in absolute control of their data. This is clearly evident when browsing online catalogs, where the catalog contents may be changing even while the user is viewing it. But it will also become more and more evident with personal data as well, as people begin to trust their computers to modify the large amounts of personal information they have, amounts so large they have no choice but to allow an automated process to organize it.

Additionally, not every problem can be solved quickly. Many of the more sophisticated things computers are beginning to do, such as learning, dealing with high dimensional data, and organizing large amounts of information, are slow processes that are not likely to be sped up even as processor speeds continue to increase and algorithms continue to improve. Being able to use the intermediate information available from these slow processes by allowing the user intelligent access to it will make these process more useful.

Previous Work: CHI '98 had several papers about visualizing dynamic information, but these papers, like most work done with dynamic information, are primarily about visualizing trends in the changes rather than effectively updating information while it is in use. Workspace issues in general, such as the work done by Cousins et al. with DLITE [1], are related, but they don't really address the specific issue. The space is so huge and the problem so difficult that no comprehensive, satisfactory solution exists. Our work is targeted at a specific approach, namely personal information management, where little has been done to date.

Approach: We are looking at the problem of displaying dynamic information within the context of Haystack. Haystack is a information retrieval system to designed to help the user manage his or her personal data. In a way, Haystack can be seen as maintaining a bookshelf of the user's personal information. It is a collection of information built up over time that reflects the needs and knowledge of its owner.

However, Haystack is a bookshelf that the user does not have complete control over. It is as if someone on the other side of book shelf were constantly trying to organize it in a better way. The system is constantly looking at the data and drawing new connections. Additionally, the user's interactions with his or her data change the manner in which the data is stored.

One specific way in which the data changes with out the user's direct control is clustering. It is often useful within Haystack to organize related information by clustering. With many clustering algorithms an initial, crude set of clusters can be arrived at quickly. These clusters are then refined over a period of time, causing documents to change clusters, and cluster names to change slightly. Additionally, the system may find, or be introduced to, new information that was not present in the set of documents being clustered but is related to a specific cluster. We are investigating ways to display these dynamic clusters to the user.

Difficulty: One of the most difficult problems with displaying information that is changing as it is being used is that users place a strong emphasis on consistency and context. Displaying new information in a consistent way, or deciding when new information is important enough to break a users context is challenging.

Also, the algorithms that are changing the information may not always be doing so for the better. Users may notice the mistakes more often then the successful changes, and perceive the changing of the information to be a burden and confusing rather than helpful.

Impact: However, this difficulty highlights a significant benefit of being able to display changing information to the user. This is that the user becomes intimately involved with the processes that are changing the data. The user can easily incorporate his or her personal opinions, and can even stop mistakes as they are happening, before they have been compounded by being built upon by other processes. Relevance feedback from the user becomes both a natural and integrated part of the user interface.

Future Work: We have developed many ideas with respect to the specific problem of clustering. These ideas must now be put before users and their opinions collected and incorporated into the design. Additionally, one can apply reasoning similar to what we have used for clustering to many different areas where information changes, to see what pieces are universal, and what are domain specific.

Research Support: This material is based upon work supported by Merrill Lynch under the Financial Technology Education Initiative, by NTT, by the Oxygen project, and by the National Science Foundation. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References:

[1] S.B. Cousins, A. Paepcke, T. Winograd, E.A. Bier and K. Pier. The Digital Library Integrated Task Environment (DLITE). *Proceedings of the ACM Conference on Digital Libraries*, pp.142-151, 1997.