HISTOGRAM-BASED VISUALIZATIONS FOR LARGE TIME DEPENDENT DATASETS

John Roberts
San Francisco State University
1600 Holloway Ave., San Francisco, CA 94132
jrob@sfsu.edu

Edward Lank
San Francisco State University
1600 Holloway Ave., San Francisco, CA 94132
lank@cs.sfsu.edu

Jim Gemmell
Microsoft Bay Area Research Center
455 Market St., Ste. 1690, San Francisco, CA 94105
JGemmell@microsoft.com

ABSTRACT

In this paper we describe the design of prototypes of histogram-based visualizations for browsing large, time-dependent collections of data. These visualizations are intended to provide an alternative to the standard hierarchical file browsing metaphor currently available, and are used to expose time-related information from the underlying dataset that would not be apparent using standard file browsers. The visualizations are also designed to provide a method of browsing that does not require knowledge of the underlying file storage architecture, as well as to be able to scale to large data stores. Two prototype visualizations are designed and presented to a small group of users to determine if the histogram-based metaphor for file browsing and visualization of large data collections is intuitive and usable.

Keywords

Graphical User Interfaces (GUI), File Browsing, Information Visualization.

1. INTRODUCTION

The MyLifeBits project is a research project dealing with the storage of large collections of data. As the name implies, the emphasis is on an individual's personal information. The MyLifeBits store is designed to be a complete archive of an individual's life. As an initial aspect of this research, the entire life of Dr. Gordon Bell, a research scientist at Microsoft's Bay Area Research Center has been scanned and archived [1]. On-going aspects of this project involve capturing even more aspects of Dr. Bell's life, including phone calls, electronic bills, and even meetings and notes.

The motivation for this project arises from the availability of large, inexpensive storage media. As storage becomes more affordable, finding space for any single piece of data is easy. Organizing that data and presenting compelling interfaces to search and browse inside that data set remain open problems [1, 2, 4, 10].

In this research, we detail our work on visualizations to understand a large set of documents and to browse into this set of documents. We base our visualizations around temporal histograms of data.

The histogram is a common form of graphical presentation of a frequency distribution. It is often used in mathematics to provide a concise visualization of frequency distributions that may expose relevant statistical data from the underlying dataset, such as modes or skew. This research seeks to determine if histograms can be effectively used as a visualization for file browsing.

The paper proceeds as follows. In Section 2, we describe related work. Section 3 outlines our visualizations. Section 4 describes the results of our early user testing. We conclude by summarizing the current status of the project and describing our on-going work on information visualizations.

2. BACKGROUND

Traditional hierarchical browsers are dependent on the underlying file system. This dependency presupposes user knowledge of the layout of the file system. Though this is not a problem for experienced users, research has shown that most users do not understand the folder/file hierarchy, and those that did had a hard time making it work for their filing needs [8]. The traditional storage system also requires that users know the location of documents that they wish to view. Given the size of current storage devices, it is difficult for most users to accurately recall the location of a document, especially as the number of files in the storage system increases [5].

In addition, hierarchical browsers provide only one visualization of file organization, displaying file location within the storage system. Though it may be possible to sort the data collection for a given folder based on various attributes, the resulting sorted list generally does not provide any additional information to the user that might expose relationships between individual members of the dataset or among data sets in different folders. As well, as the number of files gets very large, it becomes impossible to view more than a small number of those files on the screen.

Some attempts to address these usability issues have emphasized alternative visualizations of the hierarchical structure [6] or removing the hierarchical representation of files to users, while preserving the hierarchical storage system as a means of efficient archiving of files [8]. Other approaches include the creation of alternative storage approaches by going beyond the traditional file system approach. These systems include the Placeless documents project [2], the Haystack project [4], and the MyLifeBits at the Microsoft Bay Area Research [1].

The work described here is related to the MyLifeBits project. The MyLifeBits storage system stores collections of data, along with the relationships between collections, and enables access to the stored data through the use of complex queries and pivots. In order to increase the usability of such a storage system, visualizations are required to allow the user to browse large subsets of items in the store [1].

In a project such as MyLifeBits, the need for browsing is very important. Consider searching through your entire collection of digital photographs for a certain family vacation, or searching for a picture of your child on his or her first day of school. While a power-user with a degree in computer science might be able to specify a set of search terms that would retrieve that single document, a typical computer user would find specifying such a complex query challenging. However, even users adept with computer searching often do not specify complex queries. For example, Teevan et al. studied the web searching habits of a group of seven MIT Ph.D. students [10] as a component of the Haystack project [4]. They note that typical search patterns involved an "orienteering" approach to locating material on the internet. Rather than specifying a complex query, these users seek a location from which they can find data. Teevan et al. cite two primary reasons for this [10]. First, users often do not make the effort to specify formally and completely a full set of attributes to locate a specific piece of information. Second, it is often the case that users do not know **exactly** what they are looking for (consider seeking the professor who does research most closely related to HCI at a certain university…).

When browsing large sets of files, temporal-based browsers have been considered as one option for organizing data. Both the Lifestreams project [3] and the Time-Machine Computing project [7] use variants of timelines to represent files. These projects provide a one-dimensional temporal representation of the data.

In our work, we expand temporal visualizations to two dimensions, incorporating a notion of frequency in time. To accomplish this, we use temporal histogram visualizations to extend typical hierarchical file browsing in a manner independent of the underlying data store.

3. BROWSING WITH HISTOGRAMS

The histogram-based visualization is intended to allow users to immediately identify intervals of time in which portions of a dataset have been "active". This activity could reflect file access statistics, such as periods of time where file creation or file access peaked. This concept of dataset activity could also encompass time references in the files composing the dataset, allowing the user to browse collections to find files that refer to specific dates.

Using a histogram to browse a collection of files that are somehow time-dependent obviates the requirement that a user is familiar with the underlying file storage system. The histogram-based visualization itself is not dependent on the underlying file storage system, and can easily adapt to changes to the collection of data without direct user intervention. Even though the histogram-based visualization is independent of the storage system, it can easily be incorporated into a storage system that provides support for basic querying operations, such as the MyLifeBits lifetime storage system. Finally, the histogram-based visualization is able to accommodate any dataset that is time-dependent, rendering the file storage system irrelevant in this browsing context.

The basic visualization is generated by collecting a set of files that are somehow time-dependent, either by virtue of the file's creation date or by the presence of dated information within the files. The range of time covered by this dataset is calculated, and the files are distributed into bins that represent intervals within the total range of time. A basic histogram can then be generated using a count of the number of files in each bin to provide a bar height for each bar of the histogram in the visualization. An example of the basic visualization is presented in Figure 1.

3.1 Visualization Interface

Two different visualizations were created, but all of the visualizations seek to preserve a sense of context within the dataset as the user interacts with the visualization, while simultaneously presenting a focus area which allows the user to "zoom in" on portions of the dataset while browsing for a particular file. The user can click on the focus area in the visualization to view the files contained within the focus as icons. The user can then click on the icon in the standard manner in order to open the file.

The visualization was created to respond to standard mouse generated events, such as dragging, and right- and left-

clicking. In all of the visualizations, dragging causes a selection cursor to move over the context area in order to reflect the data that will be selected for display in the focus area. Left-clicking causes the magnification of the focus area to increase, while right-clicking returns the user to a previous magnification level. Some visual cues are provided to inform the user of their current location in the dataset, including the start and end times of the dataset, the start and end times of the files contained in the focus area, and, when applicable, the number of files in each area.

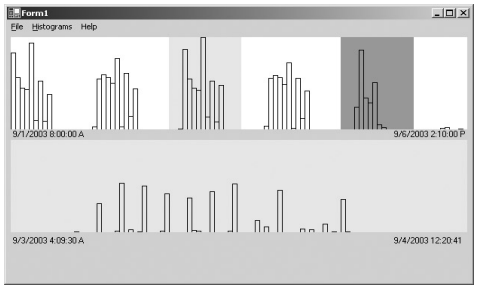

Figure 1: A pop-out histogram visualization.

3.2 Visualization Prototypes

Two different prototype visualizations were designed, each of which gives a slightly different view of the focus area and its context. These visualizations were named Pop-Out and Mound. The basic functionality of the interface was preserved across the different visualizations to maintain continuity for the purpose of user trials.

3.2.1 Pop-Out Visualization

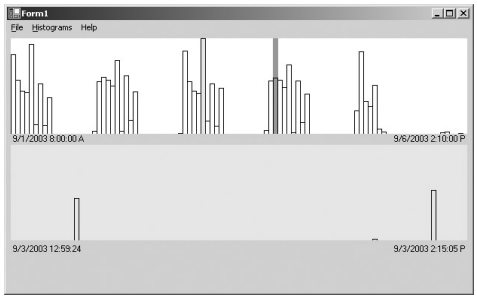

Figure 2: Zooming behavior of the Pop-Out visualization.

The Pop-Out visualization, presented in Figure 1, presents the entire dataset in the form of a two histograms. The top histogram represents the entire data set, and includes a highlighted area (shown as the darker region) indicating a focus area. This highlighted subset of the upper histogram is represented in the lower histogram. As the user zooms into the dataset presented in the Pop-Out visualization the focus area changes to represent an ever-decreasing fraction of the total dataset. This behavior can be seen in Figure 2.

The lighter highlighted region represents the cursor location. A user clicking at the indicated location would magnify that portion of the histogram. By clicking in the lower histogram, a user is able to browse further into the data set, looking at smaller and smaller subsets of the data set.

3.2.2 Mound Visualization

The mound visualization seeks to present both the focus and context in the same histogram, as can be seen in Figure 3. The entire dataset is represented in one histogram, and the area of focus is distorted to appear like a mound in the histogram.

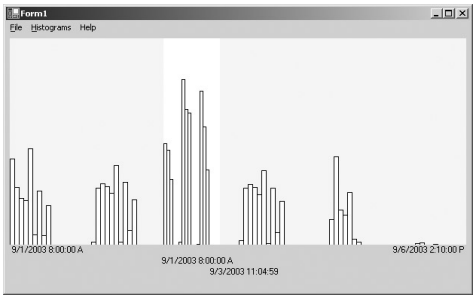

Figure 3: The Mound visualization.

4. EVALUATION METHODS

The primary goal of this research is to determine if users find the histogram-base visualizations we designed to be an intuitive method of browsing large collections of time-dependent data. We conducted an early user trial which allowed thirteen users to interact with the prototypes. A questionnaire was designed to obtain user input about the effectiveness of the histogram-based visualizations, as well as to gather user suggestions for improving the prototypes.

The dataset used for the user trial consisted of a collection of web pages, each of which contains the schedule information for courses offered at San Francisco State University during the Fall 2003 semester, as well as hyperlinks to the bulletin description of the course. The dataset was distributed over the histograms according to the start time of each course. The users were given a brief orientation describing the dataset, and asked to locate a number of courses in the data set. The questionnaire asked the user to quantitatively assess certain attributes of the visualizations, such as their ability to understand the visualization and distribution, ability to navigate in the dataset, and ability to locate information in the data collection. The user was also asked to rate the difficulty level of each of these tasks on a scale from one to five, where one indicated an easy task, and five indicated a difficult task. The quantitative results are summarized in Table 1.

The results of this task analysis indicate that users found it easiest to interact with the Pop-Out Visualization in our early-stage prototypes. The Mound Visualization was ranked second in all of the categories. In general, users stated that the integration of the focus and context in the Mound visualization was confusing, preferring the sharp

separation of the focus and context areas in the Pop-Out visualization.

	Visualization	
Ability To:	**Pop Out**	**Mound**
Visualize Position	1.71	3.07
Locate Information	1.71	2.64
Relocate	1.79	2.29
Understand Distribution	1.93	2.71

Table 1: Summary of quantitative data

Finally, users were asked to compare the various visualizations. Approximately 2/3 of the users ranked the pop out visualization first.

5. CONCLUSION

We introduce a histogram-based visualizations to access large, time-varying aspects of data sets. We describe two prototype visualizations of data. Of the two prototype visualizations designed, users indicated that the Pop-Out Visualization was most intuitive for the purpose of locating items in a simple, time-dependent data set of university courses.

A number of improvements can be made to the prototype visualizations designed. Some basic user interface improvements, such as introducing a modifiable look and feel, and smoothing some of the graphics could be performed to provide a cleaner interface. Along with these improvements, some of the more common user suggestions (colorization and additional visual clues as to current location in the dataset) can be incorporated into the next iteration of the histogram-based visualizations. Also, additional distortion methods may be designed and tested that attempt to make use of the suggestions obtained in the user trial to increase the usability of the visualizations.

Since one of the desired effects of the creation of histogram-based visualizations for browsing was to expose time-related information from the underlying dataset, the next iteration will also be designed to allow the inclusion and browsing of multiple datasets simultaneously. It is our hope that this will allow the user to draw conclusions about the relationships between seemingly unrelated datasets by observing the corresponding peaks in the histograms representing each dataset. It is also true that effective visualizations are often task-dependent. More complex tasks may require modified visualizations to support effective access.

6. ACKNOWLEDGMENTS

We would like to thank everyone who took time to participate in our user study. Funding for this research is provided by Microsoft Research.

7. REFERENCES

[1] Bell, G., Gemmell, J., and Lueder, R. The MyLifeBits Lifetime Store. ACM SIGMM 2003 Workshop on Experiential Telepresence (ETP 2003), ACM Press (2003).

[2] P. Dourish, K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. Terry and J. Thornton, "Extending Document Management Systems with User-Specific Active Properties", ACM TOIS 18:2, 2000, pp. 140-170.

[3] E. Freeman and D. Gelernter, "LifeStreams: A storage model for personal data", ACM SIGMOD Bulletin 25:1, March 1996, pp. 80-86.

[4] D. Karger and D. Quan, "Haystack: a user interface for creating, browsing, and organizing arbitrary semi-structured information", Extended Abstracts of CHI 2004, Vienna, pp. 777 – 778.

[5] Marsden, G., and Cairns, D., "Improving the usability of the hierarchical file system", Proc. 2003 South African Institute on Enablement through technology, pp. 122 – 129.

[6] Pirolli, P., Card, S., Van Der Wege, M., "The effects of information scent on visual search in the hyperbolic tree browser", ACM TOCHI 10(1), pp. 20 – 53.

[7] J. Rekimoto, "Time-machine Computing: A Time-centric Approach for the Information Environment", Proceedings of UIST 1999 Asheville, pp. 45-54.

[8] Rose, D., Mander, R., Oren, T., Ponceleon, D., Salomon, G., and Wong, Y., "Content awareness in a file system interface: implementing the "pile" metaphor for organizing information", Proc. 16th SIGIR Conf. on Research and Development in Information Retrieval (1993), pp. 260 – 269.

[9] Sullivan, K. The Windows 95 User Interface: A Case Study in Usability Engineering. Proceedings SIGCHI, (1996), 473-480.

[10] J. Teevan, C. Alvarado, M. Ackerman and D. Karger, "The perfect search engine is not enough: a study of orienteering behavior in directed search", CHI 2004, Vienna, pp. 415 – 422.