

Computational Analysis and Visualization of Electronic Records Collections

MARIA ESTEVA, JAYA SREEVALSAN-NAIR, WEIJIA XU, ASHWINI ATHALYE, AND MERWAN HADE

Abstract: The motivations for our research are analysis and visualization of collections of electronic records for archival purposes. We present two interactive approaches: a) a treemap visualization of NARA's test-bed electronic records collections and b) an application to compute and visualize relationships between electronic text records from a private organization.

Treemaps can be used to represent properties extracted from collections of electronic records and stored in a database. In a treemap representation, directories are grouped and reflect the hierarchical structure of a collection by partitions of image space to show the branches and leaves of the file structure. In turn, fill colors indicate different thresholds for size, number of distinct file types, and number of files in a given partition. As a result, information such as the collection's organizational structure, number and type of folders and files by groups and series, and location and distribution of contents according to their folder naming can be searched and visualized to obtain both an overview and a focused understanding of electronic records collections.

To identify records belonging to a same activity from collections of unstructured text, we calculate paragraph alignment. We compute a distance matrix in which the similarity of two records is the closest distance between the fragments of the two. The matrix, which also preserves existing information about the records provenance (author and or function of origin), is displayed using a graphical user interface (GUI) showing a layout of related records ranked according to their similarity. Through the GUI, an archivist can select a query record and visually identify relationships between records and their authors

These visualizations enhance understanding of the structure, patterns, and contents of electronic records collections and help in determining solutions for their appraisal, their discovery, and their long-term preservation.

About the authors:

Maria Esteva has a Ph.D in Information Science from the University of Texas at Austin with focus on digital archiving and electronic records analysis. In her research she uses text mining and visualization methods to make sense of unstructured collections of electronic information. Currently she works full time as a researcher and data archivist at the Texas Advanced Computing Center (TACC).

Weijia Xu is a full time researcher at TACC. He received his Ph.D from the Computer Science Department, at the University of Texas at Austin with focus in data management and analysis. Dr Xu has published in scientific database development, efficient proximity search methods for information retrieval, and information visualization. He currently is co-PI on a NIH funded project to develop computational foundations for comparative sequence analysis based on relational database.

Jaya Sreevalsan-Nair has expertise in the field of scientific visualization for structured and unstructured grids, scattered, and multidimensional data in scalar, vector, and tensor fields. Her recent research is directed towards various algebraic transformations that can be applied to any form of data to convert it to manageable format, for applying visualization techniques. She has experience building graphical user interfaces for various visualization applications. Her Ph.D is from the University of California at Davis.

Ashwini Athalye and ***Merwan Hade*** are graduate and undergraduate students respectively in the Computer Science Department at the University of Texas at Austin. They are student assistants in TACC's Visualization Laboratory.