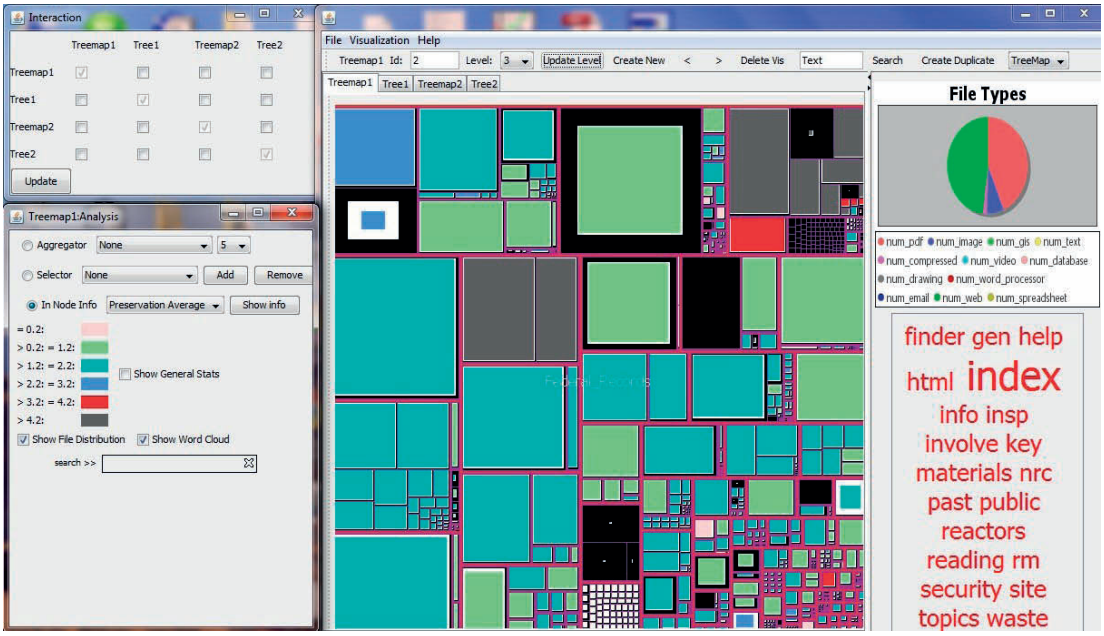
Analysis of Large Digital Collections with Interactive Visualization

Weijia Xu* Maria Esteva Suyog Dutt Jain Varun Jain

The University of Texas at Austin



ABSTRACT

To make decisions about the long-term preservation and access of large digital collections, archivists gather information such as the collections' contents, their organizational structure, and their file format composition. To date, the process of analyzing a collection — from data gathering to exploratory analysis and final conclusions — has largely been conducted using pen and paper methods. To help archivists analyze large-scale digital collections for archival purposes, we developed an interactive visual analytics application. The application narrows down different kinds of information about the collection, and presents them as meaningful data views. Multiple views and analysis features can be linked or unlinked on demand to enable researchers to compare and contrast different analyses, and to identify trends. We describe and present two user scenarios to show how the application allowed archivists to learn about a collection with accuracy, facilitated decision-making, and helped them arrive at conclusions.

KEYWORDS: Digital collections, archival analysis, visual analytics, data curation

INDEX TERMS: I.6.9.a Applications; I.6.9.c Information visualization

1 INTRODUCTION

Archivists provide services to corporations, research institutions, the government, and to the public at large by managing and providing access to varied types of digital collections. These functions require archivists to analyze collections to identify their scope and contents, determine the way in which the digital records are organized, and decide which actions are needed to make the collections accessible over the long term.

Traditionally, archival analysis tasks are undertaken in a pen and paper fashion [1]. Given the scale and complexity of modern digital collections, the archival community pointed out the need to investigate new methods to complete these tasks [2]. In response to this demand, we developed a visualization application that enables analysis tasks related to collections management and access. Throughout this project we sought feedback from archivists in iterative design fashion [3].

The application uses structural and technical metadata—automatically extracted from the collection—to represent the collection as a treemap [4], and provides analysis functionalities based on metadata aggregations, categorization, and data mining. All metadata and the analysis results are stored in a Relational Database Management System (RDBMS) through a one-time processing step and retrieved on the fly. As users interact with the visualization, diagnostic views showing the size, technical composition, and organizational structure of the collection are rendered at different levels of abstraction. These views are useful to detect general trends as well as to identify and characterize distinct groups of digital objects.. The visualization's interactivity features enable users to explore a collection from overview to details and compare and contrast multiple views.

For ease in mapping archival and computer science terms, in this paper, *digital record* is a simple or complex digital object identified by the archivists as a stand-alone unit of analysis. Given that archivists establish what constitutes a record, our visualization aims to clarify the structure of digital objects to

*email: {xwj, maria, suyog, varun}@tacc.utexas.edu

IEEE Symposium on Visual Analytics Science and Technology
October 23 - 28, Providence, Rhode Island, USA
978-1-4673-0013-1/11/$26.00 ©2011 IEEE

facilitate the decision on whether they constitute a record. In turn, we use the term *record group* to define groups of digital records as simple and or complex digital objects.

Our main contribution is that relevant analysis variables are integrated visually to enable a comprehensive understanding of the many information layers available within large collections. To the best of our knowledge, this is a first-of-its-kind visual analytics tool for archival collections management and access. The application includes improvements to the treemap visualization through pixel-based rendering and glyph-based techniques designed to show additional information dimensions.

The rest of the paper is organized as follows. Section 2 describes motivation, requirements, and design considerations. In section 3 we present related work. Section 4 details the techniques that we developed for visual analysis. Section 5 presents two user cases showing analytical workflows using the visualization. We conclude in section 6 with a discussion of the challenges and limitations of the current implementation.

2 Motivation and Design Considerations

Our motivation was to build an interactive visualization system to support digital collections analysis. The goal of archival analysis is to obtain a comprehensive understanding of collections in order to make decisions about how to store, organize, describe, and preserve them. We started our design with a series of discussion sessions with archivists to understand their information needs and to identify the visualization requirements. In the design, we present a four-information component model (See section 2.2) to analyse digital collections using corresponding visual representations.

2.1 Requirements analysis

The cognitive activities involved in collections' analysis are non-linear, sense-making processes during which archivists make different observations, and pursue, alternative thinking paths. Archivists evaluate collections in relation to their structure, size, contents, context, and provenance. Types of information required for analysis include:

***Structure and organization*:** How digital records are arranged in a directory hierarchy is understood as the collection's structure. Structure is important to identify the original order in which records were organized, to trace their provenance of authorship or the function that originated the records, to understand relationships between records by which they form groups, and in some cases to maintain their functionality. Preserving structure is one of the archivists' tenets, as it allows to maintain the authenticity of the collections and to provide access.

A collection may be formed by distinct groups of records, each of which may have large amounts of simple and complex digital objects of diverse file formats. In turn, groups of records may be organized thematically, by date, sequentially, or by naming convention. Within one collection, multiple organizational criteria may coexist. In addition, some software requires that the files be organized in specific configurations in order to render them correctly (e.g. GIS), and many archival collections are organized to map the functions of the organization that created them, which points to the collections' provenance.

Content: Beyond the intellectual content of individual files, descriptive information about the contents of a collection may be recorded in directory labels and file names. Labels may contain subject terms, proper names, time periods, and provenance information. Collections' creators manually place this descriptive metadata, which may be more or less precise.

Statistics: Information about the size and number of files of the different groups of records in a collection informs storage and long-term preservation decisions. For access purposes, these statistics indicate the scope of records that a user will have to search through in order to find specific information.

Technical composition: Knowing the file formats in the collection and their location in the structure is needed for access purposes (e.g. to establish where the videos or images are in a given collection), to identify types of simple and complex digital objects, and to preserve them over time.

Context: Refers to information about the collection's functions and creators that that is obtained from external sources. This information is typically analyzed in relation to the rest of the points that we mentioned above, and provides users with a thorough understanding of how and why the collection exists.

Influenced by a long tradition of physical archival materials, current digital collections analysis is generally conducted combining manual methods and technical tools. Reading and observing the records and or the labels that describe them allows identifying collections content [1]. Structure is explored observing and counting directories and files using file management software. In turn, file format characterization software is available to learn about the collection's technical composition [5]. During analysis, archivists document, compare, and combine these discrete pieces of information into a comprehensive assessment of a collection.

Digital technologies are changing the traditional archival analysis landscape. Increasingly, digital collections are larger, and their structure and technical composition has diversified [6]. Along with the collection, the data points to explore during analysis grow exponentially, and there is a risk that the evaluator will get buried in details or lost in generalities [7]. We apply visual analytics to facilitate, organize, and improve the cognitive processes involved in archival analysis. The challenges of this project include selecting analysis methods to narrow down large amounts of information to manageable levels, and to visually represent these many layers of information meaningfully and in an integrative fashion to improve comprehension. The visualization should allow archivists to incorporate their experience into the analysis as well as to make inferences about the collection in relation to contextual information.

2.2 Visualization model for collection analysis

For the purpose of building a visualization model to address the characteristics laid out above, the properties of a digital collection can be generalized as metadata about each digital object, as well as about the relationships between them. We characterize digital collections as having four information components:

***Hierarchical organization*:** In our application, structural information is presented so that users can distinguish the collection's components (record groups and digital objects) and their relationships. Although we expect most large collections to be organized hierarchically, our visualization can also be used for non-hierarchical ones.

Numerical metadata: Refers to any numerical value that is associated with a collection component (from a digital object to a group of digital objects), such as size and number of files. In our application we present values as aggregations, distributions, data mining scores, and any other statistical calculations.

Additional relationships among digital objects: Are relationships other than those presented by the natural hierarchical structure. These relationships may result from analysis methods such as clustering or classification.

Descriptive metadata: Refers to any type of text description associated with a digital object in the form of its naming

convention, descriptive tag, or as the label of the directory that contains it.

Our visualization system is designed to show these four information components. It includes analysis methods that summarize large data by means of aggregation, classification, data mining, and statistics. Throughout the design we made decisions about how to show relevant collection features, and about the information pieces that needed to be visually integrated for a better understanding of the collections structure, contents, size, and organization. Flexibility to pursue different analysis paths is possible by selecting different visualization interactions while at the same time organizing and keeping track of them. As a first step, archivists will use the visualization to study a collection's general trends and to obtain precise statistics. Based on a preliminary evaluation, he or she can pursue more targeted analysis workflows of specific record groups and digital objects.

2.3 Scalable architecture for data handling

For efficiency in rendering large data we adopt the Model-View-Controller architecture (MVC) [8] in which *Model* (data) is divided physically in two locations, the client machine and the database server. During a visualization session, only the data needed for viewing is transferred from the remote database to a local repository. The data is maintained locally, and used on demand until the user closes the visualization session. In this project, *View* refers to different types of visualizations which are also rendered on demand. Each *View* has its own controller and a top level controller manages all the available and the rendered visualizations. This two-tiered control model integrates different visualization libraries (Prefuse [9], OpenCloud [10] and JFreeChart [11]) into a unified framework, and provides flexibility for future expansion.

3 Related Work

Although there is abundant research on the use of information visualization to analyse texts, information hierarchies, and multidimensional data, to the best of our knowledge, ours is the first work that integrates all these components into a framework for the purposes of managing and providing access to large digital collections. Below we review work in these areas in relation to our selection of techniques.

3.1 Hierarchical information visualization

Common techniques to visualize hierarchical information may be divided in two models: node-link based visual representation, and space-filling oriented representation. A common example of the former is the File Explorer. This type of tree representation is good for exploration purposes in which branches of the tree can be hiding [12]. However, for the purposes of representing large hierarchical collections, the rendering of links requires additional space. This model also presents challenges to determine connectivity and to compare between nodes.

The space-filling oriented representation is more compact, as it does not render links between nodes. Instead, the hierarchical relationship is indicated by the arrangement of the nodes. There are two types of layout: tiled and nested. In the tiled layout, nodes are drawn next to each other without overlap. In Sunburst Tree [13], the root node is placed in the center of the display, and the children are drawn as arc blocks surrounding the parent node. The tiled layout uses a lot of screen space for rendering internal nodes. A variation of this visualization is the icicle tree [14], which places the root on one side and arranges internal nodes towards the opposite one. While this representation may be more intuitive in some applications, such as when representing disk space [15], nodes at different levels of the hierarchy may be hard to compare.

Treemaps use nested layout, which makes a more efficient use of the screen space [4]. The root is represented by a rectangle, and all the children are placed inside of the root node, presenting the problem that the hierarchical structure may be hard to recognize. In this project we use a squarified treemap algorithm [16] and multiple boundary lines with increased spacing in between to better illustrate the hierarchical structure.

Treemaps are used effectively in information visualization applications such as: threaded discussion forums [17], and Google news stories [18]. Treemaps are also scalable for large sets of data objects, for example in the display of search results [19]. Specifically related to the visualization developed in our project, a number of commercial applications use space-filling representation to display disk usage statistics. Along with the treemap showing all the information of the directory hierarchy, WinDirStat includes an explorer panel that presents aggregated information and a panel that shows file format composition [20]. The size of each square is determined by the size of the corresponding directory on disk, and the color of the square shows the type of file. To emphasize the directory structure, WinDirStat uses a cushion technique [21]. DaisyDisk, an application with a similar purpose available for MAC OS, uses a Sunburst layout to display a disk's hierarchical structure. Although our visualization has some overlapping functionalities with the ones mentioned above, its focus is to support archival analysis, specifically of large-scale collections with multiple properties. For example, large file format data aggregations are presented in relation to classes of files. In turn, the visualization allows archivists to explore the collection's hierarchical structure. As they go from general to detailed views, it is possible to observe the structure and file format composition of digital objects individually and also to see how they form patterns across a record group (See Fig. 2). To map archival principles, classes of files are presented in relation to content and to provenance information.

3.2 Text visualization

Some text visualization applications focus on visualizing text mining results and showing relationships among terms, as well as text patterns in document collections [22] [23]. Other techniques such as tag cloud [24] and WordTree [25] focus on providing visual summaries of the content and relations of text corpus. Fewer research tools focus on supporting visual investigative analysis of text corpora. Jigsaw is an application designed to help analysts discover potential terrorist threats. The application offers multiple data views such as scatter plots and network graphs [26]. Entities are extracted from texts and visualized in relation to other information dimensions such as time and social networks in multiple coordinated views. Shi et al. integrate tag cloud with stacked area graphs to help users understand text corpora through facets [27]. In their approach, a stacked area graph is used to show distributions of different categories of documents over time. The tag cloud is added directly into the area to give users a quick glance of the documents' content.

Our use of text visualization also supports investigative analysis. At different levels of the collection, hierarchy digital objects are associated with their corresponding descriptive terms as tag clouds extracted from directory labels and file names. This text visualization provides users with a summary of the contents of the selected directory, and enables them to make inferences about the types of digital records included. Also, different classes of image tags are presented in relation to provenance information

as pixel based rendering (See case study 2) to allow users to compare classes of images across directories.

3.3 Visualization of multiple attributes

Glyph and pixel oriented techniques are two well-known methods for visualizing multidimensional data. In the glyph-based visualization, each attribute of a data object is mapped to a property such as size, color, length, and orientation of a graphical object. Glyph-based rendering enables an overall visual comparison between two multidimensional objects. This technique has been used to assist in the analysis of network traffic [28] and web search results [29], but is often considered non-scalable for large volume collections. In turn, pixel-oriented visualization is effective to explore large sets of multi-dimensional data. The basic idea behind it is to map each attribute value of a digital object to a color pixel. Different attributes are then displayed in different sub-windows with informative arrangements to help users identify patterns among attributes and digital objects [30]. To facilitate the comparison of numerical attribute values and their distribution we use both horizontal and nested spiral color pixel-filling techniques within the treemap. We also use these methods to show additional relationships among digital objects.

4 Visualization Implementation

Our application provides a variety of visual analytics tools within one framework. It allows flexibility to pursue analysis workflows that can be traced back and revisited. Throughout the analysis, relevant collection features, such as size, file types, organizational criteria, and hierarchical structure, are visually integrated to improve comprehension about a collection in the context of archival practices. Visual information is presented in a consistent fashion from overview to details, and across the collection's structure. To identify and prioritize the collection's management and access actions, the visualization provides decision-making support through the possibility to discover trends, identify patterns, and compare and contrast.

4.1 Collections hierarchical structure visualization.

To represent collections we use treemap. The collection is organized as nodes and edges in a hierarchical structure. The entire visualization space is assigned to the root node, which is the first level directory. Within, child nodes (sub-directories) are rendered as nested rectangles and further offspring are rendered within their parent rectangle. This allows archivists to observe the hierarchical dependencies between sub-directories and thus to understand the relationships between digital objects. From an archival perspective, and in combination with descriptive information, such views allow archivists to identify original order and provenance.

To visually differentiate between sub-directories, we render colored border-lines for each of the levels, and add a small amount of border spacing between them to increase the visualization's readability. The space that each node occupies on the visualization is based on the number of files within. Figure 1 shows an example of how we render structure.

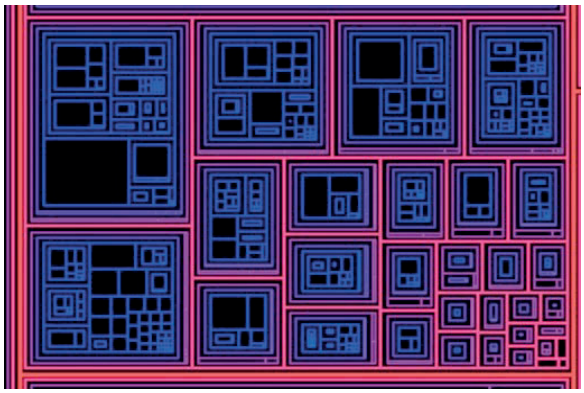


Figure 1. Structural representation of a collection as a treemap and border-spacing on the same root node.

The collection's properties are presented within the nodes. We made this decision so that classes of files, or organizational criteria of files within a directory, could be observed in relation to the size of the directory. During our design process we found that adding spacing between levels not only helps users to quickly determine the level of nesting, but also helps them to easily distinguish neighbour folders showing similar patterns (See Fig.2).

To go from general to detailed views, users are able to render one or more directory levels at a time. They can also navigate between individual directories interactively and use the zoom function to appreciate the border spacing and better understand the structure. For ease of navigation of the collection structure, we use a squarified treemap algorithm [16]. As the user is exploring the different directory levels, the treemap layout remains fixed so that the position of high-level nodes remains stable.

4.2 Numerical metadata visualization

Numerical attributes such as the number and size of files and the types and number of file classes are considered by archivists to understand a collection's scope and manage it. We use pixel based rendering to show value distributions within the treemap. As we have stated above, these views allow an understanding of the technical composition of the collection, helping archivists compare and contrast in order to detect trends and patterns across directories.

During analysis, archivists may need to know how many file classes exist in a collection and which one is the most relevant. Figure 2 shows an example of multiple-attribute rendering in the treemap. In this example, we extract structural (file path) and technical metadata (file format identification) from the collection, and we store that information on the RDBMS. To summarize large amounts of file format information, we classify file formats into 20 classes according to the file's format functions. For example, jpg and tiff files belong to the image class, and html and css files correspond to the web class. In response to user queries, we aggregate statistics at each level of the directory structure. To enable observation of all the file classes available in the nodes at the same time, we implemented horizontal pixel based rendering within nodes. We start by calculating the percentage of a given file class and then we calculate the total number of pixels by multiplying the percentage of that class to the area of the node. To render a color for that class, we start from the top left corner of the node and fill in the exact proportion of pixels. The process continues for all the classes. To highlight issues of data quality, files with unidentified types are rendered at the end of each node in black, and files with more than one possible identified type are shown in grey.

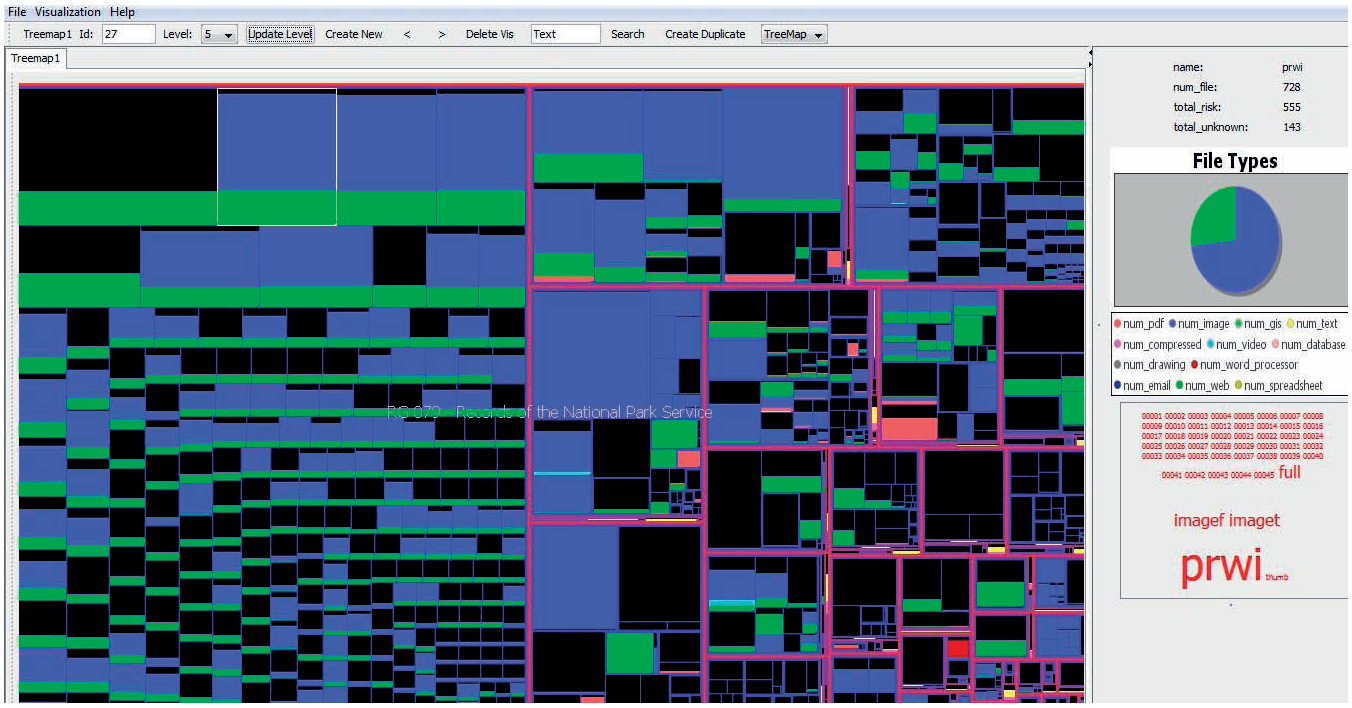


Figure 2. Color pixel rendering of file types in a directory including various sub-directories.

Figure 2 shows a directory (which may be defined by an archivist as a record group) including various sub-directories of different sizes in which existing file classes are rendered. Blue represents images, green web files, black unknown file types and red pdf files. In this view we can observe the file class trends across the directories. Given the presence of color patterns across the record group, it is possible to infer that it is composed of webpages. These webpages include a majority of images and in some cases pdf files. As the user mouses over the directories, a pie chart in the right panel shows the precise percentages of the file classes in a given directory. In turn, provenance information is available in the middle of the visualization (in contrasting white letters), and the individual directory names are shown in the statistics toolbar shown to the right.

During our iterative design process an archivist observed that, "The ability to assess varied characteristics and to compare selected file attributes across a vast collection is a breakthrough." [31] In addition, another archivist commented on how the visualization is effective to evaluate the technical composition of large collections for which there is no prior description.

4.3 Visualizing additional relations

Large digital collections may present numerous additional relationships between digital objects. For example, similarities in the way that digital objects are organized within and across directories can be determined using this method. To show relationships within the treemap we implemented a nested square rendering method. We demonstrate this type of rendering to show the results of a data mining implementation to predict organizational patterns in large collections.

Identifying the way in which records are organized in a collection is one of the activities that archivists undertake for the purpose of providing precise access. Among other criteria, collections may be organized by subject, by type of documents, or by geographical location, and their files may be ordered by a regular naming convention or sequentially by number. In turn, at different levels of the hierarchical structure, digital collections may have more than one kind of organizational criteria (e.g. by date and by geographical location). Knowing the collection's organizational criteria, a user can understand where to look for certain information and how its creator used the information. We implemented a data mining system that uses the terms in the directory labels and the names of the files as metadata to predict organizational patterns.

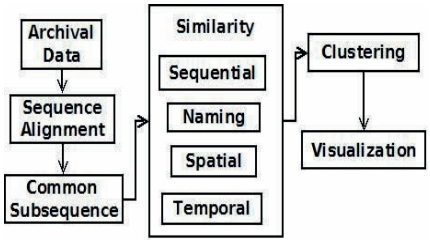


Figure 3. Architecture of the data mining system to predict the organization of collections.

Figure 3 shows the basic architecture of our data mining system which uses WordNet [32] to infer word senses, string alignment to find spatial, sequential, temporal and naming organization of documents, and clustering to form groups of documents that have the same organizational criteria.

We use Gotoh's affine gap cost alignment algorithm [33] to find similarities between the words in the directory labels and between the file names. We tokenize the file and directory names into separate words. Then, we search through the hypernym trees of these words using WordNet to find if their meaning is spatial or temporal. Spatial words such as Texas have a "location" hypernym, and temporal words like March have "calendar" in their hypernym trees. All spatial and temporal words are replaced with a string of special characters to ensure their alignment when using a sequence alignment algorithm. Numbers and words with similar naming conventions are also aligned (See Figure 4). We

find sequential, naming, spatial, and temporal patterns in the file names and calculate a similarity score for each of the patterns.

We ran our algorithm on the entire collection (> 100 high level directories) For every directory, we also derived a feature vector of the form: <Sequential, Naming, Spatial, Temporal >. Then we cluster all the directories into 4 different groups based on the similarity values of each dimension in the feature vector. Finally these groups are visually represented.

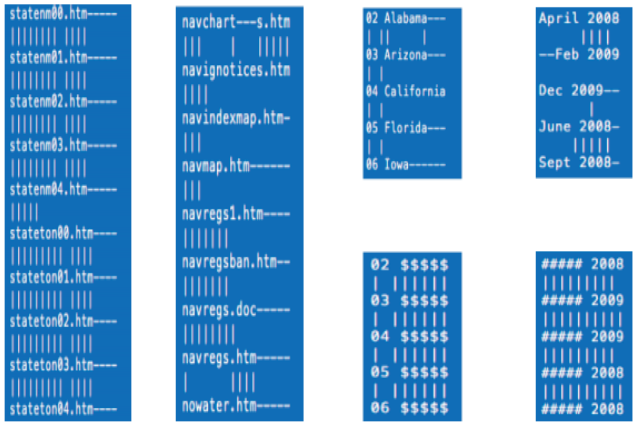


Figure 4. Four examples of pattern detection by the alignment algorithm.

From left to right, we see sequential, naming, spatial, and temporal patterns found by the sequence alignment algorithm. The spatial words are replaced by "$$$$$" allowing matching between Alabama and Iowa, and temporal words such as April and February are replaced by "#####" so that they get aligned by the algorithm.

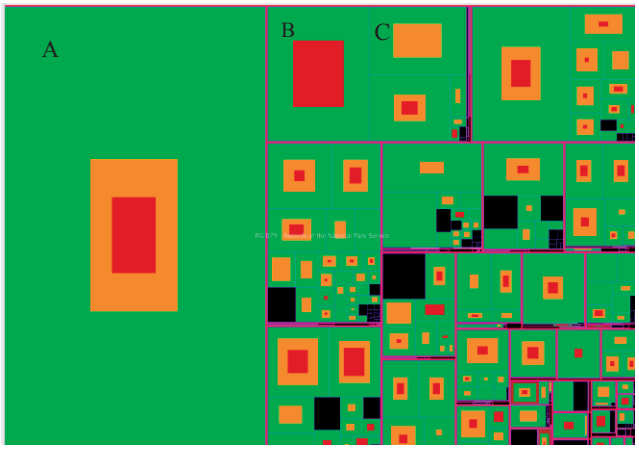


Figure 5. Data mining results shown as nested squares. Three exemplar patterns are labelled with A, B and C.

The clustering results are shown as a rectangular glyph, with the outermost rectangle representing the most shared pattern in a particular directory, and the innermost rectangle representing the least shared pattern. The width of each rectangle in the glyph shows the similarity score (a higher score corresponds to a wider rectangle). Figure 5 shows an example of the visual representation of the mining results of a collection of websites with ~38,000 files. We used the color green to denote naming pattern, orange for sequential pattern, and bordeaux for spatial pattern. It can be observed that most directories are rendered with three nested squares in order of green, orange and bordeaux, from outermost to innermost. Such pattern (e.g. A in Fig-5) indicates that all three types of organizational criteria exist in that directory in order of relevance. Other directories show that they don't have sequential (B in Fig-5), or spatial arrangements (C in Fig-5). When consulted about this representation, an archivist explained that through the glyphs it was possible to clearly identify the relevance of limited amounts of information dimensions in one directory, as well as organizational patterns across many directories [3].

In contrast to the squares representing structure, there is no spacing between the nested squares rendering additional relationships. This makes the features within the squares more distinguishable. To reduce confusion we use colors to show boundary lines between squares.

4.4 Textual context overview

To enable overview browsing over the terms in the labels of files and directories we use OpenCloud, a tag cloud visualization library. As a user navigates a particular directory, the files and directory names are inserted into the cloud. The library extracts tags from the labels and file names, and assigns weights to each tag based on its frequency of occurrence. Tag clouds are generated showing more frequent words in bigger font sizes. This information is used to learn about the directory contents.

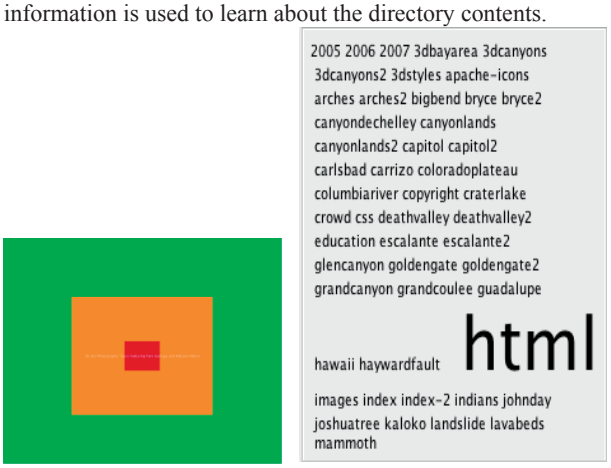


Figure 6. Tag cloud (right) rendered for a given directory (left).

The tag cloud is also used for data mining analysis results verification. In Figure 6, it is possible to see the correspondence between the tag cloud and the data mining results for a directory. In this example, sequential (orange) and naming alignments (green) exist in the same proportion. Here, 3dcanyons and 3dstyle are naming alignments, 2005 and 2006 are sequential alignments, and Carrizo and Carlsbad are spatial (bordeaux) alignments.

4.5 Interactions

The application's interactivity is fundamental to providing a rich analysis experience. Beyond basic zoom and pan functions, the interactive features implemented are:

Navigation with Dynamic Changes of Root: The visualization allows users to select any node as the root node to focus on specific areas of the treemap.

Abstraction with Dynamic Changes of Rendered Levels: Users may select the number of nested levels to be rendered at a given time. This is an important requirement to support information abstraction and syntheses. In our implementation, a high level abstraction significantly reduces the amount of data rendered, increasing performance for large-scale data. When curators need to see distributions at deeper directory levels, our interface enables users to easily navigate data vertically.

Data Selection with Search Box: We provide users with search functionality to find terms of interest in directory labels. When the term is found, the corresponding directory is highlighted with a dark pink transparent color.

Detail-on-demand: Hovering over a node with one's mouse updates the corresponding pie chart and the tag cloud with the node's specific information. For instance in Figure 2 above, the pie chart presents statistics corresponding to the selected node, and in Figure 6, the tag cloud shows the frequency of words in the directory corresponding to the selected node. To explore directories in detail, any node in a treemap can be double clicked and it expands to a new treemap with the selected node as the root. The view is presented as a tab that can be resized and moved around in the screen space. The latter enables users to organize views into different analysis workflows for observation and compare and contrast.

Linking of multiple views: To streamline analysis workflows during which users may open multiple treemap views, we devised two linking capabilities: duplication and selection. The former allows the changes in one view to reflect on the other view, but not the other way around. A more dynamic interaction, the latter, allows any of the related view instances to change the other instances. An interaction panel aids controlling and tracking duplicate instances. Duplication helps when viewing different properties (e.g. one view shows data mining results and the other file types). In turn, selection is related to detail-on-demand, which allows users to choose a section of the visualization and show its details in another view. This is useful when archivists explore the hierarchical structure in a top down mode. Summarizing the interactive visualization experience, one of the archivists contrasted its versatility to the fixed way in which collections information is traditionally presented to users.

5 Application Examples

5.1 Finding trends in large digital collections

In our project we use a testbed collection developed by The National Archives and Records Administration (NARA, http://www.archives.org). It includes publicly available data provided by Federal Agencies or harvested from their websites. Each Federal Agency corresponds to a record group and is represented as a node that includes child nodes. In turn, each record group may have more than one group of data objects bearing different arrangement and a variety of file formats. There are 1,031,118 files in 200 different formats forming different digital object types, from financial records and press releases to GIS data, CAD drawings, websites and 3D images among others. Some of the record groups have up to 12 levels of hierarchical nesting.

5.1.1 Usage scenario

Rose, an archivist working at the State Archives, has received a backlog collection with record groups from different government agencies. Her goal for the day is to use the visualization to conduct analyses and to make decisions about storage allocations and subsequent workflows with the overarching goal of making the collections promptly accessible. She first opens a general view and immediately finds out which agencies sent more records by looking at the size of the corresponding nodes. Mousing over each one, she obtains precise statistics for each agency in the form of pie charts which she will report to the storage allocation team (Figure 7). Next, she focuses on understanding general trends. For this, she selects the file type view (Figure 8) and, by looking at the color and size patterns across directories, she quickly identifies that web (green), pdf (red) and image (blue) types are predominant. Seeing this color combination within one directory, she infers that there is a high presence of webpages. With this information, she submits a ticket to the Advanced Interfaces engineers who will provide access to them through the archives portal. She also observes that some directories with mostly pdf files also contain spreadsheets (yellow) or text files (asparagus). Based on her experience, she concludes that those may be different format versions of the same record and she will have to decide whether to keep them. To her distress, she confirms that in almost all the directories there are significant numbers of files without file format identification information. This means that those files will not be functional until corresponding viewers are found. She also observes that four agencies sent compressed files, which will have to be unpacked for identification. She creates a request for the Data Analysis team to update the file format identification tool, to unpack and identify the compressed files, and to estimate a timeline to provide access to the collection.

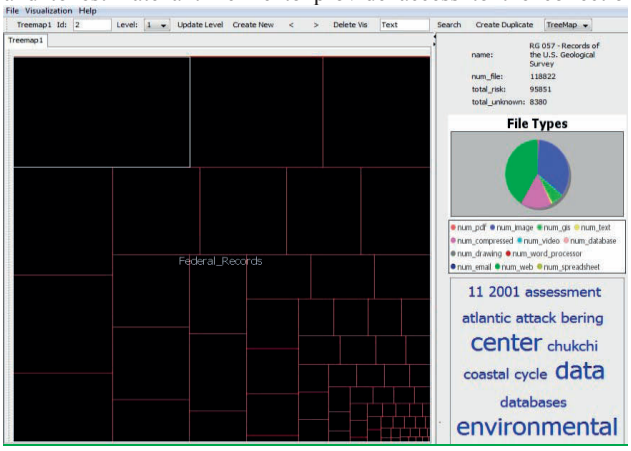


Figure 7. Statistics and word cloud views about the directories are shown on the right panel.

Figure 8. View showing file type distributions in the different record groups. A zoom out region shows the spacing between neighbouring squares to help users distinguish two groups.

Rose proceeds to learn about the collection's organization. For this, she creates a duplicate treemap and elects to render the data mining results. The view at the first directory level shows that the predominant alignment corresponds to a combination of naming (light green), temporal (brown), and spatial (wine red) alignments (Fig. 9). Knowing that most groups have more than one organizational layer, Rose creates duplicate visualizations and updates the directory level at each new instance to navigate the

hierarchical structure. While she has many views opened, she can keep track of her analysis workflow and go back to compare one level to the next. As she goes deeper in the hierarchy (Fig.10), she notices that temporal alignment starts fading out, and that sequential emerges as smaller directories are rendered. She concludes that temporal and spatial words are used at the higher directory levels to describe the records. In turn, sequential alignment becomes prominent in deeper directories containing files whose naming conventions are formed by subsequent numbering systems. After concluding the analysis, she reports that the majority of the sub-collections are organized, and that their themes can be inferred at the higher directory levels. This will guide the Cataloguing and Description team in gathering adequate access points for the collection using the tag cloud function.

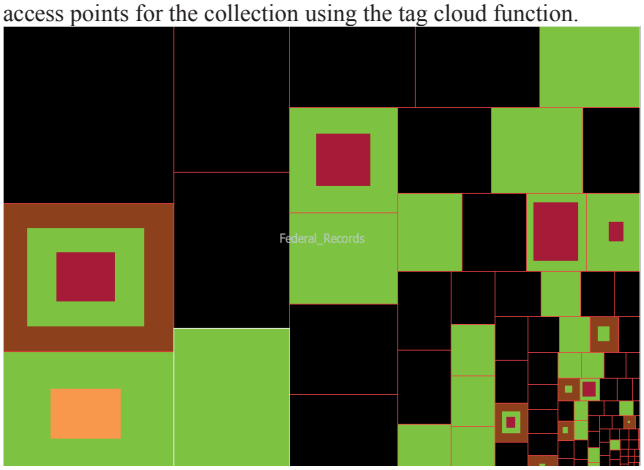


Figure 9. Visualization of inferred organizational criteria at record group levels.

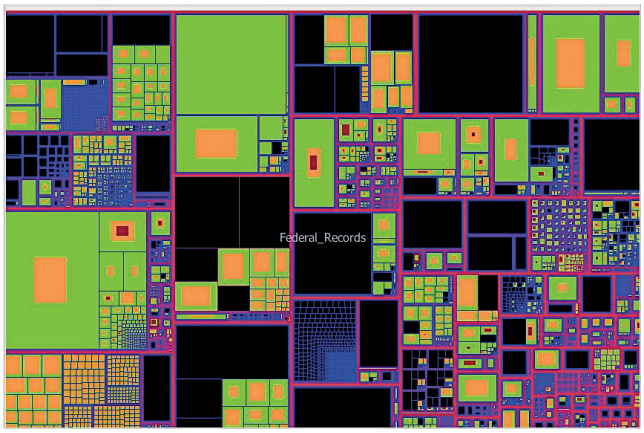


Figure 10. View of inferred organizational criteria at individual directory levels.

5.2 Visual analysis of tags in image collections

Our system was adapted to provide users with a tool to look through the organizational hierarchy of image collections in association with their tag information. Different from image visualization projects [34], here we visualize content descriptions as classes of tags from an image collection. Represented as distributions in the treemap, tags are associated with corresponding directory label information containing provenance information.

To visualize the distribution of classes of tags and to identify their relevance, we use pixel based rendering techniques. In this case, we automatically associate images to tags from their descriptions in corresponding html pages. First we parse the html files and extract the image descriptions. We then map the image to their tags by regular expression matching of their file id, which is present in the file names and on the html pages. Now every image has a set of tags, and every directory has a set of images. To do the pixel-based rendering, we find the particular tags present in a directory and render them with different colors. Given that one image may have more than one tag (e.g. images with rivers and valleys), the colors in the representation reflect the diversity of the image collection. The application allows for a richer understanding of the collection content. The method can be used to associate descriptive tags with any kind of data.

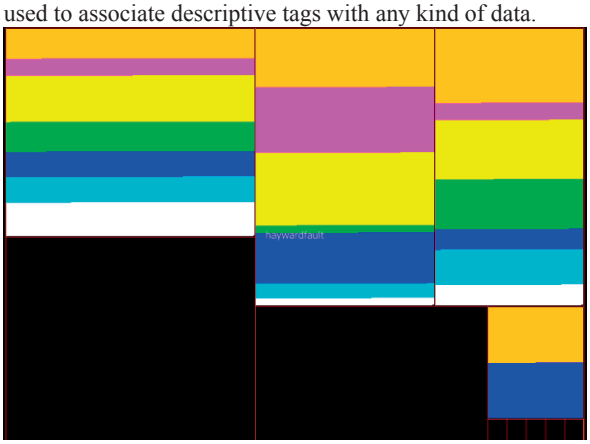


Figure 11 Color pixel rendering of image tags distribution.

Figure 11 above shows the distribution of the tags rock, park, road, valley, cliff, tree, snow and dam, each shown in a different color across six directories. Black indicates that there are no images in a directory or that there is no corresponding tag. Hence it becomes very easy to identify folders with similar types of image content.

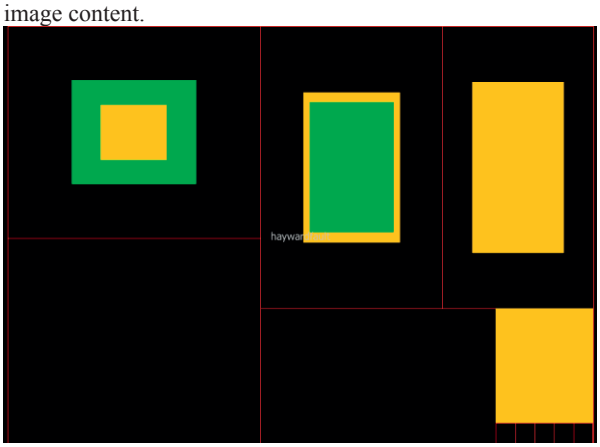


Figure 12. Nested squared rendering of image tags correlation.

To represent the correlation between two chosen tags, we use a nested square-based rendering technique. In Figure 12, a solid square indicates that all the images contain one of the selected tags. When the square has two colors, the outermost rectangle represents the dominant tag and the innermost rectangle represents the proportion of images with both the dominant tag and the other tag. For example, if tree is represented by green and snow is represented by yellow, in the square in the upper left corner tree images are dominant and the proportion of images containing trees and snow is shown.

The two pixel-based rendering representations, as bars and squared glyphs, complement each other. Representing

information as bars of different sizes and colors makes it possible to visually detect general trends for larger numbers of variables. Considering less than four tags, the visualization of tag correlations as squared glyphs allows a detailed comprehension of content and the relationships between variables. Feedback obtained from archivists indicates that the flexibility to conduct trend analyses as well as more detailed analyses maps with archival practices, and that a change in visual representation (from bars to glyphs) within one analysis workflow helps the cognitive process.

5.2.1 Usage scenario

Jane, a freelance journalist is writing an article about the diversity of landscapes in the US National Parks. She decides to find images to illustrate her article in the National Parks website archives, which provides a comprehensive image survey of the flora and fauna of the parks and corresponding descriptions. She soon realizes that it is not only time consuming to inspect each webpage, but that she will need to organize thousands of images to reflect diversity across different parks by sorting one image after another.

She decides to use the visual analysis tool offered by the National Archives to find what she needs. Using the search function in combination with the file type view and the statistics panel, she rapidly identifies a record group containing webpages with ~26,000 images corresponding to National Parks. In turn, mousing over the directories, she can read their labels and see that they are organized by park name. These steps have reduced her work significantly, and now she needs to focus on finding the right images.

Using the tags selector in the visualization interface, she can find images related to tags as well as the correlation between two tags. She first specifies the terms (e.g. mountain, river, forest, tree, water, parks etc.) that correspond to her inquiry. The system automatically associates these terms to images that have them as a tag in their html description, and presents a view showing the distribution of the various tags across directories (Figure 13).

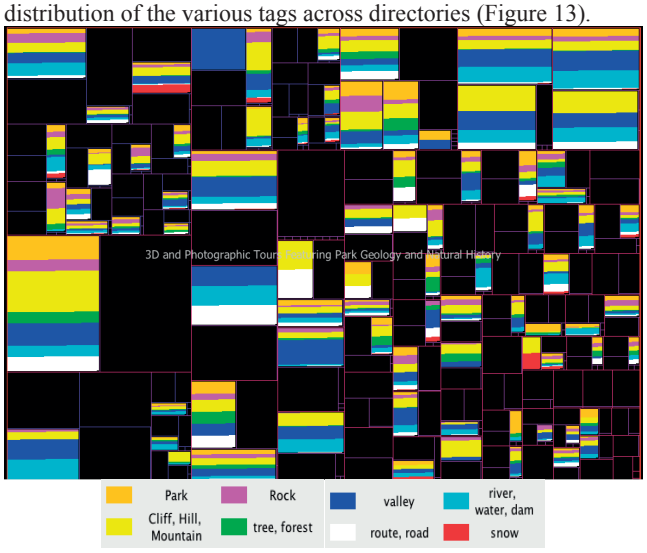


Figure 13. Visualization shows image tags in different parks.

In Figure 13, different squares represent images and html pages of different National Parks websites organized by name of the park. This view allows her to quickly learn which parks have images containing one or more of the interested landscapes. She now wants to select one image containing two landscape features present in National Parks. She notices that valleys and rivers are prominent tags in the collection and finds their correlation (Fig. 14). In the view with the results, she notes two directories (marked in red in the image) containing bigger proportions of images with both rivers and valleys. She can now select images from those. The visualization tool has narrowed her search from 26,000 to 30 images, which is the combined total of images with the selected tags contained in the two directories.

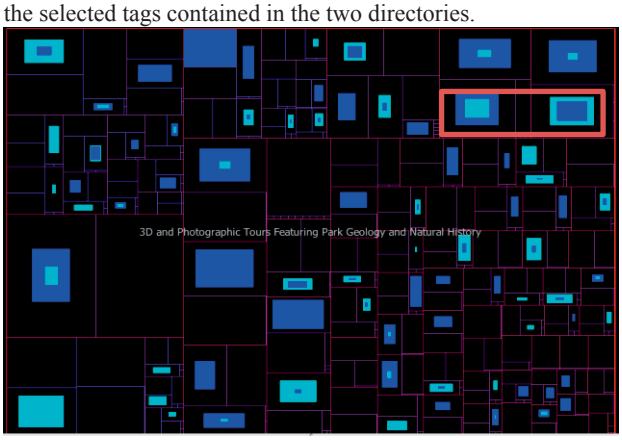


Figure 14 Correlation between river and valley tags shown as nested squares.

6 Conclusions and Discussion

To conduct digital collections analysis we designed an interactive visualization that acts as a bridge between the archivist and the data. The application enables users to identify collection trends that would be extremely difficult, if not impossible, to recognize without visual analytic methods. Throughout the design, archivists gave the team iterative feedback about the usefulness of the analysis methods and the clarity of the representations. The experience with users served as a pilot that we will further develop into a formal user experience study.

Much of our work focused on summarizing large amounts of information as meaningful visual representations. To address the accuracy of the analyses, including the completeness and quality of the data we: a) provide more than one data representation so that archivists can validate results, b) present data in context with provenance and original order to allow archivists to make informed inferences, and c) highlight what is not known or in doubt about the data.

Significant challenges still remain to present large amounts of information about digital collections. While the treemap visualization is useful to navigate the structure of a collection, for larger and nested collections the analysis may become cumbersome. In that regard, we are working on alternative visualizations that show diverse features independently from the collection's hierarchical structure. We also need to provide more precise descriptions than those derived from word frequencies in tag clouds. Currently, we are investigating the combination of graphics and NLP methods to generate high level descriptions based on aggregated label information. In addition, we learned that archivists would like to input feedback to the visualization. This work introduces new ways for users to understand, discover, interpret, and interact with the wealth of archives information.

Acknowledgement

This work was supported through a National Archives and Records Administration (NARA) supplement to the National Science Foundation Cooperative Agreement (NSF) TERAGRID: Resource Partners, OCI-0504077.

REFERENCES

[1] Jennifer Meehan, "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description," *American Archivist*, vol. 72, no. 1, pp. 72-90, Spring/Summer 2009.

[2] S. R. Anderson and R. B Allen, "Envisioning the Archival Commons," *The American Archivists*, vol. 72, pp. 383-400, Fall/Winter 2009.

[3] Maria Esteva, Weijia Xu, Suyog Dutt Jain, and Jennifer Lee, "Assessing the preservation condition of large and heterogeneous electronic records collections with visualizations," in *6th International Digital Curation Conference*, Chicago, 2010.

[4] Ben Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Trans. Graph.*, vol. 11, no. 1, pp. 92-9, 1992.

[5] The National Archives of the United Kingdom, PRONOM. (2010, April) DROID, Vers. 4.0. [Online]. http://droid.sourceforge.net/

[6] L. J. Henry, "Appraisal of Electronic Records," in *In Thirty Years of Electronic Records*, B. I. Ambacher, Ed.: Maryland: The Scarecrow Press, 2003, p. 38.

[7] J. Cox Richard, "Appraising the Digital Past and Future," in *international symposium on digital curation, DigCCur'07*, Chapel Hill, North Carolina., 2007.

[8] Martin Fowler, *Patterns of Enterprise Application Architecture*.: Addison-Wesley Professional , 2002.

[9] Jeffrey Heer, Stuart K. Card Landay, and James A., "Prefuse: a toolkit for interactive information visualization," in *SIGCHI conference on Human factors in computing systems*, Portland, Oregon, USA, 2005, pp. 421-430.

[10] opencloud. [Online]. http://opencloud.mcavallo.org/

[11] JFreeChart. (2010, Aug.) Java chart library. [Online]. http://www.jfree.org/jfreechart/

[12] John Lamping and Ramana Rao, "Visualizing large trees using the hyperbolic browser," in *CHI' 96*, 1996, pp. 388-9.

[13] John Stasko and Eugene Zhang, "Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations," in *InfoVis'00*, 2000.

[14] Jean-Daniel Fekete, "The InfoVis Toolkit," in *10th IEEE Symposium on Information Visualization (InfoVis'04)*, IEEE Press, 2004, pp. 167-74.

[15] Daisy Disk. [Online]. http://www.daisydiskapp.com/

[16] Mark Bruls, Kees Huizing, and Jarke J. van Wijk, "Squarified TreeMaps," in *Joint Eurographics and IEEE TCVG Symposium on Visualization*, 2000, pp. 33-42.

[17] Bjorn Engdahl, Malin Koksal, and Gary Marsden, "Using treemaps to visualize threaded discussion forums on PDAs," in *Proceedings of ACM Conference on Human Factors in Computing Systems 2005*, 2005, pp. 1355-1358.

[18] NewsMap : A treemap based google news aggregator. [Online]. http://newsmap.jp/

[19] Edward Clarkson, Krishna Desai, and James Foley, "ResultMaps: Visualization for Search Interfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1057-64, Nov. 2009.

[20] Windirstat. [Online]. http://windirstat.info/

[21] H.M.M. van de Wetering J.J. van Wijk, "Cushion Treemaps: "Visualization of Hierarchical Information," in *IEEE Symposium on Information Visualization 1999*, IEEE CS Press, 1999, pp. 73-78.

[22] T. Nasukawa and T. Nagano., "Text analysis and knowledge mining system.," *IBM System Journal*, vol. 40, no. 4, pp. 967-84, 2001.

[23] A. Don et al., "Discovering interesting usage patterns in text collections: Integrating text mining with visualization.," in *CIKM*, 2007, pp. 213-22.

[24] F. B. Vi´egas and M. Wattenberg., "Tag clouds and the case for vernacular visualization," *Interactions*, vol. 15, no. 4, pp. 49-52, 2008.

[25] M. Wattenberg and F. B. Viegas., "The word tree, an interactive visual concordance," in *InfoVis'08*, 2008.

[26] J. Stasko, C. Gorg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization.," in *Information Visualization 2007*, 2007.

[27] Lei Shi et al., "Understanding text corpora with multiple facets," in *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST'10),* , IEEE Press, 2010, pp. 99-107.

[28] Robert F. Erbacher, "Glyph-Based Generic Network Visualization," in *SPIE Conference on Visualization and Data Analysis*, 2002.

[29] Michael Chau, "Visualizing web search results using glyphs: Design and evaluation of a flower metaphor," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 1, pp. Article 2, 27 pages, March 2011.

[30] D.A. Keim, "Designing pixel-oriented visualization techniques: theory and applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 59-78, 2000.

[31] University of Texas at Austin. (2011, Apr.) Archives of the future. [Online]. http://www.utexas.edu/features/2011/04/11/tacc_archives/

[32] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*.: MIT Press, 1998.

[33] O. Gotoh, "An improved algorithm forc matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705-8, 1982.

[34] Hyunmo Kang and Ben Shneiderman, "Visualization Methods for Personal Photo Collections: Browsing and Searching in the PhotoFinder," in *IEEE International Conference on Multimedia and Expo (ICME2000)*, New York City, New York, USA, 2000.