Browsing the Web Using Stacked Three-Dimensional Sunbursts to Visualize Term Co-Occurrences and Multimedia Content

Markus Schedl[1] markus.schedl@jku.at
Peter Knees[1] peter.knees@jku.at
Gerhard Widmer[1,2] gerhard.widmer@jku.at
Klaus Seyerlehner[1] klaus.seyerlehner@jku.at
Tim Pohle[1] tim.pohle@jku.at

[1] Department of Computational Perception
Johannes Kepler University (JKU)
Linz, Austria
http://www.cp.jku.at

[2] Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
http://www.ofai.at

ABSTRACT

We present a novel visualization approach that extends the Sunburst technique to the third dimension, which allows for encoding additional data in the height of each arc segment. By stacking a number of such 3D-Sunbursts, we build a user interface for browsing collections of web pages. We apply techniques for web content indexing of multimedia data (text, audio, image, video) and text analysis to create a UI where each layer of the Sunburst stack represents a specific multimedia content type. We call this user interface the *Co-Occurrence Browser (COB)* and demonstrate its usability on a collection of web pages related to music artists.

Keywords: Stacked 3D-Sunbursts, User Interface, Web Content Mining, Co-Occurrence Browser

Index Terms: I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques H.5.1 [Information Systems]: Information Interfaces And Presentation—Multimedia Information Systems H.5.2 [Information Systems]: Information Interfaces And Presentation—User Interfaces

1 INTRODUCTION AND CONTEXT

The *Sunburst* as proposed in [3, 6] is a circular, space-filling visualization technique that is used to visualize hierarchical data. The center of the visualization represents the highest element in the hierarchy, whereas elements on deeper levels are illustrated by arcs further away from the center. Child elements are drawn within the angular borders of their parent, but at a more distant position from the center.

In this paper, we propose a visualization technique that makes use of the Sunburst approach, but allows for visualizing additional dimensions of the data set. This is achieved first by enhancing the visual representation of the Sunburst with a third dimension, which provides the possibility to encode continuous data in the height of each arc. Second, by organizing a set of N such enhanced 3D-Sunbursts in a stack, we are able to display, for every arc, the values of N continuous data dimensions.

We demonstrate the presented technique in our Co-Occurrence Browser (COB) application for browsing collections of web pages by means of co-occurring terms and multimedia content. Since the size of the Sunbursts of the COB is in principle only restricted by the number of possible combinations of all terms, we further introduce three constraints that limit this size. To get an impression of the appearance of a visualization employing the proposed technique, the reader is invited to take a look at Figure 1.

2 STACKED 3D-SUNBURSTS

Each arc of a fixed Sunburst in the stack can encode two different data dimensions. The first one is represented by the arc's angular extent, which obviously has to be the same for a particular arc over all layers of the stack. The second one is encoded in the height of the arc, which may vary according to the layer. Considering a Sunburst stack of N layers, it is possible to encode the values of $N+1$ data dimensions for each arc (1 in the arc's angular extent and N in its height in each of the N layers).

As for user interaction, our prototypical Java implementation currently provides the following possibilities: *rotating* the Sunburst stack around the Y-axis, i.e. the vertical axis going through the root nodes of all Sunbursts in the stack, by moving the mouse in the horizontal direction while pressing an arbitrary mouse button, *zooming* in/out (within predefined boundaries) by holding the left mouse button pressed and moving the mouse upwards/downwards, *changing the inclination of the stack* limited to angles between a front view and a bird's eye view by moving the mouse upwards/downwards while holding the right mouse button pressed.

Selecting a particular arc, i.e. moving in the hierarchy visualized by the Sunbursts, is accomplished by the arrow keys. Furthermore, two different methods for highlighting a selected arc are provided. The first one highlights only the most recently selected arc and should be used when it is important to concentrate the user's attention on this arc. Employing the second method, the selected arc at each hierarchy level is highlighted, which facilitates tracing the selection back to the root arc. We use this second highlighting method in the application presented in the next section – cf. Figure 1, where the arcs labeled "guitar" and "punk" are highlighted by coloring their labels and arc borders in white.

3 THE CO-OCCURRENCE BROWSER

We demonstrate the proposed visualization technique for the application scenario of exploring a collection of web pages about music artists. To this end, we elaborated a user interface that allows for browsing web page collections by means of co-occurring terms and multimedia content. We call this UI the Co-Occurrence Browser (COB). The COB has been implemented in Java using the *processing* environment [1] and parts of our *CoMIRVA* framework [5, 4] for music and multimedia processing, information retrieval, and visualization.

The purpose of the COB is threefold. First, it facilitates getting an overview of the web pages related to a specific topic, a music artist in our case, by structuring them according to co-occurring terms.

Second, since the terms that most often occur on web pages related to a music artist X constitute an individual profile of X, the COB is also suited to reveal various meta-information about the artist, e.g. musical style, related artists, or instrumentation. Third, by visualizing the amount of multimedia content provided at the indexed web pages, the user is offered a means of exploring the audio, image, and video content of the set of web pages.

3.1 Information Retrieval

As for the data acquisition, we use *Google* to obtain the URLs of up to 1,000 web pages related to the music artist under consideration, whose content we then retrieve via *Wget*. Subsequently, a term occurrence analysis step is performed. To this end, we use a dictionary of musically relevant terms, which are searched in all web pages of every artist, yielding an inverted file index. The dictionary was created from lists of various sources like *Wikipedia*, *All Music Guide*, or *Yahoo! Directory* and contains music genres and styles, artists, instruments, moods, and other terms which are somehow related to music.

As for indexing the multimedia content of the web pages, we first extracted a list of common file extensions for audio, image, and video files from [2]. For every web page, we then search the HTML code for links to files whose extension occur in one of the extracted lists.

Finally, we store the URLs of the found multimedia files as well as the inverted file index from the term occurrence analysis in an XML file. Using the inverted index of the web pages of an artist X, we can easily extract subsets $S_{X,\{t_1,\ldots,t_r\}}$ of the web page collection of X which have in common the occurrence of all terms $t_1, \ldots, t_r$.

3.2 Creation of the Sunbursts

Starting with the entire set of web pages $S_{X,\{\}}$ described by the XML file of an artist X, a maximum number N of terms with highest document frequencies are used to create N subsets $S_{X,\{t_1\}}, \ldots, S_{X,\{t_N\}}$ of the collection. These subsets are visualized as arcs $A_{S_{X,\{t_1\}}}, \ldots, A_{S_{X,\{t_N\}}}$ around a centered cylinder which represents the root arc $A_{S_{X,\{\}}}$, and thus the entire set of web pages. The angular extent of each arc is proportional to the document frequency of the associated term t_i, i.e. to the number of web pages containing t_i. To avoid very small, thus hardly perceivable, arcs, we omit arcs whose angular extent is smaller than a fixed threshold E measured in degrees. Furthermore, each arc is depicted in the color obtained by applying the colormap "Sun" from the CoMIRVA framework to the relative document frequency of the arc's corresponding term t_i (relative to the total number of web pages).

The described term selection with respect to document frequencies and the corresponding visualization step are recursively performed for all arcs, with a maximum R for the recursion depth. This eventually yields a complete Sunburst visualization, where each arc at a specific recursion depth r represents a set of web pages $S_{X,\{t_1,\ldots,t_r\}}$ in which all terms $t_1, \ldots, t_r$ co-occur.

Internally, the Sunburst is stored as a tree, where each arc is represented by a node. A node $A_{S_{X,\{t_1,\ldots,t_r\}}}$ at depth r in the tree thus represents the set of web pages that contain the term t_r and all other terms $t_1, \ldots, t_{r-1}$ associated with the nodes on the shortest path from $A_{S_{X,\{t_1,\ldots,t_r\}}}$ to the root node $A_{S_{X,\{\}}}$.

As for the different layers in the Sunburst stack, in each layer, we visualize the amount of a specific category of multimedia files linked to from or embedded in the web pages. The three categories are audio, image, and video. To this end, we encode the relative number of the audio, image, or video files in the height of the arcs (relative to the total number represented by the root node of the respective layer). For example, denoting the audio layer as L_A, the image layer as L_I, and the video layer as L_V and focusing on a fixed arc A, the height of A in L_I shows the relative number of image files contained in the web pages that are represented by arc A, the height of A in L_V illustrates the relative number of video files, and the height of A in L_A the relative number of audio files.

3.3 User Interface

Figure 1 depicts a screenshot of COB's user interface for 126 web pages retrieved for the artist "Bad Religion". Each arc $A_{S_{X,\{t_1,\ldots,t_r\}}}$ is labeled with the term t_r that subdivides the web pages represented by the arc's parent node $A_{S_{X,\{t_1,\ldots,t_{r-1}\}}}$ into those containing t_r and those not containing t_r. Additionally, the document frequency of the term t_r is added in parentheses to the label of each arc $A_{S_{X,\{t_1,\ldots,t_r\}}}$. In the Stacked 3D-Sunbursts visualization shown in Figure 1, the topmost layer illustrates the amount of video files indicated on the web pages, the middle layer the amount of image files, and the lower layer the amount of audio files. In this screenshot, the arc representing the web pages on which all of the terms "Bad Religion", "punk", and "guitar" co-occur is selected. For 40 out of the complete set of 126 web pages, this is the case. The following constraining parameters were used: $N = 6$, $R = 8$, $E = 5$ (cf. Subsection 3.2) In addition to the user interaction techniques described in Section 2, the COB obviously offers functions to bring up additional information about the web pages represented by the selected arc, e.g. lists of the URLs and the multimedia content. It is further possible to open the web pages and the indexed audio, image, and video files.

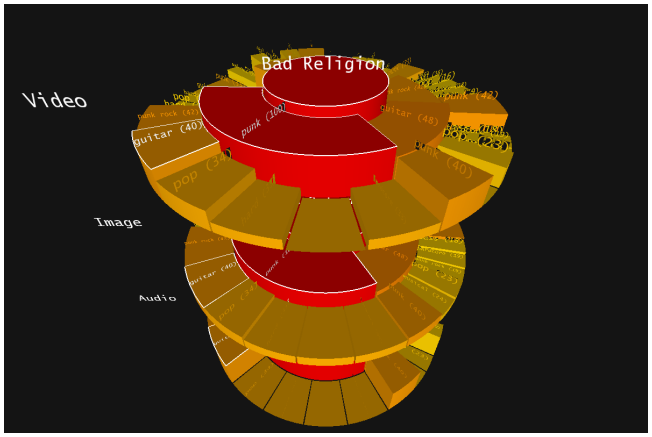

Figure 1: The user interface of the COB. Here, 126 web pages of the band "Bad Religion" are clustered according to co-occurring terms.

References

[1] http://www.processing.org, access: 2007-03-29.

[2] http://en.wikipedia.org/wiki/List_of_file_formats, access: 2007-03-29.

[3] K. Andrews and H. Heidegger. Information Slices: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs. In *Proceedings of the 4th IEEE Symposium on Information Visualization 1998 (InfoVis'98)*, pages 9–12, Research Triangle Park, NC, USA, October 1998.

[4] M. Schedl, P. Knees, K. Seyerlehner, and T. Pohle. The CoMIRVA Toolkit for Visualizing Music-Related Data. In *Proceedings of the 9th Eurographics/IEEE VGTC Symposium on Visualization (EuroVis'07)*, Norrköping, Sweden, May 2007.

[5] M. Schedl, P. Knees, and G. Widmer. Interactive Poster: Using CoMIRVA for Visualizing Similarities Between Music Artists. In *Proceedings of the 16th IEEE Visualization 2005 Conference (Vis'05)*, Minneapolis, Minnesota, USA, October 2005.

[6] J. Stasko and E. Zhang. Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In *Proceedings of the 6th IEEE Symposium on Information Visualization 2000 (InfoVis'00)*, pages 57–65, Salt Lake City, UT, USA, October 2000.