DocuBurst: Document Content Visualization Using Language Structure

Christopher Collins*
University of Toronto
Toronto, Ontario, Canada

Abstract

We present the first visualization of document content which takes advantage of the human-created structure in lexical databases. We use the WordNet hyponymy (`IS-A`) relationship as the structure for radial, space-filling trees which quickly reveal the concepts contained within a document of interest. Interactive techniques of zoom, filter, and details-on-demand support document analysis. The visualization can be generalized to multiple documents and other linguistic databases.

Keywords: Radial space-filling layout, document visualization.

1 Introduction

Comparing many documents is a common task when navigating the internet and large text databases. Analyzing and comparing the content of lengthy documents is an important task in visual analytics. Overviews of document content have been an active area of research in information visualization for many years [5, 2, e.g.], but reported works do not make use of valuable human-annotated linguistic structure in the visualization, instead providing detail on topic content using a pre-selected set of terms of interest. We provide an interactive visualization of document content based on the human-annotated `IS-A` noun and verb hierarchies of WordNet, a lexical resource created and revised by linguists over a period of 20 years [3] (see Figure 1). WordNet is a lexical database of *words*, *synsets*, and *relations* representing over 200,000 word-synset relationships and many more synset-synset relationships. Synsets are defined as sets of synonymous words, all sharing a common *sense* or definition. Synsets are related by many types of relationships, but the most widely used part of WordNet is the hyponymy (`IS-A`) partial order (e.g. '{lawyer, attorney} `IS-A` {professional, professional person}'). This relationship is particularly suited to visualizing topic content as the hierarchy offers a natural level-of-abstraction filter: the deeper in the hyponymy tree we go, the more specific the concepts become. In this work we use WordNet structure and counts of word occurrences in a document of interest to create an interactive radial space-filling graph of document content linked with full text views.

2 Linguistic Pre-Processing

In order to populate the hyponymy tree with word counts, several pre-processing steps are necessary. Starting with raw text, for example, a book, we break the text into roughly coherent topical segments [1]. For each segment, we label parts of speech and stem words (books → book/`noun`, going → go/`verb`). The full text of the segments and their associated (word, part-of-speech, count) triples are then read into the visualization. This processing was carried out on a general science textbook, used for the examples in this paper.

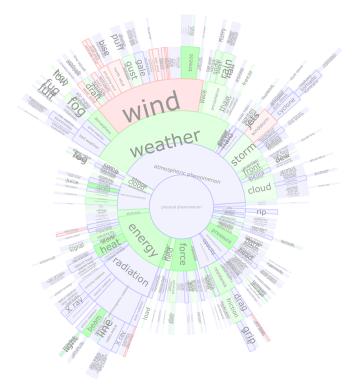
*e-mail: ccollins@cs.utoronto.ca

Figure 1: Hyponymy of single sense {*physical phenomenon*}. Nodes with words occurring in the reference document are green, nodes containing search prefix *wind* are pink. Alpha transparency corresponds to strength of occurrence in a general science textbook: {*breeze*} occurs strongly, {*gust*} weakly.

3 Layout and Shading

Radial, space-filling graph-drawing techniques have been previously reported and serve as the basis of this work [6, 7]. The combined structure of WordNet hyponymy and document lexical content is visualized rooted at a word or synset of interest (selected by the user), shown as a circle in the center of the layout. All child nodes are assigned to a sector of an annulus with angular width which is part of the parent node's width and proportional to the size of the subtree rooted at that node. Tree depth determines on which concentric ring a node appears (increasing depth corresponds to increasing radius). The width of each annulus is maximized to allow for all visible graph elements to fit within the display space. Font size is maximized and labels are rotated to allow the label to fit within the node and minimize label overlap. As senses are collections of synonymous words, they are labeled with their first word member. To improve graph clarity and visibility of document content words, for nodes with angular width < 0.25 degrees, edges, labels, and sector outlines are omitted. In the case of graphs rooted at a word, each subtree represents a synset containing that word; nodes under the same sense of the root word are assigned the same hue to ensure unrelated subtrees are distinguished. If the root is a

single synset, all nodes are colored blue.

Document content is visualized through the alpha transparency of the fill color of the nodes. Highly opaque nodes have many occurrences; almost white (transparent) nodes have none. Words and senses that are more prominent in the document of interest stand out easily against a desaturated context. Two views are provided: In the *single-node view* (Fig. 1), only nodes whose word members occur in the document are colored, revealing the precise concepts in the document. In the *subtree view* (Fig. 2), counts are propagated up to the root of the tree, using the association of synonyms into synsets and synsets into a hyponymy tree to aggregate counts for related concepts and provide a higher level view of document content. In both views, alpha is normalized so maximum counts achieve full opacity. The subtree view reveals where concepts occur when a fisheye filter has been applied to the graph and some nodes are hidden. In addition, for a dense graph, the subtree view helps us find single nodes that are too small to distinguish.

4 Interaction

We provide several techniques to visually abstract the data. First, we provide a highlight search function which visually highlights nodes whose label matches any of the given search terms. *Highlight nodes* have a pink background and a larger font size, drawing attention to even the smallest of search results. Second, we implement generalized fisheye views [4] to collapse all subtrees which are more than a user-specified distance from the central root node. The presence of non-zero word occurrence counts within collapsed subtrees is indicated by using the subtree coloring, in which counts are propagated to the root. Optionally, all highlight nodes can be exempted from the distance filter (by increasing their *a priori* importance in the DOI function), effectively abstracting the graph to all synsets within a given distance from the root or highlight nodes.

Double clicking on a node of interest restricts the visualization to the hyponyms of the corresponding synset; double right-clicking reloads the graph at the parent of the clicked node, thus providing bi-directional semantic zoom. Node angular width can be manually adjusted using the mouse wheel to increase (up) and decrease (down) the width of the node. Changes to a node's angular width affect its children equally and its siblings in an inverse manner [7].

The counts that are used to determine node alpha values are based on text segments, or automatically-determined subtopic regions of the document. The initial view is based on all segments in the document, but range selectors allow for limiting the segments from which counts are drawn (for example, an analyst may be interested in only the first half of a document).

Visual pan and zoom of the display space are also supported. Word nodes can be shown or hidden to increase detail or decrease clutter as desired. Highlighting, roll-up, fisheye filtering, pan, and zoom are provided in real time.

Because edges are not shown, it may be difficult to discern parent-child relationships using DocuBurst, especially if the graph is particularly dense. To facilitate understanding of the hyponym relations in the visualization, when the mouse pointer rests over a node it is highlighted in a saturated green and all hyperonyms of that node are highlighted in a saturated blue, tracing a path to the root. Details of the synset under the mouse pointer are provided in a linked view: the synset details window at the bottom of the interface lists the synset members and the associated sense definition. A full text details tab is also provided, in which occurrences of the synset under the mouse pointer are highlighted in the text of the document being visualized. (see Figure 2).

5 Future Work

There are several areas for further research and improvements of this prototype tool. We are interested in investigating dynamically-linked visualizations, for example, a scatter-plot of concept occur-
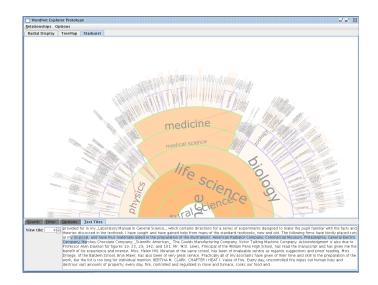

Figure 2: Subtree view rooted at {*science*}. A details window at the bottom of the interface shows the full text of the document being visualized. Controls provide filter, search, and graph-loading functions. To maximize display space, the entire control panel acts as an auto-hide task bar, lowering when not in use.

rence by text segments to quickly reveal where highly occurring words appear in the document. This work leads well into an investigation of the DocuBurst technique to view the difference between two documents, which may be useful for plagiarism detection, document categorization, and authorship attribution. Spreadsheets of many DocuBursts could also be used for comparison across document collections. Future evaluation of DocuBurst against content analysis using other tools and full-text documents will reveal the usability and power of combining radial space-filling layout with human-created linguistic structure in a document analysis tool.

References

[1] Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, pages 26–33, 2000.

[2] Philip DeCamp, Amber Frid-Jimenez, Jethran Guiness, and Deb Roy. Gist icons: Seeing meaning in large bodies of literature. In *Poster Session, IEEE Symposium on Information Visualization*, Minneapolis, USA, October 2005.

[3] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA, 1998.

[4] George W. Furnas. Generalized fisheye views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 16–23. ACM Press, April 1986.

[5] Marti A. Hearst. Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66. ACM Press, 1995.

[6] John Stasko and Eugene Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS)*, pages 57–65, 2000.

[7] Jing Yang, Matthew O. Ward, and Elke A. Rundensteiner. Interring: An interactive tool for visually navigating and manipulating hierarchical structures. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS)*, pages 77–84, 2002.