Eurographics/ IEEE-VGTC Symposium on Visualization 2009
H.-C. Hege, I. Hotz, and T. Munzner
(Guest Editors)

Volume 28 (2009), Number 3

DocuBurst: Visualizing Document Content using Language Structure

Christopher Collins[1], Sheelagh Carpendale[2], and Gerald Penn[1]

[1]University of Toronto, Toronto, Canada; [2]University of Calgary, Calgary, Canada

Abstract

Textual data is at the forefront of information management problems today. One response has been the development of visualizations of text data. These visualizations, commonly based on simple attributes such as relative word frequency, have become increasingly popular tools. We extend this direction, presenting the first visualization of document content which combines word frequency with the human-created structure in lexical databases to create a visualization that also reflects semantic content. DocuBurst is a radial, space-filling layout of hyponymy (the IS-A relation), overlaid with occurrence counts of words in a document of interest to provide visual summaries at varying levels of granularity. Interactive document analysis is supported with geometric and semantic zoom, selectable focus on individual words, and linked access to source text.

Categories and Subject Descriptors (according to ACM CCS): Document And Text Processing [I.7.1]: Document and Text Editing—Document Management; Computer Graphics [I.3.6]: Methodology and Techniques—Interaction Techniques; Information Storage and Retrieval [H.3.7]: Digital Libraries—User Issues

1. Introduction

'What is this document about?' is a common question when navigating large document databases. In a physical library, visitors can browse shelves of books related to their interest, casually opening those with relevant titles, thumbing through tables of contents, glancing at some pages, and deciding whether this volume deserves further attention. In a digital library (or catalogue search of a traditional library) we gain the ability to coalesce documents which may be located in several areas of a physical library into a single listing of potentially interesting documents. However, the experience is generally quite sterile: people are presented with lists of titles, authors, and perhaps images of book covers. In feature-rich interfaces, page previews and tables of contents may be browsable. If the library contents are e-books, users may even open the entire text, but will have to page through the text slowly, as interfaces are often designed to present a page or two at a time (to dissuade copying). Our goal in this work is to bring some of the visceral exploratory experience to digital libraries, to provide interactive summaries of texts which are comparative at a glance, can serve as decision support when selecting texts of interest, and provide entry points to explore specific passages.

Prompted by the ever increasing volume and open access to digital text, developing overviews of document content has been an active research area in information visualization for many years. However, reported works do not make use of existing richly studied linguistic structures, relying instead on simple statistical properties of documents (*e.g.*, [AC07]), or analytic methods such as latent semantic analysis (*e.g.*, [DFJGR05]), which can produce unintuitive word associations. The resulting visualizations provide detail on content without a consistent view that can be compared across documents. In DocuBurst, we provide a complement to these works: a visualization of document content based on the human-annotated `IS-A` noun and verb hierarchies of WordNet [Fel98] which can provide both uniquely- and consistently-shaped glyph representations of documents, designed for cross-document comparison (see Figure 1).

2. Related Work

2.1. Document Content Visualization

Visualizations of document content take two common forms: synoptic visualizations for quick overviews and visualizations specialized for discovering patterns within and between documents. Specialization in the type of document

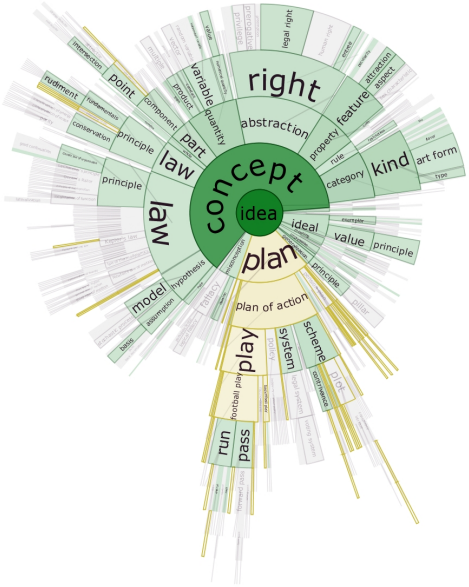


Figure 1: DocuBurst of a science textbook rooted at {*idea*}. A search query for words starting with 'pl' has been performed. Nodes matching the query are highlighted in gold.

	Synoptic					Patterns				
	Glyphs			Overview		Repetition		Features		Context
Feature	DocuBurst	Gist Icons	Starstruck	Tag Clouds	TextArc	FeatureLens	Arc Diagram	Compus	TileBars	Word Tree
1. semantic	Y	N	N	N	N	N	N	N	N	N
2. cluster	Y	Y	N	P	N	N	N	N	N	N
3. overview	Y	N	N	Y	Y	N	Y	Y	Y	N
4. zoom	Y	N	P	P	N	Y	Y	N	Y	Y
5. freq	Y*	Y*	Y*	Y	Y	Y*	N	N	Y*	N
6. compare	Y	Y	Y	Y	N	Y	N	Y	Y	P
7. search	Y	Y	N	Y	Y	Y	P	N	Y	Y
8. read	Y	P	P	P	Y	Y	P	N	P	P
9. pattern	Y+	P+	P+	P+	Y	Y*	Y	Y	Y*	Y*
10. features	P	P	P	P	N	P	N	Y	Y	N
11. suggest	P	Y	N	Y	Y	Y	Y	N	N	N
12. phrases	N	P	N	Y*	N	Y	Y	Y	P	Y
13. all words	N	Y	Y	Y	Y	Y	Y	N	Y	Y

Figure 2: Comparison of features available (Y), possible with a trivial extension (P), or not possible (N) in document visualizations. * denotes cases that only visualize a selected subset of words; + denotes a coordinated visualization.

used as input further divides the reported research: books and long documents, historical documents, multilingual texts, and computer-mediated communication archives such as emails, instant messages, and threaded discussions. In this space, DocuBurst focuses on long texts, such as books, and provides a visualization that is simultaneously synoptic, comparative, and allows for deeper intra-document analysis of occurrence patterns. In the remainder of this section we will review significant examples in each of these categories, describing how their feature sets compare.

Synoptic visualizations of text most often use a selected subset of the language to create a glyph based on word occurrence counts. Glyphs are then combined in small multiples visualizations to perform comparisons. Glyph techniques include Starstruck [HWMT98], which creates glyphs by arranging lines of varying length in a circular pattern, and Gist Icons [DFJGR05], which builds on this idea by drawing a smoothed contour around this pattern. The vocabularies used are either restricted user-selected term sets (Starstruck) or automatically selected and potentially unrelated (Gist Icons). Subtle shape perturbations can be hard to compare in side by side radial glyphs. Other synoptic visualizations of document content use space-filling techniques to provide an overview of the vocabulary a single document, such as TextArc [Pal02] and tag clouds (*e.g.*, [Fei08]). These techniques provide a quick overview of document content.

A second category of document visualizations are those created to reveal patterns within texts. FeatureLens [DZG*07] suggests repeated phrases that may be of interest, and visualizes selections, while Arc Diagrams [Wat02] provide an overview of repetition throughout a document. DocuBurst also provides views of the distribution of selected words. Other pattern-based visualizations focus on distributions of significant features in a document such as emotion (*e.g.*, extensions [OBK*08] of TileBars [Hea95]), or hand-annotated properties (*e.g.*, Compus [FD00]). The Word Tree [WV08] is an example of a pattern-based visualization focused on repetition in context. While Word Trees provide a unique view on repetition, overall word frequencies and full document overviews are not visible.

To compare document visualizations, we order a list of the types of features document visualizations have provided from most unique to DocuBurst to the most common in other visualizations:

semantic indicate word meaning
cluster generalize by clustering words into concepts
overview provide quick overviews of an entire text
zoom support varying the semantic or graphical detail
freq reveal frequency of individual words
compare compare multiple documents
search search for specific words/phrases
read drill-down to original text
pattern reveal patterns within or between texts
features reveal extracted features (*e.g.*, emotion)
suggest suggest interesting focus words/phrases
phrases can show multi-word phrases
all words can show all parts of speech

Figure 2 summarizes how these features relate to DocuBurst and other well known text visualizations. Note that only DocuBurst provides some reflection of semantics through integrated word definitions and the use of a semantically-defined linguistic structure. Only DocuBurst and Gist Icons provide word clustering into higher concepts, however in Gist Icons the groups are only one level deep

and based on statistical measures whose meaning may not be readily apparent to a reader. Note that all visualizations that provide overviews of entire text suffer from screen real estate issues with large texts.

2.2. Graph Drawing

Radial graph-drawing techniques have been previously reported and serve as the basis of this work. Of particular interest are the semi-circular radial space-filling (RSF) hierarchies of Information Slices [AH98] and the focus + context interaction techniques of the fully circular Starburst visualization [SZ00]. The InterRing [YWR02] visualization expands on the interaction techniques for RSF trees, supporting brushing and interactive radial distortion. TreeJuxtaposer [MGT*03] illustrates methods for interacting with very large trees, where nodes may be assigned very few pixels. We adapt techniques such as tracing the path from a node of interest to the root and performing interactive accordion expansion from this work.

3. Background on WordNet

Despite the growing dependence on statistical methods, many Natural Language Processing (NLP) techniques still rely heavily on human-constructed lexical resources such as WordNet [Fel98]. WordNet is a lexical database composed of *words*, *collocations*, *synsets*, *glosses*, and *edges*. *Words* are literally words as in common usage. A *collocation* is a set of words such as "information visualization" which are frequently collocated and can be considered a unit with a particular definition. For the purposes of this paper, we will use *words* to refer to both *words* and *collocations* — they are treated equally in the visualization. Sets of synonymous *words* and *collocations* are called *synsets*. *Glosses* are short definitions that the words in a synset share, thus they are definitions of synsets. An edge in WordNet represents a connection between synsets.

Synsets are the most important data unit in WordNet. Throughout this paper, we will refer to *words* in single quotes (e.g 'thought'), and synsets using a bracketed set notation (e.g. {*thought, idea*}). A *word* may be a member of multiple *synsets*, one for each sense of that word. Word senses are ranked, either by order of familiarity (a subjective judgement by the lexicographer) or, in some cases, by using a synset-tagged reference corpus to provide numerical relative frequencies.

Synsets in WordNet are connected by many types of edges, depending on the part of speech (noun, verb, *etc.*). WordNet contains 28 different types of relations, but the most widely used part of WordNet is the hyponymy (`IS-A`) partial order. An example of hyponymy is {*lawyer, attorney*} `IS-A` {*professional, professional person*}. When traversing this graph, we remove any cycles (they are very rare) by taking a depth-first spanning tree at the user-selected root. In this work we focus on the noun hyponymy relationships in English WordNet (v2.1), rooted under the synset {*entity*} having 73,736 nodes (synsets) and 75,110 edges, and a maximum depth of 14. Verb hyponymy is also supported — that hierarchy is smaller and takes a more shallow, bushier form. In addition, there is no single "root" verb. The visualizations produced can be generalized to any partial order of a lexicon.

3.1. WordNet Visualization

Many interfaces for WordNet exist, the most popular of which is the text-based WordNet Search which is part of the publicly available WordNet package. With the exception of the work of Kamps [KM02], the existing interfaces for WordNet either provide for drill-down textual or graphical interaction with the data starting at a single synset of interest or provide path-tracing between two synsets *e.g.*, [Alc04, Thi05]. We do not know of any visualization of WordNet that uses the graph structure to enhance a visualization of other data such as document content.

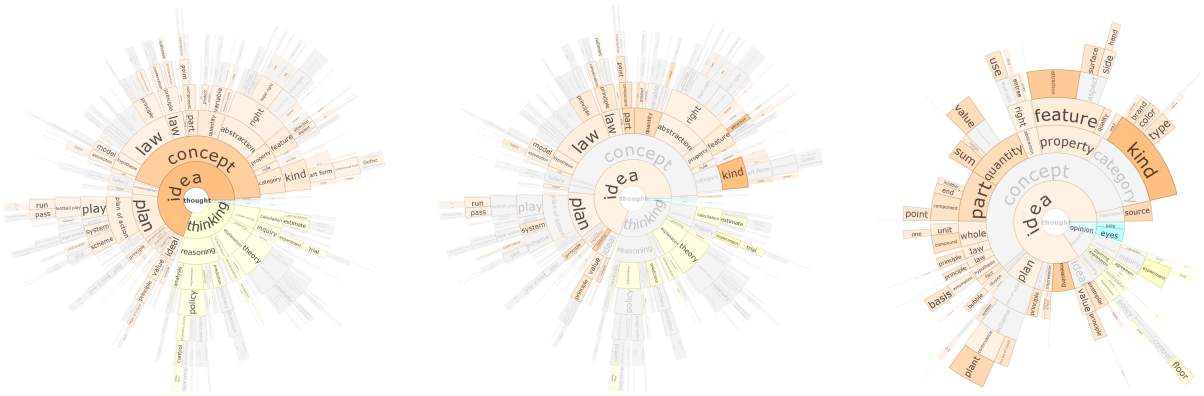
4. DocuBurst Visualization

The combined structure of WordNet hyponymy and document lexical content is visualized using a radial space-filling tree layout implemented with prefuse [HCL05]. Traversing the tree from center to periphery follows a semantic path of increasing specificity using the `IS-A` relation. In WordNet, synset members are ordered according to their polysemy count, which WordNet researchers call *familiarity*. Since more familiar words come first, we chose the first word in a synset as the node label. Label fonts are maximized, rotated to fit within the node, and overlap is minimized.

4.1. Linguistic Processing and Scoring

In order to populate a hyponymy hierarchy with word counts, several pre-processing steps are necessary. Starting with raw text, we subdivide the text into *tiles* based on the pre-existing structure, such as section headings. If no structure is detectable, we break the text into roughly coherent topic segments using a segmenter [Cho00]. For each tile, we label parts of speech (`NOUN`, `VERB`, etc.) [Bri93]. Nouns and verbs are then extracted and stemmed (*e.g.*, books → book, going → go) using a morphological processor [Did03]. Punctuation is omitted. If short word sequences, noted in WordNet, are found in the document, the words are combined into a collocation, and treated as a single word.

Next we look up in which WordNet synsets the (*word, part-of-speech*) pairs occur. Because pairs usually occur in multiple synsets, we do not perform word sense disambiguation. Instead, we divide the word count amongst the available synsets. If WordNet supplies relative sense frequency information for a word, we use this to distribute the count. Otherwise, we distribute the count weighted linearly by sense rank. This results in weighted occurrence counts that are not



(a) Node fill transparency proportional to number of occurrences of words in subtree (cumulative view). Node size based on leaf count.

(b) Node fill transparency proportional to number of occurrences of words in a single synset (single node view). Node size based on leaf count.

(c) Single node view, node size proportional to occurrence count total for subtree rooted at that node.

Figure 3: DocuBurst of a science textbook rooted at 'thought'; node hue distinguishes the synsets containing 'thought'.

integers, but the overall results more accurately reflect document content. By dividing the counts, we dilute the contribution of highly ambiguous terms. The full text of tiles and their associated (*word, part-of-speech, count*) triples are then read into the data structure of the visualization.

4.2. Visual Encoding

Node Size

Within the radial tree, angular width can be proportional to the number of leaves in the subtree rooted at that node (*leaf count*) or proportional to the sum of word counts for synsets in the subtree rooted at that node (*occurrence count*). The leaf count view is dependent on WordNet and so is consistent across documents. The word count view maximizes screen space for synsets whose words actually occur in the document of interest, thus the shape, as well as node colouring, will differ across documents. Depth in the hyponymy tree determines on which concentric ring a node appears. The width of each annulus is maximized to allow for all visible graph elements to fit within the display space.

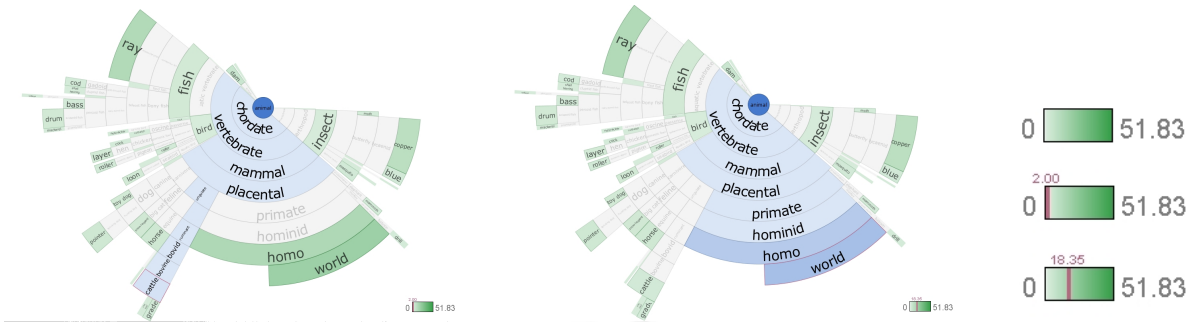
Node Colour

It is possible to look at multiple senses of a word in one view. Views rooted at a single word contain a uniquely coloured subtree for each synset (sense) containing that word. In contrast, trees rooted at a single synset use a single hue. Since luminance variation in the green region of the spectrum is the most readily perceived, it is the first colour choice [Sto03, 30]. Gray is used for nodes with zero occurrence counts, since their presence provides a visual reminder of what words are not used.

Transparency is used to visualize relative word or synset count. Similar to the concept of value, transparency provides a range of light to dark colour gradations, thus offering ordered [Ber83] and pre-attentive [War04] visuals. Highly opaque nodes have many occurrences; almost transparent nodes have few occurrences. Word senses that are more prominent in the document stand out against the more transparent context.

Two ways to visualize word occurrence are provided: single-node and cumulative. In the *single-node* visualization, only synset nodes whose word members occur in the document are coloured. In the *cumulative* view, counts are propagated up to the root of the tree. In both views, transparency is normalized so maximum counts achieve full opacity. When multiple documents are visualized, the cross-document maximum is used to set the scale. These modes support a gradual refinement of focus. The cumulative, or subtree, view uses the association of words into synsets and synsets into a hyponymy tree to aggregate counts for related concepts. Similar to the TreeJuxtaposer techniques for visualizing differences embedded deep in a large tree [MGT*03], by highlighting the entire subtree containing the node, salient small nodes can be more easily located, even if hidden from view by a filter. The single-node view reveals precise concepts in the document and supports the selection of synsets whose word members appear in the document being analyzed. In addition, for a fully expanded graph, the single node view may highlight nodes that are otherwise too small to notice. The subtree and cumulative views are compared in Figure 3.

While transparency is an effective visual method for distinguishing large differences and trends, it is impossible to read exact data values using it. To facilitate the exact reading of synset occurrence counts for the selected text tiles, we provide a dynamic legend (see Figure 4).



(a) (b) (c)

Figure 4: DocuBurst of a general science textbook rooted at {*animal*}. Single-node colouring and occurrence count sizing were used with zero-occurrence synsets hidden. Mouse hover point is revealed by blue trace-to-root colouring. **(a)** Synset {*cattle, cows, kine, oxen*} highlighted. **(b)** Synset {*world, human race, humanity, mankind, man, ...*} highlighted. **(c)** Detail of the dynamic legend, showing, from top to bottom, no selection, selection from (a), and (b).

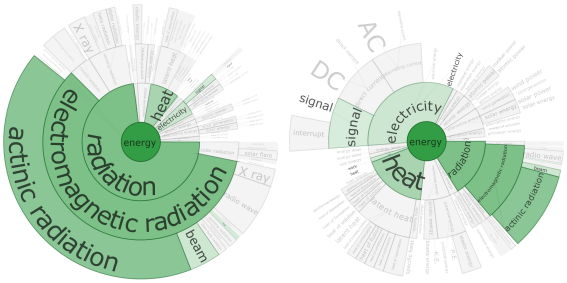


Figure 5: DocuBurst of a general science textbook rooted at {*energy*}. At right, mouse wheel interaction was used to shrink the angular width of the subtree rooted at {*radiation*} and expand the subtree under {*electricity*} exposing the previously illegible node {*signal*}.

5. Interaction

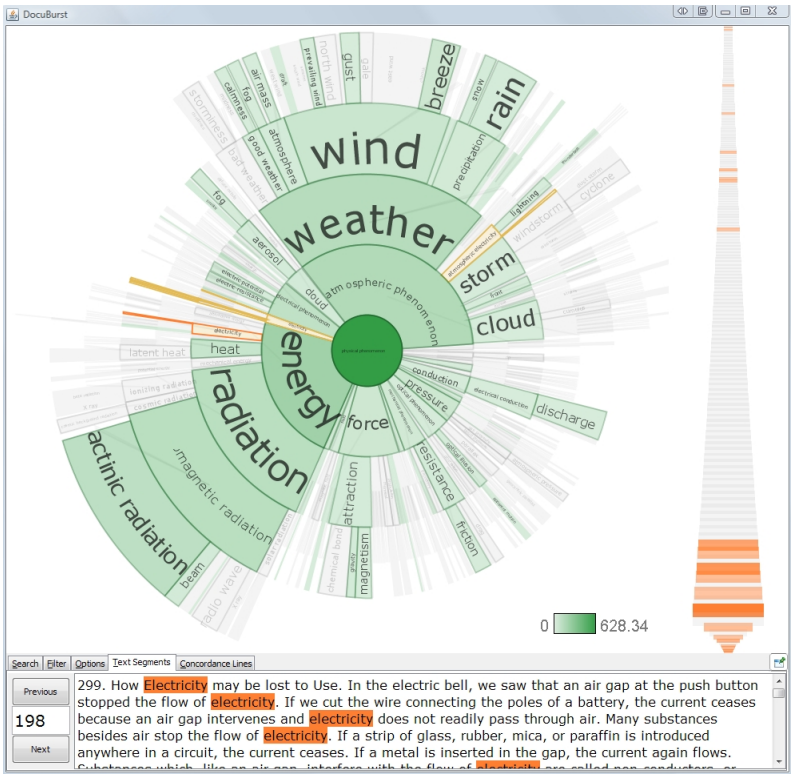
A root node can be a word, in which case its immediate children are the synsets containing that word. Alternatively the visualization can be rooted at a synset. Root nodes in the visualization are selectable by searching for either a word or synset of interest. Once a root is chosen, the visualization is populated with all its hyponyms.

As there are more than 70,000 English noun synsets in WordNet, techniques to abstract and filter the data are important. First, we provide a highlight search function which visually highlights nodes whose label matches any of the given search terms. *Highlight nodes* have a gold background and border, and a darker font colour, drawing attention to even the smallest of search results. The transparency of the highlight (gold) background is attenuated to the word occurrence counts so as to not disrupt this data-carrying value and to provide for stronger pop-out of search results with high occurrence counts.

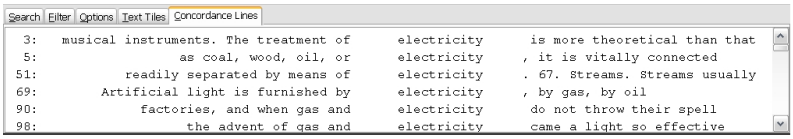
Second, we implement a generalized fisheye view [Fur86] that collapses all subtrees which are more than a user-specified distance from the central root node. Changing this distance-based filter allows for a semantic zoom, creating visual summaries of varying specificity. The presence of non-zero word occurrence counts within collapsed subtrees is indicated by using the cumulative colouring, 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 (see Figure 1).

Double clicking on a node of interest restricts the visualization to the hyponyms of the node's synset; double right-clicking reverses this action by reloading the graph at the parent of the clicked node, thus providing bi-directional data navigation through the hyponymy relation. To create more space for the details of the children of a given synset, the angular width of a node and its subtree can be manually increased using the mouse wheel. This increase provides a radial detail-in-context view which causes the node's siblings to be correspondingly compressed. Changes to a node's angular width affect its children equally and its siblings in an inverse manner (see Figure 5).

The visualization can be based on selected subsections of the document. The initial view is based on all text tiles in the document, but a selection can limit the tiles from which counts are drawn. Unrestricted visual pan and geometric zoom of the display space are also supported, as well as a zoom-to-fit control to reset the pan and zoom to a best-fit for the currently visible tree. Rendering is dependent on the zoom factor: node borders are not rendered when the nodes are very small, and labels are not rendered when they would not be legible. All highlighting, navigation, and emphasis interactions are provided in real time.



(a) Linked visualizations for details-on-demand.



(b) The details window showing the concordance lines for the selected synset.

Figure 6: A search for 'electricity' reveals synsets containing that word (gold). On selecting a node, the distribution of the word is revealed in the tile browser. Selecting a tile reveals the text with occurrences of the selected synset highlighted.

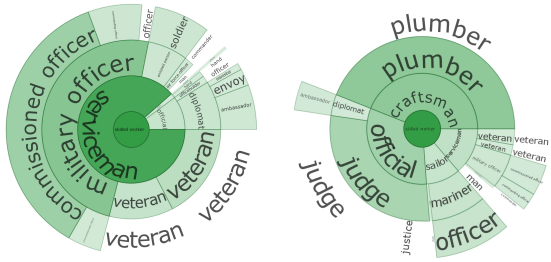
6. Drill Down

The full text can be read through accessing the text tiles at the bottom of the interface. To navigate the text, we use a linked visualization: the text tile browser. Rectangles representing the text tiles appears in a linear, vertical array to the right of the DocuBurst. A fisheye distortion [Bed00] facilitates navigation and selection of text tiles within this list. Clicking any tile brings it into view. Furthermore, this visualization can be used to see occurrence patterns in the document. By clicking nodes in the DocuBurst visualization, synsets and entire subtrees can be selected. Text tiles in which selected synsets appear show as varying intensity of orange in the text tile browser, depending on number of occurrences in the tile. Occurrences of those synsets and words are also highlighted in the full text window.

Concordance lines are a standard linguistic analysis tool in which all occurrences of a word of interest are extracted and displayed with their left and right N (usually 5) context words. Concordance lines for selections are extracted and shown in the concordance window. Patterns of orange in the tile browser can indicate how localized concepts are in the document. For example, in Figure 6, we see that {*electricity*} appears more frequently toward the end of the document. We can use the tile browser and full text window to quickly find occurrences of the terms of interest in context. By clicking the text tile rectangles in the tile browser, we find, in the tile detail window, that there is a chapter on 'electricity' at the end of the book.

7. Example: Document Comparison

DocuBurst can be used to compare multiple documents. Trees rooted at the same synset but coloured based on differ-



(a) First 2008 debate (b) Third 2008 debate

Figure 7: DocuBursts rooted at {*skill worker*} reveal the traditional US focus on military officers and veterans was eclipsed in the third US Presidential debate by discussions of plumbers.

ent texts will reveal relative frequency differences between them. While the other examples in this paper were visualization of a high school general science text book, we also can apply the technique to other forms of electronic text. In Figure 7 we applied DocuBurst to the transcripts of two 2008 US presidential debates. Note that to ensure comparability when viewing multiple documents, colour scaling is based on the maximum count for visible nodes across all documents.

A high level view of the debates rooted at {*person*} revealed strong colour for the {*leader*} and {*serviceman, military personnel, military man*} subtrees. Drill down revealed {*senator*} is a descendant of {*leader*} (both participants were senators). Attention to military issues and veterans is also expected given current conflicts. Examining the third debate showed an additional region of colour under the {*craftsman*} subtree. Further investigation, by switching to the occurrence count size function, revealed a dramatic shift in concentration within the {*skilled worker*} subtree. Military people became relatively less important compared to craftspeople — specifically, plumbers. This is the effect of Senator McCain's focus on "Joe the Plumber" in the third debate, and was the genesis point of this phrase which dominated the remainder of the campaign.

8. Challenges and Future Work

Reflecting on this work has suggested several interesting opportunities for future research in both the data and visualization realms. From a data perspective, the original goal of viewing what parts of an entire language are included in a document merits further research. As with all text visualizations, it is necessary to view a subset of language due to limited display space and computational resources with extremely large data. Views rooted at {*entity*} and covering all English nouns appear cluttered and interaction is slow. It is commonly held that WordNet sense-divisions are too fine-grained for many computational applications; investigation into other ways to abstract WordNet may help alleviate this problem. Additionally the use of uneven tree cut models to abstract the DocuBurst in a dynamic way may clarify the view. For example, if the subtree under a node has only one leaf with a non-zero count, we may omit that subtree.

Finding a place to begin exploration is a challenge with the current implementation. Providing hints for which synsets may be of interest as visualization roots for a particular document or set of documents may assist an analyst to find views of interest. Methods which may be useful include suggesting synsets with a high fraction of non-zero leaves below them, or when the cumulative count divided by the number of leaves is high, indicating an area of unusual concentration.

Word sense disambiguation is an area of active research in computational linguistics that could benefit DocuBurst. Currently, an occurrence of a word is divided among all synsets in which it appears. Thus 'river *bank*' will augment the count for {*bank, savings bank, depository financial institution*}. Incorporating word sense disambiguation into the preprocessing step would greatly enhance the value of the visualization. For example, in the general science textbook used for the examples in this paper, 'man' occurs quite often in the sense of 'people' ("man invented the wheel"). However, these occurrences are also counted towards the biological {*hominid*} sense of 'man', resulting in the incorrectly strong appearance of the {*primate*} subtree. Additionally, we currently use word occurrence count as the only available word scoring function. Other scoring functions, such as the Dunning Log-likelihood ratio [Dun93], could be used to highlight important or unusual words in a document. Other text features, such as hapax legomena (words which occur only once) could be used with the WordNet-based layout to provide special-purpose summaries of content.

Visually, the use of transparency to indicate word occurrence is useful for the intuitive mapping between data and visual appearance. However, it also introduces the possibility of misleading illusions. For instance, siblings in DocuBurst are unordered; furthermore, non-sibling nodes may be adjacent. By chance, unrelated nodes that both have high occurrence counts can appear as a large swath of strong colour. Gestalt perception may lead viewers to impart significance to this coincidence. Stronger node borders would distinguish these regions, but node borders become obstructive on small nodes. Finding an experimentally validated solution to this design trade-off could impact space-filling visualizations in general.

9. Conclusion

DocuBurst provides an overview visualization of document content which is based on a human-centered view of language whereas previous works were based on far simpler, derivative statistical analyses. The visual design is grounded in established research on human abilities in colour perception. Semantic and geometric zooming, filtering, search, and

details-on-demand provide a visual document summary, revealing what subset of language is covered by a document, and how those terms are distributed.

Initially motivated by the current lack of a digital equivalent of flipping through a book, this work leads well into an investigation of the DocuBurst technique to view the differences between two or more documents, which may be useful for plagiarism detection, document categorization, and authorship attribution. Existing digital library interfaces could be enhanced with arrays of DocuBurst icons, allowing comparison against one another or a baseline reference corpus to portray content in more pleasing and information-rich ways.

Acknowledgements

Thanks to Ravin Balakrishnan for advice and guidance. Funding for this research was provided by NSERC, iCore, SMART Technologies, and NECTAR.

References

[AC07] ABBASI A., CHEN H.: Categorization and analysis of text in computer mediated communication archives using visualization. In *Proc. of the Joint Conf. on Digital Libraries* (2007), ACM, pp. 11–18.

[AH98] ANDREWS K., HEIDEGGER H.: Information slices: Visualising and exploring large hierarchies using cascading, semi-circular discs. In *Proc. of IEEE Symp. on Information Visualization (InfoVis), Late Breaking Hot Topics* (1998), pp. 9–12.

[Alc04] ALCOCK K.:. WordNet relationship browser [online]. June 2004. Available from: `http://www.ultrasw.com/alcock/` [cited 20 February, 2006].

[Bed00] BEDERSON B.: Fisheye menus. In *Proc. of the ACM Symposium on User Interface Software and Technology (UIST 2000)* (2000), ACM Press, pp. 217–226.

[Ber83] BERTIN J.: *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983.

[Bri93] BRILL E.: POS tagger. Software, 1993. Available from: `http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z`.

[Cho00] CHOI F. Y. Y.: Advances in domain independent linear text segmentation. In *Proc. of the 2000 Conference of the North American Chapter of the Association for Computational Linguistics* (2000), pp. 26–33.

[DFJGR05] DECAMP P., FRID-JIMENEZ A., GUINESS J., ROY D.: Gist icons: Seeing meaning in large bodies of literature. In *Proc. of IEEE Symp. on Information Visualization, Poster Session* (Oct. 2005).

[Did03] DIDION J.:. Java WordNet Library [online]. 2003. Available from: `http://jwordnet.sourceforge.net` [cited 28 August, 2005].

[Dun93] DUNNING T.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*, 1 (1993), 61–74.

[DZG*07] DON A., ZHELEVA E., GREGORY M., TARKAN S., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proc. of the Conf. on Information and Knowledge Management* (2007).

[FD00] FEKETE J.-D., DUFOURNAUD N.: Compus visualization and analysis of structured documents for understanding social life in the 16th century. In *Proc. of the Joint Conf. on Digital Libraries* (2000), ACM.

[Fei08] FEINBERG J.:. Wordle: Beautiful word clouds [online]. 2008. Available from: `http://www.wordle.net` [cited 2 December, 2008].

[Fel98] FELLBAUM C. (Ed.): *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA, 1998.

[Fur86] FURNAS G. W.: Generalized fisheye views. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* (Apr. 1986), ACM Press, pp. 16–23.

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: a toolkit for interactive information visualization. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* (Apr. 2005), ACM Press.

[Hea95] HEARST M. A.: Tilebars: visualization of term distribution information in full text information access. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* (1995), ACM Press, pp. 59–66.

[HWMT98] HETZLER B., WHITNEY P., MARTUCCI L., THOMAS J.: Multi-faceted insight through interoperable visual information analysis paradigms. In *Proc. of the IEEE Symp. on Information Visualization* (Oct. 1998), pp. 137–144.

[KM02] KAMPS J., MARX M.: Visualizing WordNet structure. In *Proc. of the 1st International Conference on Global WordNet* (2002), pp. 182–186.

[MGT*03] MUNZNER T., GUIMBRETIÈRE F., TASIRAN S., ZHANG L., ZHOU Y.: Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibilit. *ACM Transactions on Graphics 22*, 3 (2003), 453–462. SIGGRAPH 2003.

[OBK*08] OELKE D., BAK P., KEIM D. A., LAST M., DANON G.: Visual evaluation of text features for document summarization and analysis. In *Proc. of the IEEE Symp. on Visual Analytics Science and Technology (VAST)* (2008), pp. 75–82.

[Pal02] PALEY W. B.: TextArc: Showing word frequency and distribution in text. In *Proc. of the IEEE Symp. on Information Visualization* (Oct. 2002), Poster, IEEE Computer Society.

[Sto03] STONE M. C.: *A Field Guide to Digital Color*. AK Peters, Ltd., 2003.

[SZ00] STASKO J., ZHANG E.: Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proc. of the IEEE Symp. on Information Visualization* (2000), pp. 57–65.

[Thi05] THINKMAP: ThinkMap visual thesaurus, Apr. 2005. Available from: `http://www.visualthesaurus.com` [cited 10 April, 2005].

[War04] WARE C.: *Information Visualization: Perception for Design*, 2nd ed. Morgan Kaufmann, 2004.

[Wat02] WATTENBERG M.: Arc diagrams: Visualizing structure in strings. In *Proc. of the IEEE Symp. on Information Visualization* (2002).

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, and interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization) 14*, 6 (Nov/Dec 2008), 1221–1229.

[YWR02] YANG J., WARD M. O., RUNDENSTEINER E. A.: InterRing: An interactive tool for visually navigating and manipulating hierarchical structures. In *Proc. of the IEEE Symp. on Information Visualization* (2002), pp. 77–84.