

Text and Document Visualization 1



CS 7450 - Information Visualization
November 11, 2013
John Stasko

Text is Everywhere



- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
 - WWW
 - Digital libraries
 - ...

Big Question



- What can information visualization provide to help users in understanding and gathering information from text and document collections?

Fall 2013

CS 7450

3

Tasks/Goals



- What kinds of analysis questions might a person ask about text & documents?

Fall 2013

CS 7450

4

Example Tasks & Goals



- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify "hidden" messages or stories in this document collection.
- How does one set of documents differ from another set?
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Understand the history of changes in a document.
- Find connections between documents.

Fall 2013

CS 7450

5

Related Topic - IR



- Information Retrieval
 - Active search process that brings back particular/specific items (will discuss that some today, but not always focus)
 - I think InfoVis and HCI can help some...
- InfoVis, conversely, seems to be most useful when
 - Perhaps not sure precisely what you're looking for
 - More of a browsing task than a search one

Fall 2013

CS 7450

6

Related Topic - Sensemaking



- Sensemaking
 - Gaining a better understanding of the facts at hand in order to take some next steps
 - (Better definitions in VA lecture)
- InfoVis can help make a large document collection more understandable more rapidly

Challenge



- Text is nominal data
 - Does not seem to map to geometric/graphical presentation as easily as ordinal and quantitative data
- The “Raw data --> Data Table” mapping now becomes more important

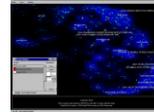
This Week's Agenda



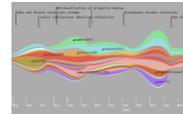
Visualization for IR
Helping search



Visualizing text
Showing words,
phrases, and
sentences



Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes



Information Retrieval



- Can InfoVis help IR?
- Assume there is some active search or query
 - Show results visually
 - Show how query terms relate to results
 - ...

Improving Text Searches



- What's wrong with the common search?
 - Is there really anything wrong?
- Visualizing the results of search queries is one potential important area of text infovis

What Hearst Thinks is Wrong



- Query responses do not include include:
 - How strong the match is
 - How frequent each term is
 - How each term is distributed in the document
 - Overlap between terms
 - Length of document
- Document ranking is opaque
- Inability to compare between results
- Input limits term relationships

TileBars



- Goal
 - Minimize time and effort for deciding which documents to view in detail
- Idea
 - Show the role of the query terms in the retrieved documents, making use of document structure

TileBars



- Graphical representation of term distribution and overlap
- Simultaneously indicate:
 - Relative document length
 - Frequency of term sets in document
 - Distribution of term sets with respect to the document and each other

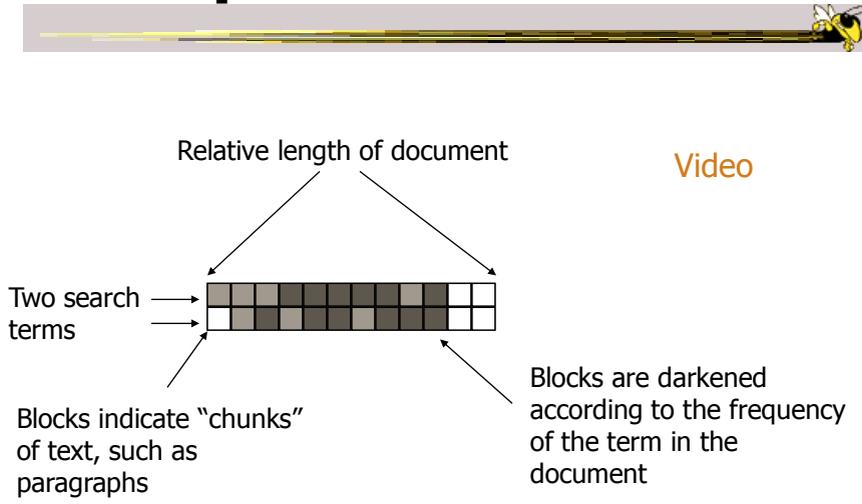
Interface

Search terms

Presentation

Fall 2013 CS 7450 15

Technique



Issues

- Horizontal alignment doesn't match mental model
- May not be the best solution for web searches
 - Non-linear material
 - Images? Apps?
- Anything else?

Fall 2013

CS 7450

17

Generalize More

- How about the "holy grail" of a visual search engine?
 - Hot idea for a while
- My personal view: It's a mistake in the general case. Text is just better for this.

Fall 2013

CS 7450

18

Search Visualization



<http://www.kartoo.com>
Defunct

Fall 2013



CS 7450

19

Sparkler



- Abstract result documents more
- Show “distance” from query in order to give user better feel for quality of match(es)
- Also shows documents in responses to multiple queries

Havre et al
InfoVis '01

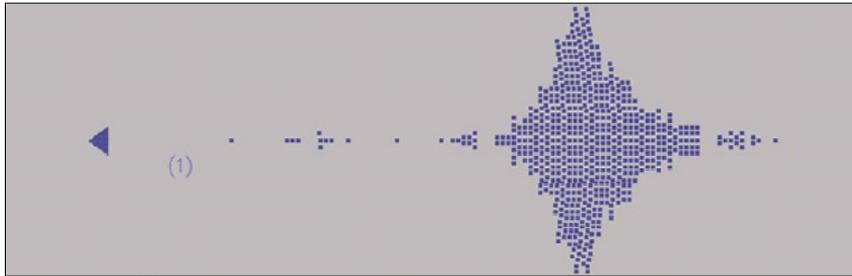
Fall 2013

CS 7450

20

Visualizing One Query

- Triangle – query
- Square – document
- Distance between query and documents represents their relevance



Fall 2013

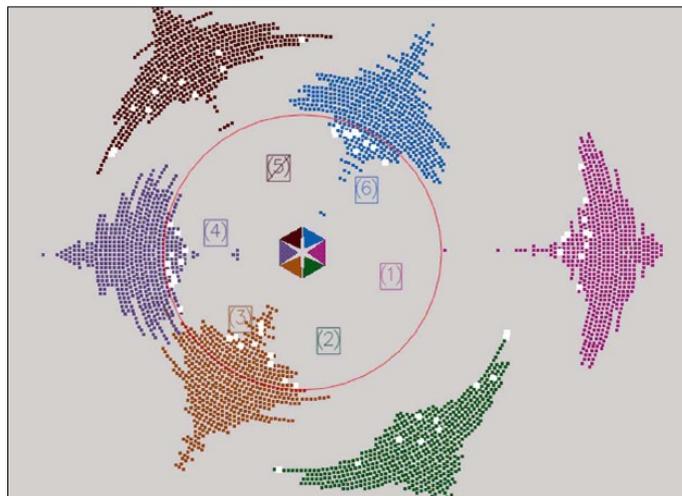
CS 7450

21

Visualizing Multiple Queries

Six queries here

Bullseye allows viewer to select quality results



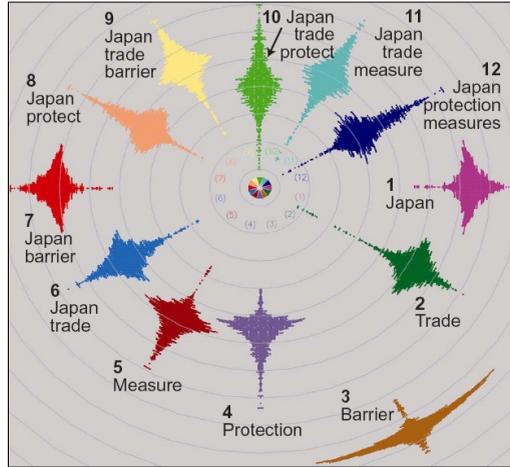
Fall 2013

CS 7450

22

Test Example

- Text Retrieval Conference (TREC-3) test document collection
- AP news stories from June 24–30, 1990
- TREC topic: Japan Protectionist Measures
- Sparkler found 16 of 17 relevant documents



Fall 2013

CS 7450

23

Another Idea



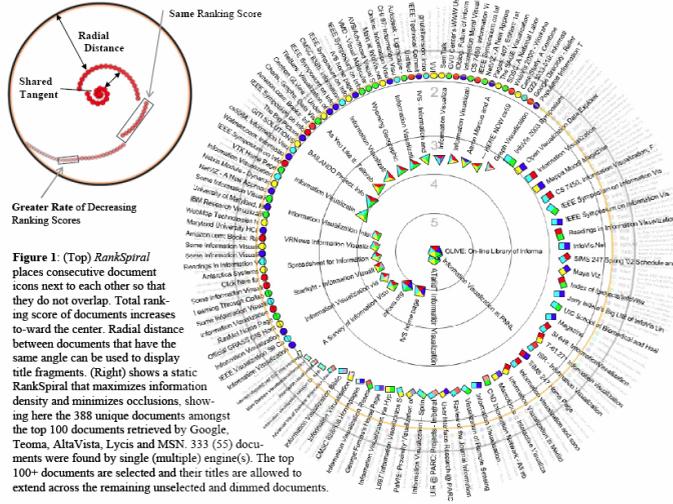
Use it to compare search results from different search engines

Fall 2013

CS 7450

24

RankSpiral



Color represents different search engines

Figure 1: (Top) RankSpiral places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases toward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral that maximizes information density and minimizes occlusions, showing here the 388 unique documents amongst the top 100 documents retrieved by Google. Teoma, AltaVista, Lycis and MSN. 333 (55) documents were found by single (multiple) engine(s). The top 100+ documents are selected and their titles are allowed to extend across the remaining unselected and dimmed documents.

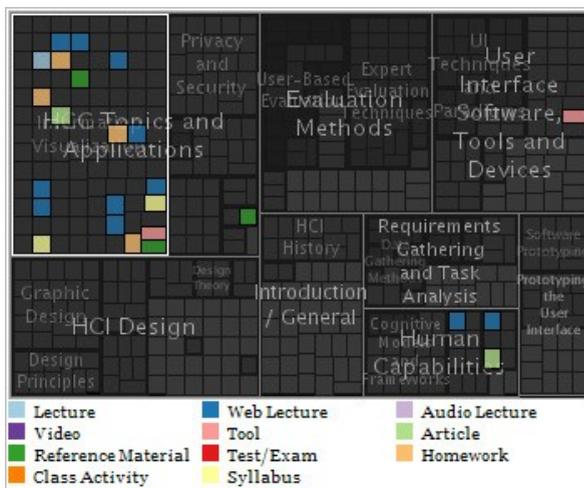
Spoerri
InfoVis '04 poster

Fall 2013

CS 7450

25

ResultMaps



Treemap-style vis for showing query results in a digital library

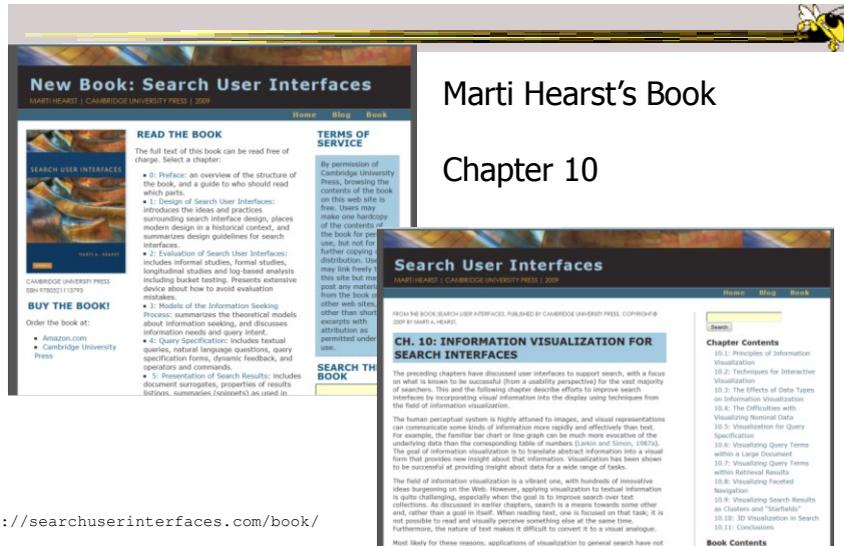
Clarkson, Desai & Foley
TVCG (InfoVis) '09

Fall 2013

CS 7450

26

To Learn More



New Book: Search User Interfaces
MARTI HEARST | CAMBRIDGE UNIVERSITY PRESS | 2009

READ THE BOOK
The full text of this book can be read free of charge. Select a chapter:

- 0: Preface: an overview of the structure of the book, and a guide to who should read which parts.
- 1: Design of Search User Interfaces: introduces the ideas and practices surrounding search interface design, places modern design in a historical context, and summarizes design guidelines for search interfaces.
- 2: Evaluation of Search User Interfaces: includes informal studies, formal studies, longitudinal studies and top-based analysis including bucket testing. Presents extensive device about how to avoid evaluation mistakes.
- 3: Models of the Information Seeking Process: summarizes the theoretical models about information seeking, and discusses information needs and query intent.
- 4: Query Specification: includes textual queries, natural language questions, query specification forms, dynamic feedback, and operators and commands.
- 5: Presentation of Search Results: includes document summaries, properties of results lists/tables, summaries (tablets) as used in...

BUY THE BOOK!
Order the book at:

- Amazon.com
- Cambridge University Press

CH. 10: INFORMATION VISUALIZATION FOR SEARCH INTERFACES

The preceding chapters have discussed user interfaces to support search, with a focus on what is known to be successful (from a usability perspective) for the vast majority of searchers. This and the following chapter describe efforts to improve search interfaces by incorporating visual information into the display using techniques from the field of information visualization.

The human perceptual system is highly attuned to images, and visual representations can communicate some kinds of information more rapidly and effectively than text. For example, the familiar bar chart or the graph can be much more evocative of the underlying data than the corresponding table of numbers (Carlin and Simon, 1974). The goal of information visualization is to translate abstract information into a visual form that provides new insight about that information. Visualization has been shown to be successful at providing insight about data for a wide range of tasks.

The field of information visualization is a vibrant one, with hundreds of innovative ideas burgeoning on the Web. However, applying visualization to textual information is quite challenging, especially when the goal is to improve search over text collections. As discussed in earlier chapters, search is a means towards some other end, rather than a goal in itself. When reading text, one is focused on that task; it is not possible to read and visually generate something else at the same time. Furthermore, the nature of text makes it difficult to convert it to a visual analogue. Most likely for these reasons, applications of visualization to general search have not

Marti Hearst's Book
Chapter 10

<http://searchuserinterfaces.com/book/>

Fall 2013

CS 7450

27

Transition 1

- OK, let's move up beyond just search/IR
- How do we represent the words, phrases, and sentences in a document or set of documents?
 - Main goal of *understanding* versus search

Fall 2013

CS 7450

28

More Word Counting



WORDCOUNT

◀ PREVIOUS WORD NEXT WORD ▶

the of and to a in that is was for on you be with by to he had one

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

CURRENT WORD

FIND WORD: BY RANK: REQUESTED WORD: THE RANK: 1

86800 WORDS IN ARCHIVE
[ABOUT WORDCOUNT](#)

WordCount™ ©2003 Jonathan Harris | Number27 | Help

<http://www.wordcount.org>

Fall 2013

CS 7450

31

Tag/Word Clouds



- Currently very “hot” in research community
- Have proven to be very popular on web
- Idea is to show word/concept importance through visual means
 - Tags: User-specified metadata (descriptors) about something
 - Sometimes generalized to just reflect word frequencies

Fall 2013

CS 7450

32

History

- 90-year old Soviet Constructivism
- Milgram's '76 experiment to have people label landmarks in Paris
- Flanagan's '97 "Search referral Zeitgeist"
- Fortune's '01 Money Makes the World Go Round

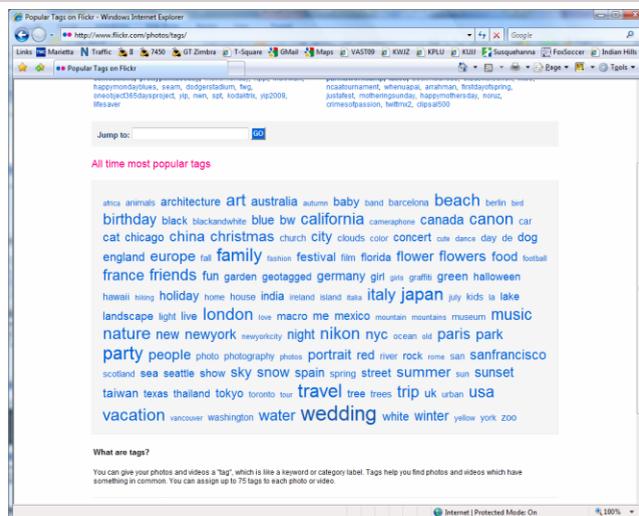
Viégas & Wattenberg
interactions '08

Fall 2013

CS 7450

33

Flickr Tag Cloud

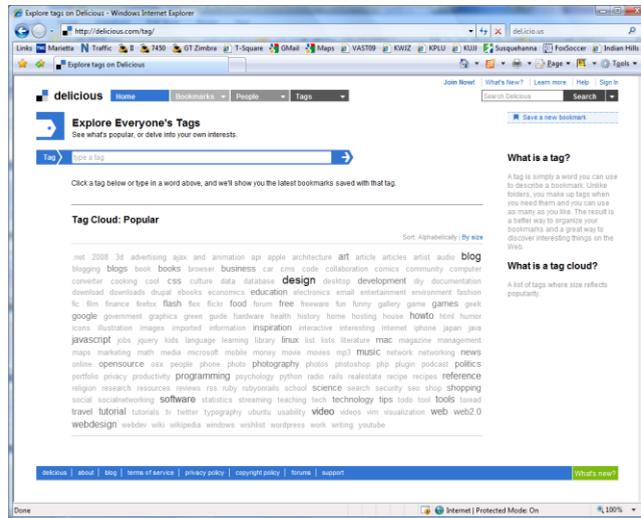


Fall 2013

CS 7450

34

delicious Tag Cloud

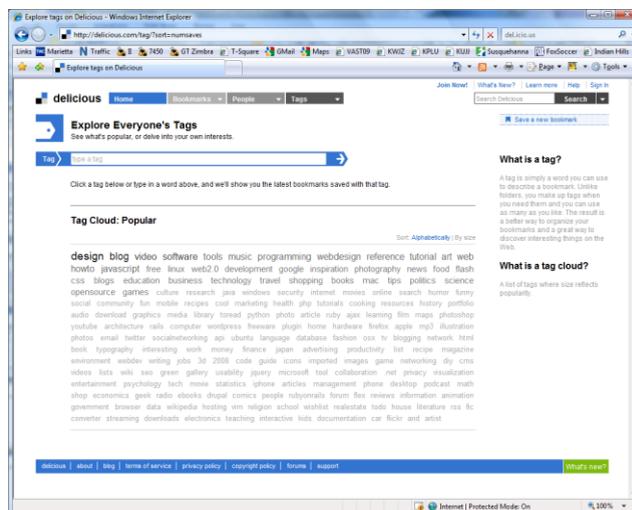


Fall 2013

CS 7450

35

Alternate Order



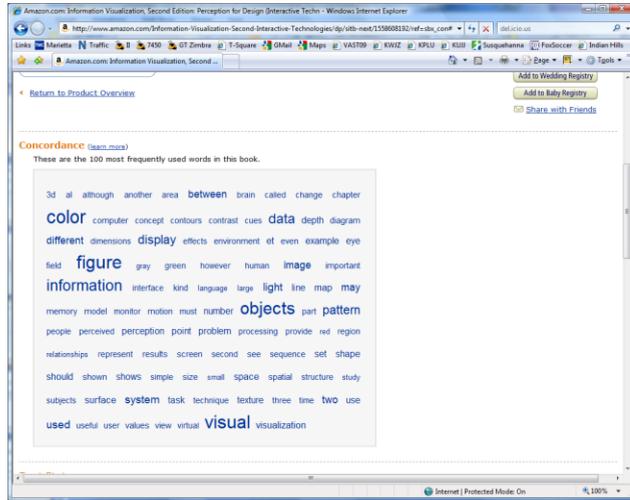
Fall 2013

CS 7450

36

Amazon's Product Concordance

Maybe now a
"word cloud"



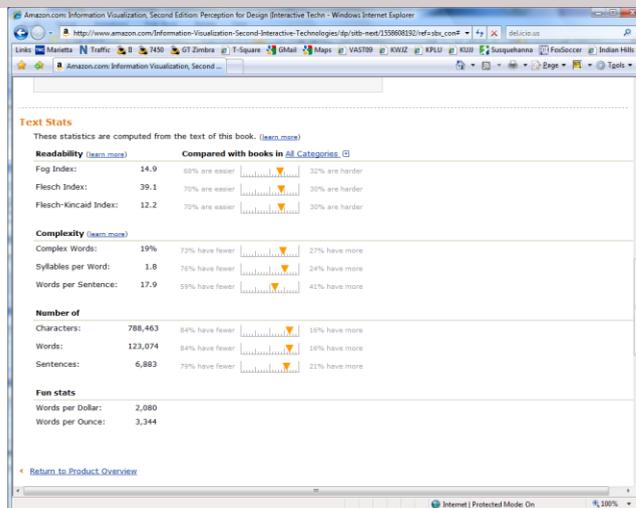
Fall 2013

CS 7450

37

Sidenote

There are
other types
of info about
a document
on Amazon



Fall 2013

CS 7450

38

Many Eyes Tag Cloud



Here, pairs of words are shown



Fall 2013

CS 7450

39

Problems



- Actually not a great visualization. Why?
 - Hard to find a particular word
 - Long words get increased visual emphasis
 - Font sizes are hard to compare
 - Alphabetical ordering not ideal for many tasks
- Studies have even shown they underperform

Gruen et al
CHI '06

Fall 2013

CS 7450

40

Why So Popular?

- Serve as social signifiers that provide a friendly atmosphere that provide a point of entry into a complex site
- Act as individual and group mirrors
- Fun, not business-like

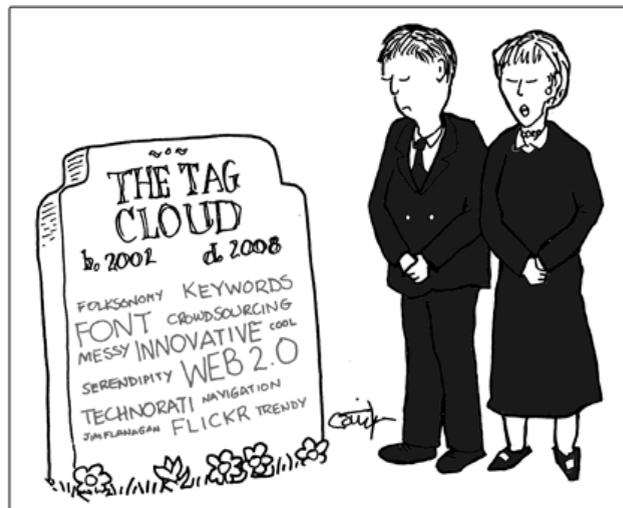
Hearst & Rosner
HICSS '08

Fall 2013

CS 7450

41

NISE TO SIGNAL
Rob Cottingham · socialsignal.com/n2s



<http://www.socialsignal.com/system/files/images/2008-08-01-tagcloud.gif>

Fall 2013

CS 7450

42

Layout Algorithm



- Details not published
- Idea:
 - sort words by weight, decreasing order
 - for each word w
 - $w.\text{position} := \text{makeInitialPosition}(w)$;
 - while w intersects other words:
 - $\text{updatePosition}(w)$;
 - Init position randomly chosen according to distribution for target shape
 - Update position moves out radially

Fall 2013

CS 7450

45

Fun Uses



- Political speeches
- Songs and poems
- Love letters (for “boyfriend points”)
- Wedding vows
- Course syllabi
- Teaching writing
- Gifts

Fall 2013

CS 7450

46

2-day Survey in Jan. 09



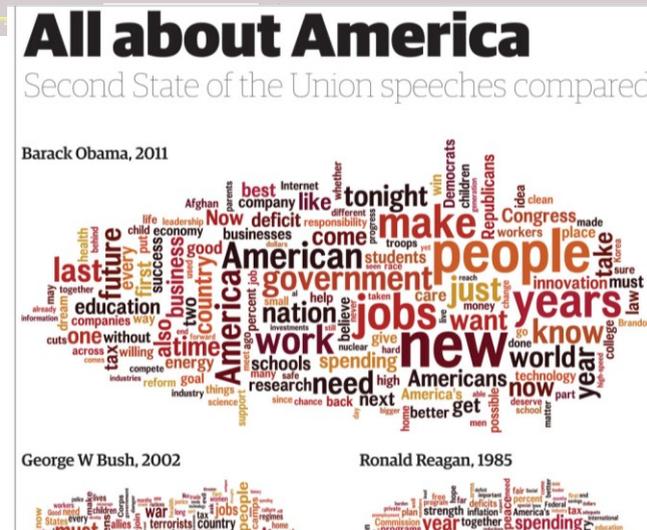
- 2/3 respondents were women
- Interest came from design, visual appeal, beauty
- Why preferred over word clouds:
 - Emotional impact
 - Attention-keeping visuals
 - Organic, non-linear
- Fair percentage didn't know what size signified

Fall 2013

CS 7450

47

SoTU Wordles



<http://www.guardian.co.uk/news/datablog/2011/jan/25/state-of-the-union-text-obama#>

Fall 2013

CS 7450

48

Mani-Wordle



- Start with nice default algorithm
- Give user more control over design
 - Alter color (within a palette)
 - Pin words, redo the rest
 - Move and rotate words
 - Smooth animation and collision detection for tracking changes

Koh et al
TVCG (InfoVis) '10

Fall 2013

CS 7450

51

Video

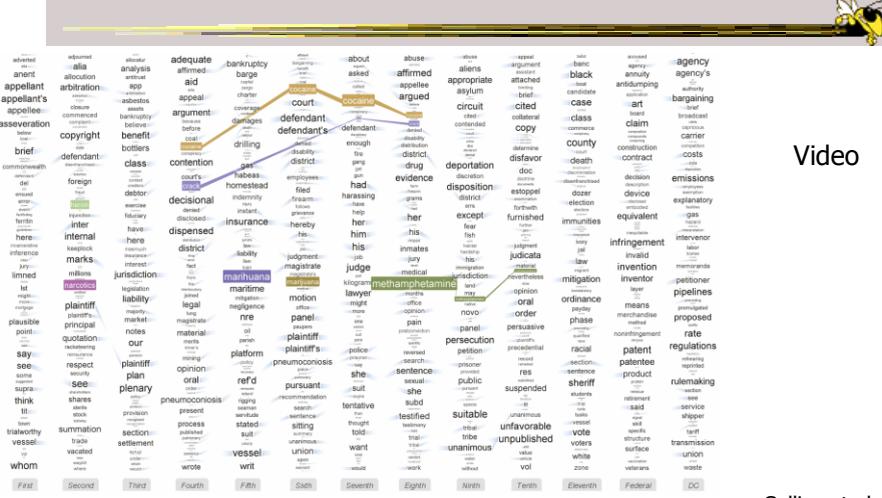


Fall 2013

CS 7450

52

Parallel Tag Clouds



Video

Collins et al
VAST '09

Different circuit courts

Fall 2013

CS 7450

55

Analytic Support

- Note: Word Clouds and Wordles are really more overview-style visualizations
 - Don't really support queries, searches, drill-down
- How might we also support queries and search?

Fall 2013

CS 7450

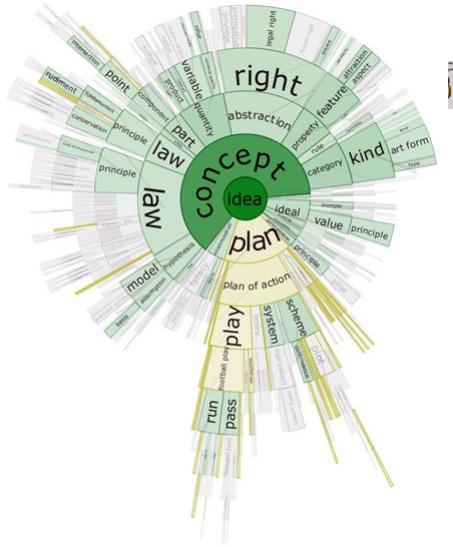
56

DocuBurst



Uses WordNet, sets of synonyms grouped together

Size – # of leaves in subtree
 Hue – diff synsets of word
 Shade – frequency of use



Collins et al
 EuroVis '09

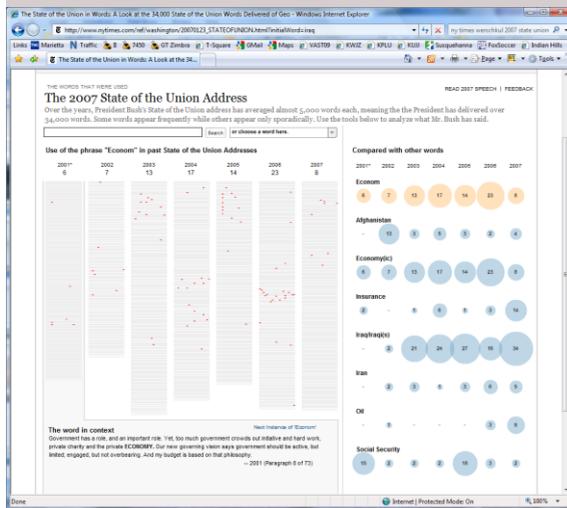
<http://faculty.uoit.ca/collins/research/docuburst>

Fall 2013

CS 7450

57

Overview & Timeline



State of the
 Union Addresses

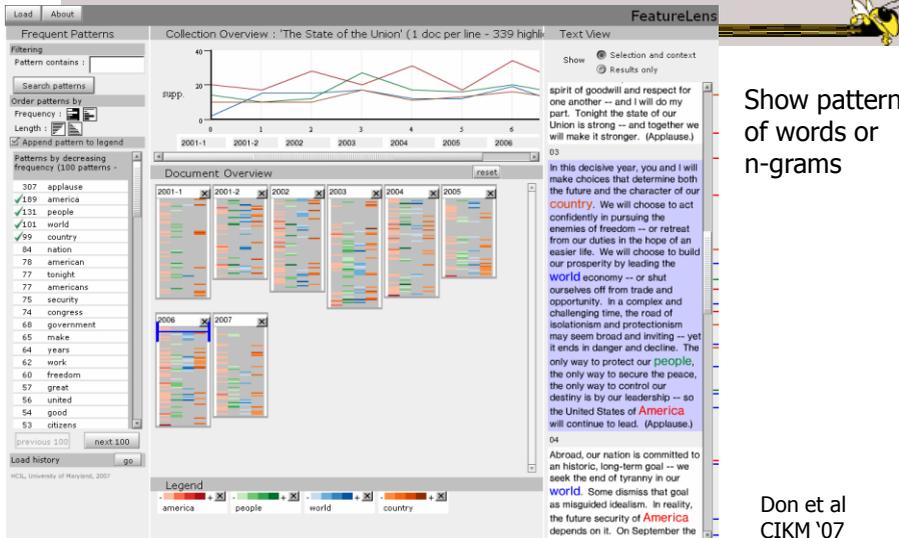
http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq

Fall 2013

CS 7450

FeatureLens

Video



Show patterns of words or n-grams

Don et al
CIKM '07

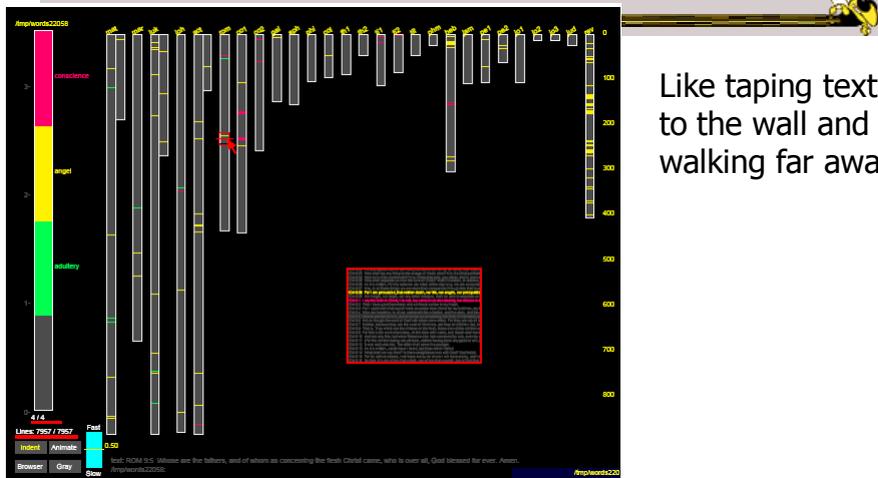
<http://www.cs.umd.edu/hcil/textvis/FeatureLens/>

Fall 2013

CS 7450

59

SeeSoft Display



Like taping text to the wall and walking far away

New Testament

Eick
Journal Comput. & Graph. Stats '94

Fall 2013

CS 7450

60

Beyond Individual Words

- Can we show combinations of words, phrases, and sentences?

Fall 2013

CS 7450

61

Concordance

Definition

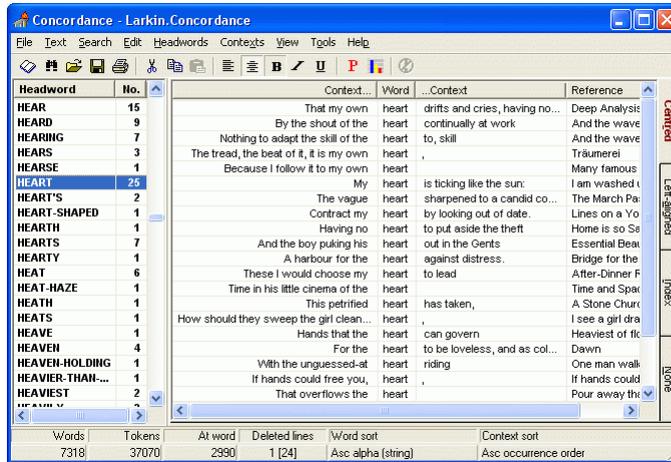
The screenshot shows the Merriam-Webster Online Dictionary page for the word "concordance". The browser address bar shows the URL "http://www.merriam-webster.com/dictionary/concordance". The page features the Merriam-Webster logo, a search bar, and a navigation menu. The main content area displays the word "concordance" with its definition: "1 : an alphabetical index of the principal words in a book or the works of an author with their immediate contexts". A red arrow points from the word "Definition" on the left to the definition text. The page also includes a sidebar with navigation links and a footer with copyright information.

Fall 2013

CS 7450

62

Concordance in Text



<http://www.concordancesoftware.co.uk>

Fall 2013

CS 7450

63

Word Tree



Fall 2013

CS 7450

From King James Bible

64

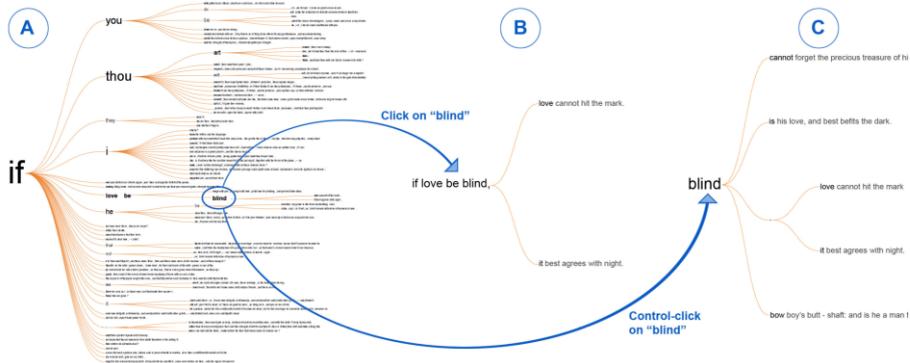
Word Tree



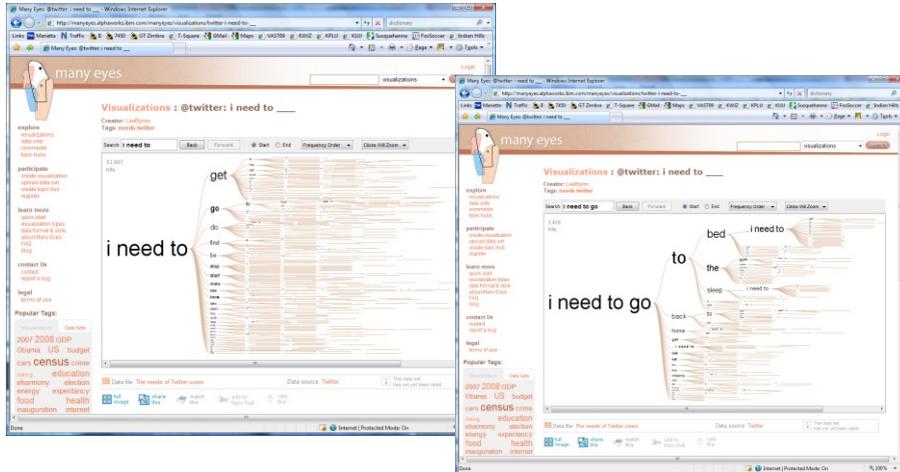
- Shows context of a word or words
 - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

Wattenberg & Viégas
TVCG (InfoVis) '08

Interaction



Many Eyes' WordTree



Fall 2013

CS 7450

67

In Many Eyes now

Phrase Nets

- Examine unstructured text documents
- Presents pairs of terms from phrases such as
 - X and Y
 - X's Y
 - X at Y
 - X (is|are|was|were) Y
- Uses special graph layout algorithm with compression and simplification

van Ham et al
TVCG (InfoVis) '09

Fall 2013

CS 7450

68

User Interface

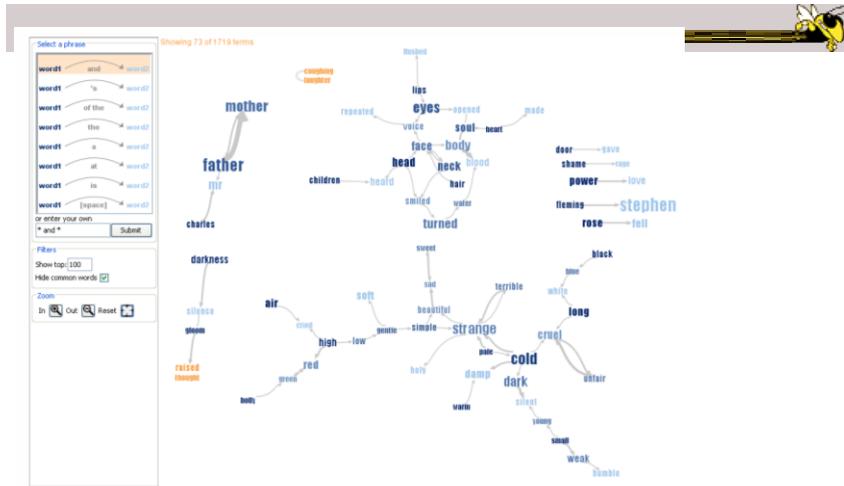


Fig 3. The Phrase Net user interface applied to James Joyce's Portrait of the Artist as a Young Man. The user can select a predefined pattern from the list of patterns on the left or define a custom pattern in the box below. This list of patterns simultaneously serves as a legend, a list of presets and an interactive training mechanism for regular expressions. Here the user has selected "...X and Y...", revealing two main clusters, one almost exclusively consisting of adjectives, the other of verbs and nouns. The highlighted clusters of terms have been aggregated by our edge compression algorithm.

Fall 2013

CS 7450

71

Another Challenge

- Visualize an entire book
- What does that mean?
 - Word appearances
 - Sentences
 - ...

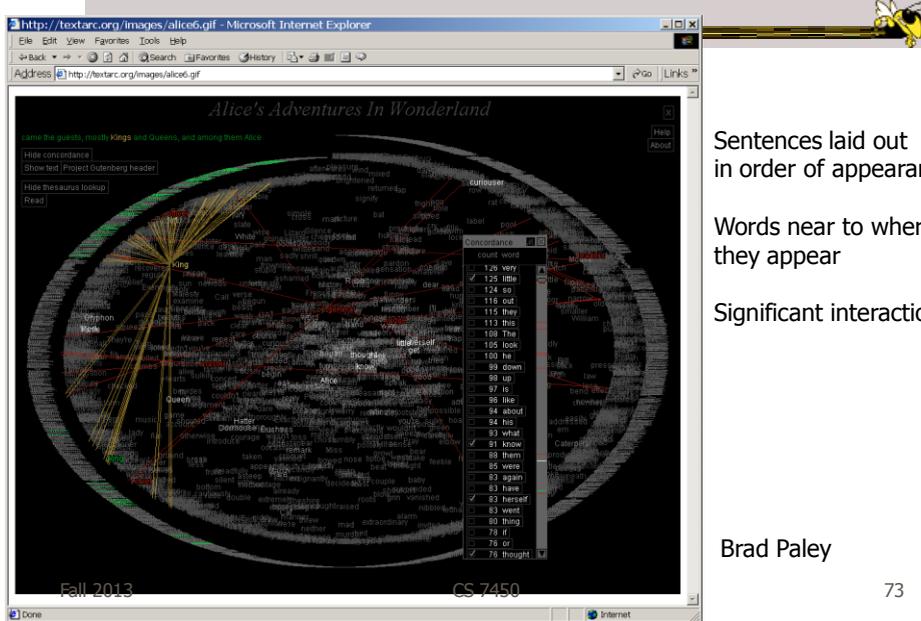
Fall 2013

CS 7450

72

TextArc

<http://textarc.org>



Sentences laid out
in order of appearance

Words near to where
they appear

Significant interaction

Brad Paley

73

Next Time

- More about collections of documents and showing other characteristics of documents
 - Analysis metrics
 - Entities
 - Concepts & themes

Fall 2013

CS 7450

74

Upcoming

- Text and Documents 2
 - Reading
Keim & Oelke '07
- Visual Analytics 1
 - Reading
Keim et al '08

Fall 2013

CS 7450

75

References

- Marti Hearst's i247 slides
- All referred to papers

Fall 2013

CS 7450

76