WordSeer: A Knowledge Synthesis Environment for Textual Data

Aditi Muralidharan
Computer Science Division
UC Berkeley
aditi@cs.berkeley.edu

Marti A. Hearst
School of Information
UC Berkeley
hearst@ischoo.berkeley.edu

Christopher Fan
English Department
UC Berkeley
cfan@berkeley.edu

ABSTRACT

We describe WordSeer, a tool whose goal is to help scholars and analysts discover patterns and formulate and test hypotheses about the contents of text collections, midway between what humanities scholars call a traditional "close read" and the new "distant read" or "culturomics" approach. To this end, WordSeer allows for highly flexible "slicing and dicing" (hence "sliding") across a text collection. The tool allows users to view text from different angles by selecting subsets of data, viewing those as visualizations, moving laterally to view other subsets of data, slicing into another view, expanding the viewed data by relaxing constraints, and so on. We illustrate the text sliding capabilities of the tool with examples from a case study in the field of humanities and social sciences – an analysis of how U.S. perceptions of China and Japan changed over the last 30 years.

1. INTRODUCTION

This paper describes a new tool for exploratory data analysis [3] that attempts to provide frictionless access to text and its metadata. Although many tools have been developed to analyze numerical and categorical data, language in its written form of text has special properties that makes it more difficult to analyze. Text has both linear and hierarchical structure, its meaning is ambiguous given its representation, it has tens of thousands or hundreds of thousands of features, and frequencies of words are usually distributed via a power law.

Even a small fragment of text does not stand alone, but evokes, in the reader's mind, other texts containing the same words, phrases, or ideas. To an analyst trying to make sense of an idea, some associations may be deeply meaningful. Transitions and associations are central to the text analysis process, and people seeking knowledge from text are engaging in *sensemaking*. They do not follow a straight path from data input to analysis output, but meander between analysis, interpretation, exploration and understanding on different sub-collections of data [2].

In this paper, we describe a text analysis tool called WordSeer that supports such transitions[1]. It allows highly flexible slicing and dicing, as well as frictionless transitions (hence "sliding") between visual analyses, drill-downs, lateral explorations and overviews of slices in a text collection. Our tool uses computational linguistics, information retrieval and data visualization, and enables scholars with no technical background to conduct analyses yielding concrete, useful and otherwise inaccessible knowledge.

The design of the tool is motivated by the desire to support the humanities and social sciences (HASS). In these fields, it is common for scholars to have hundreds, even thousands of text-based source documents of interest from which they extract evidence for complex arguments about society and culture.

We introduce the text sliding capabilities of WordSeer with a running example from a real-world case study with C.F. (author 3), a scholar at UC Berkeley's English department. He studied how U.S. perceptions of China and Japan responded to China's rise over the last 30 years using a collection of 5,715 *New York Times* editorials about China and Japan from 1980 to 2012.

A video demonstrating WordSeer through more examples from this case study is at `http://wordseer.berkeley.edu/cikm2013demo` and more videos are at at `http://wordseer.berkeley.edu`.

2. TEXT SLIDING

The paired concepts of *slices* and *views* are central to text sliding. A slice is a set of sentences – usually something meaningful, like a list of sentences containing a given term, or all the sentences in a particular document. A view is a visual representation of the data in a slice: the view can range from a simple vertical list of the sentences in the slice, to more complex linguistic processing combined with visual analytics.

We define text sliding as showing a different view of the same slice, or opening a view of an associated slice, which can consist of drilling down (narrowing, selecting), or broadening (by removing constraints), or following a new thread (moving laterally) or finding related words or sentences (also moving laterally).

2.1 Views

In WordSeer, views are window-like panels, and the user can open up any number of panels in the interface to facil-

[1]WordSeer is a web application. This is version 3.0, which has notably more flexible interactions than older versions.

itate comparison across views. Views contain the following components, as illustrated in Figure 1(b):

1) A drop-down menu for switching to a different view of the same slice (see Figure 5),
2) Breadcrumbs describing the searches and filters that define the current slice,
3) A visualization of the data in the slice. Currently, the choices are:
– A list of sentences,
– A list of documents that match the sentences in the slice,
– An interactive Word Tree [4] of the most common word in the slice, or the search term, if specified,
– Charts showing distributions of the slice's sentence counts across various metadata categories,
– A document reader,
– Bar charts showing how often different words in the slice appear in grammatical relations.
4) Summary statistics of:
– How many sentences within the slice match different metadata categories,
– The most frequent nouns, verbs, adjectives and multi-word phrases in the slice.

The simplest sliding interaction in WordSeer is creating a different view of the same slice. There is a drop-down menu at the top left corner of each view which provides this function (Figure 5). A history panel allows revisiting of earlier views.

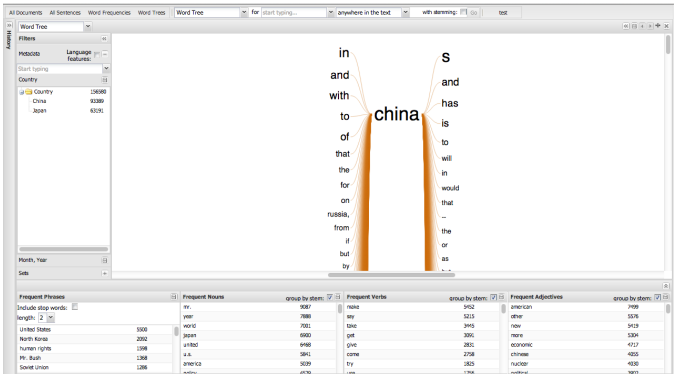
When a user opens up WordSeer for the first time, instead of showing a blank screen a requiring the user to think up a query, the tool provides summary statistics immediately, as shown in Figure 1(a) for the case study. Figure 1(b) breaks this down into its core components. This is a single *view* with a *Word Tree* visualization of the most frequent content word ('China') in the *slice* consisting of "the entire collection".

2.2 Slices

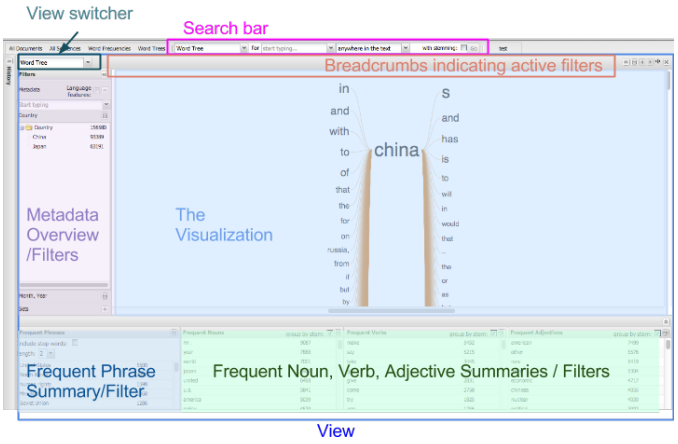
The easiest way to make slices in WordSeer is by intersecting *searches* and *filters*. A search restricts the collection to just the sentences matching the query, and a filter restricts it to just sentences matching a particular metadata value. One of C.F's first goals in the case study was to get a sense of the different ways China was discussed in the '80s, '90s and '00s. To do this, he assembled three slices, one for each decade, by starting with a *search* for "China" (Figure 2) and then filtering the 'Year' category to range over each ten-year period (Figure 6(a)). WordSeer does not require metadata to be numerical ranges, it can also work with categorical values. If he had wanted to, C.F. could have filtered these results to just editorials whose main topic tag was China or Japan, using the controls shown in Figure 6(b).

WordSeer's overviews showed clear differences between the decades. In particular, the increasing frequencies of growth-related verbs contributed to a sense of China's rise, as shown by the frequencies per decade below:

- "grow, growing": 294, 232, 421
- "rise, rising": 101, 134, 249
- "develop, developing": 274, 404, 476



(a) The initial view of 30 years of *New York Times* editorials about China and Japan, for C.F.'s case study.



(b) A breakdown of the components of the user interface

Figure 1: Views in WordSeer.

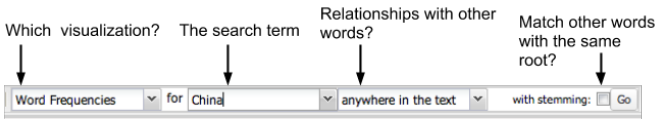


Figure 2: Searching for sentences matching "China".

2.3 Associated Slices

In the database, each sentence is indexed according to the following linguistic phenomena:

- Each word in the sentence, and its part of speech (noun, verb, adjective, etc.),
- Each consecutive two-, three-, and four-word sequence in the sentence,
- Each grammatical relationship in the sentence.

By traversing these indexes, we can compute the associations for a slice. From a slice, we can query for all the words, phrases, or grammatical relations in the sentences in that slice, and from there, to all the other sentences that contain each particular item.

Grammatical relationships are identified using the Stanford dependency parser[1] which extracts many kinds of relationships; some of the more easily understood ones include *noun compound*, where two nouns come together to signify a new concept, *adjective modifier* where an adjective describes another word, and *direct subject* in which a word is the agent of a verb.

Each view automatically presents the most common nouns, verbs, adjectives, and phrases (Figure 7), along with their counts in a panel at the bottom.

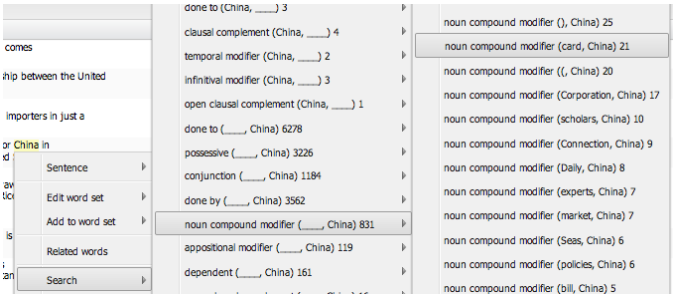


Figure 3: The word menu for 'China'. After selecting *China* > *Search* > *noun compound modifier*, the noun-compound relationship "China card" stood out to C.F.

Individual words are jumping-off points. They can be acted upon wherever they appear via the Word Menu (Figure 3) which enables *lateral movement*. Any time a user sees a word, they can follow up on it by examining the grammatical relations in which it occurs, seeing related words, and creating visualizations of the slice of sentences that contains the word, as well as the slices containing various relationships to other words.

The related words option in the Word Menu shows the nouns, verbs, and adjectives that co-occur most frequently with the clicked-on word. For example, if we click on 'Japan' and open up the related words (Figure 8) the pop-up shows the words that co-occur most frequently with 'Japan' in this collection. Each of these related words can be clicked in turn, opening up a new Word Menu. These menus have the additional option to 'See co-occurences', as shown in the new Word Menu for 'exports'. Selecting that option opens up a new view showing just those sentences in which the two words appear together (in this case, 'Japan' and 'exports').

The word menu reduces friction in both discovery and search. It only takes one menu click to discover that 'exports' occurs frequently with 'Japan', and only one more to see all the sentences in which 'Japan' and 'exports' are mentioned together.

2.4 Custom Slices with Sets

Searches and filters are useful, but cannot always express specific analysis goals. WordSeer therefore allows users to construct custom slices through Word-, Sentence-, and Document Sets. These custom slices behave like any other slices, which means that they can be summarized in views, analyzed, filtered and searched. But they are more powerful that other slices because they also behave like metadata, transforming them into *categorical filters*.

Word Sets are well-illustrated by an example from Case Study 1. One concept of interest in this study was "growth". The scholar wished to confirm his intuitions about China's rise by checking whether growth-related words became more frequent over time in editorials about China. First, he created a new Word Set and typed in some growth-related words "growing, develop, developing, grow, rise, rising" (Figure 9(a)). The result was a Word Set representing a new slice of sentences, those containing at least one of those words.

After the Word Set is created, the entire user interface responds to its presence. The search box now shows a dropdown option for the set (Figure 9(b)). The Word Menu shows the option to add a new words (Figure 9(c)) and the metadata overviews (Figure 9(d)), previously restricted to pre-defined categories, now show this new "category", and allows C.F. to filter based on it.

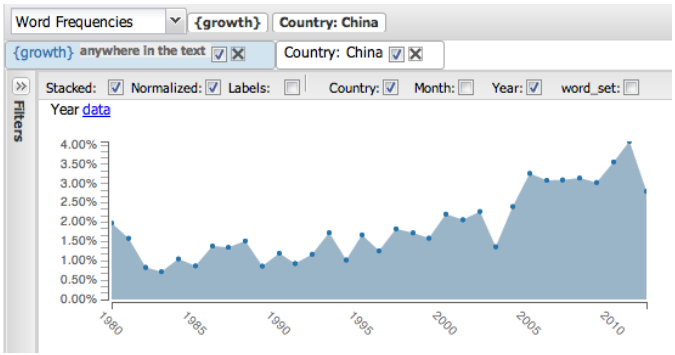


Figure 4: The `{growth}` Word Set used as a Word Frequencies query.

To verify that China was indeed described as rising, C.F. selected the `{growth}` Word Set as his search query, and opened a Word Frequencies view with the `country = China` filter. The resulting visualization is Figure 4, which shows almost a doubling of the frequency of these words in editorials about China over the 30-year period from 1980 to 2012. Satisfied that WordSeer was capable of reproducing this widely accepted fact, he was able to move on to deeper questions.

For sentences and documents, the idea is the same. Users can hand-pick collections of sentences from the reading view, and from search results view, or collections of documents from the document search results view. Once created, all these sets can be overviewed and analyzed like any other slice, and additionally used as filters.

3. REFERENCES

[1] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. ACL '03*, volume 1, pages 423–430, 2003.

[2] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. International Conference on Intelligence Analysis*, volume 1, pages 2–4, MacLean, VA, USA, 2005.

[3] J. W. Tukey. Exploratory data analysis. *Reading, MA*, 231, 1977.

[4] M. Wattenberg and F. B. Viegas. The word tree, an interactive visual concordance. *IEEE Visualization and Computer Graphics*, 14(6):1221–1228, 2008.

APPENDIX

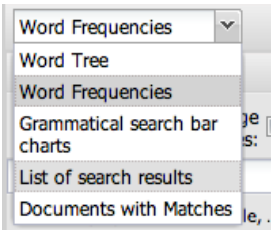
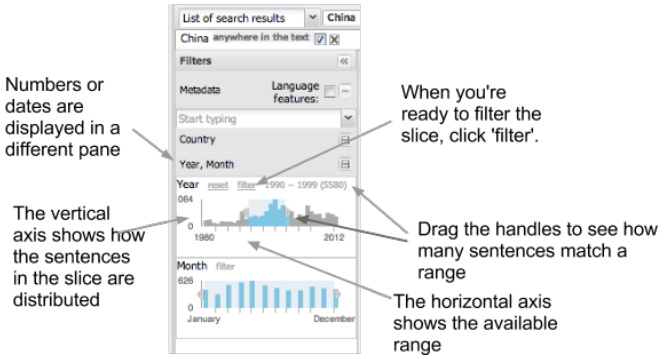
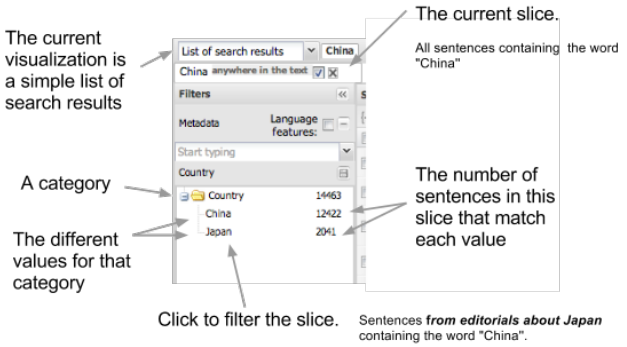


Figure 5: The view-switcher drop-down menu.



(a) C.F. used the date-range overview to select the slice "all sentences from the 1990's matching *China*"



(b) Clicking on a value will filter the slice to match. For example, clicking on 'China' would create the slice "all sentences containing *China* from editorials about China".

Figure 6: Both types of overviews double as filters .

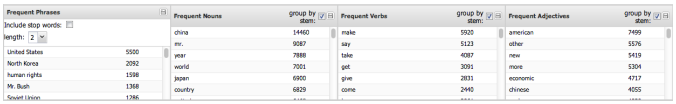


Figure 7: The most frequent phrases, nouns, verbs, and adjectives in the *New York Times* editorials for C.F.'s case study.

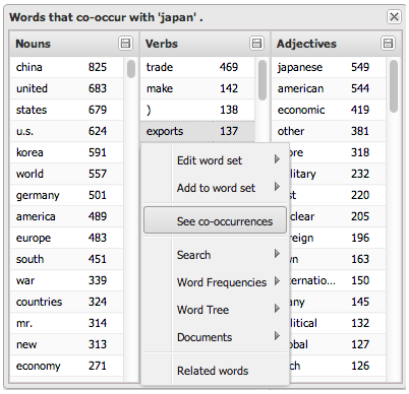
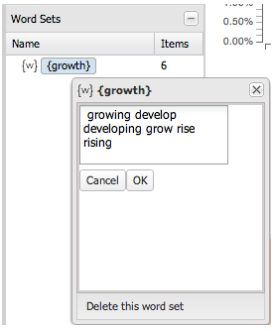
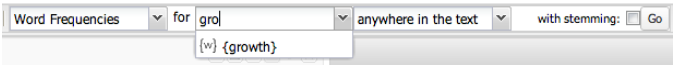


Figure 8: The words that co-occur most frequently with 'Japan'. Clicking on any of these words opens up a Word Menu, this time with the option to see the sentences containing the co-occurrence.

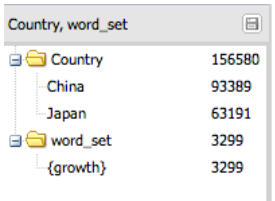


(a) Creating a "growth" word set with 6 words in it.



(b) The set now appears in the drop-down menu in the search box.

(c) The set also appears in the word menu



(d) The set also appears in the metadata overview

Figure 9: Word Sets in WordSeer.