Supporting exploratory text analysis in literature study

Aditi Muralidharan and Marti A. Hearst
University of California at Berkeley, USA

Abstract

We present WordSeer, an exploratory analysis environment for literary text. Literature study is a cycle of reading, interpretation, exploration, and understanding. While there is now abundant technological support for reading and interpreting literary text in new ways through text-processing algorithms, the other parts of the cycle—exploration and understanding—have been relatively neglected. We are motivated by the literature on sensemaking, an area of computer science devoted to supporting open-ended analysis on large collections of data. Our software system integrates tools for algorithmic processing of text with interaction techniques that support the interpretive, exploratory, and note-taking aspects of scholarship. At present, the system supports grammatical search and contextual similarity determination, visualization of patterns of word context, and examination and organization of the source material for comparison and hypothesis building. This article illustrates its capabilities by analyzing language-use differences between male and female characters in Shakespeare’s plays. We find that when love is a major plot point, the language Shakespeare uses to refer to women becomes more physical, and the language referring to men becomes more sentimental. Future work will incorporate additional sensemaking tools to aid comparison, exploration, grouping, and pattern recognition.

1 Introduction

To date, text analysis systems for humanities scholars have focused on aiding interpretation (Fekete and Dufournaud, 2000; Plaisant et al., 2006; Clement, 2008; Llora et al., 2008; Vuillemot et al., 2009; Rockwell et al., 2010). First, they apply some form of natural language processing (NLP) to extract aggregate statistics about word usage, topics, named entities, and parts of speech. Second, they display the extracted information with visualizations such as word clouds, node-and-link diagrams, and lists of word contexts. Such systems make patterns of style, form, and theme visible, and interpretable by people.

Several such text analysis systems exist, but there is a comparative lack of convincing applications of such systems to real humanistic questions. By this, we mean an example of text analysis and visualization applied to a question of literary interest in which it contributed useful insights not previously known to the final scholarly argument. In our present research, we set about creating exactly such a situation. Instead of beginning with a set of computational techniques and attempting to demonstrate their use in literature study, we began with a literary question and attempted to design computational tools to answer it. By observing our scholars’ analysis needs, we combined computational linguistics, visualizations, and interaction techniques into software tools that could help satisfy their requirements.

WordSeer, the system we built, successfully demonstrated that these techniques can be useful to literature scholars. The subject of study was the...
prevalence of stereotypes and conventions in a collection of North American pre-civil war slave autobiographies. Using WordSeer, scholars were able to find initial evidence that certain so-called stereotypical events in the slave narratives were not as common as implied by scholars. These results were presented at a panel at the 2012 Modern Language Association convention (Muralidharan, 2012). Along the way, the tool developed tools for powerful linguistic search and word similarity exploration, and visualizations of word context and prevalence.

However, our experiences with the slave narratives case study made it clear that, even from a computational standpoint, literature study was not a cut-and-dry hypothesis testing activity. Instead, it is a form of exploratory analysis: a cycle of reading, interpretation, exploration, and understanding. As useful as they are, current digital humanities text analysis systems leave the exploration and understanding part of the cycle unsupported. Our work is motivated by the literature on 'sensemaking', an area of computer science devoted to supporting open-ended investigation on large collections of data.

Computer scientists have built up models of the exploratory analysis process by observing professional analysts (people such as business- or government-intelligence analysts, venture capitalists, and lawyers) engaging in open-ended investigations tasks on large document collections. Observational studies found that the process, called sensemaking, tends to be a cycle between two classes of activities (O’Day and Jeffries, 1993; Russell et al., 1993; Pirolli and Card, 2005). The cycle starts with ‘information foraging’, searching for documents believed to be relevant, reading them, and otherwise extracting information from them. This leads to ‘synthesis’, during which sensemakers try to understand the information they have just encountered.

O’Day and Jeffries (1993) investigated the synthesis process with an ethnographic study of fifteen professionals. These professionals regularly engaged in sensemaking tasks on finance- and business-related topics and came from fields such as venture capitalism, statistics, product marketing, and engineering. They were interviewed about their search and information seeking processes, and the interviews audio taped and transcribed. The various reported behaviors were coded into different groups by two independent groups of researchers, and then reconciled into consistent categories. The researchers found that, during the synthesis part of the cycle, sensemakers spent roughly 20% of their time cross-referencing, summarizing, and creating visual representations of information, and the remaining 80% of their time doing the following:

1. Looking for trends or correlations.
2. Making comparisons of different pieces of the data set.
3. Experimenting with different aggregates and/or scaling.
4. Identifying a critical subset of relevant or unique items.
5. Making assessments.
6. Interpreting data to find meaning in terms of domain or problem concepts.

Therefore, according to Hearst (2009), tools that support exploratory analysis should give overviews of the contents of the collection, help users keep track of what they have seen, and suggest what to look for next. They should also encourage the user to try new queries, help find documents similar to those already found, and allow for aliasing of terms and concepts.

Pencil-and-paper tools that we create to deal with sensemaking activities work: it is intuitive and convenient to mark physical books with pencil, to underline their passages, to compare them side-by-side, and to interleave their pages with our own notes and interpretations. As attractive as these ‘physical’ sensemaking activities are, however, they are (at present) simply not translatable to digital text. Even tablet computers, which allow note taking and other types of marking, fall short in that they represent a single page at a time. They cannot be stacked into piles or open multiple documents to different pages at once for side-by-side comparison. Further, these physical sensemaking methods are not scalable to large collections, cannot be shared easily, are difficult to replace if lost or disturbed, and are hard to cross-reference, reorganize, and search.
In this article, we present a new incarnation of the WordSeer text analysis system: our effort to create a sensemaking environment for literature and language study. Like other systems for the humanities, it has search and visualization capabilities, but it also supports sensemaking activities such as collecting and reorganizing information, exploring related words, and annotating and tagging items. At present, it is being used by three groups of literature scholars to analyze the North American slave narratives, the works of Stephen Crane, and the complete works of William Shakespeare. It is available on the web at http://wordseer.berkeley.edu.

We demonstrate the system’s current capabilities by using it to explore the following open-ended question on the Shakespeare collection:

How does the portrayal of men and women in Shakespeare’s plays change under different circumstances?

As one answer, we show how the tool suggests that when love is a major plot point, the language referring to women changes to become more physical, and the language referring to men becomes more sentimental.

In the following section, we give an overview of related text analysis systems in the humanities. Then, we describe our current system and finally conclude with a discussion of future directions for the project.

2 Related Systems

One of the first visual text analysis systems for the humanities was Compus (Fekete and Dufournaud, 2000), which allowed users to visually explore a collection of 16th century legal documents. Each document was represented as a long rectangular column, side-by-side with all the others. The user could search for items of interest and have them highlighted as smaller sub-rectangles in the long document columns. This created a visualization showing the prevalence of that item in the entire collection. Users could filter documents and sort them by date, title, and other metadata.

At present, there are several well-known text analytics efforts in the digital humanities. However, none of them performs computational linguistic analysis of sentence structure, and therefore does not offer related word exploration or grammatical search such as WordSeer. The first is the now inactive MONK project (Unsworth and Mueller, 2009) incorporating the SEASR analysis toolkit (Llora et al., 2008). This project offered two computational linguistics tools in addition to word distribution and frequency statistics: tagging words with their parts of speech and extracting named entities. Users could visualize occurrence patterns of word sequences within a chosen text and plot networks of how often named entities occurred near each other. This research led to visual text-mining analyses of Emily Dickinson’s correspondence (Plaisant et al., 2006), and of Gertrude Stein’s ‘The Making of Americans’ (Clement, 2008). It also resulted in the development of an interface for exploring the parts of speech used near query words of interest (Vuilleumot et al., 2009).

The second is Voyant (Rockwell, 2003; Rockwell et al., 2010), which operates entirely at the word level. It allows users to plot word frequencies, see concordances (contexts in which words occur) and create tag clouds. For visualization of word prevalence, there is a single scatter plot. It shows users whether query words occur toward the beginnings or the ends of documents, but not how common the words are overall, or whether they are more prevalent in certain documents than others.

Third, WordHoard (Mueller, 2008), and the associated MorphAdorner program (Mueller, 2009) focus on morphology and word forms. Using side-by-side tables, but no visualizations, WordHoard allows users to compare the frequencies and contexts of words across parts of speech, and various different categories of text. It relies on rich metadata, such as (in Shakespeare) speaker gender and sentence type.

Finally, DocuScope (Ishizaki and Kaufer, 2011) is software that was developed to assist in ‘corpus-based rhetorical analysis’. Linguists can compile patterns of language use into hierarchical categories and then visualize their prevalence and distribution within a text collection. It has been used to study...
Shakespeare (Hope and Witmore, 2004) and Arab-American news (Al-Malki et al., 2012).

3 Current System

3.1 Input
The tool is run on a collection of text documents. The input is a set of XML files in a directory, each representing a document in the collection, and the output is a web application with search, visualization, and annotation capabilities. We chose XML because TEI (TEI Consortium, 2011), an XML specification for encoding documents, is a widely adopted digitization standard in the humanities. Many documents of interest to literature scholars are encoded as TEI-XML files.

3.2 Search
A scholar might begin an analysis of the portrayal of men and women in Shakespeare with a question, ‘what are some things that are his and some things that are hers?’ With a keyword-search system, an initial query might be his for men, or her, and hers for women. Such a system would return an unstructured list of results. The word his is always a possessive pronoun, so word sequences containing his would nearly always be relevant. But her can also be a third-person pronoun and will yield constructions such as ‘I told her that X’ and ‘I gave her the Y’.

A standard approach in the literature study is to view search results in a ‘concordance’: a list of all the sentences in which a word occurs, with the target word aligned in the center of the view, exposing the

Fig. 1 Word tree (Wattenberg and Viegas, 2008) for the word ‘her’ generated by WordSeer on Shakespeare's complete works
contexts to its left and right, sorted in some manner. WordSeer uses the word tree visualization (Wattenberg and Viegas, 2008) which makes common contexts in a concordance easier to view by grouping them in an arced tree-like structure. The word tree for her is shown in Fig. 1. Some words like beauty stand out, but constructions like her own muddy the picture.

With the WordSeer project, we make headway on this problem by providing an easy interface to view the results of ‘grammatical search’ (Fig. 2). The system uses NLP to extract relationships between words—see Jurafsky and Martin (2009) for an overview—and allows users to specify ‘both keywords and’ relationships between them. With this feature, a scholar can, for example, take advantage of the fact

![Fig. 2 Grammatical search results for ‘possessed by’ his, filtered on the word father. Individual search results, corresponding to matching sentences, are highlighted to show the words in the relationship](image-url)
that possessive relationships between words can be automatically detected, to express the question precisely: ‘what are all the words with which her has a possessive relationship’.

In the tool’s search interface, pairs of words are specified using input boxes, and the relationship between them is selected from a drop-down menu (Fig. 3). Leaving a word-input box blank returns all matches. We know of no other text analysis system in the humanities that applies this technology.

By comparing the results of various searches, a scholar can begin assemble a picture of language use around a concept. For example, Fig. 2 shows search results for all words for which his has a possessive relationship. The top few words are head, hand, father, life, and heart. Comparing these words with those for her (Fig. 4) reveals immediate differences. The word father is most common for her, with husband, and son close behind. Several body parts enter the picture: eyes, hand, face, tongue, lips, and cheek. A patriarchal picture begins to suggest itself: women’s most commonly mentioned possessions are their male relatives and their bodies.

### 3.3 Collections

Upon seeing the physical, patriarchal portrayal of women revealed by grammatical search, a scholar might wish to investigate whether this picture is consistent, or whether it changes in different types of plays. WordSeer’s collections feature (Fig. 5) supports this mode of analysis by allowing users to collect and organize documents into hierarchical sub-collections. Collections can be moved, renamed, merged, and modified. The document listing (Fig. 6) is used to add the appropriate plays, which are sortable and filterable by date, title, full-text search, grammatical search, and length.

In our example, the collections feature can be used to divide the plays into ‘comedies’, ‘tragedies’, and ‘histories’—the three most commonly accepted categorizations of Shakespeare’s plays. Collections are created using the ‘collections’ bay, a collapsible window at the bottom of the screen. Temporal differences can be investigated by creating the ‘pre-1600’, and ‘post-1600’ categories.

### 3.4 Visualization

For comparison and visualization of word prevalence, we have created the tool’s ‘newspaper-strip’ visualization (Eick, 1994) (Fig. 7a). Each play is represented as a long column. Within each column, small, colored horizontal blocks (corresponding to ten sentences each) highlight the presence of a match. Hovering over a column (e.g. ‘Much Ado About Nothing’ above) darkens it and displays the title. Hovering over a highlighted block displays the matching sentence.

Comparing the prevalence of different search words and grammatical relationships across
different collections can give thought-provoking results. For example, our scholar might compare references to her body parts with references to her relatives in the different collections. The resulting visualizations would show that the ‘comedies’ (Fig. 7a) look much like the ‘tragedies’ (Fig. 7b) with regard to the number of references to her body parts (dark dashes) and her relatives (light dashes). In the ‘histories’, however, an interesting pattern emerges (Fig. 7c). The visualization reveals that references to her body parts (dark dashes) are somewhat less prevalent in the historical plays, when compared with the comedies and tragedies, but references to her family (light dashes) are about as prevalent.

Our scholar can investigate why her body parts seem less prevalent in the historical plays by hovering over a few highlighted blocks corresponding to body-part results (dark dashes). A rough sample will find that many of the sentences in which body parts occur are romantic in nature. The scholar can use the reading and annotating interface to follow up on this phenomenon by clicking on the highlighted blocks in the newspaper-column visualization.

3.5 Close reading
WordSeer supports quick, large-scale analysis through search and visualization, but in all cases maintains links back to the source text. The reading interface (Fig. 8) loads up the full text of the document, and the system automatically scrolls to the relevant sentence and highlights it.

For example, hovering over a blue or orange highlighted block in Fig. 7a or c brings up a popup displaying the matching sentence. Clicking opens the reading interface to that point. The example scholar could thus quickly find interesting mentions of women’s body parts or relatives and...
3.6 Annotation

WordSeer allows sections of text to be highlighted as ‘snippets’ (Fig. 9) to which notes and tags can be attached. This functionality is designed to help scholars keep track of interesting sections of text and to organize them into conceptual categories.

In our example, WordSeer’s annotation facilities can be used to tag speeches referring to female body parts by the topics the speeches seem to contain. In this particular case, it soon becomes apparent that many of the mentions are speeches by a lover.

Fig. 7 Comparing the prevalence of body parts ‘possessed by’ her (eyes, lips, cheeks, and face) (dark dashes) and relatives ‘possessed by’ her (husband, father, sons, daughters, and children) (light dashes) in the a) comedies b) tragedies and c) histories. Each column is a play, represented in alternating shades of gray.
Our scholar might therefore reasonably form the hypothesis that female body parts are predominantly referred to in romantic contexts.

### 3.7 Exploration

For exploration of style and language, WordSeer uses computational linguistics to calculate ‘related words’. These words are either commonly used in similar contexts (Lin, 1997) or commonly used within a ten-sentence window of each other. Clicking on any word while reading brings up a small window showing related words.

In our example, the related words for body parts (Fig. 10) help strengthen the hypothesis that female body parts are predominantly referred to in romantic contexts.
body-part mentions are predominantly associated with romance. The popup shows that other body parts are frequently mentioned, along with love, fair, and sweet.

3.8 Assembling evidence
The process of constructing a final argument involves verifying hypotheses and collecting supporting evidence. In our example investigation, the hypothesis that female body-part mentions are associated with romance can be tested by creating a final two categories of plays. These are ‘love-stories’ for plays in which love is a major plot point and ‘not-love-stories’ in which love is not. When the plays are reorganized along these lines, the results are immediate.

In the love-stories collection (Fig. 11a), we see ‘both’ body parts (dark dashes) ‘and’ male relatives (light dashes). In contrast, the not-love-stories collection (Fig. 11b) shows ‘predominantly’ male relatives (light dashes), and hovering over the occurrences of body parts (dark dashes) reveals a gloomy picture of her tear-stained cheeks and her sorrowful eyes.

Grammatical search results agree with the newspaper-strip visualizations and related words. We see more physical attributes ‘possessed by’ her in the ‘love’ collection than in the ‘not-love’ collection (Fig. 12a and b).

The grammatical search results show that the language around men changes as well (Fig. 12c and d). In the not-love case, the only woman to appear is mother, at number 20, but in the love case, wife takes first place, followed by favor. Compared with the physical language for women, these words have a more sentimental quality.

Thus, a scholar can gather evidence to suggest that, while a male-dominated picture of both men and women is always present, the physical aspects gain prominence for women in plays about love. For men, the more sentimental aspects come to the fore.
Discussion

The previous section demonstrated WordSeer’s capabilities as an exploratory analysis environment through an investigation of gender portrayals in Shakespeare. By comparison, this analysis would be impossible or extremely time-consuming in other tools developed for humanities scholars such as MONK (Unsworth and Mueller, 2009) and Voyant (Rockwell, 2003; Rockwell et al., 2010). For example, in either of these systems, searching the plays for items possessed by women would entail

Fig. 12 Comparison of grammatical search results for possessed by his and her in the love and not love plays. Possessed by her: a) not love and b) love and possessed by his: c) not love plays and d) love plays.

4 Discussion

The previous section demonstrated WordSeer’s capabilities as an exploratory analysis environment through an investigation of gender portrayals in Shakespeare. By comparison, this analysis would be impossible or extremely time-consuming in other tools developed for humanities scholars such as MONK (Unsworth and Mueller, 2009) and Voyant (Rockwell, 2003; Rockwell et al., 2010). For example, in either of these systems, searching the plays for items possessed by women would entail
keyword search with her, and guesses such as eyes, face, and beauty. These systems return an unstructured list of results that is fairly uninformative. They provide no way to discover the vocabulary of the collection—what are all the things possessed by her?—and no way to visualize, and compare, the collection-wide prevalence and predominant usage contexts of a set of terms.

A caveat to such an analysis, however, is that even state-of-the-art language processing algorithms are not 100% accurate. The bigger the disparity between their training data and the input text, the more their accuracy degrades from the reported figures of around 85% (Jurafsky and Martin, 2009).

5 Future Work

Our tool will be improved and evaluated through extensive case studies of real questions with our literature-scholar collaborators. These, along with the exploration, collection, and comparison features are the subjects of ongoing work. For instance, our collaborators say that when studying patterns of imagery, theme, and rhetoric, it would be useful if the system suggested other passages expressing similar concepts. A number of NLP algorithms for calculating text similarity and grouping words are available (Lewis et al., 2006; Mihalcea et al., 2006; Metzler et al., 2007; Islam and Inkpen, 2008). We plan to incorporate these into the system with relevance feedback so users can mark relevant passages and refine the matches the system delivers to suit their particular interest (see Ruthven and Lalmas (2003) for a review).

Second, in order to compare visualizations and grammatical search results, users currently have to switch between collections using a drop-down menu. A side-by-side interface would be less cumbersome. It would also be useful to highlight differences in frequency and distribution between data represented with the various visualizations.

Finally, we plan to make WordSeer applicable to any text collection. The three versions of WordSeer—Shakespeare, Stephen Crane, and Slave Narratives—are publicly accessible through wordseer.berkeley.edu. This website also links to open-source repositories containing code for both the front-end interface and the back-end text-processing system. We plan for an open version of WordSeer—one that can be used on any collection—to be released in 2013. A scholar will be able to go to wordseer.berkeley.edu, upload a collection of text files, and (after the text-processing algorithms finish) be presented with a WordSeer instance for that collection.

Funding

This work was supported by NEH Digital Humanities start-up (HD-51244-11).

References


