Visual Exploration of Text Co-Occurrences

Brandon Liu
University of California, Berkeley
bran_liu@berkeley.edu

Figure 1: co-occurring terms of 'Israel' in the first 30 books of KJV.

ABSTRACT
This paper investigates the design of visualizations for exploring co-occurrences. Co-occurrences are defined in terms of ordered, discrete events. Examples of such corpora are Twitter (unit = tweet, order = time, event = appearance of hash tag) or the King James Bible (unit = line, order = book, event = appearance of word). Standard approaches to exploring co-occurrence are simple frequency counts. This paper presents a new, intuitive approach to displaying co-occurrences that combines elements of area-proportional (Venn) diagrams and stacked bar graphs. The technique enables human analysts to better understand the nature of co-occurring events. Detailed design considerations are presented with justification. The technique is integrated into an application for exploring real-world data.

Author Keywords
Information visualization, natural language

ACM Classification Keywords
H5.2. [Information Interfaces] : User Interfaces.

INTRODUCTION
A common operation in exploratory data analysis of text is finding terms that co-occur with a certain term, or co-occurrence. One example is asking for the most frequent co-terms (what we will call a co-occurring term of a given term) and receiving a list. Given this information, it is easy to construct a node-link diagram or tag cloud based on scalar frequencies.

However, there are many reasons why the scalar value of frequency does not tell the whole story about a co-term. One reason is that if the text is ordered, the co-occurrence may vary over time. To represent this change over time with tag clouds or node-link diagrams, one would have to animate the visualization, which usually faces usability problems. Another reason why simple frequency counts on co-terms is potentially misleading is because some co-terms are just more frequent. In natural language processing, the metric for determining this ‘importance’ is called term frequency-inverse document frequency (tf-idf). For example, our co-term summary may tell us that ‘lord’ is a frequently occurring co-term of ‘Israel’ in the King James Bible (abbreviated KJV), but this fact is not relevant to the analyst, since ‘lord’ is a frequent co-term of nearly anything. One common way to determine this is by ‘drilling down’ into event B: that is, perform the same analysis you just did for event A. Having to drill down on each co-term is a time-consuming process: rather, one would want to have all of this co-term information immediately.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2011 ACM 978-1-4503-0267-8/11/05...$10.00.
Thus, our goal is to develop a way to visualize not only that event B (e.g. ‘lord’) occurs often with event A (e.g. ‘Israel’), but also to visualize how often event A occurs when event B does. Furthermore, we are interested in creating a visualization system to enable exploratory analysis – that is, finding interesting points in the data that are not known a priori. Finally, our interface should be intuitive in the sense that the metaphors used should make sense to an end user who does not have significant experience working with visualization.

We spend a majority of the paper considering design choices towards these goals.

RELATED WORK
Node-Link diagrams are one of the dominant ways of visualizing co-occurrence. Many Eyes provides a visualization for the King James Bible [2]. These are an effective method for co-occurrence, showing the existence of co-occurrence and also the degree, but does not show historical data.

Don et al. [1] present an application to explore interesting uses of language. The benefit of this paper is it specifically addresses how to make co-occurrence data more accessible to those who do not have a lot of visualization experience. Part of this system enables of tracking co-occurring patterns in a corpus, and indicates this by encoding the degree of co-occurrence using value. This paper specifically addresses intuition by using a consistent metaphor of overlapping areas that is easy to understand.

Figure 2: Phrase Nets (from Wattenberg)
Viegas and Wattenberg[4] describe popular techniques for visualizing text such as presidential speeches. Their work is related because the techniques they describe are for exploratory analysis: to help the viewer gain an understanding of which terms are used in the text, without actually displaying the text in its entirety. A similar technique is presented by Ham, Wattenberg and Viegas [5] is Phrase Nets, which specifically address co-occurring words text. These are similar to node-link diagrams.

Figure 3: Tag Clouds.
Figure 4: Sparklines in Google Analytics.
Sparklines are designed to avoid recency bias: Instead of showing the value of a variable at the current time, the analyst can also look into the past, and see trends side-by-side in a dense display.

Heer et al. [3] explore design principles for data-dense displays of time series data – in particular, layout techniques mirror charts and horizon graphs. Some of the same principles also apply here. This paper takes a similar approach in that data density is maximized, but towards a different goal: we look for ‘interesting’ patterns, rather than attempting to perceive numbers correctly. However, perception is still not impacted due to some of the design decisions that we make.

Byron, Wattenberg in [8] investigate a related class of graphs that we investigate here that they call streamgraphs. They specifically look at the ordering, color, and layout decisions to help the viewer best perceive trends in the data, such as shifting the baseline. We consider many of the same approaches in finding the best design for our specific task.

METHODS
Layout
The specific layout of area-proportional charts was the subject of refinement. In the beginning, we took the most basic approach, and attempted to optimize on three points:

1. Ability to perceive the area of interest – the overlapping sections.
2. ‘intuition’, or how well the concepts expressed visually correspond to the abstract notion of co-occurrence.
3. Aesthetic considerations related to how prominent colors and features are.

We describe the layout design process, starting from the beginning. In each case, the second and third most frequent co-terms of ‘Israel’ in the KJV are visualized (they are “king” and “children”, respectively, with “king” appearing in light blue and “children” appearing in purple-pink). The first layout considered was a parallel display where the term of interest (‘Israel’) is displayed on top, and the co-term is displayed below, with darker areas corresponding to the intersection.

The problems with this layout are 1. The area of intersection is repeated across graphs 2. It does not make efficient use of space, and 3. It is difficult to compare the height relative to 0 of both the term and its co-term.

The second layout considered was a stacked area chart in which each frequency was added. Some algebra is required here: namely, the area for the intersection is subtracted from both the term and co-term areas, so the intersections are not double-counted.

The benefits of such a layout are that it takes up half the space of the previous layout, and relative frequencies of the term and co-term can be easily compared. The drawbacks are the same as for general streamgraphs: areas, such as in the bottom graph, are difficult to judge when they are highly sloped. Also, The wide variations of the topmost layer of the chart are highly affected by the bottom layer: these two data series are unrelated to each other.

Figure 5: Streamgraphs (from Byron & Wattenberg)
Benson, Lafleur et al. explore visualizing text co-occurrence using Euler, or Venn diagrams. They devise a heuristic for approximating area-proportionality in co-occurring data. One novel aspect of their research is extending co-occurrence to beyond that of pairwise co-terms: for example, one word may overlap with multiple areas. This points to a larger area of research in creating area-proportional diagrams of overlap.
The next layout considered was flipping the co-term area around the X-axis. The benefits to this layout are that it is just as dense as the previous layout, and also allows for easy comparison of positive (term) and negative (co-term) values. The area of intersection is in the middle, and since this is the area we are most interested in, it is easy to compare it to 0. Finally, we avoid the erratic layering of the previous layout, since we are always dealing with no more than 2 layers (excluding the intersection).

Commenters on this last layout noticed that it became hard to interpret the overlapping areas as 2 shapes: in order to preserve the metaphor suggested of intersecting areas like in a Euler Diagram, the straight edge of the top area led to the perception of 3 separate areas.

Thus, a small modification was made where the area of intersection itself is symmetric around the X axis:

This layout technique best preserves the perception of the blue and purple/pink areas as overlapping in the center. The drawback to this modification is that no area chart is now based against the X axis.

Figure 6: The final layout strategy.

The layout of this stacked chart is extremely simple to compute and can be drawn with any basic curve routine. By drawing the term and co-term areas first, we can draw the area of intersection above them and avoid any complex curve math. Our implementation uses basic Catmull-Rom splines.

Concept and Examples

Commenters on the layout techniques presented suggested that an aid to a new user of the visualizations is a ‘legend’ showing each possibility of overlap. A legend for the visualization technique would look as follows:

Figure 7: Term and Co-term always occur together. Chart is symmetric around the X-axis.

Figure 8: Term and Co-term have considerable overlap.

Figure 9: Term and co-term have no overlap.
Why might the chart in Figure 5 occur, if there are no co-occurrences, in a visualization designed to explore co-occurrence? It is important to note that the situation in Figure 5 occurs in an application either only at one point in a timeline, or if the intersection area is so small that it is not graphed at all at the scale the chart is being displayed.

Color
The color of areas in the diagram was the focus of detailed investigation. In particular, three design principles were discovered:

1. The hue of the area of intersection should be distinct from either that of the term and co-term areas.
2. The area of intersection should be a darker color than either of the single-term colors.

In an intuitive sense, the areas of the term and co-term can be interpreted as partially transparent. This corresponds to our intuitive sense of ‘overlapping’ data. This is suggestive of physical objects that let some light pass through, such as colored glass. Thus, the color of the intersection can be determined programmatically by making one layer partially opaque:

After experimenting with this approach, we found that a better principle with which to determine the color of the intersection area was with the ‘multiply’ blending mode: this can be described by:

\[
\text{component(intersection)} = \frac{\text{component(term)} \times \text{component(co-term)}}{255.0},
\]

where component is each of Red, Green Blue and we are working where each channel is represented in 8 bits. This mode of blending is popular in the design of Venn Diagrams. This produces colors such as the following:

Even with this principle in mind, care must be taken not to choose colors in which the hue of the intersection is too close to that of either the term or co-term areas:

Since the area of intersection is usually small relative to the other two areas, such a color combination can be difficult to perceive.

What does Real Data Look like?
After exploring some real datasets, we discovered several aspects of language data that significantly affect the design of visualizations:

1. The area of intersection is always a small fraction of either the area of the term or the co-term:
2. There are always very common co-terms that dominate the graphs, even after eliminating common stop-words.

Thus, two other areas where design choices were investigated were how to scale the displays, and what mapping to use.

Choice of Scales
In the KJV corpus, the following were the top 3 co-terms of ‘israel’:

For the purpose of comparison, all of the charts are scaled to the same axis. However, the drawback is that the terms that co-occur very frequently, such as ‘Lord’ in the KJV, dominate the charts: thus, all the charts are normalized to these high frequencies, and the areas of intersection become less prominent.

We investigated using a logarithmic scale for the data, as this would help emphasize smaller areas of intersection, and compress large peaks. However, this leads to problems in interpreting the absolute ratio of intersection to total (these terms co-occur about half the time). Another possibility is using horizon graphs as described in Heer et al [3], but the layer of values would conflict with the values used to indicate overlap area. Another possible solution is percentage: however, this would cap the chart out at 100% and not indicate the total frequency at some point. Thus, in our first implementation we simply normalize the entire system to the frequency of the term being investigated (‘Israel’) and truncate the chart of the co-term: the rationale is that this best emphasizes the intersection area relative to the total area of the term, and if the co-term has a prominent area, this can be inferred:
RESULTS

The application was implementing in the Processing framework. The default Catmull-Rom splines were used to draw smooth charts. The architecture of the application was such that a “Area Time Series” was a first-class data type in the system that responded to algebraic operations: thus, time series could be “added” and “subtracted” from each other, simplifying the code.

The backend for the search system was MySQL. An appropriate text corpus was found, and full-text search indices were built over the text columns. The queries for intersections were performed directly in the database using MySQL’s Boolean mode. Other relational databases support similar features. Stopwords were eliminated by ignoring them in the fulltext search index: this also included all words less than 3 characters long.

The visual layout of the application was made to be data-dense, facilitating easy comparison across all terms.

One point that should be made about the design of the application is the way it issues queries. For each co-occurring term of a word, the application issues a separate database query. This order-of-magnitude increase is marginal for a database available locally, but unsustainable for applications that send queries via an API. For example, finding 30 co-terms of a hashtag on twitter would result in issuing 30 API requests: the API has a limit of 350 requests per hour, so it would be impossible to return results at interactive speeds.

The performance of the system running locally was very good. To retrieve the top 30 co-terms of a given word over the KJV corpus, the average time to display all data was 1400 milliseconds, or 1.4 seconds. The speed at which the system operated allows one to be curious about the data and do fast, iterative searches.

There were numerous dimensions on which we could have done a user study. These include

- How well is area perceived using our color system?
• How well are intersecting shapes perceived using our color system?
• Are numbers able to be estimated from the data quickly?

In the interest of time, a simple method was used to determine if the layout considerations actually produced ‘insights’. Inexperienced participants were invited to evaluate the system as a whole; for the discovery of interesting co-occurrences. Our evaluation method was, in a set amount of time, be able to point out “interesting” areas of data co-occurrence. One flaw with this method is that since what makes a co-term ‘interesting’ has not been quantified, only positive results could be found: People using the new method could find an ‘interesting’ term that the old method did not deem ‘interesting’, but participants never marked a term as ‘uninteresting’.

Each participant began with a search on the term ‘Israel’. They were first asked to comment on the relationship between the top 3 co-occurring terms, which were ‘lord’, ‘children’, and ‘king’. All of them were able to deduce the difference between the three terms:

While ‘Israel’ co-occurs the most with ‘lord’, this is mostly due to chance. ‘Children’ and ‘Israel’ appear to be correlated in the earlier parts of the Bible, while ‘King’ and ‘Israel’ are correlated in the books after those.

Participants were then asked to look for other interesting data in the application. One of the participants pointed out that ‘david’ and ‘moses’ had interesting areas of overlap: thos was moreso then ‘thou’, ‘thee’, and ‘land’. ‘Moses’ was the 11th most frequent co-term, and ‘David’ the 14th most frequent. Another participant investigated ‘saith’, a verb that they weren’t familiar with. Each participant was able to navigate using the query interface successfully.

DISCUSSION
One task the system was very good at was finding areas of co-occurrence that appeared deep down in the list of co-terms. This showed that the system was good at finding interesting data points that may not even have appeared on a list of the most frequent co-terms.

However, the application was hindered by the property of real data: often, many of the first few terms are extremely high in frequency, and the rest follow a ‘long tail’ distribution. Since the data scale is normalized around the query term frequency and not the intersection term frequency, areas farther down the list quickly become very small.

A positive aspect of this design is that for the most part, the data was represented faithfully and users could estimate the absolute ratio of intersections to non-intersections easily. This was positive feedback, given that none of the areas were based on a flat axis.

Participants were interested in extending the technique beyond pairwise interactions: for example, exploring three terms as they co-occur. This is not directly possible currently without adding another dimension to the charts.

Finally, many of the participants and commenters at the poster session expressed that the charts were both intuitive and pleasing to look at. The metaphor of overlapping, transparent areas for mapping co-occurring events had a lot of support.

FUTURE WORK
One area of future work is in developing heuristics to complement the visualizations. Even in our interactive system, the order in which we displayed the results (and the concept of having a “top 30” referred to term co-occurrence frequency counts. As shown by our user study, this does not
have a one-to-one relationship with how “interesting” the co-occurrences are. Thus, a possible way to augment the system is to provide summary statistics related to natural language metrics alongside the small charts. In general, the ordering of visualizations in small multiples when only a “top” slice of the data is to be returned is a rich area of future research.

Another area of future work for these time-series charts is how to best normalize the scale of each chart. The technique of banking to 45 degrees helps to maximize the discriminability of trends: when three time series are being plotted at the same time, what is the best way to show trends?

Finally, using the visualization technique in real time may also present other challenges. For example, a marketing analyst may want to understand co-occurrences of hash tags on Twitter, and filter out what they deem irrelevant. Thus, there should be an interface in which co-terms can be eliminated interactively, and the rest of the terms can be scaled.

One interaction technique that was not explored is that of a magnifying glass. By providing a cursor to zoom in on small areas, intersection areas can be made more visible while still preserving the absolute scales of the entire visualization.

REFERENCES


