Intersection testing

The pseudocode in Example 3-4 breezily suggests that you move a word while it intersects other words, but it does not suggest how you’d go about determining such a thing. Testing spline-based shapes for intersection is expensive, and a naïve approach to choosing pairs for comparison is completely unaffordable. Here are the techniques that Wordle currently uses to make things fast enough:

Hierarchical bounding boxes

The first step is to reduce the cost of testing two words for intersection. A simple method for detecting misses is to compare the bounding boxes of two words, but it’s not uncommon for two such boxes to intersect when the word glyphs do not. Wordle exploits the cheapness of rectangle comparisons by recursively dividing a word’s bounding box into ever-smaller boxes, creating a tree of rectangles whose leaf nodes contain chunks of the word shape (see Figure 3-18). Although it’s expensive to construct these hierarchical bounding boxes, the cost is recovered by an order of magnitude during the layout. To test for collision, you recursively descend into mutually intersecting boxes, terminating either when two leaf nodes intersect (a hit) or when all possible intersecting branches are excluded (a miss). By taking care with the minimum size of leaf rectangles and by “swelling” the leaf boxes a bit, the layout gets a pleasing distance between words “for free.”

Figure 3-18. Hierarchical bounding boxes

Broadphase collision detection

In choosing pairs of words to test for intersection, the simplest approach is to test the current candidate word against all of the already-placed words. This approach results in a hit test count around the order of $N^2$, which is far too slow once you get up to 100 words or so. Therefore, Wordle does some extra work to avoid as much collision testing as possible.

Caching

One very simple improvement stems from the observation that if a word A intersects some other word B, it’s very likely that A will still intersect B if A is moved slightly. Therefore, Wordle caches a candidate word’s most recently intersected word and tests it first.
Spatial indexing
To further reduce the number of hit tests, Wordle borrows from computational geometry the *region quadtree*, which recursively divides a two-dimensional space (in this case, the Wordle playing field) into four rectangular regions. Here, a quadtree serves as a *spatial index* to efficiently cull shapes from the list of words to be compared to some candidate shape. Once a word is placed on the playing field, Wordle searches for the smallest quadtree node that entirely contains the word, and adds the word to that node. Then, when placing the next word, many already-placed words can be *culled* from collision testing by querying the quadtree.

There’s an entire research field around efficient collision detection, much of which is very well summarized in Christer Ericson’s (2005) book *Real-Time Collision Detection*. I recommend that book to anyone who wants to play with randomized graphics algorithms like Wordle’s; my own quadtree implementation is based on my understanding of its discussion.

**Is Wordle Good Information Visualization?**

If you consider Wordle strictly as an information visualization tool, certain aspects of its design could be criticized for their potential to mislead or distract its users. Here are some of my own Wordle caveats.

**Word Sizing Is Naïve**

Wordle does not take into account the length of a word, or the glyphs with which it’s drawn, when calculating its font size. The result is that, given two words used the same number of times, the word with more letters will take up more space on the screen, which *may* lead to the impression of the longer word being more frequent.

On the other hand, I don’t know of any studies on how relative word size corresponds to perceived relative weight. What’s more, the commonly used trick of scaling by the square root of the word’s weight (to compensate for the fact that words have *area*, and not mere length) simply makes a Wordle look boring.

**Color Is Meaningless**

In a medium—your computer screen—that provides precious few dimensions, Wordle is shockingly free with its use of color. Color means absolutely nothing in Wordle; it is used merely to provide contrast between word boundaries and for aesthetic appeal.

Color could be used to code various dimensions, such as clustering (indicating which words tend to be used near each other) or statistical significance (as in the inaugural address word clouds—see Figure 3-19). Wordle could also use color to let two or more different texts be represented in the same space.
“Government” was used a lot in this speech, but not much more than in the other speeches; “pleasing” was used only a couple of times but is an unusual word in the corpus; “people” was used a lot and is unusually frequent.

It should also be mentioned that Wordle makes no provision for colorblind users, although one can always create a custom palette via the applet’s Color menu.

**Fonts Are Fanciful**

Many of Wordle’s fonts strongly favor aesthetics and expressiveness over legibility. This has to do, partly, with the design of the Wordle website—the gallery pages would be monotonous without fairly broad letter-form diversity. Most importantly, a font has to look good *in a Wordle*, which may mean that it wouldn’t necessarily work well for body text.

For applications where legibility is paramount, Wordle provides Ray Larabie’s Expressway font, which is modeled on the U.S. Department of Transportation’s Standard Alphabets.

**Word Count Is Not Specific Enough**

I have seen Wordle used to summarize each book of the New Testament, leading to one page after another of “Lord,” which tells you nothing about how the chapters are *distinct* from one another. Merely counting words does not permit meaningful comparisons of like texts. Consider, for example, a blog post. It might be most revealing to emphasize how the post differs from other blog posts by the same author, or to show how it differs from posts on the same topic by other bloggers, or even to show how it differs from the language of newspaper reporting.
There are plenty of statistical measures that one may apply to a “specimen” text versus some “normative” body of text to reveal the specific character of the specimen, with proper attention paid to whether some word use is statistically significant. Given a more nuanced idea of word weight, beyond mere frequency, one could then apply the Wordle layout algorithm to display the results.

I explored this idea in an analysis of every presidential inaugural address,* in which each speech was compared to the 5 speeches nearest to it in time, the 10 nearest speeches, and all other inaugural addresses. Such an analysis has the advantage of revealing the unexpected absence of certain words. For example, Figure 3-20 is a visualization of Harry Truman’s 1948 inaugural address. On the left is a Wordle-like representation of the words he used, and on the right are the words that his contemporaries used more than he did. This visualization reveals Truman’s emphasis on foreign policy.

* See http://researchweb.watson.ibm.com/visual/inaugurals/.

Figure 3-20. Harry Truman’s 1948 inaugural address: the words in red were conspicuously absent from Harry Truman’s speech, relative to those of his contemporaries.
How Wordle Is Actually Used

Wordle was not designed for visualization experts, text analysis experts, or even experienced computer users. I tried to make Wordle as appliance-like as possible.

As of this writing, people have created and saved over 1,400,000 word clouds in the Wordle gallery. They have been used to summarize and decorate business presentations and PhD theses, to illustrate news articles and television news broadcasts, and to distill and abstract personal and painful memories for victims of abuse. Wordle has also found an enthusiastic community in teachers of all stripes, who use Wordles to present spelling lists, to summarize topics, and to engage preliterate youngsters in the enjoyment of text.

As the survey results in Table 3-1 (Viégas, Wattenberg, and Feinberg, 2009) illustrate, when people use Wordle they feel *creative*, as though they’re making something.

**Table 3-1. How people feel when they make a Wordle**

<table>
<thead>
<tr>
<th></th>
<th>Agree %</th>
<th>Neutral %</th>
<th>Disagree %</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt creative</td>
<td>88</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>I felt an emotional reaction</td>
<td>66</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>I learned something new about the text</td>
<td>63</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>It confirmed my understanding of the text</td>
<td>57</td>
<td>33</td>
<td>10</td>
</tr>
<tr>
<td>It jogged my memory</td>
<td>50</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>The Wordle confused me</td>
<td>5</td>
<td>9</td>
<td>86</td>
</tr>
</tbody>
</table>

So, by one traditional academic measure of a visualization’s efficacy—“I learned something new about the text”—Wordle can at least be considered moderately successful. But where Wordle shines is in the creation of *communicative artifacts*. People who use Wordle feel as though they have created something, that the created thing succeeds in representing something meaningful, and that it accurately reflects or intensifies the source text. This sense of meaningfulness seems to be mostly intuitive, in that many people do not realize that word size is related to word frequency (guessing, instead, that the size indicates “emotional importance” or even “word meaning”).

The special qualities of Wordle are due to the special qualities of text. Simply putting a single word on the screen, in some font that either complements or contrasts with the sense of the word, immediately resonates with the viewer (indeed, there have been many thousands of single-word Wordles saved to the public gallery). When you juxtapose two or more words, you begin to exploit the tendency of a literate person to make sense of words in sequence. Wordle’s serendipitous word combinations create delight, surprise, and perhaps some of the same sense of recognition and insight that poetry evokes intentionally.
Using Wordle for Traditional Infovis

Notwithstanding Wordle’s special emotional and communicative properties, the analytic uses of information visualization are certainly available to the expert user. To serve those who want to use Wordle as a visualization for their own weighted text, where the weights are not necessarily based on word frequency, the Wordle website provides an “advanced” interface, where one can enter tabular data containing arbitrarily weighted words or phrases, with (optional) colors.

Still more advanced use is possible through the “Word Cloud Generator” console application, available through IBM’s alphaWorks website.*

The ManyEyes collaborative data visualization site also provides Wordle as a text-visualization option beside its innovative Phrase Net and Word Tree visualizations (and a more traditional tag cloud).†

Conclusion

People often want to preserve and share the Wordles they make; they use Wordles to communicate. A beautiful visualization gives pleasure as it reveals something essential.

Acknowledgments

I would like to thank Martin Wattenberg and Irene Greif at IBM CUE for making possible my participation in this book. I am very grateful to Ben Fry, Katherine McVety, Fernanda Viégas, and Martin Wattenberg, who each read this chapter with great care and suggested many improvements. Please see http://www.wordle.net/credits for information about the many people who have helped me create and improve Wordle.

References


