Word Cloud
What is Word Cloud?

- Word clouds provide a concise yet fun way to summarize the content of websites or text documents.

- In a typical word cloud, tags from a website (or words from a document) are packed into a rectangular region in which font size indicates tag popularity (or word frequency) and font color indicates other useful information.

- http://www.wordle.net
- http://www.tagclouds.com
- http://tagcrowd.com/
Visual Appearance

- The tags can appear in alphabetical order, in a random order, they can be sorted by weight, and so on.

- Most popular is a rectangular tag arrangement with alphabetical sorting in a sequential line-by-line layout.

- Some prefer to cluster the tags semantically so that similar tags will appear near each other.

- Heuristics can be used to reduce the size of the tag cloud whether or not the purpose is to cluster the tags
Related Word


Context-Preserving, Dynamic Word Cloud Visualization

Weiwei Cui and Yingcai Wu ■ Hong Kong University of Science and Technology
Shixia Liu, Furu Wei, and Michelle X. Zhou ■ IBM China Research Lab
Huamin Qu ■ Hong Kong University of Science and Technology
Motivation

- Existing layout techniques are inadequate at balancing semantically meaningful clusters with visually appealing layouts

- Multidimensional scaling to project words onto 2D space results in wasted empty space

- Arranging clusters in lines to more efficiently use screen space sacrifices relationships between words, and aesthetically appealing word clouds that pack tags more tightly lose meaningful word positions
Proposed Method

- The proposed method uses context-preserving, dynamic word clouds to illustrate content evolution.

- It generates a sequence of word clouds in which related words are grouped together.

- This sequence is then coupled with a trend chart that summarizes content changes so that users can better explore large collections of documents.
Figure 1. An overview of our system for generating dynamic word clouds. (a–e) Our method creates five word clouds for five selected time points with high significance values. (f) The top center box presents a significance trend chart whose significance curve is extracted from a collection of documents with different time stamps.
Figure 2. The pipeline for creating a semantic and stable word cloud layout. (a) Extracting an initial set of words from documents with different time stamps. (b) Placing extracted words on a 2D plane using multidimensional scaling. (c) Filtering out unrelated words for a specified time point. (d) Triangulating the remaining words. (e) Optimizing the layout by a force-directed algorithm.
Word Extraction

- Consider $n$ documents $T = \{T_1, T_2, \ldots, T_n\}$ with different time stamps
  - Remove unimportant words

- The system builds a histogram $Hist_i$ to indicate the frequency of all unique words in $T_i$

- Words remaining in $Hist_i$ with frequencies higher than a user-specified threshold are selected as candidate word set $W_i$, which creates the word cloud for $T_i$. Finally, we obtain a set of extracted words $W = \{W_1, W_2, \ldots, W_n\}$ for $T$
Initial Word Placement

- With the extracted word set $W$, we place all important words ($\cup W$) on a 2D plane to create an initial word layout that semantically groups words.

- Three semantic-coherence criteria to generate different layout styles:
  - Importance criterion
  - Co-occurrence criterion
  - Similarity criterion
The importance criterion creates layouts that cluster words on the basis of importance values at different time points (font sizes represent importance values).

It groups words with similar variations in font sizes over time.

The corresponding feature vector is \( V_i = \{v_1, v_2, \ldots, v_n\} \), where \( n \) is the number of time points in the documents and \( v_j \) is the importance value (the font size) of word \( w_{dp} \) at time point \( j \).
Co-occurrence criterion

- The *co-occurrence criterion* ensures that words with similar appearances or disappearances over time are clustered together.

- The corresponding feature vector is \( V_a = \{v_1, v_2, \ldots, v_n\} \), with \( v_j = 1 \) if \( wd_p \) is visible at \( j \), and \( v_j = 0 \) otherwise.
Similarity criterion

- The *similarity criterion* creates layouts in which semantically similar words are clustered.

- Hinrich Schütze’s method suggests that semantically similar words share similar neighboring words.

- The feature vector is $V_s = \{v_1, v_2, \ldots, v_m\}$, where $m$ is the number of words in $\cup W$. The element $v_q$ represents the number of times $wd_q \in \cup W$ occurs close to $wd_p$ (in a sentence or larger context) in the documents.
Initial Word Placement

- The similarity between vectors $V_p$ and $V_q$ by the cosine measure:

$$\cos(\theta) = \frac{V_i \cdot V_j}{\|V_i\| \cdot \|V_j\|}.$$ 

- Dissimilarity matrix $D$ is created, where element $\alpha_{p,q}$ represents the similarity ($\cos(q)$ in the previous equation) between words $p$ and $q$.

- **Multidimensional scaling** is employed to reduce each high-dimensional vector to a 2D point.
Delaunay Triangulation

- **Delaunay triangulation** on word positions is performed to reduce wasted space.

- In the graph $G = (V, E)$, word positions are arranged on the 2D plane flexibly to reduce empty space while keeping the semantic relations between the words.
Force-Directed Model

- Follows three design principles:
  - Overlapping (top priority): repulsive force prevents a word from being occluded with other words

\[
f_r(a,b) = \begin{cases} 
    k_r \min(\Delta x, \Delta y) & \text{if word } a \text{ overlaps word } b \\
    0 & \text{otherwise},
\end{cases}
\]

- where \( k_r \) is a given weight and \( \Delta x \) and \( \Delta y \) are the overlapping region’s width and height
Figure 3. Two overlapped words exert a repulsive force on the connected edge. As long as these two words are overlapping, the repulsive force will always exist and finally make sure they are separated.
Planar: attractive force

During layout adjustment, if a mesh triangle is flipped (if one vertex in the triangle goes to the other side), the mesh will become nonplanar. In this case, the attractive force between the edge and the vertex takes effect and flips the triangle back

\[ f_a(a,l) = \begin{cases} \frac{k_a \Delta d}{0} & \text{if word } a \text{ is flipped} \\ 0 & \text{otherwise} \end{cases} \]

where \( k_a \) is a given weight and \( \Delta d \) is the distance between word \( a \) and its edge \( e \).
Figure 4. The attractive force ensures layouts are stable and semantically meaningful. (a) The force between edge $e$ and word $a$ is zero if the mesh is planar. (b) The force takes effect if $a$ is flipped to the other side of $e$. 
Force-Directed Model

- Compact (lowest priority): spring force removes empty space and packs words

\[ f_s(a,b) = w_a w_b \Delta l, \]

- \( k_r << k_a << w_{\text{max}}^2 \), where \( w_{\text{max}} \) is the words’ maximum importance value.
Figure 5. Two separated words exert a spring force on the connected edge. The spring force will try to draw the words to each other, and finally make a compact layout.
User Interaction

- **Creating a Storyboard**
  - Users can selectively visualize specific word clouds by sliding a bar at the bottom of the trend chart

- **Anchoring Words**
  - The system can anchor words in the same position across all word clouds

- **Displaying Frequency Changes**
  - A line chart is overlaid on each keyword in the word cloud such that users can easily perceive the frequency change
Case Studies

- Artificial City Data
- AIG News Data
  - 13,828 news articles related to American International Group from 14 Jan to 5 Apr 2009
- CG&A Abstract Data
- Apple News Data
  - 1993 news articles related to Apple from August 1989 to August 2009
Figure 7. The word cloud layout process: the (a) initial layout and (b) its mesh generated by Delaunay triangulation, a (c) sparse word cloud and (d) its mesh captured at a step during layout adjustment, and the (e) final word cloud layout and (f) its mesh. The figures illustrate key adjustments during the process. We can see that, after the adjustments, city names are packed together while their relative positions are kept.
Case Studies

- Artificial City Data
- AIG News Data
  - 13,828 news articles related to American International Group from 14 Jan to 5 Apr 2009
- CG&A Abstract Data
- Apple News Data
  - 1993 news articles related to Apple from August 1989 to August 2009
Figure 8. Word cloud layouts created by (a) our method and (b) Wordle. The clouds on the left are generated for documents at the same time point, as are the clouds on the right. Users must search a larger space in the Wordle layouts.
Figure 9. Word cloud layouts generated by the (a) importance criterion and (b) co-occurrence criterion. As we expected, all big words in (a) are placed in the center, while all words with the same background color are generally grouped together in (b).
Figure 10. How positions change as a result of our system’s word anchoring. (a) The upper right of Figure 8a. (b) The upper right of Figure 8b. (c) How the upper right of Figure 8b changes when “money” and “funds” are anchored in (a). We can see that the positions of these two words are exactly the same in (a) and (c). Therefore, users should have less trouble tracking them in (c).
Case Studies

- Artificial City Data
- AIG News Data
  - 13,828 news articles related to American International Group from 14 Jan to 5 Apr 2009
- CG&A Abstract Data
- Apple News Data
  - 1993 news articles related to Apple from August 1989 to August 2009
Figure 11. Word clouds summarizing five years of IEEE Computer Graphics and Applications article abstracts: (a) 1989, (b) 1994, (c) 1999, (d) 2004, and (e) 2009. In the sequence, some words appear, some words disappear, and some words change in size. By tracking those behaviors, several interesting patterns are observed.
Case Studies

- Artificial City Data

- AIG News Data
  - 13,828 news articles related to American International Group from 14 Jan to 5 Apr 2009

- CG&A Abstract Data

- Apple News Data
  - 1993 news articles related to Apple from August 1989 to August 2009
Figure 1. An overview of our system for generating dynamic word clouds. (a–e) Our method creates five word clouds for five selected time points with high significance values. (f) The top center box presents a significance trend chart whose significance curve is extracted from a collection of documents with different time stamps.
Questions?