Three-level Visualization of Internet Discussion with Extruded Word Clouds

Pavol Fabo
Comenius University
pavol.fabo@fmph.uniba.sk

Matej Novotný
VIS GRAVIS, s.r.o.
novotny@visgravis.sk

Abstract

Visualization of discussions is an intensively explored information visualization area. Many existing approaches explore the discussions from the point of view of individual participants, often omitting interesting information about the discussion itself.

We present a visualization technique that centers around the discussion, its dynamics, intensity and topic changes. The flow of the text is divided into uniform time spans that aggregate the data and where the text analyzer runs. Then, the discussion is visualized using three different levels of details. The topmost shows the global view of the discussion development over time, the middle shows an inside view on discussions' main topics, and the bottom level displays individual word clouds.

The technique is demonstrated on data from an Internet Relay Chat (IRC) but it can be used for visual analysis of any time-dependent text data such as email communication, search terms, keywords and similar.

Keywords—**Text visualization, text analysis, word cloud**

1 Introduction

Text visualization has become an interesting and vidid area of information visualization, mostly due to the development and popularity of the Internet. Electronic media, blogs, emails, instant messaging, social networks and other services generate an enormous volume of text-based data. This data is a valuable source of information about the world it describes or about the various sides participating in the communication.

The analysis of text goes beyond simple text reading. The underlying trends, patterns or features have abstract nature and their discovery and comprehension require approaches on a higher level than words or sentences. Moreover, the text might involve a temporal aspect, such as in a sequence of emails or instant messaging. This allows us to also observe the dynamics within the data.

The proposed technique allows for visual analysis of online discussions in three levels of detail, each addressing a different aspect of the data within the temporal context:

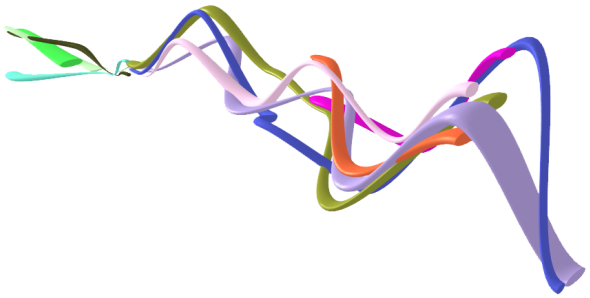

Figure 1: An example of the introduced technique visualizing the most frequent words in an Internet discussion.

the development of overall volume of the discussion, the emergence, evolution and possible decline of individual keywords, and relative occurrence of keywords. While our technique is primarily oriented on discussions, it can be used on any time-referenced text data. This extends the area of our interest from online discussions to email communication, news reports, blogs or popular search terms.

2 Related work

Text visualization has been a part of information visualization for quite some time, starting with the early works such as Perspective Wall [8] or Seesoft [2]. The origins of the now popular word clouds (or tag clouds) can be traced back to Milgram's Psychological maps of Paris [5] which used geographic position as the reference grid for word placing in 1976.

However, the word clouds were not very popular until the development of the Internet and online societies. Web 2.0 sites such as Flickr or del.icio.us required a way to form structure out of user-contributed content and the word clouds became a regular asset among online visualizations. Following this success, various modifications and inspired visualizations started to emerge, often incorporating some temporal aspect. Dubinko et al. [1] created an animated and interactive visualization of tags with associated images. Nguyen et al. [7] and Lee et al. [4] add temporal

information to the word clouds by changing the visual appearance of the words or by embedding new artifacts into the display of an individual word.

Visualization of discussions, on the other hand, has been evolving in a different fashion. Centered around the people partaking in the communication, the observed relations are often social patterns [11], individual habits [6] or a combination of both [10].

We approach the Internet discussions from the textual point of view rather than the social point of view. The word clouds are extruded into the 3^{rd} dimension to record their temporal evolution in a manner inspired by ThemeRiver [3]. Thus, we create a three-level visualization that captures over time the discussion dynamics, changes in discussion topics and relative word frequency.

3 Basic idea

The presented visualization technique is based on the extrusion of two-dimensional word clouds into the third dimension and creating a tubular surface. Thus, the shape of the surface represents the temporal characteristics of the discussion – the changes in the popularity of a certain term, in the volume of the discussion or in the number of users participating in the discussion.

We start by dividing the time-referenced text (e.g. an online discussion) into successive time intervals. For each interval, a word cloud is generated based on the most frequent words and their relative population. These word clouds are then placed perpendicular on the z-axis of the three-dimensional space. We create a bounding rectangle for every word in every cloud and use the vertices of the rectangles to create a Bezier surface that connects the instance of a certain word with its instances in the neighbor word clouds.

The same principle is then repeated to create bounding rectangles for each word cloud as a whole. This creates a discussion surface, wrapping around the word surfaces and creating a visual representation of the overall discussion volume. We encode additional information – the number of discussion participants within a certain time interval – as the y-coordinate of the word cloud representing the time interval.

The result is a three-level visualization with different levels of abstraction and different analysis goals related to each level. These levels are further explained in the Section 4. Section 5 explains in detail the generation of the surfaces and Section 6 demonstrates our technique on a discussion in the Internet Relay Chat (IRC) environment and on popular search terms gathered by Google Trends.

4 Three-levels of discussion data

A discussion visualization is considered a special subset of text visualization where the data is augmented with time stamps. Because of such data augmentation, a text visualization strategy needs to be revised in order to include the new information as well. We divide the temporal characteristics of the discussion into three levels of information.

This allows us to provide overview first, zoom and filter second, and details on demand as suggested by Ben Shneiderman's visual information seeking mantra [9]. Another important advantage of this approach is that we get simultaneous visual access to both the temporal and the textual information.

4.1 Single time spans

The first level includes the individual words organized into word clouds. We compute the word frequencies for each individual time span and select only the most frequent words. In order to preserve the information value of the text analysis, the words are first filtered either by a blacklist (to filter out meaningless words) or by a whitelist (to only include a selected set of words). Word clouds for consecutive time spans are then placed equidistant on the z axis (which represents time). The height placement of the word cloud is derived from the number of the users contributing to the discusion in the respective time span.

4.2 Discussion topics

The second level of the visualization connects the words in adjacent word clouds. The connection is done using Bezier surfaces defined by the word's bounding box within each word cloud. These surfaces form a three-dimensional representation of the evolution of a word's relative frequency over time

It happens that a word disappears from the selection of the most frequent words and does not appear in the subsequent word cloud. Such "death" of a word is represented by a round "cap" on the end of the surface. Alternatively, a word can become popular at a later stage and does not appear in a word cloud before a certain time. This event is not depicted by a "cap". Instead, the surface starts in its full elliptic extent in order to make place for the type should the user decide to combine word cloud with a surface display (Figure 5.)

4.3 Discussion volume

At the third level of visualization, we create a bounding box around each word cloud and build a single bounding surface over the whole discussion. The Bezier surface control points are derived from the word cloud bounding boxes, thus the width of the surface depends on the size of the word cloud. Varying surface extent depicts the overall discussion activity in a particular time span. Since the word clouds are offset on the y-axis based on the number of participants, the discussion volume surface displaced is vertically shaped according to the number of participants.

This helps to identify large-scale relations and patterns such as "hot hours" where large number of people create a

vivid discussion, ”cold hours” where large number of people is present but the discussion is sparser or temporal patterns such as periodicity or abrupt changes.

5 Extruded word clouds

We developed our own word placing algorithm to generate word clouds in the 2d plane: The most frequent word is placed in the center of the word cloud plane. The bounding box of the word divides the rest of the plane into eight neighbor areas where the next 8 frequent words are placed in a counter-clockwise order, starting from the upper left area. After filling the 8 neighbor cells, each of the neighbor cells creates their own neighbor cells and new words are placed in the same way unless the area is already occupied. This process is iterated until the word list is depleted or there is no more room. However, our extrusion technique will work with word clouds generated by any other algorithm.

The word surface is generated as a sequence of Bezier surfaces between instances of the same word in adjacent word clouds. We use two cubic Bezier surfaces for each time span. One for the top and one for the bottom part of the surface. Cubic Bezier surfaces allow us to blend smoothly between the top and bottom surfaces as well as between surfaces in adjacent time spans.

The generation of a word surface is illustrated in Figure 2. We start with two instances of the same word in adjacent parallel word planes. Their mutual distance is marked as Δz. First, we divide the word bounding boxes to top halves (solid red/blue lines) and bottom halves (dotted lines) which will be processed separately. The corner points of the top parts of the red box and the blue box form the first and the fourth set of control vertices of the Bezier surface. The second and third set of control vertices are generated by perpendicular offsetting the first and the fourth set off their word planes by $\Delta z/3$ in positive or negative z-direction respectively.

The Bezier surface computed on these 16 control vertices (Fig. 2, bottom) has its tangent vector on the borders perpendicular to either the word planes or the horizontal plane. This ensures C^1 continuity between the top/bottom surfaces as well as between subsequent surfaces.

A special case happens when a word does not appear in the subsequent word plane. In that case, we create a Bezier surface ”cap” that gradually shrinks towards a single point generated by offseting the center point of the word's bounding box by $\Delta z/4$ in the direction of the subsequent word plane.

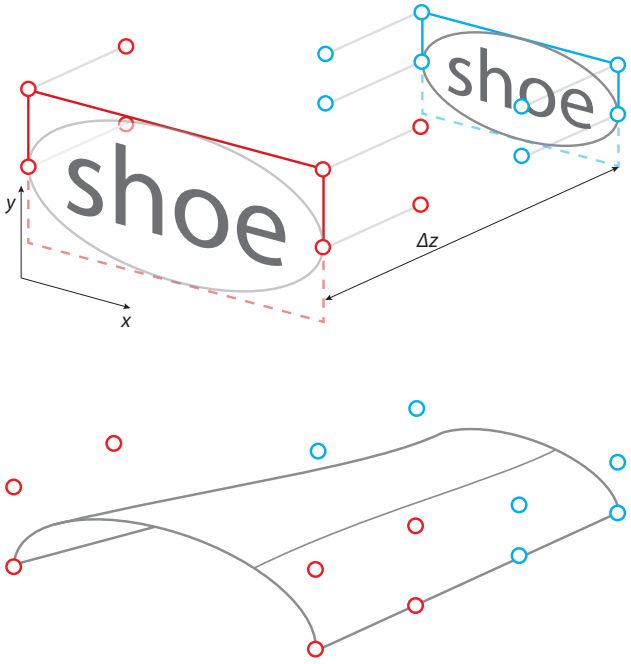

Figure 2: Generating the upper half of the word surface for a single time span. The two adjacent instances of the same word (top). The generated control vertices and resulting surface (bottom).

5.1 Visual enhancements

We employ several visual effects to enhance the comprehensibility of our visualization. A shadow is created by projecting the whole discussion volume surface onto the ground plane. This way, the overall discussion volume is always visible. It provides context when the focus is on individual word surfaces or word clouds. Furthermore, the shadow serves as a time axis reference where the individual time spans are marked and it improves the spatial perception of the 3d surface too. We also added specular light component to surface shading in order to improve the perception of the surface shape.

We use color coding for individual word surfaces. The colors are currently chosen randomly upon the word clouds creation. But we keep in mind the contrast against the background and only generate saturated colors.

While the implementation itself is interactive in terms of camera position or data filtering, we also provide the user with the option to run an animation that shows how the surface evolves from the earliest towards the latest time span. This is done by translating a perpendicular clipping plane along the z-axis.

6 Demonstration and results

We demonstrate our technique on a three days long Internet relay chat (IRC) discussion. The input data was divided into time spans of 6 hours each. The text analysis used a blacklist disregarding meaningless words such as ”and”, ”the” etc.

Figure 3: Overall view of the discussion development in time. The time axis augmented with surface shadow shows the discussion volume development.

Figure 3 shows the overall discussion volume, i.e. the third level of the visualization – the overview. The shadow beneath the surface supports the perception of the volume and makes it easy to detect the peaks of the discussion activity: first day’s noon and third day’s 6 a.m. Judging by the relative height of these two segments, we can also say, that there were less people contributing to the third day’s discussion than the first day.

The second level of visualization is shown in Figure 4. The individual words in time span’s word clouds connect to the corresponding words in the subsequent word cloud. From this point of view, we are able to see the development of the words in time, emergence of new words, extinction of weak words (the visual mapping of the end words in form of a cap) or the existence of words that survive even during weak night times. The shadow in the ground plane serves in this level as a context to the discussion topics.

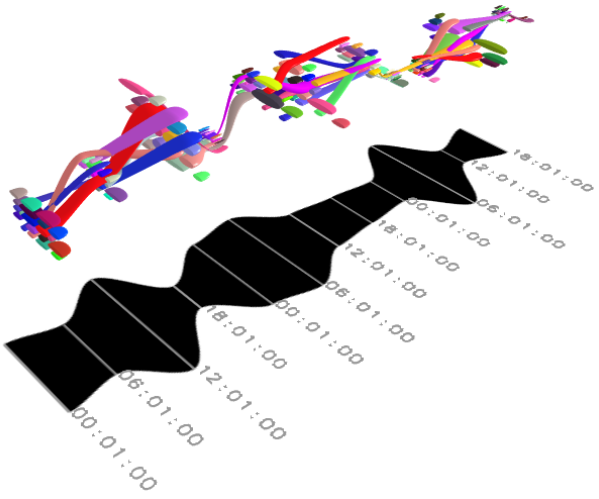

Figure 4: More detailed view showing individual words connected between the adjacent word clouds.

The base level – word cloud for a single time span – is shown in Figure 5. The words are enclosed within an approximation of an ellipse created with Bezier surfaces. If a word repeats itself in a subsequent time span, the word’s elliptical enclosing extrudes and connects these corresponding word instances.

Figure 5: A section through the surface shows the word cloud for the particular time span. Notice new words emerging in the background.

An interesting analysis option is the ”word life” filter. We can specify the minimum and maximum limit for the number of subsequent occurences of words in the word cloud sequence. This way we can focus on short-, medium- or long-living words. Figure 6 shows a word life filter set to 4 to 6 consequent occurences.

Figure 6 also shows how the number of users contributing to the discussion affects the height displacement of the word clouds and the generated surfaces. From this point of view, the most people were contributing to the discussion around the midnight of the second day.

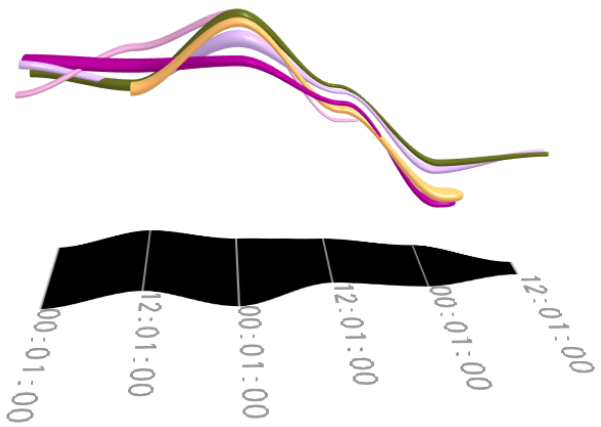

Figure 6: Words filtered by the number of consecutive occurences. Also, the varying height depends on the number of users in the particular time span.

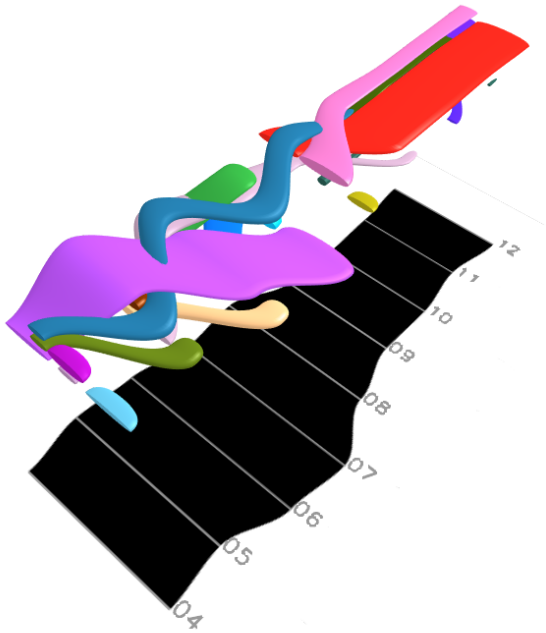

Figure 7: Visualization of Google Trends from 2004 to 2012. Notice the red word starting in 2009, which corresponds to ”facebook”.

The presented visualization technique can also be used on any other time-referenced text data. In Figure 7, we tested our technique on Google Trends data. The visualization presents the top Google-searched terms from France between the years 2004 and 2012. The turquoise snake-like surface represents the term ”jeux” (games) and it dies between 2008 and 2009 which is (probably not by coincidence) the period when the large red surface appears, representing the term ”facebook”.

7 Conclusions and future work

We present a novel approach to visualization of Internet discussions based on word clouds and inspired by ThemeRiver. We combine three levels of visualization, each corresponding to a different layer of information about the discussion. The proposed technique is further improved by focus+context combination of individual words and overall discussion. We also combine the discussion density and discussion population into a single 3d shape giving the user the option to switch between these two variables by simply looking at the shape from a different angle.

The presented concepts are not limited to Internet discussions but can as well be used on any other text data as long as it refers to a temporal domain. An even more abstract extrapolation of the technique could involve any time-dependant data with weighted elements such as clusters.

Our next research will include investigation of color semantics that can be used to give an even better overview of the content and the reduction of surface overlapping by employing a word cloud algorithm that involves temporal coherence of word placement.

Acknowledgements

This research was partially supported by VEGA grant No. 1/1106/11.

References

[1] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. In *Proceedings of the 15th international conference on World Wide Web*, WWW ’06, pages 193–202, New York, NY, USA, 2006. ACM.

[2] Stephen G. Eick, Joseph L. Steffen, and Eric E. Sumner, Jr. Seesoft-a tool for visualizing line oriented software statistics. *IEEE Trans. Softw. Eng.*, 18(11):957–968, November 1992.

[3] Susan Havre, Paul Whitney, and Lucy Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8:9–20, 2002.

[4] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelash Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, November 2010.

[5] Stanley Milgram. Psychological maps of paris. In Harold M. Proshansky, William H. Ittelson, and Leanne G. Rivlin, editors, *Environmental Psychology: People and Their Physical Settings, 2nd ed.*, pages 104–124. Holt, Rinehart and Winston, 1976.

[6] Petra Neumann, Annie Tat, Torre Zuk, and M. Sheelagh T. Carpendale. Keystrokes: Personalizing typed text with visualization. In *EuroVis*, pages 43–50, 2007.

[7] Dinh Quyen Nguyen, Christian Tominski, Heidrun Schumann, and Tuan Anh Ta. Visualizing tags with spatiotemporal references. In *Proceedings of the 2011 15th International Conference on Information Visualisation*, IV '11, pages 32–39, Washington, DC, USA, 2011. IEEE Computer Society.

[8] George Robertson, Jock D. Mackinlay, and Stuart Card. The perspective wall: Detail and context smoothly integrated. In *Proceedings of CHI '91 Conference*, pages 137–139, 1991.

[9] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[10] Dana Sean Spiegel. Coterie : a visualization of the conversational dynamics within irc. Master's thesis, Massachusetts Institute of Technology, 2001.

[11] Annie Tat and Sheelagh Carpendale. Crystalchat: Visualizing personal chat history. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), IEEE Computer Society*, pages 58–67, 2006.