Kurt Bognar

Date Due - 4/19/17

Capstone 1

Research Paper

D3 Powered Visualization

Existing methods of visualizing information are sophisticated; however this has yielded

many different paid visualization tools. D3 is an open source visualization that allows many data

operations and helps to make data interactive and dynamic. Using data to tell a story would

require that temporal data be used. Since databases containing scraped article data from various

websites have been built for Capstone earlier in the semester, a temporal visualization would

enhance the usability of this project and exploit all elements of the data. The problem addressed

in this paper is twofold: what is the best solution to use to visualize data and what are good

design principles/elements that should be included in that visualization.

D3

Data Driven Documents (D3) is an open source JavaScript library that promotes easy

visualization. [3] Even though Tableau makes visualizing data more simply, since it is not free it

does not promote the wide adoption of this projects output. Infographics also could be a good

way to tell the story attempted by this project, but since they are not typically dynamic they again

do not allow for wide spread application and adoption. The ability to dynamically show what an

article contains and tie that statistically to a word cloud is ideal. Adding interaction on top of that

through D3 allows the user to increase their relationship to the output of this product. For

example, a user may want to see how CNN and FOX have differing relations towards the topic of "alt-right" during the months leading up to the US's 2016 presidential election while another user wants to see a NBC's views on "healthcare" for all of 2016. While these users activities are very different, both can be done in D3. D3 achieves this by binding data elements to scalable vector graphics (SVG) before rendering these objects. This allows dimensionality to by dynamically modified before display. Figure 0 shows how a user could view data in two different ways at the click of a box using D3 which prevents additional page loads and wait time while increasing visibility into the data. Having this freedom both in price and customization of views proves why D3 is a good choice for visualizing news articles.

Figure 0:
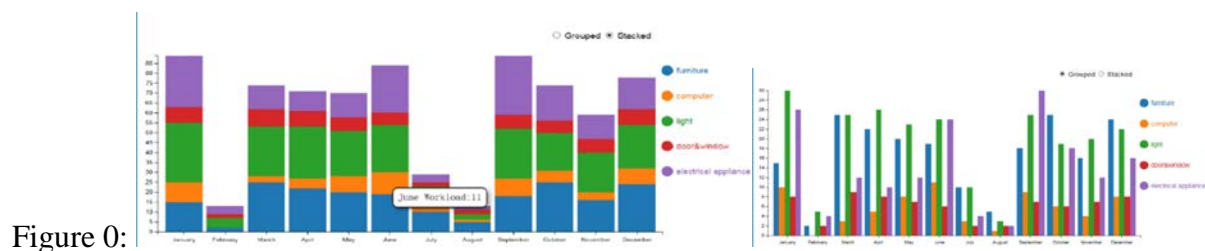
Visualization

User interface design has six main principles: structure, simplicity, visibility, feedback, tolerance and reuse. [6] By applying these principles to interactive data visualizations allows for more intuitive use of software. The easiest way to show these principles is through negation and so the following figures are used to show why these principles are necessary. The structure should be intentionally designed and nothing should be random. Figure 1 shows random placement of word, negating any ability to show relationships between words. The interface should be simple and be task oriented. Figure 2 violates this principle by incorporating too many dimensions into each unit it displays. Visibility means that a user should get the things they need on-demand and no sooner. No materials should be redundant and overlapping is discouraged.

Figure 1 violates this but could've avoided this design flaw by rotating objects or using a packing algorithm. Since infographics do not incorporate interaction, they do not incorporate a feedback loop unlike visualizations. The visualization needs to have sufficient error checking so that it is stable and accounts for times when the article extraction tool is down or if a website went down. Since the visualization will dynamically pull data from a database, the reuse principle will also be fulfilled.
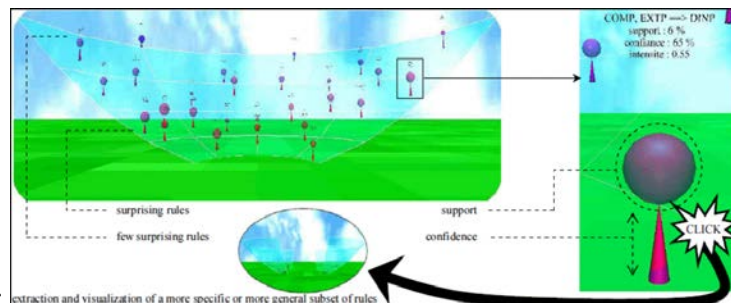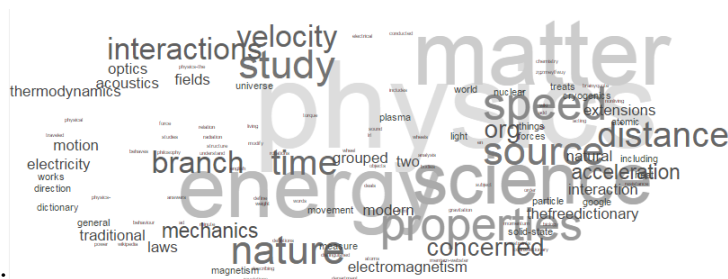
Figure 1:

Figure 2 [1]: extraction and visualization of a more specific or more general subset of rules

A word cloud is made up of six dimensions: width, length, word width, length, color and orientation. The words themselves have statistics about them that come from the various natural language processing that will be applied in this particular application. These NLP stats will drive the value and normalization of the dimensions of the word cloud. Width and length of the page will be kept static as to not imply anything to the user that comes from the data, however changing the words size, color and orientation will allow the user to get the most out of the data present on the screen while not over-whelming them. The CDF or a cumulative statistic of a

word could be used to show changes between specific time periods in regards to one author or site. Doing this with a timeline slider would be intuitive to current mental models a user has and would allow users to mark milestones in the timeline that are relevant to them or to show how an opinion of a site towards a specific topic changes. For example, the alt-right, the libertarian movement, and the socialist party have recently grown roots in the US. If certain news sites are dedicated towards the promotion of these ideals, the settling of these beliefs systems platforms would manifest in some way in the language that those sites use for specific time periods.

Tying Data to Visuals

Data streams from public social media sites have been used to classify hotels. [8] Applying the same concept to news data would be a way to evaluate the quality of journalism on a site. Yelp reviews from over a thousand users was used to determine the quality of 5-1 star hotels. Since the product of a news site is the content that they publish, being able classify <u>and</u> visualize why your classifier makes its decisions would allow consumers better insight into the bias of their news. Objectivity of the classifier is important which is why it will be purely statistic. Since NLP can be used to generate statistics about words and sentences, it would be informative to use this initial processing step to drive visualization. Figure 3: shows the results of passing the contents of the paper this section was based on through a word map generator. It shows a lot of information about its topic; combining that with more sophisticated NLP libraries would allow researchers to skim articles in less time in order to verify the topic and sentiment of a paper or article. If a reader thinks an article is biased and they could pass the URL to an API that would classify and visualize the language used in an informative way, they would have a better grasp of the high level contents of said article.

Figure 3:

Interactivity

Interactivity allows a user to discover information that is particularly relevant to a specific user. [5] The overuse of interaction can make a visualization into a game and not be driven to promote discovery. It can also become an expensive and time consuming way for developers to show their skills as in figure 4. There currently is not enough adoption of virtual and augmented reality to display data in a way that it can be consumed. Also, the human brain typically does not respond well to 3D input that is not directly tied to length width and depth and so does not fit the type of data being used in this application. It is much more important for a user to be able to see changes over time in the case of news because populations change the way they think about certain topics and to be able to see that is more valuable than distracting them with extraneous dimensions.
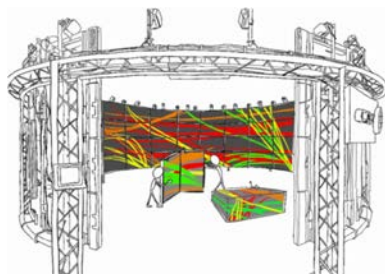

Figure 4 [9]:

Conclusion

      Providing an objective way to show users news data must be available to increase transparency into journalism. Doing this in a dynamic way increases the number of users that could potentially benefit and doing it in a visual fashion makes instruction of analysis minimal. Keeping interaction to a minimum decreases distraction and lets the data speak for itself. Taking all things into account, D3's word cloud is the best method to show a user the reduced dimensionality of articles in browser, a medium that is already widely adopted.

Works Cited

[1] Blanchard, J., Guillet, F., & Briand, H. (2003). A user-driven and quality-oriented visualization for mining association rules. *Third IEEE International Conference on Data Mining*. https://doi.org/10.1109/ICDM.2003.1250960

[2] Byrne, L., Angus, D., & Wiles, J. (2016). Acquired Codes of Meaning in Data Visualization and Infographics: Beyond Perceptual Primitives. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2015.2467321

[3] Chen, L., & Zhou, H. (2016). Research and application of dynamic and interactive data visualization based on D3. *2016 International Conference on Audio, Language and Image Processing (ICALIP)*. https://doi.org/10.1109/ICALIP.2016.7846608

[4] Fu, T. c., Sze, D. C. M., Leung, P. K. C., Hung, K. y., & Chung, F. l. (2007). Analysis and Visualization of Time Series Data from Consumer-Generated Media and News Archives. *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*. https://doi.org/10.1109/WI-IATW.2007.104

[5] Jacucci, G. (2016). Resourceful interaction in information discovery. *2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)*. https://doi.org/10.1109/ICDEW.2016.7495639

[6] Naidoo, J., & Campbell, K. (2016). Extended abstract: Best practices for data visualization. *2016 IEEE International Professional Communication Conference (IPCC)*. https://doi.org/10.1109/IPCC.2016.7740509

[7] Setlur, V., & Stone, M. C. (2016). A Linguistic Approach to Categorical Color Assignment for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2015.2467471

[8] Suzuki, T., Gemba, K., & Aoyama, A. (2013). Hotel classification visualization using natural language processing of user reviews. *2013 IEEE International Conference on Industrial Engineering and Engineering Management*. https://doi.org/10.1109/IEEM.2013.6962540

[9] Thomas, B. H., Marner, M., Smith, R. T., Elsayed, N. A. M., Itzstein, S. Von, Klein, K., … Suthers, T. (2014). Spatial augmented reality — A tool for 3D data visualization. *2014 IEEE VIS International Workshop on 3DVis (3DVis)*. https://doi.org/10.1109/3DVis.2014.7160099