Deep Models and Artificial Intelligence for Military Applications:
AAAI Technical Report FS-17-03

Using D3 to Visualize Lexical Link Analysis (LLA) and ADS-B Data

Quinn Halpin, Ying Zhao, Anthony Kendall
Naval Postgraduate School, Monterey, CA
qmh4@cornell.edu, yzhao@nps.edu, wakendal@nps.edu

Abstract

The objective of this research is to create effective visualization tools to display Big Data, meaning data incomprehensible to the mind in raw form, in order to extract high-value information from Deep Analytics. This paper examines visualizations tested on Lexical Link Analysis (LLA) and Automatic Dependent Surveillance-Broadcast (ADS-B) data sets. One challenge that shaped this project is satisfying users' needs for Smart Data because the definition of high-value information varies by user and by data set. To address this challenge, a variety of tools and visualization strategies were implemented for the same data set to analyze the strengths and weaknesses of each design. These visualizations were created with D3.js, a JavaScript visualization library. The preliminary findings of this research is that force-directed visualizations are currently the best visualization for LLA results.

Introduction

Military applications need Big Data that are distributed, disparate, multi-sourced and real-time to reveal abnormalities. Deep Analytics, which includes deep learning (DL) and artificial intelligence (AI), transform Big Data into Smart Data, i.e., the resulting knowledge repository could be considered actionable. One defining problem in the Big Data research field is displaying results in a comprehensible manner to big decision makers. One answer is to use Lexical Link Analysis (LLA).

LLA is a deep learning method (Zhao, MacKinnon, and Gallup 2015). In a LLA, a complex system can be expressed as a list of attributes or features with specific vocabularies or lexicon terms to describe its characteristics. For example, for text documents, word pairs or bi-grams are extracted as lexical terms. Figure 1 shows an example of such a word network discovered from data. Clean energy and renewable energy are two bi-gram word pairs. For a text document, words are represented as nodes and word pairs as the links between nodes. A word center (e.g., energy in Figure 1) is formed around a word node connected with a list of other words to form more word pairs with the center word energy. The Bi-gram method allows LLA to be extended to structured data. LLA outputs Smart Data such as social and semantic networks, patterns such as associations, themes and topics grouped into popular, emerging and anomalous information. LLA has been used in many different application domains to facilitate the discovery of high-value information. Authoritative information from network nodes were used to discover leadership and archetypes in a social network(Zhao, Kendall, and Young 2016). Emerging information can be used to discover innovation from crowdsourcing(Zhao, MacKinnon, and Zhou 2017). Anomalous associations were used to identify fraudulent behavior and imposters(Zhao, Kendall, and Young 2016).

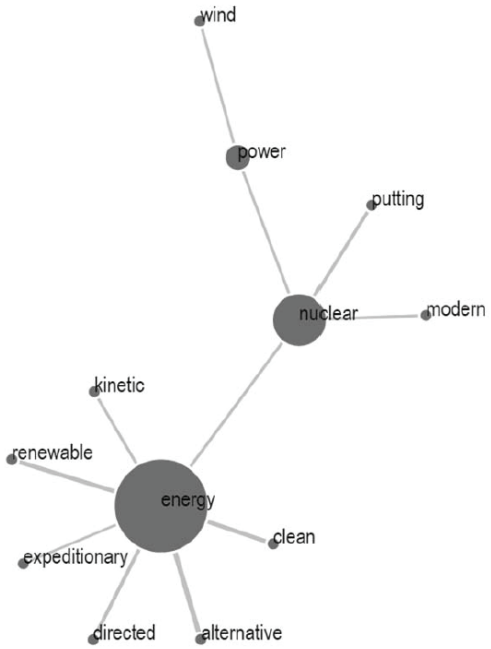

Figure 1: An example of a theme discussed in LLA.

A geo projection visualization was also implemented to visualize Automatic Dependent Surveillance-Broadcast (ADS-B). This data is intended for Naval Common Tactical Air Picture (CTAP) and Accurate Combat Identification (CID). CTAP collects, processes, and analyzes data to provide situational awareness to decision makers. CID enables warfighters to locate and identify airborne objects as

friendly, hostile or neutral. CID plays an important role in generating the CTAP and other combat systems(Zhao, Kendall, and Young 2016).The NPS team downloaded historical JSON files (taken every minute) of 4TB for a whole year (ADSBexchange.com 2017). We use LLA to analyze patterns and anomalies in ADS-B data. The results can be also used to improve CTAP and CID.

Methodology

Our main strategy for project development is as follows. First, research relevant visualization techniques. Next, implement and create customize features to enhance usability. Finally, assess how the implementation matches users' needs and provides enough axes of freedom while remaining easy to use.

Visualizing Lexical Link Analysis (LLA)

Force-Directed Graphs

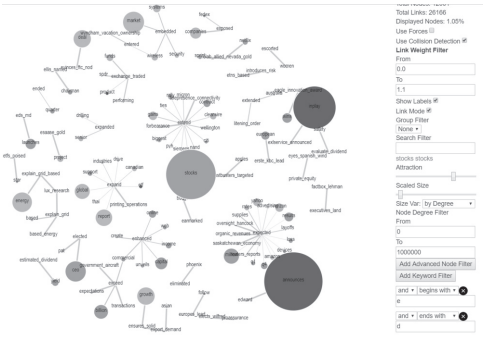

Figure 2: News articles data set filtered by words that begin with 'e' and end with 'd' with no latency issues.

The force-directed diagram attempts to maximize the relationship between user inputted filters while keeping a clear interface. It utilizes an auto clustering force that push nodes with strong links close together. A user can make choices like displaying only anomaly items that begin with "g" and have link strengths above 0.7. Every word that matches this description will pop up with its association that meet the strength requirements. This functionality is unmatched by any other visualization mentioned in this paper.

# Nodes Displayed	Quality
$\leq$ 500	No Latency, Mostly legible
1000	No Latency, hard to read everything
2500	Small Latency
4000	Moderate Latency
7500	Moderate-Severe Latency (hard to type)
10000+	Terrible Latency(little functionality)

Table 1: Stress test results for force-directed graph

On a force-directed graph (Bosack 2017b), each word is displayed as a node and a word pairing is displayed as a link. The nodes are colored by their type. Anomalous words are yellow, emerging words are blue, and popular words are green. One useful example of a good purpose for this model is to find all the words in a group of articles that have a strong ($\geq$ 0.8) association with the word "intelligent". A strong association means that two words frequently appear together in sentences in the inputed documents. The word "intelligent" is placed in the center and all associations are connected to it with a strength percentage.

One key weakness of the force-directed graph visualization is that it does not show large unfiltered data sets well. The screen will begin to lag with too many nodes are displayed at once. The nodes also cannot leave the svg boundaries, so as more nodes are featured they will begin to overlap and block other nodes making some illegible and leading to a cluttered and disorienting view.

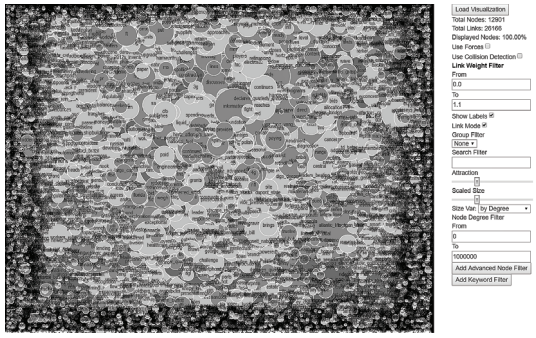

Figure 3: Same data as Figure 2 with the data unfiltered. Application is facing extreme latency issues.

Table 1 shows the results of stress tests to give a more qualitative view of the current performance of the visualization. The obvious trend is that performance degrades as the number of nodes displayed increases. As shown in Figure 3, visualizing an unfiltered data set is not only slow but also incomprehensible. To combat this issue, more filtering tools are provided to filter the data set to the user's needs. Tools include filtering by group, source, link weight, node degree, keyword, or words that begin with 'x', end with 'x', and include 'x'. 'x' can be any string of alphanumerics. The nodes can also be dragged around and forces can be turned off.

Dynamic Time Series

This visualization as shown in Figure 4 was used on LLA output to display dates on the x-axis. A circle's size represents a number of actual interests (i.e., ground truth) for the information that had their own type (popular, emerging, anomaly) analyzed from LLA. The visualization is intended to show if and how the types analyzed by LLA are correlated with the ground truth of interests in real-life, and how the correlations spread out over time. An example for this visualization is if you had a transcript of a meeting and mapped every word to time uttered in a file and then wanted to see LLA breakdown of which words were labeled as anomalous, emerging, or popular.

The benefit of this visualization is that it can use the same data file and change the y-axis and radius dynamically depending on the attribute the user chooses to compare. The

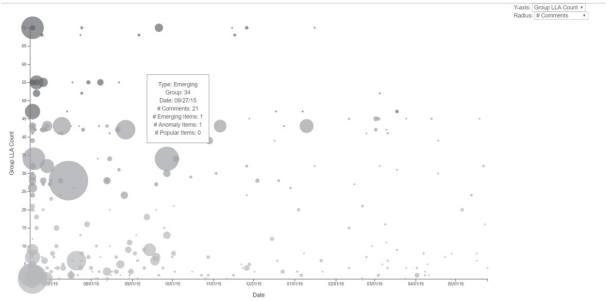

Figure 4: Time series visualization. Radius by number of popular items. Y-axis by number of Emerging items.

user can also hover over any data point to see all the information displayed at once. Another benefit of this visualization is that the user can quickly change attributes and follow the data point as it moves to a new position. This visualization also faces fewer latency issues than a force-directed diagram because the design is more simplistic.

Dynamic Donut Charts

Figure 5: Current stage of dynamic pie charts.

This data visualization is meant to display percentage information. In figure 5, the data is the same from the force-directed graphs. The user can add an infinite number of dynamic pie charts to the screen. For any pie chart, the user can choose what attribute the pie chart is displaying from the data such as percent from each group or percent from each source. Pies can be deleted and added. Currently, the donuts are static but are created with a custom class that can place them anywhere and create as many as desired at any size. In Figure 5, the outer donut showcases the percentage of word associations that came from each of the two sources. LLA fuses together many data sources/files so some users have requested seeing which link comes from which sources. The largest portion of the donut is the percentage of word associations that came from both sources (i.e., R55687.txt and R55688.txt), meaning one word of the link came from one source and the other word came from the other source. The inner donut is showing the percentage of words that belong to each LLA group. The donut pieces are arranged in order of decreasing percentage as one reads clockwise. You could also place the same two pies next to each other and view the same data set information at the same time. The predicted strengths of this visualization is that it will be good for showing percentage statistics in a dynamic fashion. Weaknesses include that it has the same limitations as any pie chart other than it can transition between attributes quickly.

Chord Diagrams

This visualization (Figure 6), created with the help of Bosack's template(Bosack 2017a), functions as a different way of displaying LLA information. Each tiny line along the perimeter of the outer circle represents a word. Each line in the center represents an association between two words. This visualization is not an excellent way to view the information quantitatively but it does present the data in a visually pleasing manner.

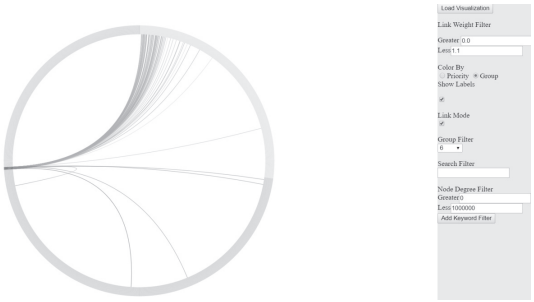

Figure 6: Chord Diagram colored by group and only showing links that have at least one node in group 6.

One of the many weaknesses of this visualization is that it can't present each word well when there are thousands of words. In Figure 6, there are 12000 words being displayed. When searching for a specific word, it's difficult to maneuver over the correct node to display the related information. The visualization was improved over the original design of the D3 template by employing a filter in the side bar as shown in Figure 6. On the side bar, the user can filter by many attributes such as the LLA group and keyword. An example of a strong use case is if displaying 30 emerging themes from several documents and displaying how these themes relate to each other.

Visualizing the ADS-B Data

The goal of the ADS-B visualization is to be able to view the trajectories of airplanes and then filter out different vehicles, study deltas in velocity, study flight patterns, filter destinations and arrivals, etc. This project is in its preliminary stages. It was created with the help of the D3 projection API (Bosack 2017c). It currently can display the signal of every plane at any given moment and animate the planes movement from minute to minute. Another option currently available is displaying all planes that took off at the selected time and showing where that plane went in the

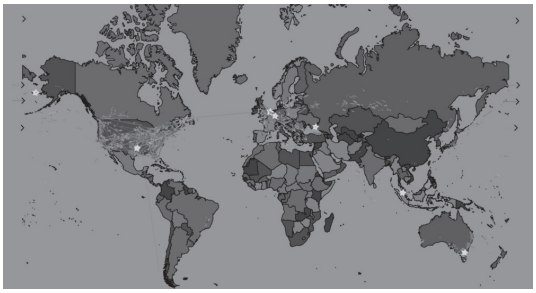

Figure 7: ADS-B visualization of flight data from 6/20/16. Each line represents the trajectories of all planes that were in flight at some point over a 40 minute period.

scope of x minutes. Currently this visualization is limited by the scope of the data set. To play 10 minutes of flights, 10 5KB files need to be loaded in and looped through to show all of the data. Therefore, as more time is displayed, latency increases. Work has already began to generate smaller data files to decrease latency. The ADS-B data came from ADS-B Exchange (ADSBexchange.com 2017). Current features of this visualization include zooming, panning, and playing aircraft movement.

Future Work

According to Ying Zhu, assessing data visualizations includes using heuristic evaluation and user studies (Zhu 2007). Future work includes creating new metrics to assess the quality and efficiency of these visualizations for understanding the data compared to traditional methods, and performing user studies that measure task completion time, error rate, and user satisfaction.

To improve specifically the Force-Directed Diagram, which is most likely to help the user, efforts need to be made to improve the latency effect. Although, latency is somewhat dependent on browser performance limitations. Some suggest canvases perform slightly better than svgs on data visualizations, but the cons of canvases are that it's harder to create dynamic features for the user like the ability to drag nodes.

The ADS-B visualization still needs filters implemented to allow the user to improve analysis. Future filters will include by arrival, destination, plane type, manufacturer, and ICAO. The visualization will also display the tracks of the flights in a more visually appealing way.

Conclusion

Many of these visualizations are still in the process of being implemented and refined. These tools extend the user's understanding of Smart Data results from Deep Analystics on LLA data and ADS-B data. Once fully completed, the visualizations will need to be tested for effectiveness as mentioned in Future Work. Redesign will be required as users offers more feedback. The visualizations vary in complexity and offer a variety of viewpoints for the same type of data set. However, the user would most likely only require the force-directed Diagram for most cases. The ADS-B visualization is a work in progress and many other visualizations should be implemented to supplement the one geo-projection visualization to provide more statistics for the data. The dynamic donut and dynamic time series can easily be re-purposed to help present statistics related to the ADS-B data set. In closing, D3 is a wonderful tool for these specific data sets and once fully implemented will likely extend the user's capacity and comprehension of large data sets.

References

ADSBexchange.com, L. 2017. Ads-b exchange.

Bosack, M. 2017a. Chord diagram.

Bosack, M. 2017b. Force-directed graph.

Bosack, M. 2017c. Geo projections api.

Zhao, Y.; Kendall, W. A.; and Young, B. W. 2016. Leveraging lexical link analysis (lla) to discover new knowledge. In *Military Cyber Affairs*.

Zhao, Y.; MacKinnon, D.; and Gallup, S. 2015. Big data and deep learning for understanding dod data. In *Journal of Defense Software Engineering, Special Issue: Data Mining and Metrics*.

Zhao, Y.; MacKinnon, D.; and Zhou, C. 2017. Discovering high-value information from crowdsourcing. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Zhu, Y. 2007. Measuring effective data visualization. In *Proceedings of the 3rd International Conference on Advances in Visual Computing - Volume Part II*, ISVC'07, 652–661. Berlin, Heidelberg: Springer-Verlag.