Research Report

Visual Analytics

Visual analytics is the science of deriving information from vast amount of data through the process of visual modeling, interactivity, and/or dynamic feedback. This allows the marrying of both computers processing power along with human analytical capabilities to extract details that would otherwise be inaccessible to each technique alone, or being able to make conclusions much faster while expending fewer resources.

Datapresso aims at using this concept to facilitate the processing of large amounts of documents from the traveling/homeland security domain. This is in hopes to identify, amongst other details, persons of interest, location activity, and organization sentiment from the selected corpus. It is targeted to run on a web browser to increase accessibility.

Overview

Most of the research centered on the implementation of three main principals; geospatial, sentimental, and relational. The documents within the corpus where manually taken from various online sources and processed through Datapresso's entity extraction framework. The unstructured text was converted to structured entities which were used as the base points for all the visualization. These entities fall within a categories ranging from people, organizations, and locations to name a few. By doing this, Datapresso allows individuals to quickly gain a sense of what the document's general idea, feel, and sentiment. To further simplify this process visualization where used by modeling the extracted entities in different graphical layout.

Software

The graphics for the visualizations are handled by Protovis, a JavaScript graphics library. The simplicity required to implement the library proved an important factor since no plugins or additional installation is required, asides from a web browser, which allows access to a wide variety of users. The ability to draw simple shapes from arrays of data without the need to write complicated loops also made Protovis very attractive. The time that would have been required to come up with such loops was instead invested elsewhere, and this simplification reduces debug time. Very importantly, Protovis also includes a built in ability to interact with the draw shapes using the mouse cursor. This opened the chance for interactive visualizations, and helped stream line the process from raw data to its visual representation on the screen.

Initial efforts were made to pre-process the data to make loading quicker, but ultimately led to issues since the options available to the user became limited. This was overcome by directly using MySQL's query capabilities. It not only allowed for quick retrieval of the data, but also to opened dynamic changes to data during the execution of the visualization.

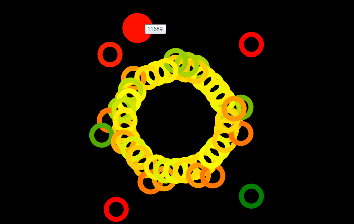
Visualizations



Different approaches to visualizing the data where taken. One of these was to take a desired organization and look for regional influence and the sentiment associated with it. The documents where the organization is mentioned are searched for occurrences of locations. They can range from specific facilities, to cities, to countries. These location are then plotted on a world map with the aid of Google maps api which allows the translation of a location's name to its global position. This geospatial placement of markers allows for the interpretation of the global influence the organization.

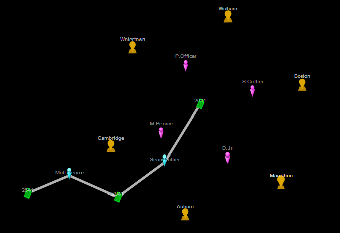
Along with the geospatial details the visualization also takes the sentiment data from within the documents and displays it alongside. Datapresso detects certain phrases within the document's text and labels them as either positive or negative. The visualization then only focuses on the documents that mention the organization and tallies both the positive and negative sentiment entities. The final count is displayed using a color scale centered at yellow, neutral sentiment, and gradually turns to either green, net positive sentiment, or red, net negative sentiment.

Each sentiment entity is assumed to carry the same weight and future development could at considering different phrases having different degrees. Also, the sentimental data could be further merged with the geospatial markers to display each specific area's towards the organization. This will allow the user to make an assertion of region specific influence as well as its distribution across the board at the same time.



Another visualization approach aims at evaluating the entire corpus on a sentimental aspect. The entity extraction capabilities of Datapresso allow for the extraction of certain phrases and the labeling of the sentimental inclination. This visualization then calculates a net sentiment as a whole by adding those entities for the entire document. The results are familiarly displayed in a color range where positive, neural, and negative sentiment correlate to green, yellow, and red respectively. The user can very quickly grasp the overall feeling of specific documents. To further aid, the documents are positioned towards the center, and those with greatest sentiment, in either direction, are moved outward. This is not only for easier selection purposes, but allows quicker identification of those documents that poses extreme cases in the entire corpus.

This visualization again does not account for variable sentiment in phrases, and this as a result does not make a normalized comparison between them. However, being able to distinguish the sentiment inclination does allow for concentration of further processing of only a subset of documents which could ultimately reduced time and resources need to evaluated the key points of the entire corpus.

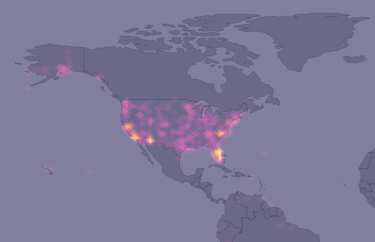


The last visualization takes a different approach from the previous ones. While they concentrate on highlights specific aspects this one simply allows the user to interact with the visual representation of the data as it dynamically changes. The hope is that that through the visual manipulation of the entities, certain patterns or links can be recognized. It allows the traversal of the structured data in a way much like a detective following leads and making connection between them. These links are currently limited to the person, location, and document entities. The basic correlation between these entities is their mentioning within their respective documents.

The user is initially presented with all the documents in the corpus to start from. After making a selection the related entities to that one are made available for further selection. Once one of these new entities is selected the selections change to those entities correlated to it and the process repeats. Those entities that have been picked are tethered, on a flexible line, in the order of their selection. The user can make changes to the line or can retrace their steps and start over from a previously selected entity. This design allows the user to narrow down the number of documents that must be thoroughly read.

Further enhancements can be made to include the text of the selected document alongside the visualization, and for it to change as the user progresses though their search. This would allow the user to gauge the strength of the correlations, assumed by the software, between the entities themselves. For example, if the names of four individuals are mentioned within a document the user can read a highlighted portion of text, which mentions an individual of interest, and makes the conclusion that one of the other entities is sibling, or a mentor, or business partner. Further progression of this enhancement could add the ability to label the connection within entities to clarify the exact nature of the links between entities.

Side Visualization



The project's main focus was to take advantage of the structured data created by Datapresso. However, additional areas relating to the traveling/homeland security domain were explored. These included aviation incident reports for the United States ranging back ten year. This is to see the possibility of making comparison between the entities to data from other sets. Currently with the limited size of Datapresso's data set, relative to the aviation data as well as the structure of the entities, a definite comparison cannot be made. However, change to this can occur as the size of the corpus increases.

This visualization takes advantage of both the timestamps and latitude/longitude coordinates of the aviation incidents to create a heat map. The heat map is animated to reflect the passage of time showing concentration of incidents clearly by bright yellow areas. Future developments could include the automatic detection of these concentration areas. Entities would be made from these detections, and would be compared with other types of entities to hopefully extract meaningful information that may be hidden.

Conclusion

Based on the capabilities of Datapresso, to structure data, and the use of Protovis, for visual manipulation, It was possible to show the strengths of visual analytics within the travel/Homeland Security domain. Each visualization highlights a different aspect of the same data, and allows a much quicker understanding of that information without having to read each of the individual documents. This has the potential to reduce and prevent threats while traveling. As well as reducing the resources necessary to provide such security.