Information Network Navigation

Ryan W. Blankemeier

Dartmouth College

Follow this and additional works at: https://digitalcommons.dartmouth.edu/senior_theses

Part of the Computer Sciences Commons

Recommended Citation
https://digitalcommons.dartmouth.edu/senior_theses/160

This Thesis (Undergraduate) is brought to you for free and open access by the Theses and Dissertations at Dartmouth Digital Commons. It has been accepted for inclusion in Dartmouth College Undergraduate Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
Information Network Navigation

Senior Thesis
Department of Computer Science
Ryan Blankemeier '20
Advised by Professor Dan Rockmore
Abstract

In this paper, we develop an interactive system to navigate information networks as a space with geometry, assigning each node in the network to geographical coordinates, and with that the ability to navigate as if on a map. A map-based rendering of the network gives the user the ability to understand meta-relationships (i.e., non-link-based relationships) that exist in the dataset that are lost with a traditional web search and (hyper-)link navigation. This requires first being able to represent the information corpus in such a way as to enable a quantifiable notion of similarity between the information nodes. A t-SNE (t-distributed stochastic neighbor embedding) model then finds an optimal embedding of the nodes in two-dimensional space such that the pairwise distances (dissimilarities) between points are best preserved. With this there are many opportunities to enhance the exploration of the space such as visualizing exploration paths and a compass displaying the orientation of the information space.

1. Introduction

Rarely when we interact with information networks do we have any notion of a space or geometry. Most interactions with networks allow us to move through them, but allow little intuition on where we lie in relation to the network as a whole. Importance metrics like PageRank [13] guide such interactions. PageRank and its variants [13] suggest important pages that relate to our current page, but generally present options in a prioritized list, giving no sense of the ambient “space” -- i.e., no larger context in which to

---

1 Special thanks to Faraz Dadgostari (Department of Computer Science, University of Virginia) for supplying the dataset and assisting throughout the development of this project.
consider the options, much less the parts of the information corpus (network) that the search engine has deemed irrelevant.

If a user is navigating a network in the traditional way, they may follow hyperlinks to new topics but miss the related pages (non-link relationships) because they did not appear in the narrow path of hyperlinks [4]. Some papers have suggested the idea of direction when orienteering a network, moving through a network towards broad concepts that make up that information space (e.g., [1]) and how to represent this orientation to the user to assist their navigation of the space [12]. This paper builds on these ideas by creating a system that visualizes the information network as a map, provides an orientation for the map in terms of its aggregated subject matter, and displays an interactive interface that allows the user to realize the non-link relationships in the space while keeping the link relationships in the data available.

2. Data

In this paper we focus on a specific dataset from the open source legal platform Bending the Law [2]. This dataset consists of 21,893 Supreme Court text opinions. We use the topic model representation of this data derived by [11] in order to represent the semantic content of these documents. In a topic model, the opinions are transformed into a bag-of-words representation where the opinion is represented by word frequencies. Because word order is ignored in the bag-of-words representation, the dimensionality of the corpus is greatly reduced. This dimensionality is reduced even more by representing the opinions as distributions of topics instead of words. A topic is a probability distribution over a set of words. Topic modeling is the process of deriving a set of word distributions
that can be used to represent the corpus as a whole and there are a range of such models [14]. This paper uses the topic information from [11] generated by running a latent Dirichlet allocation model (topic modeling that assumes a Dirichlet distribution over the corpus). This model produces two matrices which this paper will use: PHI and THETA. PHI is a matrix with each row \( i \) representing Topic \( i \), and column \( j \) representing the frequency of word \( j \). Entry \( W_{ij} \) in PHI represents the frequency of word \( j \) in topic \( i \). The \( W_{ij} \) weights are non negative and must add to one for each row.

| \( T_1 \) | \( F_1 \) | \( F_2 \) | \( F_\ldots \) | \( F_m \) |
| \( W_{11} \) | \( W_{12} \) | \( W_{1\ldots} \) | \( W_{1m} \) |
| \( T_2 \) | \( W_{21} \) | \( W_{22} \) | \( W_{2\ldots} \) | \( W_{2m} \) |
| \( T_\ldots \) | \( W_{31} \) | \( W_{32} \) | \( W_{3\ldots} \) | \( W_{3m} \) |
| \( T_n \) | \( W_{n1} \) | \( W_{n2} \) | \( W_{n\ldots} \) | \( W_{nm} \) |

We can assign meaning to the topics by analyzing the top word frequencies. The following are a few of the topics, their top words, and resulting (human-) assigned label.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top (7) words</th>
<th>Assigned Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic043</td>
<td>['alien', 'aliens', 'united', 'deportation', 'states', 'immigration', 'citizenship']</td>
<td>immigration</td>
</tr>
<tr>
<td>Topic045</td>
<td>['property', 'bankruptcy', 'debtor', 'lien', 'claim', 'creditors', 'trustee']</td>
<td>credit/lending</td>
</tr>
<tr>
<td>Topic051</td>
<td>['prison', 'inmates', 'inmate', 'prisoners', 'prisoner', 'officials', 'conditions']</td>
<td>prison</td>
</tr>
<tr>
<td>Topic052</td>
<td>['speech', 'amendment', 'public', 'free', 'government', 'ed', 'expression']</td>
<td>speech</td>
</tr>
<tr>
<td>Topic058</td>
<td>['abortion', 'state', 'woman', 'medical', 'physician', 'life', 'health']</td>
<td>abortion</td>
</tr>
</tbody>
</table>

THETA is a matrix with each row \( i \) representing Opinion \( i \), and column \( j \) representing the frequency of Topic \( j \). Entry \( W_{ij} \) in THETA represents the frequency of topic \( j \) in opinion \( i \). As in PHI, the \( W_{ij} \) weights are non negative and must add to one for each row.
With PHI we can represent the topics as a distribution of the words in the corpus and of a dimension equal to the size of the vocabulary. With THETA, we can represent opinions numerically in 100 dimensions as a distribution of 100 topics. Our motivation was to take this high-dimensional data and represent it as an interactive map of opinions in two dimensions for a user to navigate [2]. For more information on the topics and how they were created see [2].

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_\ldots$</th>
<th>$T_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1$</td>
<td>$W_{11}$</td>
<td>$W_{12}$</td>
<td>$W_\ldots$</td>
<td>$W_{1m}$</td>
</tr>
<tr>
<td>$O_2$</td>
<td>$W_{21}$</td>
<td>$W_{22}$</td>
<td>$W_\ldots$</td>
<td>$W_{2m}$</td>
</tr>
<tr>
<td>$O_\ldots$</td>
<td>$W_{31}$</td>
<td>$W_{32}$</td>
<td>$W_\ldots$</td>
<td>$W_{3m}$</td>
</tr>
<tr>
<td>$O_n$</td>
<td>$W_{n1}$</td>
<td>$W_{n2}$</td>
<td>$W_\ldots$</td>
<td>$W_{nm}$</td>
</tr>
</tbody>
</table>

3. Data Analysis

Using the PHI and THETA matrices, we explore different visualization and clustering algorithms to display the data clearly. This requires reducing the dimensionality of the opinions in the THETA matrix from a 100-dimensional space (100 topics) to the two-dimensional cartesian coordinate system. Once the opinions are given coordinates in two-dimensional space, we can fill them in with different colors. Our first choice for coloring the opinions was a color-coding of the topics and then using (the color of) the largest topic in the opinion's topic distribution. We then apply various clustering methods to opinions in the THETA matrix to determine if there are larger groupings of opinions that might suggest a coarser, but easier to understand "map" - continents, as opposed to
countries, if we continue the analogy. Additionally, we attempt to cluster the topics themselves to understand the broader topic clusters within the corpus and assign meaning to the directions of the axis in the visualization.

3.1 t-SNE

In order to assign coordinates to each one of the opinions, we ran scikit-learn’s Python implementation of the t-SNE model on the opinion topic distributions [6]. t-SNE (t-distributed stochastic neighbor embedding) is a way to visualize high-dimensional data. A brief explanation for how the model works is as follows. First it calculates the euclidean distance between all pairwise opinions. Then for each opinion it creates a probability distribution over its all opinions in the corpus. This means that for each opinion, it gives the probability that another opinion would be selected as a neighbor, which would be high for related opinions, and low for unrelated opinions. Then it creates a similar probability distribution for the low-dimensional space, and performs gradient descent on the Kullback-Leibler divergence between the high-dimensional probability distribution, and the low-dimensional probability distribution [6].

Kullback-Leibler divergence is a way of comparing two probability distributions, and reveals how much information is lost when one probability distribution is used to approximate another [8]. From the low-dimensional embedding of the t-SNE model we get x and y coordinates for each opinion. Using the rows of the THETA matrix, we find the maximum topic for each opinion. Then we color each point according to its majority topic. We also include the topics in the t-SNE model as topic vectors with only one topic.
We add the topics to the THETA matrix by specifying a weight of one for its topic, and zero for all others.

Figure 1: Opinions visualized in two dimensions, colored by their majority topic. Topics are labeled by topic number, given the x/y coordinates returned from the t-SNE model.

If we zoom in on this visualization, it becomes clearer that opinions are generally surrounded by opinions that share the same majority topic from their topic distributions in the THETA matrix.
Figure 2: Zooming in on the visualization shows that the topics are separated well.

Now that we have the opinions fixed in two-dimensional space, colored according to their majority topic, we want to determine if there are better ways to represent the opinions by reducing the number of clusters from the original 100 topics.

3.2 k-means Clustering

In the images above, the opinions appear in clusters according to their majority topic. *K-means* is a clustering algorithm that attempts to create *k* clusters in a dataset of *n* samples such that the WCSS (within cluster sum of squares) is minimized [7]. This error metric describes the summation of the squared distance from each point the center of its cluster. The equation for WCSS is shown below, where *C* is the set of clusters and *μ* is the mean of a cluster.
Several K-means clusterings of the opinions were attempted to see if the topics should be grouped into more general clusters. K-means was run testing from 2 clusters to 100 clusters, and calculating the WCSS for each number of clusters.

\[ \sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2) \]

Figure 3: WCSS decreases steadily for each increase in the number of clusters. We do not see a clear "elbow" where an increase in the number of clusters results in only a small decrease in the error.

As shown in the graph, there is no clear "elbow". This would be a point of interest because if the rate of change went from highly negative to slightly negative as we increase the number of clusters, we would see that increasing the number of clusters did not give a big reduction in error. Because WCSS did not yield much useful information a similar analysis was run with the silhouette score, which measures how similar a point is
to its cluster compared to other clusters. For each sample, the silhouette score is the difference between the distance from sample to its own cluster and the distance from the sample to the next nearest cluster, divided by the maximum of the two. The silhouette score is averaged over all samples, so it ranges between 0 and 1.

**Figure 4**: Silhouette Score is maximized from around 40 to 60 clusters, suggesting that the optimal clustering of opinions from the 100-dimensional THETA matrix may lie in this range.

The silhouette score implies that the best clustering might be around 50, but is also not definitive. We visualized the data with a clustering k=50 on the 100-dimensional THETA matrix of opinions. We found that while the clusters on the outer regions of the map stayed generally the same, the clusters around the middle of the map kept expanding to form one large middle cluster. The documents in the middle tend to be noisy even when colored with their majority topic, largely because they do not fit into a single topic, their distributions are very scattered. This problem exposes the difficulty of displaying very
high-dimensional data in a two-dimensional space. The points in the middle of the map tend to have much more complicated distributions and therefore are rendered in a position more equidistant to the other clusters. This causes the smaller, more extreme clusters to stay intact, while the more complex clusters all combine in the middle. Cluster 21, in the middle of the visualization, is the cluster that contains opinions spread over all the topics. The takeaway is that rather than create new clusters across the space of topics, the clustering preserved the clusters around the outside of the space, and combined the ones in the middle.

**Figure 5:** The light green cluster in the middle (21) swallows up many of the surrounding clusters, while the clusters on the outside stay intact.

We then ran k-means with a much smaller k, hoping to obtain even more general clusters. We found the same result but more extreme. The center catch-all cluster kept getting bigger, and the specific border clusters did not change, adding little information in terms of generalizing the clusters further. This illustrates the difficulty of rendering such
high-dimensional data in two dimensions, and reinforces the idea that the best coloring of the opinions is the majority topic from their topic distribution. The results of k-means with k=20 are shown below. The clusters are dominated by topics on the outside of the visualization, and the center topic keeps expanding to contain all the topics in the middle.

*Figure 6*: Clustering over the topics in the 100-dimensional space of the THETA matrix does not produce helpful clusters of topics. It simply clusters the topics in the middle of the visualization that have more variation in their topic distributions.

Because we did not create new uniform clusters containing groups of topics, but rather preserved the clusters with less diverse topic distributions and combined all the ones with more complicated distributions, we decided to color the opinions by majority topic. In the future, clusterings could be applied to the 2-dimensional representation of the opinions from the t-SNE model to create more general colored clusters.
4. Orientation and Navigation

Once the opinions are given fixed coordinates in the information map, organized by pairwise similarity, we see that the space has an inherent orientation and direction.

Although the topics are represented in the PHI matrix as word distributions, they can also be represented just as the opinions are in the THETA matrix as topic distributions. They would just be a topic distribution of 100% their identity topic and 0% all other topics. These observations were included in the t-SNE model to assign coordinates in space to the topics as well as the opinions. Hierarchical clustering is an algorithm to form clusters by iteratively merging or splitting previous clusters. We used scikit-learn’s Agglomerative Clustering implementation, which starts off with each observation in its own cluster, and iteratively merging clusters to form the final clusters [7]. This sort of clustering on the topics coordinates could reveal a more general cluster of topics. These general clusters can be used to label large sections of the map, which give meaning to moving in a direction across the map.

4.1 Compass

As described above, we first visualized just the topics. Topics can be represented like the opinions, as a pure topic vector with a distribution consisting of one topic frequency in the distribution set to one and the rest set to zero.
We wanted to take these vectors and transform them in a way that displayed some information on how the topics change when you move in each direction. We normalize each vector to lie on the unit circle, producing a circle where every topic vector has a magnitude of one, in the same direction as it is pointed in the visualization. The result is shown in Figure 8, with the spaces between each vector colored in for clarity. This image can act as a compass because it describes which topic vectors point in which directions. This compass has a few limitations in its interpretability. First, there are so many topics around the circle that they are hard to comprehend all at once. Second, for topics that are very close to the center of the space, it is misleading to give them an orientation on the unit circle with the same weight as topics near the edge of the space.
Figure 8: First rendering of the compass. The two-dimensional coordinates of the pure topic vectors from Figure 7 are normalized and projected onto the unit circle. It is hard to read because the high number of topics clutters the circle.

For these reasons, we perform a clustering of the topics to form more general topic clusters. These clusters, with their aggregated topics, better represent orientation in the space, and create a compass that is more readable for the user while navigating the space.

4.2 Hierarchical Clustering of Topics

Hierarchical clustering is a form of clustering where nested clusters are created by splitting or merging clusters recursively [7]. A hierarchical clustering was performed on the topic coordinates generated by the t-SNE model to generalize the topics. A dendrogram is a diagram which shows how clusters are formed by showing where each
cluster merged. The y-axis of the dendrogram shows the value where clusters that have a distance larger than that value will not be merged [15]. The dendrogram is shown below. The general topic labels were qualitatively labeled based on the original topic labels: government, bias, freedom, rights, economy, and legal. For each general cluster, we calculated the average $x$ and $y$ coordinates of each topic in the cluster and then normalized those vectors to be on the unit circle. Because these generalized clusters were dominated by the more specific topics on the outer portions of the map, these generalized clusters have more intuition on how the topics change as you move across the map. They are also much more readable.

**Figure 9:** The topics are clustered into six clusters. The dendrogram shows how the topics are split to form the final clusters. Each colored cluster in the dendrogram corresponds to a column in the chart below that shows which topics make up each cluster.
The dendrogram shows that the topics are divided fairly evenly into these generalized clusters. The topic labels are shown below for each cluster.

<table>
<thead>
<tr>
<th>Bias</th>
<th>Freedoms</th>
<th>Rights</th>
<th>Economy</th>
<th>Legal</th>
<th>Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Legislation</td>
<td>Precentent</td>
<td>Economy</td>
<td>Legislation</td>
<td>Justice</td>
</tr>
<tr>
<td>Evidence</td>
<td>Litigation</td>
<td>Attorney</td>
<td>Retail</td>
<td>Employment</td>
<td>Residency</td>
</tr>
<tr>
<td>Commercial</td>
<td>Jobs</td>
<td>Contracts</td>
<td>Court</td>
<td>International</td>
<td>Budget</td>
</tr>
<tr>
<td>Illegal</td>
<td>Discrimination</td>
<td>Misdemeanor</td>
<td>Felony</td>
<td>Appeals</td>
<td>Government</td>
</tr>
<tr>
<td>possession</td>
<td>Prison</td>
<td>bail</td>
<td>Antitrust</td>
<td>Tax</td>
<td>spending</td>
</tr>
<tr>
<td>Environment</td>
<td>Speech</td>
<td>Damages</td>
<td>Media</td>
<td>Corruption</td>
<td>Abortion</td>
</tr>
<tr>
<td>Property</td>
<td>Documents</td>
<td>Rights</td>
<td>Law</td>
<td>Social programs</td>
<td>domestic</td>
</tr>
<tr>
<td>Time</td>
<td>Real estate</td>
<td>Land</td>
<td>Procedure</td>
<td>Punishment</td>
<td></td>
</tr>
<tr>
<td>Immunity</td>
<td>Smuggling</td>
<td>States</td>
<td>Workers</td>
<td>Plea</td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>Constitutional</td>
<td>Natural resources</td>
<td>Defamation</td>
<td>Jury</td>
<td></td>
</tr>
<tr>
<td>Interrogation</td>
<td>Utilities</td>
<td>Loyalty</td>
<td>Trade</td>
<td>Cases</td>
<td></td>
</tr>
<tr>
<td>credit/mortgage</td>
<td>Investment</td>
<td>Court</td>
<td>Waters</td>
<td>Interpretation</td>
<td></td>
</tr>
<tr>
<td>States</td>
<td>church/state</td>
<td>Trade</td>
<td>Trade</td>
<td>Protections</td>
<td></td>
</tr>
<tr>
<td>immigration</td>
<td>segregation</td>
<td>Irs</td>
<td>Irs</td>
<td>Military</td>
<td></td>
</tr>
</tbody>
</table>

To form the compass, we first plotted the normalized vector for each cluster on the unit circle. Then we drew a line halfway between each cluster point on the edge of the circle. Filling in these regions creates our generalized clusters regions of the compass. Now we can intuit that if we move right in the visualization we are moving in the direction of bias, but if we move left, we are moving in the direction of the economy.
Figure 11: The final compass showing the areas on the circle which correspond to the normalized vectors of the average of the two-dimensional coordinates of the topics over each cluster.

Combining the points and coordinate system with the colors by topic and the compass, we are now ready to create an interactive visualization for the user to navigate.
5. User Interface

**Figure 12:** Screenshot of the interactive user interface. The user sees their current path displayed on the right. They see their path displayed in yellow with a star at the current opinion. Dotted lines from the current case point to any cases that the current case cites.

The goal of this project is to create a visualization where a user can interact with an information network as a space with geometry and direction. This is achieved in the website at [https://information-network-navigation.herokuapp.com/](https://information-network-navigation.herokuapp.com/). The website allows a user to click from opinion to opinion creating a path connected by edges. The user can see where they have been in relation to the other opinions in the space. The user has the option to clear the path at any time. Everytime the user clicks a new opinion, edges appear showing which cases are cited by the current opinion. This preserves the underlying linkage of the network of opinions, while still showing the user related opinions in space. The user can pan and zoom around the space to understand their current location in reference to the bigger space.
5.1 Visualization

The open source python library plotly was used for the visualization [9]. Plotly allows me to draw all the points in space, show the opinion titles on hover, draw edges between citations, and do everything related to the displaying the graph. Dash is a python framework built on top of plotly for developing responsive web applications. Dash wraps a web app around the visualization that allows it to be interactive, responding to user clicks and functions, and rerendering the graph to allow path visualization, panning, zooming, and anything else related to the interactive display. Heroku is used for management and deployment [10].

5.2 Path/Caching

To visualize the user's path we needed to develop some sort of persistent data. If the path was just stored locally, any new user would first see the path of whoever had changed it most recently, and then the new path would overwrite any other user’s path. To solve this we created a cache in the form of a csv. When a new user opens the website we assign them a session id, which is written to the csv file. When a user clicks on points to form a path, the points are appended to the column in the csv file that corresponds to their session id. The cache is set up to handle 50 unique users, if there were 51 users, the 51st would overwrite the 1st users position in the cache. This would be better handled using a database that would scale much easier but for the purpose of this demonstration the csv cache is lightweight and sufficient.
6. Use Cases for the Map and Navigation of Court Opinions

6.1 Legal
Imagine a lawyer reading opinion after opinion taking notes on precedent and how they connect. If the same lawyer could interact with these opinions as a space, this process would be intuitive. They could move through the space following citations, and understand where their current opinion fits within the space of precedent.

6.2 Internet
This idea could be generalized to an internet search. If the user is just looking for the answer to a simple question, then an internet search would suffice. But if the user is looking to understand an idea, or search a space, then the ability to move through the space and understand how the topics relate on a higher level is important. The user could click through the results of their query as a space, reading the web pages while moving through the space and understanding where they are within the larger topics.

7. Future Work

7.1 Map
Instead of each opinion being represented by a point, we could expand each point to be a region, making the space resemble even more an actual map. A voronoi diagram could be generated from the map of points to create a space in which a user is always in a region that represents a search result.
7.2 Interactive pages
   As a user moves through the space, clicking on new results, they could be displayed in many ways. The page could be displayed as a thumbnail, allowing the user to see a preview of its content. Clicking on the result could link to the actual page and allow the user to leave the visualization to read its contents, without losing their bearings within the space.

7.3 3D/VR
   We visualize the data in two dimensions, but this could be increased to three-dimensional space. With the increased usability of VR and AR technology, the user could actually move through the space and see the page rendered in space while they also navigate the geometry of the space. The user could save their path at any point to revisit their train of thought later.

Acknowledgements
   We would like to acknowledge the support of Professor Dan Rockmore (Department of Computer Science, Mathematics, Dartmouth College) for his guidance and support throughout the entire process. We would also like to acknowledge Faraz Dadgostari (Department of Computer Science, University of Virginia) for supplying the dataset and assisting in the data analysis.

REFERENCES


Accessed 2020 May.