JOHN HERRICK
7/1/17

NLU RELATIONS

Data is everywhere. With massive amounts of data coming from several outlets at once and industries striving hard to pick out the bits that matter, understanding big data has become an integral part of the success of businesses around the world. It comes to no surprise then that there is a huge market for the technologies that can accomplish this and for the people that do it. The real devil is in the detail: how can we make sense of the billions of bits of data we can accumulate, filter it down to the information that is valuable in a certain context, form meaningful connections with this data, and derive insight quickly by analyzing the results. This is known as digesting the data. By digesting we can find the valuable bits, but the challenge beyond this point as a developer is: how can I display this data to executives and clients with no formal technical education? This is the end goal for the Cognitive Environments Team at IBM Watson in New York: to clearly show the capabilities of Watson services to process the billions of bits of information into something that makes sense and prove Watson's company value to these big industry figures.

One such big industry of interest is news. Each day, thousands of news stories are published, some with similar stories and concepts, others with inter-story relationships that are not as obvious. By using a Watson service known as Alchemy (Now used as Natural Language Understanding), a new project by the Cognitive Environments Team makes connections between these individual articles, gauging popular article concepts by overall global sentiment at the time, analyzing popular emotions, and entities. They are able to slice data to extract insights and turn seemingly unstructured data into structured data. In essence, a user is able to start with a mass collection of news stories displayed to them, filter it down to a few topics of interest, then to a collection of stories that relate to those chosen topics, and finally to an individual article view. Here the user may view an individual article out of the thousands they started with and analyze it for top concepts, entities, and even strong assertions. This Article Detail View will be the topic of discussion here.

But now the question is: how can I take this further? How may I use Watson services and the Natural Language Understanding capabilities to both enrich the user experience by showing meaningful insights into the data present in the Article Detail View and prove the value of Watson Services by clearly showing clients that Watson is processing the information it is given? Well once a user is given an individual article and the general sentiments and concepts of interest, it only makes sense to process this information further by somehow summarizing the contents of that article. Being able to exhibit the Natural Language Processing capabilities of Watson is accomplished by using the "Relations" feature of the Natural Language Understanding service. Relations picks out the top mentioned concepts of the article by sentence subject and object and links them by a defined relationship: (Located At, Is a Member Of, etc.). This relations service feature, when filtered down to only the important bits that matter can offer that summarization feature that we desire and offer a high level overview of the article events without requiring the user to actually read the article. Why is this valuable? Relations offers an objective view of the news through the basic relations of the elements in an article, and not relying on subjective and biased sentiment. Here users may digest data without being swayed by how other people or news outlets feel about a certain topic. Relations analysis is pure in that sense and removes the risk of polarizing the user base simply because of how partisan news outlets may frame a certain concept. It is not hard to set up an API to retrieve

the raw data of a particular article by link and display that information to a screen, but the real challenge again, is to create a visualization that is the best possible representation of the link between each Relation object's subject and object. The answer is a Knowledge Tree.

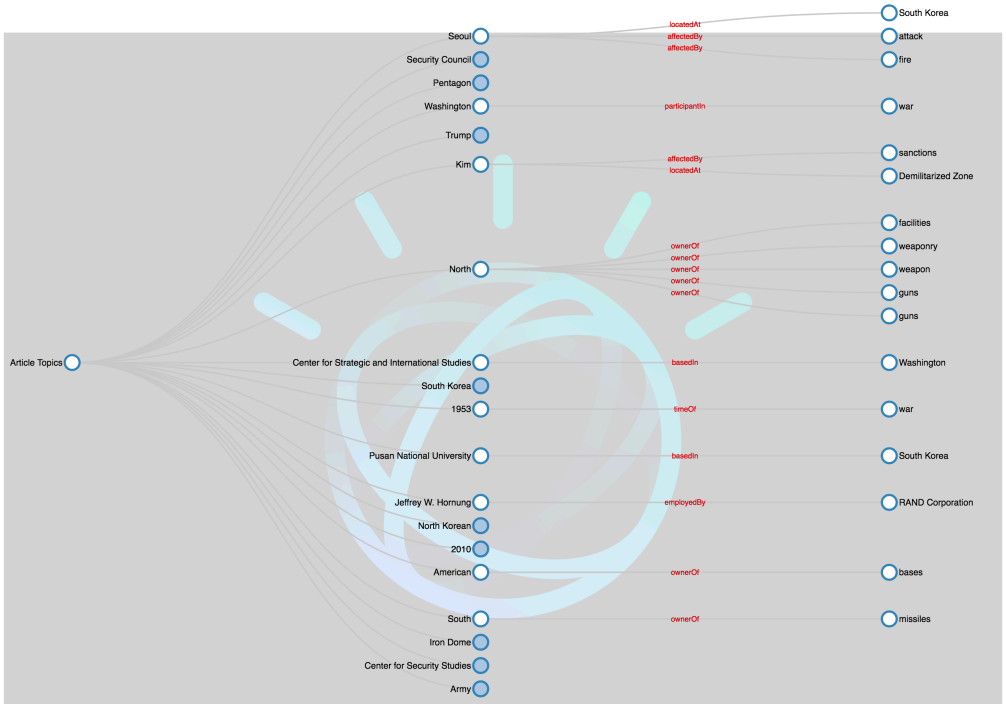
After careful consideration of hundreds of knowledge tree variations, drawing on my knowledge and research on HCI (Human Computer Interaction) concepts, and an effort to sketch and implement a few of the best candidates, the best graph that hones the best qualities that we look for and has the least amount of perceivable flaws is a Dendrogram. Essentially it is a flipped tree with elements indicated by nodes and linked through varying levels with each terminal node on the same level. This type of tree diagram is used typically to show taxonomic relations between species of animals, but it is equally valuable in showing the relations between article concepts. The difficult part of finding the ideal visualization is noting that the links between node elements are just as important as the nodes themselves due simply to the fact that this is a graph to display unique relationships and focus in on "how" two concepts are related not just if they are or aren't. Using this fact, I was able to eliminate a good amount of diagrams that did not accomplish this. As well, there were many that made the relationships more ambiguous and harder to understand. The value in the Dendrogram is its simplicity. But to accommodate for the extra needed feature of labeling links with the defined relation, I had to make a modified Dendrogram with labeled links and a limited height of 2, as well remove any scales that define relevance of node height because we do not care about that information. A complete list of this new modified Dendrogram graph's pros and cons in the overall design are listed below.

First it is worth noting the challenges we face when implementing any visualization that uses the NLU Relations feature, as well the nature of derived visualizations. NLU Relations is actually quite bad at detecting some sentence structures, such as titles and attributes for target sentence subjects (e.g. "King James" will be interpreted as two different subjects: "King" and "James", which requires filtering to avoid showing on the graph, but the risk of losing a few meaningful relations is there). It is required that I not only filter out relations that are not useful to the understanding of the overall article, but also find some way to keep subject attributes connected in the final version. This is not an easy task at all. As well, the nature of graphs requires user interaction with a responsive design, and this means that it may not be immediately clear to an audience how these relations form. These challenges we will face are consistent across any visualization presented below. Thus it is essential that we: find the visualization that handles these risks effectively, enriches the user experience the most without, presents the data with the least amount of ambiguity, and follows basic principles of HCI to ensure longevity of the design for us and our posterity.

JOHN HERRICK
7/1/17

NLU RELATIONS

Dendrogram



Pros:

- Is consistent with the goal of showing how Watson processes data and makes sense of unstructured data: shows a progression of high level data (Article Topics node), expands to several topics (clustered data), and eventually to a terminal node (structured data).
- Pulls upon user heuristics on how two elements are related as well uses a popular and familiar tree shape that is learned in most early education systems (Taxonomy Trees)
- Consistent with designs already implemented in the News Disco Project (Concepts Filter)
- Offers a new mode of exploration once a user reaches a terminal node (Does a search of subject and object). This emulates a “rabbit hole” mentality of discovery that allows users to continuously find information on topics that interest them.
- Simplicity: There are no unnecessary shapes or elements that take away from the important data
- Minimizes the cognitive load
- Enhances a user’s understanding of the events of an article by sequential relations
- Enhances a user’s understanding of how article concepts are related by basic links
- Scales easily for large datasets
- Compatible with filters to only show the important data
- By limiting the tree height to 2, you can get a wholesome view of how one subject is related to several objects (one-to-one or one-to-many)
- Fully responsive so the user has the choice to pick only their topics of interest without distraction from other node relations

- Links are labeled so users can easily understand how concepts are related

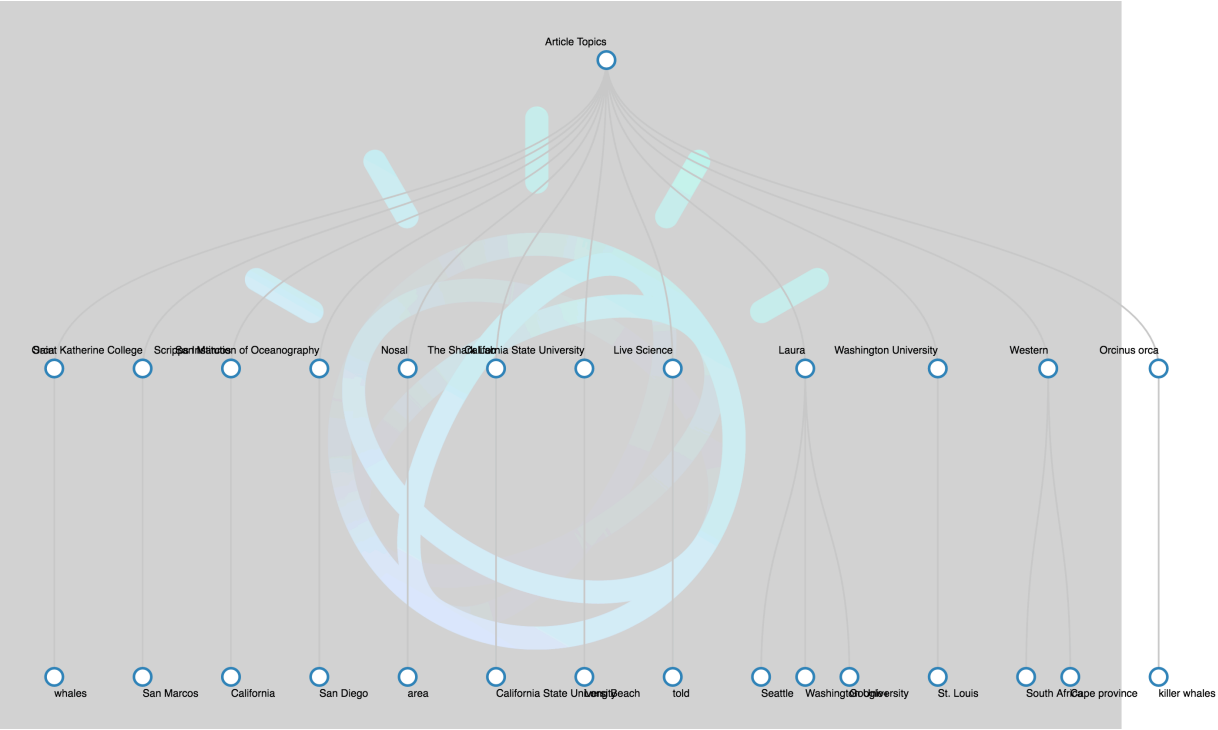
Cons:

- Screen Real Estate: Because these relations need to be seen, ample space on the screen needs to be made to allow users to comfortably view
- Nodes need to be adequately sized to be targeted by the user wand
- Screen Bezels may block the view of article topics if this eventually takes up multiple screens
- Due to filtering of relations that aren't deemed important, some information that is critical to the understanding of the article may be omitted.

Experimenting with other brands of knowledge graphs revealed many types that were not conducive to making a visualization easily understandable to clients. They only succeeded in making information more ambiguous. However, some stood out as possible candidates for the final deliverable (one being the Dendrogram). Some of the most promising candidates are below.

Unworthy Graphs

Pre-expanded Tree/Top Down Orientation

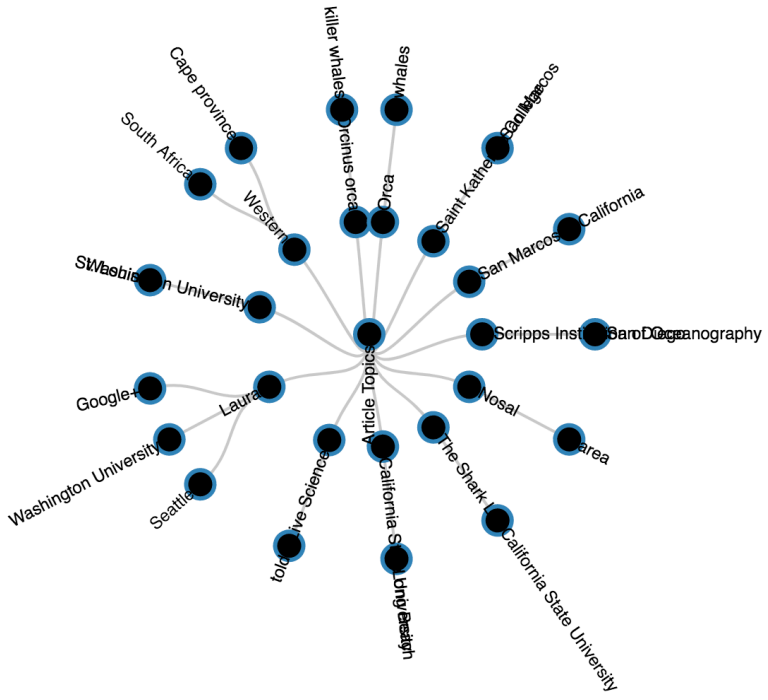


As stated before, the tree orientation was deemed best for portraying data. However, when deciding on the orientation of the tree structure, we were faced with the decision to follow a number of different orientations of hierarchy: top-to-bottom, bottom-up, or left-to-right. When looking at the Top-Down Orientation, it seems that this tree structure serves the ability to transcribe information well. The derivation of a top item, into several branches of knowledge and eventually to a terminal node that completes the relation, flows in the same logic as the left-to-right Dendrogram before. But the basic English language paradigm is to read left to right, and this flow of knowledge pulls upon that natural human response. It feels more intuitive and pulls on learned heuristics to read left-to-right, rather than top-to-bottom. The same can be said for a bottom-up orientation.

As well, we had to decide whether to allow the user to expand node relations or have all topics pre-expanded upon the initial loading of the screen. Just looking at a fully expanded graph is tiresome. Immediately there is an overload of information on the screen, which defeats the purpose of this knowledge tree: to reduce the amount of information the user must consume to understand the premise of the article. At this point, they might as well read the entire article.

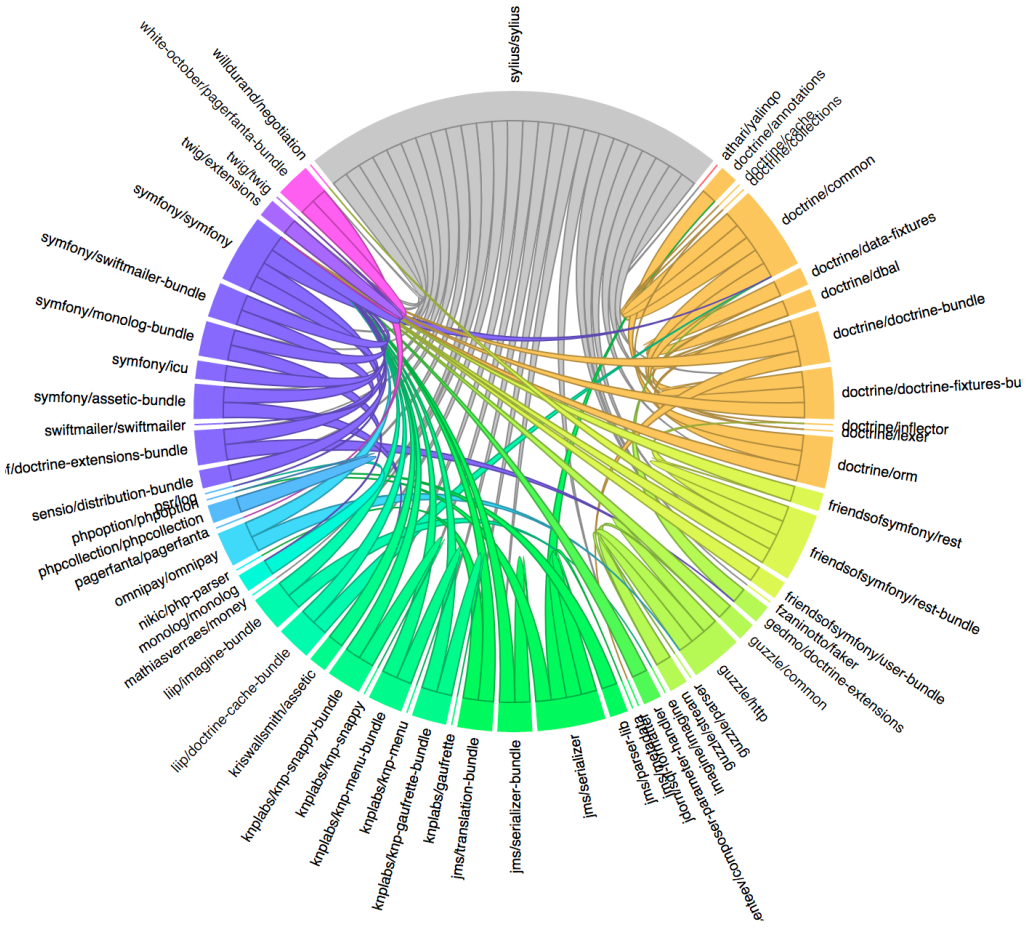
To reduce the cognitive load on the user, it would be wise to make these connections expandable/collapsible. Then users can pick the article topics that most interest them and expand upon them further. Or collapse and hide them to remove the distraction if they choose to delve into a separate topic on the tree. This builds on the idea of discovery and gaining insight through structured and targeted data, which is what News Discovery does well with so far.

Circular Orientation/Circular Graphs



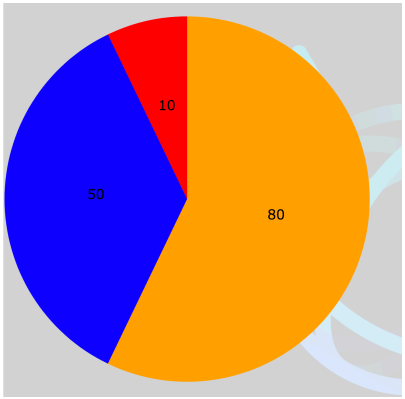
Building on the Tree structure, radially oriented graphs do well to portray a central topic and their derived topics as one navigates from the center of the circle outward. But like the flipped top-down tree, there is a severe amount of ambiguity with this design when showing to a casual user. Symmetry was one of the core strengths of the original tree diagram, but once a tier 2 topic has multiple children, the lines become less intuitive to follow. Moreover, even though this is a quick fix, the labels of lines on the wheel for both nodes and links are tilted to the angle of the link, which make it hard for a user to read immediately. If I rotate these labels so they are upright, we run into the problem of long labels overlapping with the links of other relations. This problem is unavoidable and not only destroys the readability of the design, but

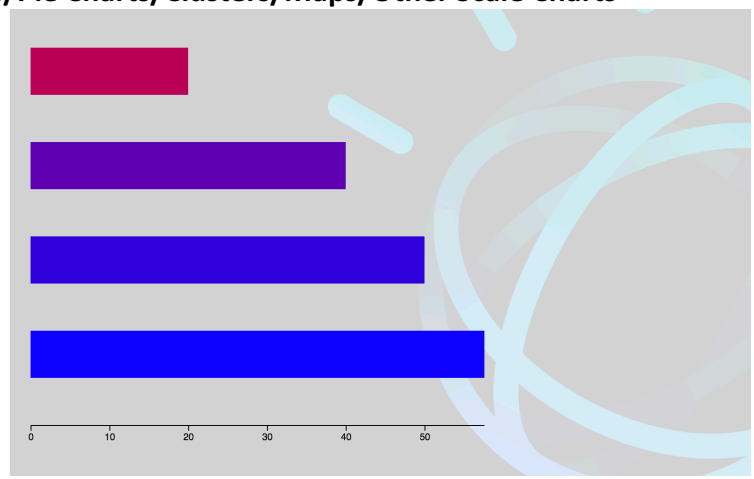
also fogs up the relations between topics, which fails our overall mission. Furthermore, the same problem of directional logic flow follows from previous designs: The derivation of a central item, into several branches of knowledge and eventually to a terminal node that completes the relation, flows in the same logic as the left-to-right Dendrogram before. But English paradigms teach humans to read naturally from left to right, and it is in our best interest to capitalize on that predisposition. Having the user move in multi directional patterns skews their attention to different links.

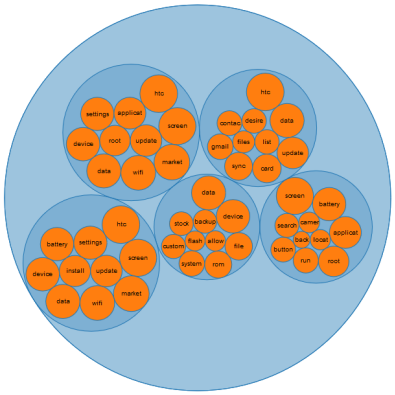


Variations of this circular design all suffer the same flaws mentioned before: ambiguity in traversing relations, tilted labels make visibility harder, and labels that if straightened horizontally would overlap other DOM elements. Also by eliminating the feature of links, you lose the ability to clarify the type of relation between elements. That is the real challenge of making this knowledge tree: The connections between topics are as important as the terminal node themselves. With this type of design, the user will not be able to understand how elements are related.

Histograms/Line Graphs/Pie Charts/Clusters/Maps/Other Scale Charts







The other well known graphs were considered, but to my knowledge, there is no real good way to form connections between node points. This is the nature of scaled graphs, which is only to compare statistics such as frequency, percentage, time progression, etc. The only real option at this point is some tree or tree-derived structure that shows relationships clearly though the linking of desired targets.

Keeping in mind all of these things, we need to look back at the nature of Natural Language Processing: to successfully process natural human linguistics and to produce something of value that leads to the understanding of this language data. One of the big goals of NLP is to effectively summarize large entries of human language. To date, no real service can do this, but the core of summarization is to take the main elements of the story and understand how they are related, eventually formulating a basic high level overview. This visualization and Watson Services NLU is the next step towards reaching that goal by providing these analysis capabilities. And to do so with the Article Detail View of the News Discovery Project will only enrich the Watson user experience even further.