TopicLens: A Visualization Design for Topic Model Exploration and Social Networks

Laura Devendorf, John O'Donovan, Brynjar Gretarsson, Tobias Höllerer

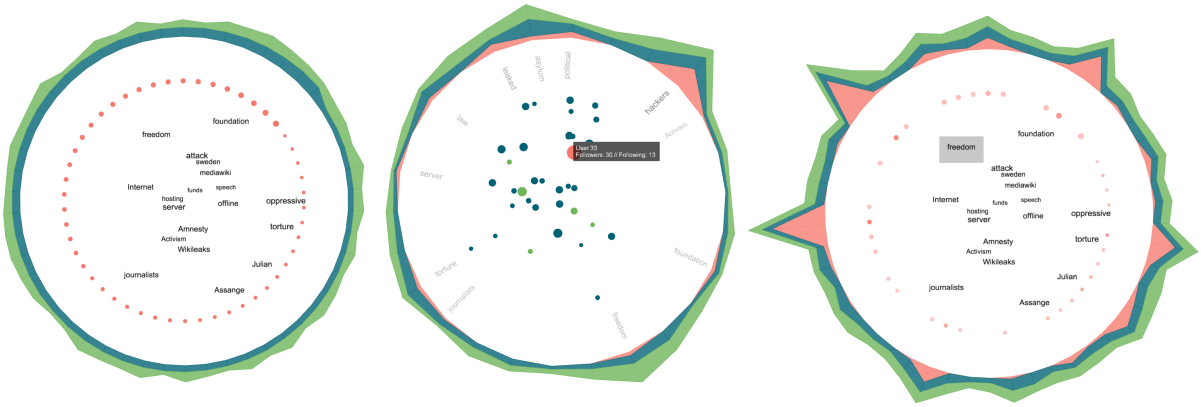

Fig. 1. Three TopicLens visualizations showing user-topic relations and social network information.

Abstract—Topic modeling is a powerful technique allowing users to quickly glean insights into large text corpora. The manner in which topic modeling results are communicated to end users presents an interesting visual design challenge. In this paper we examine the specific challenges associated with representing both statistical information about a document as well as its relation to the broader corpus, and the network of people associated with the document.
To address this challenge, we present and evaluate a novel interactive visual model combining river and graph based approaches to illustrate frequency and connectivity within a shared space. The model is applied to social network data from Twitter and results of an LDA algorithm run over "bag-of-tweets" representations of 34K users. To evaluate the model, we present various use cases to highlight the versatility of the approach and explain design decisions for the system.

Index Terms—Topic modeling, text visualization, graph visualization, network visualization

1 Introduction

Topic modeling is a statistical method for extracting relevant topics from a large corpus of text. Visualization of connections formed through topic modeling can enable the user to quickly identify trends and other insightful details from a large data set. Successful visualizations are especially effective at highlighting patterns within high dimensional data through intuitive, simplistic interfaces. These visualizations may also allow the user to navigate and dynamically filter information in order to extract specific and relevant data. Such cases are:

- To augment the users ability, beyond keyword based search and navigation, to discover relationships in texts.
- To highlight popular trends and conversations within social networks.

The focus of this paper is on the design and interaction techniques used to represent connections formed over large text datasets via automated text analysis algorithms such as topic modeling. The key elements in our visual representations include:

- *Items*: Abstract entity which can translate to either a text document, or a user within a social network. These are conceptually grouped because they are both represented by collections of terms. For example, in the Twitter data set, a user is represented as a collection of Tweets.
- *Topics*: Multinomial distributions over a set of terms, which can be associated with items.

While given models, such as word clouds and tree maps [28] can be ideal for visualizing frequency in topic-item relationships, we describe a model that also preserves and represents relationships at the metadata level. This allows users not only to see which topics arise, but also how they arose and under what conditions. The approach enables more informed reasoning about documents a user wishes to investigate, while highlighting trends over a number of different types of networks with respect to a particular investigation.

Microsoft's "Twahpic" approach to visualizing topics in conjunction with meta-data leverages a composite view that optimizes its visualization strategy for each different facet of the data. This strategy is effective for illustrating and highlighting multifaceted nature of the data, but is difficult to navigate due to the separation of each frame and the segregation of the data networks. In short, the interaction model helps a user form impressions of the data rather than supporting investigations into the data. Past research in [7] shows a benefit to using multiple approaches to visualizing the different facets of the data and in this paper, we will present a model that takes a hybrid approach rather than a segregated approach in order to facilitate navigation and interaction with the data. The key features of the proposed technique are as follows:

- Enables a user to filter topics in relation to the pre-existing networks in the data within a clean, simplistic interface.

- *Department of Computer Science, University of California, Santa Barbara. Contact: (ldevendorf, jod, brynjar, holl)@cs.ucsb.edu*

- To explore a dataset as a map, traversing and isolating regions of particular interest in order to extract relevant items.
- Caters to diverse topic modeling scenarios, including additional data such as social and information networks.
- Presents a choice of view modes and controls for navigation and dynamic filtering.

In the remaining sections, we will discuss the related research and provide a brief background of topic modeling before describing in detail the design decisions made when developing the TopicLens interface. The design decisions include those related to overall structure and the mapping of formal elements to relational information. Novel aspects of the interface are also discussed, particularly in relation to view inversion and an optional 3D mode. We will then present three applications of the system, one of which uses data that does not contain topic-based relations, thus highlighting a more generalized application of the design.

2 Related Work

Due to the proliferation of data available on the web, there is an increasing need for better techniques for exploration of large amounts of text data. This is commonly known as addressing an information overload problem [17]. Ongoing research has produced proactive, query-based solutions in the fields of search [10] and reactive or filter-based approaches in the field of recommendation [17, 6]. In the context of this work however, we are interested in approaches that employ visual methods to tailor an information space to a users individual needs. The novel approach presented in this paper employs a statistical method known as Latent Dirichlet Analysis (LDA) or "topic modeling" [4, 3] to discover useful linkages between documents upon which visualizations are built.

While there has been a significant amount of research on this topic from a variety of perspectives, from early approaches such as [23, 32, 16] to more recent work in [29, 30, 31, 18, 21], visual techniques for exploring large sets of documents have not yet been widely adopted.

For our discussion of related work, we now detail the LDA technique and its applications for visualization of large text collections. This is followed by a discussion of research on the visualization techniques we have chosen, and discussion our design decisions and alternative approaches.

2.1 Topic Modeling

LDA or "topic modeling" is a statistical technique introduced by Blei et al. [4] that computes focused probability distributions over the words in a set of documents. The algorithm functions by mapping documents onto a smaller number of "topics". In this sense, a topic consists of a multinomial distribution over words in a document set. For example, as $p(w|t)$, for $t \in 1 \ldots T$, where T is the number of topics [4, 14]. In many cases, topics are displayed as a list of the top n words with the highest probability in the set. Table 1 from [13] shows some example topics produced by an LDA algorithm. In this case, the words "theorem, lemma, proof, follow, constant..." seem to relate to the topic "Mathematical Theory". Recent research in [8, 22] has shown that although LDA topics can be misinterpreted, they are generally well understood by users. Techniques for the automatic labeling of topics have been presented in [19].

In TopicLens, topics are leveraged to form associations between items in a large corpus, and these associations are used to produce informative and highly flexible representations of the broader item space, using novel layout and interaction techniques. Before describing our approach to visualizing a topic space, we now present a discussion of existing approaches to visualization of large document sets.

2.2 Visualizing Text Corpora

Since humans find it inherently difficult to process large amounts of text quickly [12], a variety of tools and techniques have been researched to address this problem. Generally, these systems support some form of visual overview and then a filtering mechanism is applied to produce a detail view [27]. For the purpose of our analysis we can loosely classify these systems by their reliance on, or need for preexisting metadata.

Mathematical Theory	theorem lemma proof follow constant bound exist definition
Software Engineering	software process tool project development design system developer
Gene Expression	protein genes expression network motif interaction pathway genome
Politics and Society	political social policy economic china law government national
Business and IT	business firm services customer technology management market product
Fluid Dynamics	flow velocity wall fluid turbulence reynold pressure channel

Table 1. Examples of LDA topics learned on a corpus of research papers

Some successful commercial tools such as Matheo Analyser [1] perform filtering in a visual manner, while others use text-based approaches, as with Questel's q-pat system [24], for instance. In contrast with TopicLens, which forms and filters document networks based on topic relations, both systems leverage supporting metadata to refine result sets based on a user's search. When such metadata is available, it can play an important role in generating good visualizations. For example, InfoSky [2] uses heirarchially structured knowledge spaces in a 2D visualization, using the analogy of stars in a galaxy to represent the document corpus. FacetAtlas [7] also works with pre-existing metadata, in that it requires pre-computed similarity values between documents. As an example, the system used common terminology across a set of rich text medical documents to form a network that a user can visually analyze. Other systems elicit connection information between visual elements from the user on the fly. For instance, Van Ham's PhraseNets allows users to specify relations of interest from within a text and generate a representative graph.

In such cases where documents contain informative metadata, or provenance trails are available, this information can be leveraged to provide high-level semantic context, which can form the basis of an effective visualization. However, in many cases, rich metadata is not available and automated analysis techniques such as LDA can play an important role in "generating" data to form the basis of an effective visualization. Along this line, ScatterGather [9] is a system that employs clustering techniques to generate links between documents. TopicIslands [20] uses a similar technique but represents the document corpus in terms of wavelet transforms. The wavelet based method was successful in defining thematic 'channels' to visualize, but had less success with complex writing styles. Rushall's DEPICT system [26] is another example of a lower dimensional representation of a document space. However, in this case the system uses neural network techniques to compute the representation. IN-SPIRE [33] employs clustering analysis and depicts hot topics in large document collections using the mountain and galaxy metaphors.

To summarize, there have many approaches in the literature that deal with representation of large text collections, ranging from traditional static representations, e.g. [16], to more recent and highly interactive representations which use advanced methods to relate documents together, e.g. [13, 5]. They can rely on pre existing meta data, or can compute relations on the fly. In this paper we present a novel interactive design and layout for exploring topic based and social network relations in large document sets. Before presenting the prototype system in detail, the following section provides a brief account of the design choices for using a combination of river and graph-like visual representations in the system.

2.3 The Need for a Hybrid Model

As shown in Figure 2, we are supporting exploration of multi-faceted data in a variety of ways. Specifically, examples are demonstrated on three different network types: social network data with bi-directional connections from Facebook; social network data with unidirectional edges (followers and followees) from Twitter, augmented with topic relations, and a topic modeled network of news articles from the New York Times. Across all examples, the goal is to use simple interaction and novel layouts to facilitate user comprehension of complex

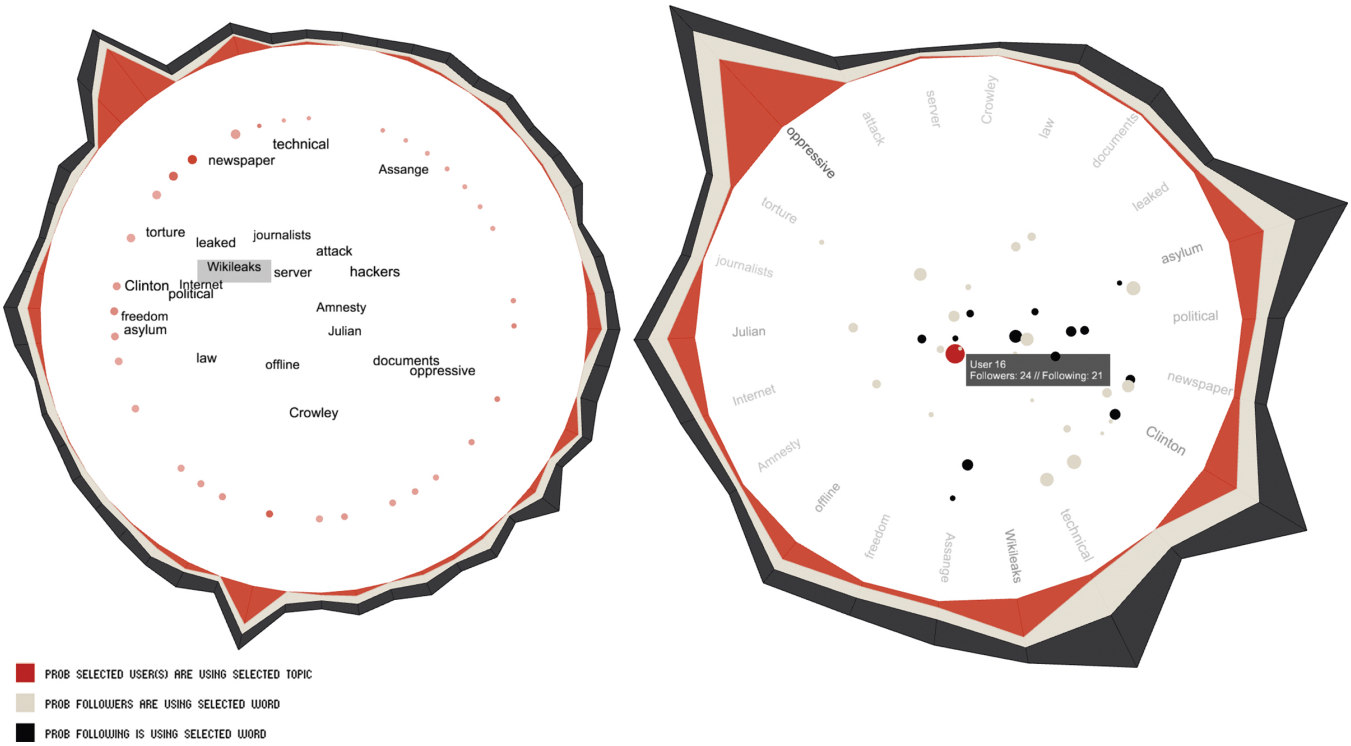

Fig. 2. Two detail views of the TopicLens visualization, each showing connections between items and related topics. In this case the data is from the Twitter social network, so our generic "items" represent Twitter users. Frequency measures are shown on the outer river-like component. The two views are of the same data, with items and topics inverted.

data, particularly to communicate the "credibility' factor of peers in a network with respect to particular topics of interest. This complexity would be inherently difficult to communicate in a single visualization technique such as a river or graph visualization. Accordingly, we have opted for a hybrid approach which uses a graph-like mechanism similar to TopicNets [13] for highlighting relations between document and topic nodes, and a river-like view similar to ThemeRiver [15] overlaid to communicate frequency or "credibility" of different sets of peers within the context of a topic selection. This approach has been successful in application such as Freire's ManyNets [11].

3 Visualization Design

The most prominent feature of the visualization, shown in Figure 2, is its use of the wheel to structure information. The decision to use the wheel was largely chosen to accommodate variability in the size of the datasets. The wheel dynamically expands to fit the data and contracts upon filtering. This keeps all information present within the visualization space, regardless of how much there is to display. Zooming allows the user to hone in on particular regions of interest.

The visualization is designed to fit within a rectangular window with width larger than height. The exact dimensions can vary and in our examples, we found it most effective to use a full screen view, especially when dealing with large sets. The left side of the screen, shown in Figure 4, contains the controls and legends and the wheel rotates on an axis in the center of the screen. A static camera is also positioned at the center of the screen. This allows the user to zoom in and out of the wheel and to scale the view to fit the entire space. The river is positioned along the outer edge of the wheel and protrudes in different directions depending on selection. At times, the areas that expand the most are cut off by the controls or the bottom or top of the page. The ability to spin the wheel allows the user to fill empty space by fitting protruding regions into empty spaces on the right side of the screen.

3.1 Organization

After studying the data and the relationships formed, we found the type of relationships present within the topic modeled system could be classified into three types: primary, secondary and ternary. By splitting the wheel into three concentric regions, we were able to map each type relationship to its own position on the wheel. Dividing the relationships in such a way provides clarity and a visual hierarchy which serves to illustrate the type of connection formed. This supports the user in detecting patterns and allows them to make wise decisions when browsing and extracting information. The following paragraphs provide a detailed explanation of the relationship types and the regions they map to.

3.1.1 Primary Relations: Center

Primary relations are formed between items. Prior to topic modeling, each item has its own meta-data, whether it be a list of friends and followers or a category specification. A primary relation connects items based on this meta-data. With this in mind, two items in the same category form a primary relation. These are mapped to the center as it isolates the given node type from the rest of the data, allowing the user to visualize connections in a shared space. Figure 2 shows primary relations through coloring in the view on the right. In the view on the left, topics are featured in the center. Since primary relations don't exist within topics, no explicit color mapping is represented. A discussion about the mapping of color to primary relations will take place in section 3.3.

3.1.2 Secondary Relations: Center & Inner Ring

Secondary relationships occur as a result of the topic modeling and define the relationships between topic and item nodes. Each of these relationships occurs with a given probability as defined by the LDA algorithm. These relationships as well as their respective probabilities are represented by interactions between the center and inner ring. While the nodes in the center are not bound to an axis, the nodes in

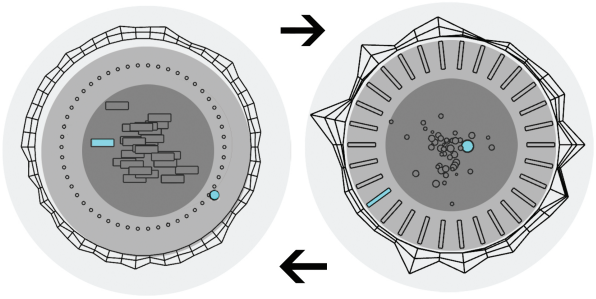

Fig. 3. This a wireframe depicting the inverted views of the same data. The rectangles represent topic nodes, and circle, item node. The colored nodes show how each element maps to each view. This also shows how the skeleton is preserved though the inversion. The darkest circular represents the center, mid-value circle represents the inner-ring and lightest circle represents the outer ring / river.

the inner ring have a defined equally spaced position. This is primarily due to the fact that the inner ring also functions as the axis points for the river visualization but also reinforces simplicity by defining only one type of data to be related spatially. On the left side of Figure 2, highlighting Wikileaks changed the opacity of the nodes on the inner ring in order to indicate how related each item is to this topic. On the right, highlighting User 16 changed the opacity of the topics in the inner ring, similarly showing the strength of the connection.

3.1.3 Ternary Relations: Outer Ring / River

Ternary relationships are formed between the topic modeled results and the meta-information of the items related to those results. Using the river visualization to graph these relationships allows us to see an overall frequency of the node in addition to the meta-information frequencies within the same space. For instance, when modeling a large text corpus with each item belonging to a given subject, the river clearly shows the overall frequency as a component of the frequencies for each subject. Depending on the data and filtering, the river model can be customized to show any particular facet of the meta-information. Figure 2 is showing average probabilities over each facet of item meta-data in relation to the selected item. The colors in the river match the colors of the meta-data in the center, reinforcing this relationship.

3.2 View Inversion

Visually navigating the results of topic modeling can be approached in two ways, by topic or item. The choice between the two depends on the usage scenario. The structure defined in the previous section and illustrated in Figure 3 supports both approaches, for instance:

- Modeling topics in the center, items in the inner ring and topic frequencies with respect to item in the river is optimal for browsing by item, filtering topics based on items and visualizing items for a particular subset of topics. See the left side of Figure 2.

- Modeling items in the center, topics in the inner ring and item frequencies with respect to topic in the river is optimal for browsing by topic, filtering items based on topics and visualizing topics for a particular subset of items. See the right side of Figure 2.

While each view is presenting the same data and unique qualities are elucidated depending on organization. Rather than force the user to think in terms of a single view mode, we allow the user to invert, or switch from one organizational mode to the other. We call this transition ”inversion” because you are essentially turning the visualization inside out, grabbing information from the center and putting it to the edge and vice versa.

One of the factors we considered when developing this inversion technique is whether or not the user will understand how the inverted views relate to each other. Using the same framework to represent different facets of the data could potentially confuse the user. To avoid this confusion, we offer an animation that clearly depicts each data point as it transitions to the new mode. As a result of findings in the Robertson study [25], we use a one second animated transition to visually guide the user through the inversion. Confusion can also be avoided by keeping the interactive actions and formal design qualities consistent from one view to the other. As a result, each mode is distinctive in look, further distinguishing between the two. Full descriptions of visual mappings and interaction techniques are presented in the following sections.

3.3 Formal Elements

As mentioned above, consistency is paramount when translating information into multiple view modes. In order to maintain simplicity we map objects and relationships to specific formal elements.

Shape corresponds directly to node type: topic or item. Items are mapped to circles and topics are mapped to rectangles with the text label of the topic in the center. Depending on application, the topic rectangle can be present at all times or only on mouse-over while the text inside the rectangle is viewable at all times. In some cases, the rectangle's color provides insight into the documents that comprise this topic. In such cases, the rectangle is shown more often than on mouse-over.

Color is used to communicate item meta-information. For instance, if the meta-information contains the item's category, each category type would map to a unique color. This mapping was chosen partly because it enables a quick visual grouping of items and extends to any number of categorizations. Another reason was the ability for color to create a visual connection between primary and ternary connections. Both types of relationships concern item meta-data and the color of the band in the river directly correlates to the color of the node in the center of the graph. This offers the user two levels of understanding by illustrating how the meta information is connected to the item as well as the topic.

Given the explanation above, color only is used in relation to item nodes but there are specific cases when it is beneficial to use color on topic nodes as well. Another way to represent the information communicated in the river is to average all of the colors of each of the items contributing to the topic and displaying it in the topic rectangle. This provides another way to view a topics classification and visually groups topics with similar meta-data distributions. When more explicit information is needed, the river can be used to reference the exact breakdown between meta-information.

In order to improve the aesthetic experience of the visualization, we defined a custom palette to use when selecting colors. This was accomplished by defining an image from which we randomly grab a pixel. The image is of ”gobelin,” a quilt by Bauhaus artist Gunta Stölzl created in 1926-7. The Bauhaus school was known for their experiments in color theory and we felt that this subtle detail increased the novelty of the interface by picking new interesting and attractive color palettes of any size.

Opacity is used to illustrate secondary relationships. These relationships occur with a probability specified by the LDA algorithm. Opacity is an effective means of illustrating these connections and probabilities by using the opacity range to indicate connection strength. Darker nodes have strong relations, lighter have weaker. If a node is unrelated, it is removed from the space. Secondary relations are highlighted upon interaction as you must specify a single node in order to view its connections. If multiple nodes are selected, then the opacity value is determined by the average probability from all nodes in the selected set.

Position and order are used together to illustrate patterns from the data. Patterns are created by using the ordering of the inner nodes to position the values in the center. Each center node begins in the center of the circle and is pulled towards all of its related nodes in the center ring. The strength of attraction depends on the probability of the connection. The result is a spatial grouping of nodes that share similar relationships. This spatial grouping is much more interesting when the

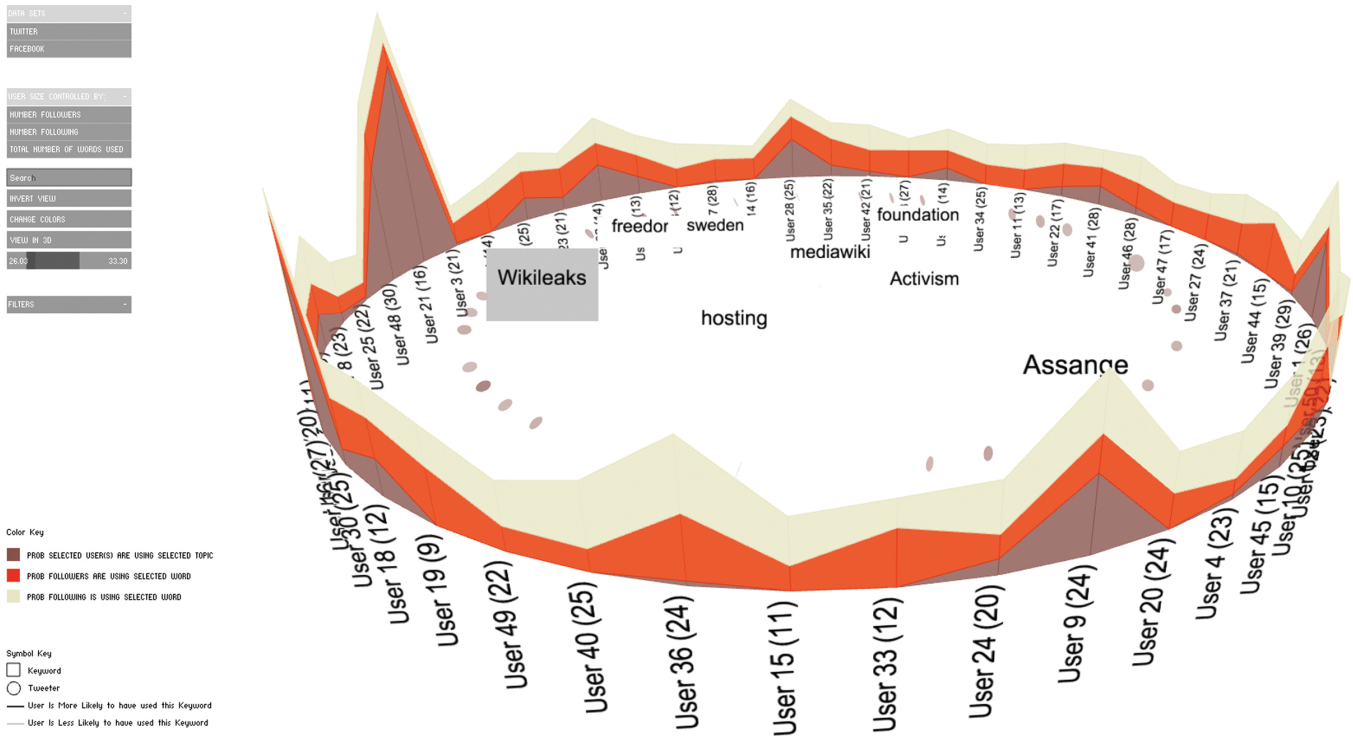

Fig. 4. Full view of the TopicLens interface showing a 3D visualization of the Twitter data, with topics in the center and users on the inner ring.

ordering of the nodes in the inner ring is significant. Naturally, some patterns are more interesting than others and modifying orderings has the potential to find the more interesting patterns in the data. For this reason, the ordering in the inner nodes can be assigned in a number of different ways. In our examples below, we give the user the option to select the ordering criteria. Future implementations will allow the user to also manually position nodes in the inner ring to fit special cases. Since the river is organized with the inner nodes as axis points, reordering the inner ring reorders both the center nodes and the river which increases potential for pattern recognition.

Another implementation of ordering exists in the ordering of topics and items to circular regions of the wheel. We discuss the implications of this ordering in the inverted view section but it is of note that it is consistent in its aid for pattern recognition and holds to the mappings defined here.

Size is used to illustrate measures of numerical magnitude such as frequency or number of relations. Similar to position and ordering, some mappings of attributes to size can be more informative than others. For this reason, we allow the user to select the node attribute that determines node size rather.

4 Navigation Tools

Navigating the visualization space is highly interactive and the system is designed to empower the user to modify the space while maintaining a simple and clean interface. A shared trait of all of the interaction methods is their ability to dynamic modify the visualization space with smooth transitions. Any action that results in the repositioning of nodes does so with a transition time of one second.

4.1 Details on Mouse-Over

Hovering the mouse over any node will reveal information about the entire node's network of information. Opacity changes illustrate relevance to the selected node and the river dynamically adjusts to illustrate ternary relationships specific to the selected user. Slight variations for how and what information is revealed exist between topic and item nodes and depend on the underlying data and use-case scenario. Revealing information on mouse-over allows the user to visualize what would happen if they were to filter based on this node to support fast browsing by limiting the number of mouse clicks it takes to see a change to the system.

In some cases, lines can drawn on interaction to explicitly define secondary relations, connecting the item node to the topic nodes. Careful consideration must be taken into account when adding lines to the visualization as they tend to obscure information in the center. In our examples, lines are only used to express a type of relationship that isn't already being defined. In many cases, adjusting the opacity is suitable enough to convey information, Section 6.2 will discuss a case when lines are used in order to convey information.

4.2 Selection and Multi-Selection

Selecting a node will filter the visualization space, showing only nodes associated with the selected node. If multiple nodes are selected the resulting networks can either aggregate, showing all nodes that share an association with at least one selected node in the space, or exclude, showing only nodes that have an association with all selected nodes. Depending on the usage scenario, one single mode can be pre-defined or the system could contain a control that allows the user to determine which mode it should operate in at a given filtering step. Adding this functionality allows for advanced filtering, yet, may intimidate and or confuse novice users. For our examples, this control is pre-determined in order to keep the interface clean and manageable.

4.3 2D to 3D Space

The visualization can be represented in both two and three-dimensional space. Figure 4 shows a three dimensional view from TopicLens. This is accomplished by developing the entire model in three dimensions and simulating two-dimensionality with a stationary camera that points to the center of the wheel. In the 2D mode, the only functionality the user has over the camera is zooming and rotation. All of the text in the visualization is corrected to be upright so rotating the camera has the effect of spinning the wheel. Because of the upright realignment of the elements in the center, rotating the camera can also

function to reveal text that is being obscured by another element.

When a user selects 3D mode, the camera is no longer held constant and the user is free to manipulate its position in any direction. The main distinction between the two modes is the river visualization since it extends towards the camera. When the camera is held stationary, it foreshortens the depiction of objects in the z-plane creating a logarithmic-like scaling as the user looks down into a 3D map. When the camera is released, the user can rotate the plane and see the linear representation of the river visualization and gain more accurate insights into the specific datasets. The primary goal of the 3D visualization is to support users who prefer interacting within a 3D space and to increase the novelty of the interface.

4.4 Filter History

Navigating or detecting patterns within the topic modeled networks often requires the user to filter based on many different nodes. Filtering in this way often makes it difficult to locate your place within the dataset or remember what filters have already been added. To resolve this, present the user with an ordered list of references to filters. Upon selecting a node and filtering the system, a button is added to the list. If the user wants to remove this filter, they can do so by clicking the selected node again or clicking the button in the filter history. In both cases, the node is removed from the history and the visualization is restored to its previous state.

4.5 Control Panel

A control panel on the left side of the screen offers the user a number of tools with which to manipulate the data. While some tools are constant across all implementations, others cater to specific use cases.

All visualization models contain controls for text search and threshold adjustment. Text search is implemented in order to allows the user to select nodes based on their label. Search offers quick navigation to users who are targeting a specific piece of information in the dataset. Thresholds are controlled by a range slider. With the range slider, the user can set the minimum and maximum amount of items that must be associated with a topic in order for it to appear in the space. While they adjust the range, the data is dynamically changed to show the effects of the range modification. This feature also makes the visualization of large datasets more manageable since it allows the user remove outliers and hone in on the most meaningful sections of the data. At times, the outliers of a dataset are equally as interesting as the common topics. The range slider also provides users with the ability to filter into these regions as well.

Optional controls allow users to map specific data to formal elements. We also allow the user to explicitly define particular mappings. As we mentioned in Section 3.3, size of an element can depend on any type of frequency such as number of associations or number of size of a particular piece of meta-data like number of friends.

5 Implementation

This visualization evolved through a number of design iterations. Using Processing to program the design and interaction allowed us to easily make these design changes and instantly see the results. The program also made it simple to program animations and transitions between states. A number of libraries exist that extend the of Processing. The PeasyCam library provided the basic camera functionality, the ControlP5 library was used to implement the controls such as text boxes, range sliders and list boxes and an OpenGL library was used to add custom functionality into the system such as smoothing and alpha blending.

The program creates node and edge objects by parsing and XML on load. During the execution of the program, nodes and edge objects are referenced in order to create dynamic links. Links are the elements that are drawn to the canvas and much of the code is devoted to maintaining those links and dynamically updating their values to indicate relationships. The smooth transitions were created using an integrator class that allow the user to specify characteristics such as mass, position, damping and attraction. When a link targets a given position, the integrator dynamically updates its position depending on its physical characteristics.

6 Use Cases and Discussion

In order to show the various applications of this visual model, we present three use cases that explore varying datasets. Each use-case will discuss the design decision made to specifically cater to the data as well a usage scenario to illustrate its potential for a variety of applications.

6.1 Visualizing Credibility In Twitter Networks

Preserving social network relations in topic modeled systems allows us to glean insights into the networks and salient topics therein. This example is catered specifically as an attempt to visualize credibility in Twitter networks. Our definition of "credibility" relates to the probability which which a user is connected with a particular topic, based on LDA analysis over a bag of words representation of all of that user's tweets. In analyzing credibility, we also examine that user's followers and followees and their respective associations with the given topic.

In this visualization, which is represented earlier in Figure 2, each topic node contains a label that represents the list of words in a mined topic. Primary relationships are formed between a user and their followers and followees. Secondary relationships occur between the users and extracted keywords and the ternary represent probabilities over the meta data.

As noted in Section 4.1, ternary relations are mapped to the river and the nodes in the inner ring form the axis points. The river displays specific information depending on the organization of the nodes within the space. This approach affords the user an opportunity to uncover potentially interesting relations in the following 6 view configurations.

With topic nodes in the center, and user nodes on the outer ring:

- *Upon selection of an individual user* the river view shows that user, their friends and their followers' probabilistic association with each topic on the outer ring.
- *When no user is selected* the river shows the average probability for each topic across all users.
- *When a topic is selected* the river shows each user's association with that topic.

With user/item nodes in the center, and topic nodes on the outer ring:

- *When a topic is selected* the user's friends and follower's opacity is varied to represent association with that topic.
- *When a user is selected* their association with each topic on the outer ring is shown in the river view.
- *When no user is selected* the probability of each topic in the global space is shown on the outer ring

For this scenario, the river represents three probabilities for each node, the average probability of the user using the term, the average probability of the user's followers using the term and the average probability that the people following this user are using the topic.

Since topics are represented along the inner ring, this information is available for every topic. Each of the probabilities is represented on the river uses color matching to indicate the group or single user it applies to. To further explain what the river is visualizing, a key on the bottom left of the interface dynamically changes text in order to explain the model. In this case it of particular importance as the river maps different values through the life of the visualization.

When a user is highlighted in the space interactions take place at each of the three levels. In the center the primary relationships are presented through colors. All users who don't belong to this users network are removed and the remaining users are color coordinated to indicate whether they are a follower of the selected user, or someone the selected user is following. Spatially, each user is attracted to the

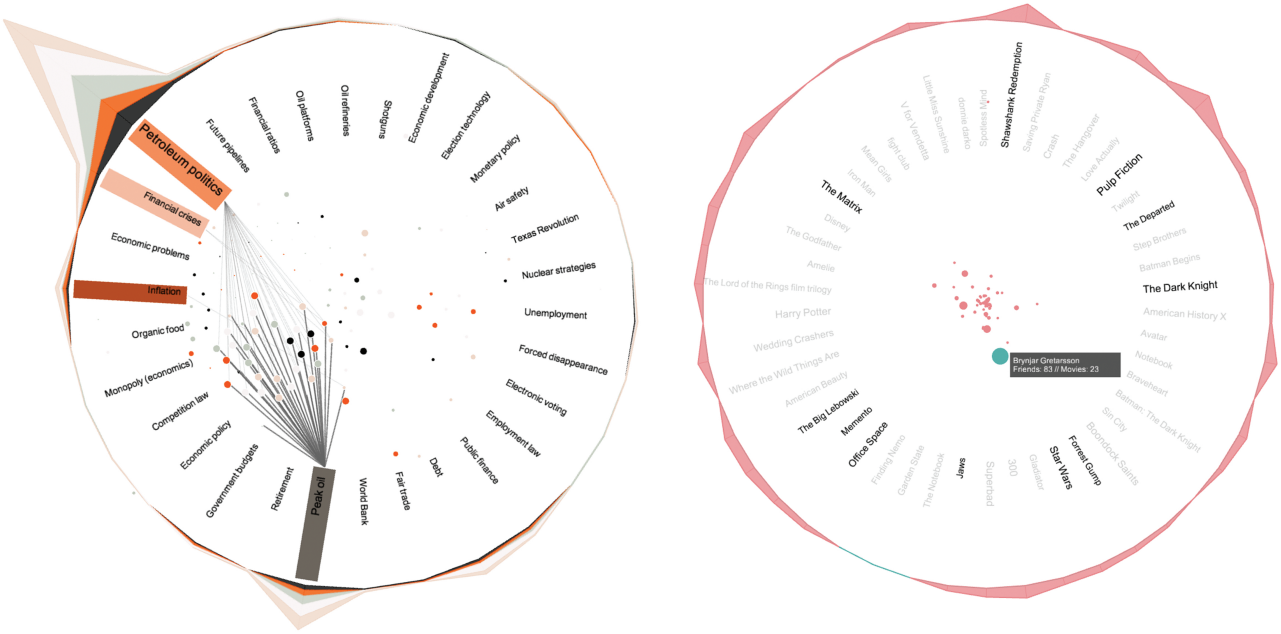

Fig. 5. TopicLens view details for varying data sets. Left: TopicLens view of news articles from the New York Times showing topics on the inner ring and articles/items in the center. The view shows selection of an individual topic (peak oil) and edges linking to related articles. Right: TopicLens view of a Facebook social network showing an individual's friend network and associated item preferences.

topic nodes in the inner ring by the positioning algorithm mentioned above. Topics related to the selected node vary by opacity in order to indicate the strength of connection.

The probability mapping were specifically designed to investigate credibility or trust. The top left of Figure 2 shows a network with two people selected. All of the nodes in the set represent both of the selected nodes' networks. On the outer river, you can see the probability distributes for this network over each topic in the network. From the river you can conclude that these two users are using the word "Crowley" quite a bit, however their friends and followers aren't. For this reason, they may not be a trusted source for this keyword since their followers don't appear to be interested in similar topics. On the other side of the visualization is the term "asylum" which is being used largely by the network and not so much by these users.

Drawing conclusions at this level is skeptical at best but better information can be introduced by selecting the right network. For instance, if you know the terms "Assange", "Julian" and "Wikileaks" are all terms related to Wikileaks, then you could select those terms from the visualization and view the results over the given network of users associated with those terms. By investigating the probability of these three words occurring together across the social network you may be able to visualize trends about who is followed, by whom and for what reasons.

6.2 Topic Modeling on the New York Times

In the example shown on the left of Figure 5 topic modeling was performed on a set of New York Times articles and is used for investigation and discovery of related articles that may not have been discovered through traditional search models. Each document node represents an article and topic nodes represent the topics extracted from those articles. Each article contains information about the section of the paper which it belonged to, such as opinion, world or national news.

Two unique design features were included in this visualization to improve the functionality in relation to the underlying data. The first is colored rectangles on topics. These colors are used to reinforce ternary connections though the use of color averaging. The color of the rectangle is determined by the category of each of the articles associated with it. Should a color tend heavily towards a single categories color, on could deduce that the topic tends to appear most frequently within that category. The actual distribution of the categories is explicitly represented in the river.

The second unique feature is the use of lines. When hovering over a topic, darkened lines extend form the topic itself to all related documents. Lighter lines then extend from each of those related documents to all of the other topics they are related to. This conveys information to the user about other topics related to their selection. The user is able to specifically locate the documents that contributed to this relationship by following the lines or selecting multiple topics and browsing the filtered document space.

The lines are particularly useful for illustrating how two topics are related to each other and upon what criteria. This is helpful when browsing for articles associated with a given theme. Let's say a researcher is looking for references on "peak oil." Searching for and selecting "peak oil" from the space would show the researcher other related topics as well as articles specifically contain the relation. If one of the related articles contains a topic that is also of interest to him or presents a particularly interesting comparison, he or she can easily isolate and obtain information about the articles containing both topics by filtering the space and hovering over the document node, revealing information about the article such as title, date and author.

TopicLens could also be used visualize trends associated with a subset of articles. Say a user read a few articles in the Times Opinion section and they would like to find other articles about similar and related subjects. This could be accomplished by typing each article name into the search field. This would select each of the entered articles in the space and illustrate the topics associated with them. In order to remove outliers, they would adjust the slider to specify the amount of documents that need to be associate a topic in order for it to appear in the space. After this filtering, they are presented with a number of related topics, the most popular being the largest and darkest. By selecting that topic, the space is reorganized to show all the the arti-

cles related to that topic. The user could visually browse these articles and quickly identify which one appeared in the opinion section based on color. Hovering the mouse over a document node would reveal its specific information.

6.3 Visualizing Systems without Topic Modeling: Facebook Networks

In this example, shown in the right of Figure 5, we use the existing framework to visualize data that is not topic modeled in order to show how the interface also operates on similarly structured datasets. Reinterpreting the definitions of item and topic allows us to use the existing visual model for this dataset. In this example, a single Facebook user takes the place of an item and a movie takes the place of a topic. Since movies can be related to any number of Facebook users and Facebook users can be associated with any number of movies, this dataset can function similarly to the topic model examples. Each item-topic, or rather user-movie, combination is assigned the probability of 1 since the user has specified explicitly that they like the given movie.

This visualization provides is able to provide exploratory view of the most popular movies within a Facebook friend network as well as the least popular movies. It can also isolate pockets of users that are fans of these most or least popular moves.

7 Conclusion

In summary, this paper has presented a novel design and interaction mechanism for visual analysis of topic-based relations in large document collections. The design is a hybrid which combines river and graph-like data representations. Details of our design choices and methodology have been outlined over three example applications including social network data from Twitter augmented with a topic modeling over users' tweets, social network data from Facebook, including item preferences, and a topic modeled set of New York Times news articles. In each example case, we have discussed ways in which the approach facilitates discovery of relevant information which may go undiscovered in traditional analysis tools.

References

[1] M. Analyzer. Matheo analyzer, database analysis and information mapping. http://www.matheo-analyzer.com/, 2010.

[2] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1:166–181, December 2002.

[3] D. Blei and J. Lafferty. Correlated topic models. In *Advances in NIPS 18*, pages 147–154. MIT Press, Cambridge, MA, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[6] R. Burke. Knowledge-based recommender systems. In *Encyclopedia of Library and Information Systems*, volume 69, 2000.

[7] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16:1172–1181, November 2010.

[8] J. Chang. Reading Tea Leaves: How Humans Interpret Topic Models. 2009.

[9] D. R. Cutting, D. R. Karger, and J. O. Pedersen. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 126–134, New York, NY, USA, 1993. ACM.

[10] B. D. Davison, T. Suel, N. Craswell, and B. Liu, editors. *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*. ACM, 2010.

[11] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. Manynets: an interface for multiple network analysis and visualization. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 213–222, New York, NY, USA, 2010. ACM.

[12] B. Fry. *Visualizing data - exploring and explaining data with the processing environment*. O'Reilly, 2008.

[13] B. Gretarsson, J. O'Donovan, A. Asuncion, D. Newman, S. Bostandjiev, T. Hllerer, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on the Web: Special Issue on Intelligent Text Visualization*, 16:1172–1181, November 2011.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.

[15] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[16] M. A. Hearst. Tilebars: visualization of term distribution information in full text information access. In *CHI '95: Proc. of the SIGCHI Conf.*, pages 59–66, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[17] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22:5–53, January 2004.

[18] S. Koch, H. Bosch, and T. Ertl. Towards content-oriented patent document processing. *IEEE Symposium on Visual Analytics, Science and Technology*, pages 203–210, 2009.

[19] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *KDD*, 2007.

[20] N. E. Miller, P. Chung Wong, M. Brewster, and H. Foote. Topic islands - a wavelet-based text visualization system. In *Proceedings of the conference on Visualization '98*, VIS '98, pages 189–196, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[21] D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2-3):169–175, July 2010.

[22] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224, New York, NY, USA, 2010. ACM.

[23] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: the vibe system. *Inf. Process. Manage.*, 29(1):69–81, 1993.

[24] Questel. Qpat, intellectual property patent and trademark searching. http://www.qpat.com/, 2010.

[25] G. Robertson, S. K. Card, and J. D. Mackinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4):57–71, Apr. 1993.

[26] D. A. Rushall and M. R. Ilgen. Depict: Documents evaluated as pictures. visualizing information using context vectors and self-organizing maps. In *Proceedings of the 1996 IEEE Symposium on Information Visualization (INFOVIS '96)*, INFOVIS '96, pages 100–, Washington, DC, USA, 1996. IEEE Computer Society.

[27] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations, 1996.

[28] B. Shneiderman and M. Wattenberg. Ordered treemap layouts. *Information Visualization, IEEE Symposium on*, 0:73, 2001.

[29] W. Spangler, J. Kreulen, and J. Lessler. Mindmap: Utilizing multiple taxonomies and visualization to understand a document collection. *Hawaii International Conference on System Sciences*, 4:102, 2002.

[30] J. T. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[31] Wanner. Towards content-oriented patent document processing. *World Patent Information*, 30(1):21–33, 2008.

[32] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In N. D. Gershon and S. Eick, editors, *IEEE Information Visualization '95*, pages 51–58. IEEE Computer Soc. Press, 30–31 Oct. 1995.

[33] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. In-spire infovis 2004 contest entry. *Information Visualization, IEEE Symposium on*, 0:r2, 2004.