Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis

Jason Chuang, Daniel Ramage, Christopher D. Manning, Jeffrey Heer
Computer Science Department
Stanford University
{jcchuang, dramage, manning, jheer}@cs.stanford.edu

ABSTRACT

Statistical topic models can help analysts discover patterns in large text corpora by identifying recurring sets of words and enabling exploration by topical concepts. However, understanding and validating the output of these models can itself be a challenging analysis task. In this paper, we offer two design considerations — *interpretation* and *trust* — for designing visualizations based on data-driven models. Interpretation refers to the facility with which an analyst makes inferences about the data through the lens of a model abstraction. Trust refers to the actual and perceived accuracy of an analyst's inferences. These considerations derive from our experiences developing the Stanford Dissertation Browser, a tool for exploring over 9,000 Ph.D. theses by topical similarity, and a subsequent review of existing literature. We contribute a novel similarity measure for text collections based on a notion of "word-borrowing" that arose from an iterative design process. Based on our experiences and a literature review, we distill a set of design recommendations and describe how they promote interpretable and trustworthy visual analysis tools.

Author Keywords

Visual analysis, text, statistical models, design guidelines

ACM Classification Keywords

H.5.2 Information Interfaces: User Interfaces

INTRODUCTION

To make sense of complex data, analysts often employ *models*: abstractions (often statistical) that represent data in terms of entities and relationships relevant to a domain of inquiry. Subsequent visual representations may depict a model, source data, or both. A central goal of visual analytics research is to augment human cognition by devising new methods of coupling data modeling and interactive visualization [47].

By suppressing noise and revealing structure, model-driven visualizations can greatly increase the scale of an analysis. However, unsuitable or unfamiliar abstractions may impede interpretation. Ideally, model abstractions should correspond to analysts' mental models of a domain to aid reasoning. Reliable discoveries arise from analysts' ability to scrutinize both data and model, and to verify that a visualization shows real phenomena rooted in appropriate model assumptions.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

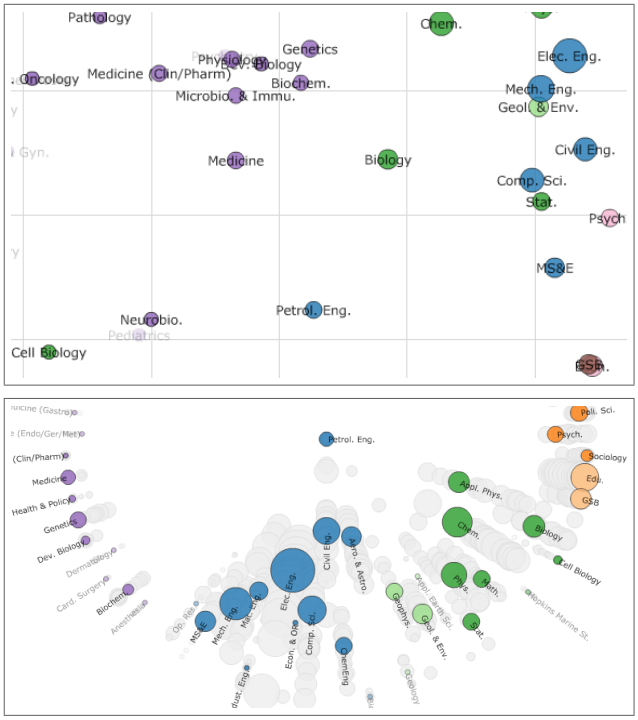

Figure 1. The curious case of Petroleum Engineering. The top visualization shows a 2D projection of pairwise topical distances between academic departments. In 2005, Petroleum Engineering appears similar to Neurobiology, Medicine, and Biology. Was there a collaboration among those departments? The bottom visualization shows the undistorted distances from Petroleum Engineering to other departments by radial distance. The connection to biology disappears: it was an artifact of dimensionality reduction. The visual encoding of spatial distance in the first view is *interpretable*, but on its own is not *trustworthy*.

Consider the visualizations in Figure 1, which depict "topical similarity" between university departments in terms of their published Ph.D. theses. We fit a statistical topic model (latent Dirichlet allocation [4]) to the text and compute similarity using the angle (cosine) between departments' topic vectors. In the top view, we project departments to 2D via principal component analysis of the similarity matrix. Using this visualization we note an unexpected trend: over the years Petroleum Engineering pulls away from other engineering departments, and in 2005 it is situated between Neurobiology, Medicine, and Biology. This observation comes easily, as the visualization is readily *interpretable*: pixel distance on the screen ostensibly represents topical similarity. However, the display is the result of a chain of transformations: topic modeling, similarity measures, and dimensionality reduction. Can an analyst *trust* the observed pattern?

The bottom view instead shows undistorted distances from Petroleum Engineering to the other departments. The relationship with Biology evaporates: it is an artifact of dimensionality reduction. Stripping a layer of modeling (projection) enables validation and disconfirms the initial insight.

In this paper we introduce interpretation and trust, two design considerations for model-driven visual analytics. We define **interpretation** as the *facility with which an analyst makes inferences about the underlying data* and **trust** as the *actual and perceived accuracy of an analyst's inferences*. Designs lacking in either restrict an analyst's ability to generate and validate insights derived from a visualization.

Our understanding of these issues is shaped by our experiences designing the Stanford Dissertation Browser and refined via a survey of text analysis and visualization research. The Dissertation Browser is a visual analysis tool for investigating shared ideas and interdisciplinary collaboration between academic departments. We initially envisioned an interface using existing statistical models. However, we quickly arrived at a working visualization that revealed unexpected shortcomings in the underlying model. Our design work instead involved close collaboration among HCI and NLP researchers to develop and evaluate models that better supported our analysis goals. In a subsequent literature review, we observed that many tools lack consideration of how model abstractions align with analysis tasks; iterative design often focuses on the visual interface alone, not modeling choices.

In this paper, we first present selected examples from prior work, drawing attention to issues of interpretation and trust as well as highlighting successful design decisions. We then describe our experience of building the Dissertation Browser. In the process, we contribute a novel similarity measure for text collections based on the notion of "word-borrowing" and show how it arose from our iterative design process. Finally we contribute a series of design process recommendations for constructing interpretable and trustworthy visual analysis systems. While we focus on the domain of exploratory text analysis, we believe our recommendations can help inform the design of a wide range of model-driven visualizations.

RELATED WORK AND CASE STUDIES

A rich and growing literature considers the use of modeling methods to drive text visualizations. Many, such as tag clouds [50], analyze documents by their constituent words to support impression formation [56], augment search [43], reveal language structure [49, 52], or aid document comparison [16, 17]. Other analyses infer latent topics [22, 23, 24, 25, 34, 46, 53], sentiment [5, 36, 51], or word relationships (e.g., overlap [44], clustering [26, 28], or latent semantics [18, 29]) from text. For large corpora, a common approach is to model document similarities, and visually convey patterns in the corpus via dimensionality reduction [9, 10, 13, 30, 38, 54, 55]. A related literature concerns "science mapping" [6, 7, 8, 32, 40, 42], often via 2D projection of academic citation networks.

Here, we review in greater detail a subset of this prior work. We choose three classes of visual analysis tools due to their widespread use and significant research attention: summaries via word clouds, document visualization using latent topic models, and investigative analysis of entity-relationship networks. We pay particular attention to visual designs and model abstractions, and discuss how they relate to analysis tasks.

Text Summarization with Word Clouds

Word clouds are a popular visualization method used to summarize unstructured text. A typical word cloud shows a 2D spatial arrangement of individual words with font size proportional to term frequency. Despite documented perceptual issues [39], word clouds are regularly found both in analysis tools and across the web [50]. Though simple, a word cloud rests on a number of modeling assumptions. Input text is typically treated as a "bag of words": analyses focus on individual words ignoring structures (e.g., word position, ordering) and semantic relationships (e.g., synonym, hypernym). Most implementations assume raw term counts are a sufficient statistic for indicating the importance of terms in a text.

The ostensible goal of most word clouds is to provide a high-level summary of a text. Is the visualization well suited for the task? A strength of word clouds is that they are highly *interpretable* and directly display the *units of analysis*, words and word-level statistics. Users can readily assess word distributions and identify key recurring terms. Studies found summary information provided by a word cloud can help form meaningful impressions [14] and answer broad queries [43].

To enable more specialized tasks, however, changes are required to the underlying language model. For decades, researchers have anecdotally noted that the most descriptive terms are often not the most frequent terms [31]. Significant absence of a word can be a distinguishing indicator of a document's content relative to a corpus. To better support document comparison, Parallel Tag Clouds [17] apply G^2 statistics to surface both over- and under-represented terms. Others note that single words account for only a small fraction of descriptive phrases used by people [48]. To better capture sentiment in restaurant reviews, Review Spotlight [56] extends the bag-of-words model to consider adjective-noun pairs ("great service" vs. "poor service", instead of just "service"). By modifying the unit of analysis, the tool improves impression formation while retaining a familiar visual design.

In-depth analyses may require more than inspection of individual words. Analysts may want additional context in order to *verify* observed patterns and *trust* that their interpretation is accurate. For example, does the presence of the word "matrix" indicate an emphasis on linear algebra, the use of matrices to represent network data, or a scatterplot matrix for statistical analysis? Interactive techniques can provide *progressive disclosure* across modeling abstractions, e.g., selecting a word in a cloud can trigger highlighting of term occurrences in a view of the source text. In other tools, changes in visual design are accompanied by corresponding changes in the model. WordTree [52] discloses all sentences in which a term occurs using a tree layout. Taking into account the frequency of adjacent terms, WordTree expands branches in the tree to surface recurring phrase patterns. DocuBurst [16] applies radial layout to show word hierarchy; the tool infers word relationships by traversing the WordNet hypernym graph.

Document Visualization using Latent Topic Models

A growing body of visual analytics research attempts to support document understanding using topic modeling. Latent Dirichlet allocation (LDA) [4] is a popular method of discovering latent topics in a text corpus by automatically learning distributions of words that tend to co-occur in the same documents. Given as input a desired number of topics K and a set of documents containing words from a vocabulary V, LDA derives K topics β_k, each a multinomial distribution over words V. For example, a "physics" topic may contain with high probability words such as "optical," "quantum," "frequency," "laser," etc. Simultaneously, LDA recovers the per-document mixture of topics θ_d that best describes each document. For example, a document about using lasers to measure biological activity might be modeled as a mixture of words from a "physics" topic and a "biology" topic.

Latent topics are often presented to analysts as a list of probable terms [12], which imposes on the analysts the potentially arduous task of inferring meaningful concepts from the list and verifying that these topics are responsive to their goals. In this case, modeling abstraction increases the gulf of evaluation [27] required to interpret the visualization.

Evaluations of existing visualizations indicate that an analysis of "topical concepts" can provide an overview of a collection [19], but that the value of the model decreases when the analysis tasks become more specific [28]. Beyond "high-level understanding," many existing systems (e.g., [23, 53]) stop short of identifying specific analysis tasks or contexts of use. This omission makes it difficult to assess their utility.

Notable issues of *trust* arise in the application of topic models to specific domains. Talley et al. [46] examined the relationships between NIH-supported research and NIH funding agencies. To characterize research output, the authors applied LDA to uncover 700 latent topics in 110,000 grants over a four-year period. To *verify* that the topics accurately capture significant research fields, the authors manually rated individual topics and noted the presence of a large number of "junk" or nonsensical topics. The authors *modified* the model by removing 1,200 non-informative words from the analysis and inserting 4,200 additional phrases. The authors then performed extensive parameter search and removed poor topics from the final model before incorporating model output into their analysis. Hall et al. [25] studied the history of Computational Linguistics over forty years. The authors applied LDA on 14,000 papers published at multiple conferences to analyze research trends over time, and recruited experts to verify the quality of every topic. The experts retained only 36 out of 100 automatically discovered topics, and manually inserted 10 additional topics not produced by the model. In many real-world analyses, extensive research effort is spent on validating the latent topics that support the analysis results.

Investigative Analysis of Entity-Relation Networks

One particularly successful class of visual analysis tools uses entity-relation models to aid investigative analysis. In the context of intelligence analysis, "entities" may include people, locations, dates, and phone numbers; "relationships" are modeled as connections between them. Example systems include FacetAtlas [11], Jigsaw [45] (a VAST'07 challenge winner), and Palantir [35] (a VAST'08 challenge winner).

In contrast to other text visualization systems, these tools exhibit clearly-defined units of analysis and provide strong support for model verification, model modification, and progressive disclosure of model abstractions. First, the units of analysis (people, places, events) are *well-aligned to the analysis tasks*. The entity-relationship model provides an interpretable analytical abstraction that can be populated by statistical methods (e.g., using automated entity extraction [21]) and modified by manual annotations (e.g., selecting terms in source text) or other override mechanisms (e.g., regular expressions). Jigsaw uses a simple heuristic to determine relations among entities: co-occurrence within a document. This model assumption is readily interpretable and verifiable, but might be revisited to infer more meaningful links. To foster trust, Palantir provides an auditable history for inspecting the provenance of an observed entity or relation.

Progressive disclosure, particularly in the form of linked highlighting, is used extensively by both Jigsaw and Palantir to enable scalable investigation and verification. According to Jigsaw's creators, the "workhorses" of the tool are the list view (which groups entities by type and reveals connections between them) and the document view (which displays extracted entities within the context of annotated source text). In contrast, Jigsaw's cluster view receives less use, perhaps due to the interpretation and trust issues inherent in assessing an arbitrary number of automatically-generated groupings.

Summary

Across these examples, we note that successful model-driven visualizations exhibit relevant *units of analysis* responsive to delineated *analysis tasks*. However, we also find that many text visualizations fail to align model abstractions with real-world tasks; iterative design often considers interface elements, but not modeling choices. These observations emphasize a recurring lack of attention to model design and a need for principled approaches. In the remainder of this paper, we share both a case study exploring these issues and a set of process-oriented design guidelines for model-driven systems.

THE DESIGN OF A DISSERTATION BROWSER

Our interest in model-driven visualization stems from our experiences working on an interdisciplinary team involving social scientists, NLP and HCI researchers. We were tasked with investigating the impact of interdisciplinary collaboration at Stanford University. Our approach adopted the idea that we could identify influences and convergent lines of research across disciplines by detecting shared language use within university-wide publications. Manually reading the document collection is infeasible due to both the size of the corpus and the expertise required to discern topical overlap between papers. The project also receives the attention of university administrators who wish to evaluate the effectiveness of various research institutes on campus. Do multi-million dollar collaborative centers return suitable intellectual dividends? Our collaboration has resulted in the Stanford Dissertation Browser, a visual analysis tool for exploring 16 years of Ph.D. theses from 75 departments.

Identifying the Units of Analysis

The social scientists hypothesized that interdisciplinary collaborations foster high-impact research, and wanted to identify ideas that might bridge disciplines. For example, they posited that statistical methods are topically situated at the center of the sciences and engineering. What data, models and representations would enable rapid assessment of such hypotheses? We began by collecting 16 years of dissertation abstracts, for which text and metadata were readily available.

Early conversations with our collaborators emphasized the need to examine large scale patterns in the university's output. A first step toward that goal is to survey the research at a "disciplinary" level. Such a survey might suggest areas of horizontal knowledge transfer — such as application of theory, methodology, or techniques across domains — that could be verified as interdisciplinary collaborations. Because each department approximately acts as its own discipline, the university's 75 *academic departments* were suggested as a sensible baseline unit of analysis. Each department's school (such as Engineering or Medicine) provides further organizational context that is meaningful to our collaborators and target audience within the university. A visualization that demonstrates which departments share content would allow our collaborators to discover unexpected areas of inter-disciplinary collaboration and verify known ones.

Our collaborators also emphasized the need to assess the impact of interdisciplinary initiatives, which requires tracking the topical composition of involved groups over time. Our collaborators want to correlate change in research output to the formation of academic ties that cross disciplinary boundaries, such as the creation of research institutes, joint grant proposals, and co-authorship. *Time*, in this case the year of filing, is therefore necessary for the analysis tasks.

Textual similarity provides one means of identifying which disciplines are sharing information. Because each dissertation is associated with one or more departments, the content of these dissertations was seen as a reasonable basis for inferring whether two departments are working on the same content as seen through the words in their published dissertations. We thus explored various text-derived similarity measures as the basis of these similarity scores.

Data and Initial Models

Our dataset contains abstracts from 9,068 Ph.D. dissertations from Stanford University published from 1993 to 2008. These dissertations represent over 97% of all Ph.D. degrees conferred by Stanford during that time period. The text of the abstract could not be recovered for the remaining 263 dissertations. The advisor and department of each dissertation are included as metadata as well as the year of each publication. The abstracts average 181 words in length after tokenization, case-folding, and removal of common stop words and very rare terms (occurring in fewer than five dissertations). The total vocabulary contains 20,961 word types.

These words serve as the input to our models, from which we derive scores of departmental similarity based on the text of each department's dissertations. We initially constructed two models, each representing a common approach to textual similarity in the literature. The first metric is based on **word similarity**, measuring the overlap of words. The second is **topic similarity**, in which we measure similarity in a lower dimensional space of inferred topics.

TF-IDF Word Similarity

We can compute the **word similarity** of departments as the cosine similarity of TF-IDF vectors representing each department, a standard approach used in information retrieval [41]. Each component i of the vector for a department v_D is computed by multiplying the number of times term i occurs in the dissertations from that department (TF) by the inverse document frequency (IDF), computed as $\log(N/df_i)$ where N is the number of dissertations in the dataset and df_i is the number of dissertations that contain the term i. We define the word similarity of two departments D_1 and D_2 as the cosine of the angle between their corresponding TF-IDF vectors v:

$$\cos(v_{D_1}, v_{D_2}) = \frac{v_{D_1} \cdot v_{D_2}}{\|v_{D_1}\|\|v_{D_2}\|}$$

LDA Topic Similarity

While TF-IDF is effective for scoring similarity for documents that use exactly identical words, it cannot assign a high score to the shared use of related terms (e.g., "heat" and "thermodynamics") because each term is represented as its own dimension in the vector space. To address term sparsity issues, we apply latent Dirichlet allocation (LDA) [4] to infer latent topics in the corpus, and represent documents as a lower-dimensional distribution over the topics.

We compute the **topic similarity** of two departments D_1 and D_2 as the cosine similarity of their expected distribution over the topics θ_d learned by LDA. This expectation is the average distribution over latent topics for dissertations in that department, and is computed simply as $\mathbb{E}[\theta_D] = \frac{1}{|D|}\sum_{d \in D} \theta_d$.

Accounting for Time

In both of the models above, we quantify the similarity of departments over time by computing a time-aware signature vector. To compute the vector for a department D within a year y, we sum across all dissertations in D either in the year y or in the preceding two years $y-1$ and $y-2$, weighting the current year by $\frac{1}{2}$, the preceding year by $\frac{1}{3}$ and the remaining year by $\frac{1}{6}$. The extra years are included in the signature to reduce sparsity and account for the influence of a student's work prior to completing a dissertation.

Visualizations: Landscape, Department & Thesis Views

The first visualization we created is the *Landscape View* (Figures 1 & 2). The intention of the view was to reveal global patterns of change in department's topical compositions. We encode academic departments as circles, with areas proportional to the number of dissertations filed in a given year. Distance between circles encodes one of the similarity measures, subject to PCA projection. We ensured visual stability by limiting the amount of movement between adjacent years under the projection. Time is controlled by a slider bar that enables analysts to view an animation of temporal changes or immediately access a specific year.

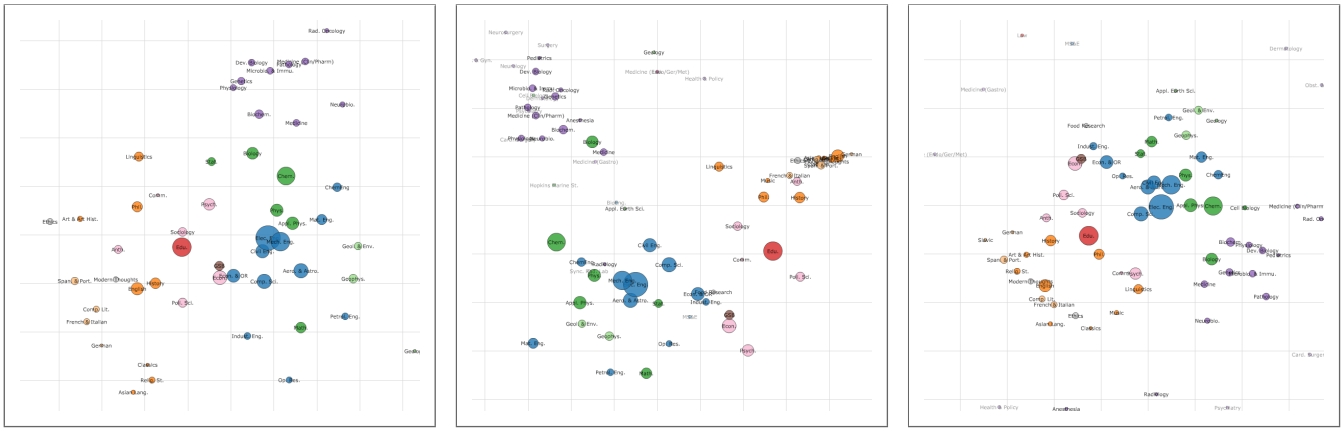

Figure 2. Departmental relationships seen in Landscape View. From left to right: (a) TF.IDF similarity, (b) LDA topic similarity, and (c) Department mixture proportions. These overviews seem plausible, but each makes different predictions and offers little guidance in choosing a model.

Consider the landscapes in Figure 2. Word similarity suggests a relatively uniform landscape, while topic similarity predicts tight overlap of research topics in Medicine (purple) and Humanities (orange) with a relative diverse set of topics in Engineering (blue) and Sciences (green). Which measure best characterizes the university's research output? Without an interactive validation mechanism or an external ground truth, we were left with no way to choose between the similarity measures, nor to trust that the projection faithfully represents the similarity scores derived from each model. The social scientists were unable to confirm whether the observations (in any of the views) correspond to interdisciplinary work, nor to gain insight about the nature of potential collaborations.

In response to these issues of trust, we designed the *Department View* to focus on a single department at a time. This view explicitly shows the distance from a focused department to every other department (i.e., a single row in the similarity matrix). Similarities are encoded as radial distances from the focused department at the center of the display. The remaining departments are arranged around the circle, first grouped by school, then alphabetically within school. A circular representation was chosen to avoid a false impression of rank-ordering among departments and to fit in a single display without scrolling. By restricting the data visible at a single time, the department view avoids projection artifacts.

This view enabled our collaborators to observe expected patterns (e.g., connections between economics and business) and discover surprises. For example, contrary to their expectations, they found that statistics and computer science were not becoming consistently more similar: indeed, they were most similar in 1999. This surprise suggests the need for an even deeper level of verification: to examine the dissertations that contribute to the high (or low) similarity scores of two departments in a given year.

The department view also reveals peculiarities in the underlying models. Figure 3 centers on English, and corresponds to the landscape view in Figure 2(b). This figure immediately suggests a fundamental issue in the topic similarity score derived from latent topic models: how to appropriately select

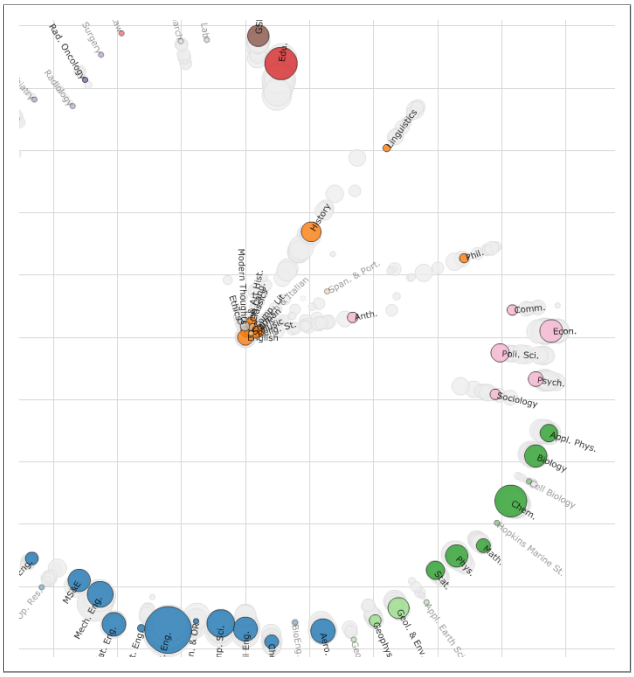

Figure 3. Department View using LDA topic similarity, focused on the English department. While the overview (Fig. 2(b)) seems plausible, we now see that the humanities have been clustered far too aggressively.

the number of topics K used to model the corpus. For the model in Figure 3, we chose the topic count that maximizes the perplexity of held-out data — the technique most commonly used to select the number of topics. However, the visualization demonstrates that the model clearly has too few topics to adequately describe variation within the humanities. A larger number of topics may mitigate this effect, but we lack data-driven metrics for making a principled selection.

As a result, we added the *Thesis View* (Figure 4) to support validation and exploration of observed similarity scores. The thesis view is presented in response to a click on the centered department in the department view. Every thesis from

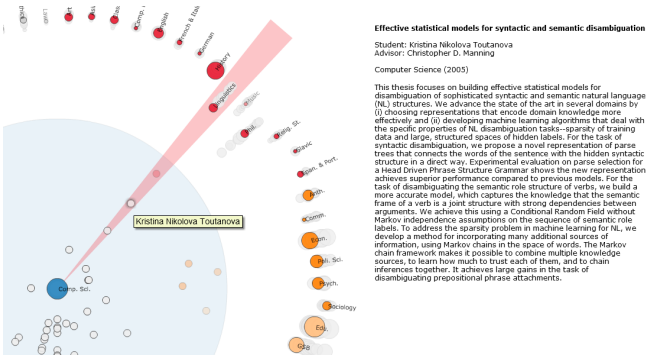

Figure 4. The Thesis View shows individual dissertations as small circles placed between the focus department and the next most similar department. Reading the original text of the dissertation enables experts to evaluate observed dept-dept similarities, and confirm the placement of three computational linguistics Ph.D.s that graduated in 2005.

the focused department, as well as the most similar theses from other departments, are added to the visualization within a concentric circle between the focus and the other departments. The angular position of a thesis aligns with the most similar department, excluding the focus; the radial position is a function of the ratio of the dissertation's similarity to those two departments. This encoding provides a simple means to note theses that might connect two departments.

Upon mouse-over, the text of the thesis abstract is shown, enabling analysts to read the source text and judge whether the two departments are sensible anchors for the dissertation. This view enables users to explore the relationships between departments at a fine-grained level, providing texture and context to the observed department-level similarities.

Evaluating the Models

To assess our modeling options, we conducted an expert review. We invited academic domain experts (professors and graduate students) to use the interface and recorded their responses. We found that the visualizations benefit from being *model agnostic*: they display departmental similarity, but otherwise are not constrained by other modeling assumptions. Thus, we can use the visualizations to compare the results of different modeling approaches.

Using the landscape view, participants could not fully justify their observations. Many potentially interesting patterns turned out to be projection artifacts, ultimately leading us to remove this view from the tool. Using the department view, participants were adept at noting similarities that violated their assumptions. Both word and topic similarity led to many such instances. Rather than identify a preferred model, we became increasingly skeptical of both approaches.

The successes and mistakes of each similarity model were revealed by the thesis view through the (mis)placement of individual dissertations with respect to the other departments. Participants were able to discover systematic errors made by topic similarity. For instance, several biology dissertations were spuriously linked to computer science and vice versa because of the existence of a computational biology topic that connected the dissertations, even though many dissertations made use of only the biology or computer science words in the computational biology topic. The TF-IDF measure used for word similarity, on the other hand, often assigned documents very high similarity to departments that happened to heavily use a common rare word.

We also used our own domain knowledge to examine the relationships between dissertations and departments. The placement of three computational linguistics Ph.D.s that graduated in 2005 provides an illustrative example (Figure 4). We expected these dissertations to fall on the line between computer science and linguistics. In the latent topic model's similarity function, two of them did, but several unrelated dissertations were deemed substantially more similar to linguistics than the computational linguistics dissertations. We discovered this was due to a shared latent topic that covered both linguistics and information retrieval. While the TF-IDF model succeeds in placing these three dissertations between computer science and linguistics, it failed to accurately describe the relationship between the two departments: a year with only one dissertation (2000) is the year of maximum similarity even though the dissertation is not computational in nature.

Revising the Model: Department Mixture Proportions

The high frequency of "mismatch" between experts' mental models and our similarity scores led us to revisit our modeling assumptions. First, we wished to avoid arbitrary parameters such as the number of latent topics (K) and realized that we might better exploit the available metadata. Second, we had implicitly assumed that our similarity measure should be symmetric, as required by the mathematical definition of a metric. However, this need not be true of analysts' views of departmental similarity. In response, we formulated a novel similarity score that we call the **department mixture proportion**. This measure uses a supervised machine learning approach to directly represent the contents of each department, our primary unit of analysis. We estimate the similarity of two departments by measuring how often dissertations from one department "borrow" words from another.

To compute the department mixture proportion, we use the machinery of Labeled LDA[1] [37], which models each document as a latent mixture of known labels. In a two-step process, we first learn latent topics using the departments associated with each dissertation as labels. In a second inference step where labels are subsequently ignored, we infer department mixtures for each thesis.

We train a Labeled LDA model using the departmental affiliations of dissertation committee members as labels. Thus the departments themselves are the "topics". Each dissertation may have one or more labels. During training, we learn both the per-topic term distributions (β_k) and initial label-based topic mixtures (θ'_d). In Labeled LDA, topical term distributions are allowed to take on any word, as in normal LDA training. However, per-document topic mixtures are restricted to only labels associated with the document. For example, the

[1]Our Labeled LDA implementation is available online at `http://nlp.stanford.edu/software/tmt/`

topic mixture for a thesis labeled "Biology" and "Chemistry" is zero for all topics except the two labeled departments.

Using the learned topical term distributions (β_k), we next ignore all labels and perform standard LDA inference on each dissertation (as if we were seeing it for the first time). This results in a new topic mixture (θ_d) in which the dissertation can "borrow" words from *any* department, not just the ones it was initially labeled with. We average the distributions for all dissertations in a given department to construct the department mixture proportion. The values of this averaged distribution are the desired similarity scores.

In short, we first determine the term distributions of each department, and then use these distributions to answer a simple hypothetical: if we let each dissertation borrow words from *any* department, what mixture of departments would it use? The resulting mixture proportion tells us the fraction of words in each dissertation that can be best attributed to each department. The similarity of a department D_1 to D_2 is now simply the value at index D_2 in θ_{D_1}. Unlike the previous measures, this score need not be symmetric. For instance, Music may borrow more words from Computer Science than Computer Science does from Music, which indeed we find in several years where computational music Ph.D. dissertations are filed. We find that this new similarity score ameliorates many of the "mismatches" identified by our earlier expert review.

System Deployment & Use

We first deployed the Dissertation Browser[2] outside of our research team in March 2010, as part of a presentation to the University President's Office. For convenience, we launched the tool on the web, where it remained available after the presentation. Our collaborators found the primary value of the tool to be in validation and communication. They noted the start of a large-scale Biophysics project connecting Biology and Physics in 2006. Several finer stories were discovered that exhibit interdisciplinary collaboration and knowledge transfer. In one case, the visualization demonstrated a strong connection between two departments driven by a small number of individuals centered around the Magnetic Resonance Systems Research Lab. This lab graduated a series of Electrical Engineering Ph.D. students in the 1990's who worked on EE-aspects of various MRI techniques. Around the same time, a hire in Radiology held a courtesy appointment in Electrical Engineering. For the next decade, the influence of these groups strongly connected the two departments until both eventually moved onto other research areas.

As we made no effort to publicize our tool, we were taken by surprise when the system gained public attention from users on the web (e.g., in hundreds of Twitter comments) beginning in December 2010. The majority of tweets expressed interest or enjoyment in the use of the tool (*"geeky and cool"*, *"i could spend hours on this site"*). Several pointed to specific patterns (*"In 2003 Edu was closer to PoliSci than English", "Watch Psychology and Education PhD theses doing the hokey-pokey over time"*). Later, over a dozen science and tech blogs (including Hacker News, Discover Magazine and Flowing Data) posted articles about the tool. We observed commenters interpreting specific patterns of interest: *"I was not surprised to see the link between Computer Science and Philosophy. Heartened by a slight connection between dissertations in Computer Science and Genetics."* and *"Aha, so there are terms that are common between civil engineering and biology but not between civil engineering and religion or art history."* We also observed issues of trust: *"[browser] thinks neurobiology is closer to electrical engineering than to biology. It is easy to see why that might be so based on key vocabulary terms (voltage, potential, conductance, ion), but ..."*. From these and similar comments, we note that the ability to transition between levels of model abstractions enabled users to interrogate the model and assess unexpected correlations.

[2]The Stanford Dissertation Browser is available online at `http://vis.stanford.edu/dissertations/`

DESIGN GUIDELINES

To facilitate interpretation and trust in model-driven visualizations, we distilled a set of guidelines from both our experiences and literature review. Along with illustrative examples, we now present process-oriented recommendations for model and visualization design:

- **Align** the analysis tasks, visual encodings, and modeling decisions along appropriate *units of analysis*.
- **Verify** the modeling decisions: ensure that model output accurately conveys concepts relevant to analysis.
- Provide interactions to **modify** a model during analysis.
- **Progressively disclose** data to support reasoning at multiple levels of model abstraction.

Model Alignment

We use the term *alignment* to describe the correspondences among modeling decisions, visual encoding decisions, and an analyst's tasks, expectations, and background knowledge. We consider a visual analysis system to be well-aligned when the details surfaced in the visualization are responsive to analyst's tasks, while minimizing extraneous information that might confuse or hamper interpretation. Alignment does not result from interface design alone; both the visualization and model may require iterative design.

Identify Units of Analysis

Alignment requires a sufficient understanding of users, their tasks, and the context of use. Such *domain characterization* [33] relies on methods familiar to HCI researchers (e.g., interviews, contextual inquiry, participant-observation). However, these techniques may be foreign to model designers in fields such as statistics or machine learning. To facilitate communication among stakeholders with varying backgrounds, we found it useful to frame insights in terms of *units of analysis*: entities, relationships, and concepts about which the analysts reason. These units serve as a resource for evaluating models and their fitness to the analysis task.

With the Dissertation Browser, we engaged in participatory design meetings with our collaborators to determine the units of analysis. This process led us to realize that changes in inter-department similarity could provide answers to the social scientists' research questions. In turn, we were led to

depict similarity data in the visualization and avoid the potentially confusing route of trying to convey topical composition. In later iterations we further aligned our model with this unit of analysis: we reduced the number of abstractions by computing similarity directly as the department mixture proportion. This eliminated the need to set model parameters such as the number of topics and freed analysts from unnecessarily assessing and classifying latent topics.

Assess Reliability vs. Relevance Tradeoffs

Selecting the appropriate units of analysis often involves a balance between how reliably a concept can be identified, and how relevant the concept is to the analysis task. The final units of analysis reflected in a visual analysis tool may result from a compromise: the units should correspond to the analysts' questions but must also be practical to model.

In the Dissertation Browser, we quantify "units of research" as academic departments. While our social science collaborators would ideally like to assess research at a finer granularity (e.g., trends in microbiology or evolutionary systems), we lacked reliable means to quantify such units of research. LDA models have the potential to discover unnamed research activities, but in our case collapsed all of the humanities into a single topic. Similarly, while investigating historical trends using LDA models, Hall et al. [25] found that only 36 out of 100 automatically inferred "topics" were judged relevant by experts in the field. Named organizations such as departments can be identified reliably, and correspond to concepts that the analysts can comprehend and verify during analysis. More generally, we recommend leveraging available metadata to provide reliable and relevant units of analysis.

Enumerate Model Assumptions

To assess alignment, it is valuable to explicitly enumerate the assumptions implicit in a modeling approach. Common assumptions in quantitative statistics are that data values are independently and identically distributed according to a known probability distribution (e.g., Gaussian, Poisson, etc.). Within text processing, many models are predicated on a bag-of-words assumption that ignores word ordering and relations. Understanding such assumptions is important for determining if a model is appropriate for the given units of analysis. Enumerating assumptions also provides a resource for design, suggesting potential starting points for alternative models.

While designing the Dissertation Browser, we assumed that similarity must be based on a proper metric, and hence symmetric. Once we identified this assumption, it freed us to consider the possibility of asymmetric similarity scores, ultimately leading to a "word borrowing" model based on the department mixture proportion. In Review Spotlight [56], the mismatch between the bag-of-words model and sentiment perception was resolved by making adjective-noun pairs the units of analysis, yielding improved performance.

Model Verification

Once candidate models have been identified, we need to assess how well they fit an analyst's goals. An analytical abstraction based on identified units of analysis can often be realized by different modeling approaches. Verification may require collaboration among designers and domain experts to assess model quality and validate model output.

Assess Model Fit

In domains with objective accuracies, one can take a quantitative approach to verification: common evaluation measures include precision (e.g., comparing model output to known ground truth data) or internal goodness-of-fit statistics (e.g., information criteria such as AIC and BIC). However, one should ensure that such metrics correlate with analysis goals. Domains such as text interpretation may be subjective in nature and so difficult to quantify. For LDA topic models, quality is typically measured in perplexity, which describes the "distinctiveness" of the learned topics. While perplexity is a sensible measure of encoding quality in an information-theoretic sense, in our case it did not correspond to our task: identifying concepts representing coherent "research topics."

Conduct End-User Evaluations

HCI evaluation methods can enable verification. For example, task-based user studies or real-world deployments may be used to assess how well a system aids analysis tasks. Walkthroughs with representative users can help designers gauge analysts' familiarity with a presented analytical abstraction. A potential trade-off is that if analysts don't fully understand the model (e.g., higher gulf of evaluation) but gain more useful and verifiable insights, a less familiar model may be preferred. In our case, we found that expert review was a relatively lightweight means to assess model quality by cataloging instances in which users believed the model to be in error. These "mismatches" became points of comparison across modeling options. An interesting challenge for future work is to better correlate the results of user-centered evaluation with less costly model quality metrics: Can we identify or invent better metrics that reliably accelerate verification?

Enable Comparison via Model-Agnostic Views

Another method for verification is triangulation: comparing the output of multiple models or parameter settings and gauging agreement. To enable cross-model comparison in a model-driven visualization, the visualized units of analysis should be stable across modeling choices. We use the term *model-agnostic views* to describe visualizations that use a single analytical abstraction to compare the output of various underlying modeling options. To be clear, such views rely on a stable abstraction; what they are "agnostic" to is the inferential machinery of the models. For example, the Dissertation Browser uses inter-department similarity as the shared unit of analysis, enabling comparisons with any model that can generate suitable similarity scores. Interactive comparison of parameter settings and modeling options can be invaluable to model designers when assessing choices. Providing similar facilities to end users is also helpful, but might best be treated as a "last resort" when an accurate, well-aligned model can't be found.

Model Modification

Even with careful attention to alignment and verification, a model's output may be incorrect or incomplete. Whether due to limited training data or inaccurate yet pragmatic modeling assumptions, analysts often require mechanisms to modify a

model abstraction over time. The approaches listed below constitute ways to interactively improve model alignment.

Modify Model Parameters

A simple form of model modification is to adjust free parameters. Examples include setting the number of topics in an LDA model or adjusting threshold values for data inclusion (e.g., weights on edges in a social network). We have found that this ability is critical for early stage model exploration. While ideally this would not be necessary in a final analysis tool, in practice one rarely finds a "perfect" model. Consequently it is important for analysts to be able to assess various parameterizations. One challenge is to support real-time interactivity, as changes of model parameters may require expensive re-fitting or other operations. For such cases, visual analysis tools might provide facilities for scheduling offline, batch computation across a range of parameter values.

Add (Labeled) Training Data

Another approach to model modification is to introduce additional training data. For example, an analyst might add new text documents labeled as positive or negative examples of a category. In the context of the Dissertation Browser, new inference procedures might incorporate expert annotations into the model fitting process. To avoid costly re-fitting, designers might leverage techniques for online, interactive machine learning [1, 20]. An important research challenge is to design reflective systems that elicit the most useful training data from users, perhaps using active learning methods [15].

Adjust The Model Structure

Analysts familiar with a modeling method may wish to directly edit the model structure. An analyst might add new latent variables or conditional dependencies within a Bayesian network, or add a new factor to a generalized linear model. In this case, the model itself becomes a unit of analysis, requiring that users possess sufficient modeling expertise.

Allow Manual Override

An alternative approach is to bypass the modeling machinery entirely to override model output. For example, to correct modeling mistakes or impose relations outside the scope of the model or source data. Analysts may wish to delete or modify inferred LDA topics. Hall et al. [25] removed 64 topics and inserted 10 hand-crafted topics in order to complete their investigation; Talley et al. [46] removed poor topics and flagged questionable topics in their visualization. Similar to model agnostic views, manual override benefits from an analytical abstraction decoupled from any inferential machinery. However, overrides may prove problematic with dynamic data: should overrides persist when modeling incoming data?

Progressive Disclosure

By abstracting source data, models can improve scalability, surface higher-order patterns and suppress noise. However, they might also discard relevant information. To compensate, model-driven visualizations can enable analysts to shift among levels of abstraction on-demand. *Progressive disclosure* is the strategy of drilling down from high-level overview, to intermediate abstractions, and eventually to the underlying data itself. Progressive disclosure balances the benefit of large-scale discovery using models with the need for verification to gain trust. A tool can support reasoning and improve interpretation by displaying the right level of detail when it is needed. The critical concerns are that detailed data (1) is revealed on an as-needed basis to avoid clutter and (2) highlights the connections between levels of abstraction to aid verification. We identify two primary interaction techniques for achieving progressive disclosure: semantic zooming [3] and linked highlighting (a.k.a. "brushing and linking") [2].

Disclosure via Semantic Zooming

Semantic zooming changes the visible properties of an information space based on the current "zoom" level, exposing additional detail within an existing view. Using semantic zooming for progressive disclosure entails incorporating elements across different levels of modeling abstraction. The Dissertation Browser uses semantic zooming to move from department view to thesis view: individual dissertations are visualized in relation to the higher-level departmental structure. We hypothesize that semantic zooming is particularly effective for facilitating interpretation if it can show the next level of abstraction within the context of an established abstraction. Semantic zooming relies on a hierarchical organization of relevant model abstractions or metadata.

Disclosure via Linked Highlighting

Another option is to present different levels of analytical abstraction in distinct visualizations. Linked selection and highlighting between views can then enable investigation: given distinct visualizations at different levels of abstraction (e.g., a network of extracted entities and a document viewer) highlight the cross-abstraction connections (e.g., the occurrences of the entity in the document). Perhaps the simplest case is showing details-on-demand. The Dissertation Browser shows the source text of a dissertation abstract in a separate panel when a thesis is selected. Linked highlighting is desirable if the different levels of abstraction are more effectively presented using disjoint visual encodings — that is, when combining levels via semantic zooming is either impossible or inadvisable. When faced with non-hierarchical relations or simultaneous inspection of three or more levels of abstraction, linked views are likely to be preferable to semantic zooming.

Choosing Levels of Analytical Abstraction

A primary design challenge for progressive disclosure is to select the proper levels of abstraction. We consider this an instance of (vertical) model alignment that depends on the identified units of analysis. Another outstanding question is how "deep" progressive disclosure should go. For example, comments from Dissertation Browser users suggest that our design would be further improved by incorporating word-level details to aid verification of thesis-level similarities (e.g., what words does Civil Engineering "borrow" from Biology?). In most instances, we find that progressive disclosure should terminate in the original source data, enabling analysts to connect model abstractions to the raw input.

CONCLUSION

Text visualization research has traditionally focused on improving the effectiveness of a visualization without considering how the underlying model itself affects or can be adapted

towards an analysis goal. This oversight constitutes a limitation in the face of big data applications and the growing need for models. Moreover, machine learning research has normally been content with formal measures of model quality, with less emphasis on user- and task-centric evaluations, even though the limited effectiveness of formal measures has become increasingly evident. In this paper, we proposed *interpretation* and *trust* as criteria to guide the design of model-driven visualizations. We described the design of the Stanford Dissertation Browser, and demonstrated how a novel "word-borrowing" similarity measure arose through an iterative design process that considered task analysis, visualization design, and modeling choices in a unified fashion. We contributed strategies (*align*, *verify*, *modify*, *progressive disclosure*) as practical aids for designers to achieve interpretability and trustworthiness in visual analysis tools. With these strategies, HCI methods can play an important role in the formulation of new interfaces, algorithms, evaluations, and models to enable productive analytic reasoning with massive data sets.

ACKNOWLEDGMENTS

This research was part of the Mimir Project, and was supported by the President's Office at Stanford, the Boeing Company and National Science Foundation Grant No. 0835614.

REFERENCES

1. Amershi, S., Lee, B., Kapoor, A., Mahajan, R., and Christian, B. CueT: human-guided fast and accurate network alarm triage. In *CHI* (2011), 157–166.
2. Becker, R. A., and Cleveland, W. S. Brushing scatterplots. *Technometrics 29* (1987), 127–142.
3. Bederson, B. B., and Hollan, J. D. Pad++: a zooming graphical interface for exploring alternate interface physics. In *UIST* (1994), 17–26.
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *J Machine Learning Research 3* (2003), 993–1022.
5. Bollen, J., Pepe, A., and Mao, H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM* (2011).
6. Börner, K., Maru, J. T., and Goldstone, R. L. The simultaneous evolution of author and paper networks. *PNAS 101* (2004), 5266–5273.
7. Boyack, K. W., Börner, K., and Klavans, R. Mapping the structure and evolution of chemistry research. In *ISSI* (2007), 112–123.
8. Boyack, K. W., Mane, K., and Börner, K. Mapping Medline papers, genes, and proteins related to melanoma research. In *InfoVis* (2004), 965–971.
9. Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., and Börner, K. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE 6*, 3 (2011), e18029.
10. Boyack, K. W., Wylie, B. N., and Davidson, G. S. Domain visualization using VxInsight for science and technology management. *JIS&T 53* (2002), 764–774.
11. Cao, N., Sun, J., Lin, Y.-R., Gotz, D., Liu, S., and Qu, H. FacetAtlas: Multifaceted visualization for rich text corpora. In *InfoVis* (2010), 1172–1181.
12. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *NIPS* (2009), 288–296.
13. Chen, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JIS&T 57* (2006), 359–377.
14. Clough, P. D., and Sen, B. A. Evaluating tagclouds for health-related information research. In *Health Info Management Research* (2008).
15. Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *J Artificial Intelligence Research 4* (1996), 129–145.
16. Collins, C., Carpendale, S., and Penn, G. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum 28*, 3 (2009).
17. Collins, C., Viégas, F. B., and Wattenberg, M. Parallel tag clouds to explore and analyze faceted text corpora. In *VAST* (2009), 91–98.
18. Crossno, P., Dunlavy, D., and Shead, T. LSAView: A tool for visual exploration of latent semantic modeling. In *VAST* (2009), 83–90.
19. Cutting, D. R., Karger, D. R., and Pedersen, J. O. Constant interaction-time scatter/gather browsing of very large document collections. In *SIGIR* (1993).
20. Fails, J. A., and Olsen, Jr., D. R. Interactive machine learning. In *IUI* (2003).
21. Finkel, J. R., Grenager, T., and Manning, C. D. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL* (2005), 363–370.
22. Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. The Topic Browser: An interactive tool for browsing topic models. In *NIPS (Workshop on Challenges of Data Vis)* (2010).
23. Gretarsson, B., O'Donovan, J., Bostandjiev, S., Llerer, T. H., Asuncion, A., Newman, D., and Smyth, P. TopicNets: Visual analysis of large text corpora with topic modeling. *Trans on Intelligent System and Technology 3*, 2 (2011).
24. Griffiths, T. L., and Steyvers, M. Finding scientific topics. *PNAS 101* (2004).
25. Hall, D., Jurafsky, D., and Manning, C. D. Studying the history of ideas using topic models. In *EMNLP* (2008), 363–371.
26. Hearst, M. A., and Pedersen, J. O. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *SIGIR* (1996), 76–84.
27. Hutchins, E. L., Hollan, J. D., and Norman, D. A. Direct manipulation interfaces. *Human-Computer Interaction 1* (1985), 311–338.
28. Ke, W., Sugimoto, C. R., and Mostafa, J. Dynamicity vs. effectiveness: studying online clustering for scatter/gather. In *SIGIR* (2009), 19–26.
29. Landauer, T. K., Laham, D., and Derr, M. From paragraph to graph: Latent semantic analysis for information visualization. *PNAS 101* (2004), 5214–5219.
30. Lin, X. Visualization for the document space. In *Vis* (1992), 274–281.
31. Luhn, H. P. The automatic creation of literature abstracts. *IBM J of Research and Development 2*, 2 (1958), 159–165.
32. Moya-Anegón, F. d., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., and Herrero-Solana, V. Visualizing the marrow of science. *JIS&T 58*, 14 (2007), 2167–2179.
33. Munzner, T. A nested process model for visualization design and validation. In *InfoVis* (2009), 921–928.
34. Newman, D., Asuncion, A., Chemudugunta, C., Kumar, V., Smyth, P., and Steyvers, M. Exploring large document collections using statistical topic models. In *KDD (Demo)* (2006).
35. Palantir technologies. `http://www.palantirtech.com`, 2011.
36. Procter, R., Vis, F., Voss, A., Cantijoch, M., Manykhina, Y., Thelwall, M., Gibson, R., Hudson-Smith, A., and Gray, S. Riot rumours: how misinformation spread on Twitter during a time of crisis. `http://www.guardian.co.uk/news/datablog/2011/dec/08/twitter-riots-interactive`, 2011.
37. Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. In *EMNLP* (2009), 248–256.
38. Risch, J. S., Rex, D. B., Dowson, S. T., Walters, T. B., May, R. A., and Moon, B. D. The STARLIGHT information visualization system. In *InfoVis* (1997).
39. Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. Getting our head in the clouds: toward evaluation studies of tagclouds. In *CHI* (2007).
40. Rosvall, M., and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *PNAS 105* (2008), 1118–1123.
41. Salton, G., Wong, A., and Yang, C. A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (1975), 613–620.
42. Sandstrom, P. Scholarly communication as a socioecological system. *Scientometrics 51*, 3 (2002), 573–605.
43. Sinclair, J., and Cardew-Hall, M. The folksonomy tag cloud: when is it useful? *J of Information Science 34* (2008), 15–29.
44. Smalheiser, N. R., Torvik, V. I., and Zhou, W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer M&P in Biomedicine 94*, 2 (2009), 190–197.
45. Stasko, J., Görg, C., Liu, Z., and Singhal, K. Jigsaw: Supporting investigative analysis through interactive visualization. In *VAST* (2007), 131–138.
46. Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., Leenders, A. G. M., and McCallum, A. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods 8*, 6 (2011).
47. Thomas, J., and Cook, K., Eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.
48. Turney, P. D. Learning algorithms for keyphrase extraction. *Info Retr 2*, 4 (2000).
49. van Ham, F., Wattenberg, M., and Viegas, F. B. Mapping text with phrase nets. In *InfoVis* (2009), 1169–1176.
50. Viégas, F. B., and Wattenberg, M. TIMELINES: Tag clouds and the case for vernacular visualization. *Interactions 15* (2008), 49–52.
51. Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D., and Keim, D. A. Visual sentiment analysis of RSS news feeds featuring the US presidential election in 2008. In *VISSW* (2009).
52. Wattenberg, M., and Viégas, F. B. The Word Tree, an interactive visual concordance. In *InfoVis* (2008), 1221–1228.
53. Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., and Zhang, Q. TIARA: a visual exploratory text analytic system. In *KDD* (2010).
54. Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *InfoVis* (1995), 51–58.
55. Wong, P. C., Hetzler, B., Posse, C., Whiting, M., Havre, S., Cramer, N., Shah, A., Singhal, M., Turner, A., and Thomas, J. IN-SPIRE contest entry. In *InfoVis* (2004).
56. Yatani, K., Novati, M., Trusty, A., and Truong, K. N. Review Spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *CHI* (2011), 1541–1550.