TopicViz: Interactive Topic Exploration in Document Collections

Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology
jacobe@gmail.com

Duen Horng "Polo" Chau
Machine Learning Department
Carnegie Mellon University
dchau@cs.cmu.edu

Aniket Kittur
Human Computer Interaction Institute
Carnegie Mellon University
nkittur@cs.cmu.edu

Eric P. Xing
Machine Learning Department
Carnegie Mellon University
epxing@cs.cmu.edu

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Existing methods for searching and exploring large document collections focus on surface-level matches to user queries, ignoring higher-level semantic structure. In this paper we show how topic modeling — a technique for identifying latent themes across a large collection of documents — can support semantic exploration. We present TopicViz: an interactive environment which combines traditional search and citation-graph exploration with a force-directed layout that links documents to the latent themes discovered by the topic model. We describe usage scenarios in which TopicViz supports rapid sensemaking on large document collections.

Author Keywords
Topic analysis, interactive visualization

ACM Classification Keywords
H.4 [Information Systems Applications]: User Interfaces.

General Terms
Design

Introduction
While interaction techniques for navigating document collections continue to improve [1, 9], they are generally hampered by a view of language that is restricted to the

surface level, and oblivious to the semantic meaning behind the text. At the same time, machine learning and natural language processing have developed powerful statistical methods for recovering latent semantics [2], but there has been little investigation of how to present the output of these methods to users.

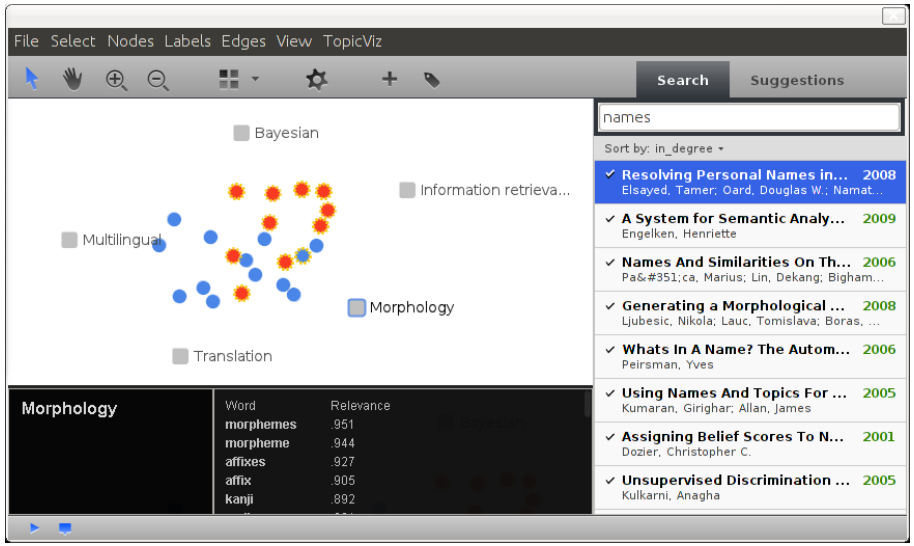

Figure 1: The TopicViz environment. The main panel shows the initial presentation for two sets of documents (differentiated by color), which are arranged in a force-directed layout controlled by the best matching topics. The right panel shows the search results in list form, and the lower-left panel describes the selected topic, "**morphology**".

In this paper, we introduce TopicViz, a new tool for searching and navigating large document collections (Figure 1). TopicViz infers a set of *topics* that summarize the latent semantic organization of a collection, and exposes this semantic organization using a force-directed layout. This layout offers a range of interactive affordances, allowing users to gradually refine their understanding of the search results and citation links, while focusing in on key semantic distinctions of interest.

The analytic engine of our approach is the topic model – a statistical method that identifies latent themes in a document collection [2]. Topic models extract sets of semantically-related words, and describe each document as a mixture of these themes. For example, this paper might be characterized as 75% human-computer interaction, and 25% machine learning. The ability to assign multiple themes to a document distinguishes topic models from more coarse-grained techniques that treat each document as a member of a single cluster [5].

One of the main strengths of topic models is their flexibility. Topics need not correspond to any predefined taxonomy; rather, they capture the themes inherent to the document collection [3]. But this flexibility comes at a price: the distinctions between topics — and their relationship to documents of interest — must somehow be conveyed to the user. A recent survey paper notes that after nearly a decade of progress on the statistical methodology of topic modeling, one of the most important unsolved research problems is to develop user interfaces and visualizations capable of leveraging this statistical power to support knowledge discovery [2].

The TopicViz approach is rooted in interaction: the user can manipulate the visualization by adding, removing, and visually rearranging topics, and by controlling the set of documents to visualize. This design is motivated by our focus on local information exploration: we aim to provide a deep understanding of a local area of the information landscape that is relevant to the user's goals, rather than a surface-level static view of thousands of documents or dozens of topics.

"Morphology"	"Multilingual"
morphemes	bilingual
morpheme	english-chinese
affixes	bitext
affix	english-french
kanji	monolingual
endings	melamed
inflections	cognates
suffixes	hansard
inflectional	japanese-english
katakana	systran

Figure 2: The top ten keywords from two topics in a dataset of research papers on computational linguistics. The topic names were assigned manually.

Background

A topic model is a hierarchical probabilistic model of document content [2]. Each topic is a probability distribution over words; every word in every document is assumed to be randomly generated from one topic. In a given document the proportion of words generated from each topic is indicated by a latent vector θ_d, providing a succinct summary of the document's main themes. The number of topics typically ranges from ten to several hunded, and the number of documents can range from a few hundred to millions. TopicViz is not currently designed to support training new topic models; rather, we address the setting where a topic model is trained in advance, but must be understood by users who may not be familiar with the document collection.

Much of the existing work on visualizing large document collections focuses on using spatial location to reveal properties of either the document content [6, 13] or the citation graph [12]. These presentations are usually static and display the entire collection at once. This can successfully convey the high-level structure of a collection, but it is poorly suited for more specific tasks for at least two reasons. **First**, while there may be thousands or millions of documents in a collection, tasks like reviewing the scientific literature or searching for legal precedents require the user to quickly focus in on a few dozen relevant documents [5]. **Second**, language is multifaceted and inherently multi-dimensional; if a single static visualization were truly sufficient, then we would require only two or three topics to represent the range of subjects covered in any collection. In reality, any projection down to a lower-dimensional representation causes information to be lost, and the question of which information can be safely ignored depends on the specific sensemaking task at hand. For these reasons, TopicViz offers a dynamic visualization designed to reveal distinctions and relationships among the topics and documents that are relevant to the user's goals.

System and Scenarios

The key idea behind TopicViz is to integrate a force-directed layout for topic models with a set of affordances for expanding and refining document and topic lists. As in conventional search, the entrance point is the query; but rather than simply listing the search results, we present an interactive force-directed layout inspired by the "dust-and-magnet" visualization [14]. Both topics and documents are nodes, and their locations are determined by applying Hooke's law, with $1 - \theta_{di}$ as the force of the spring between topic i and document d.

In one view (Figures 1 and 3), the most relevant topics are pinned, while the documents float between them. A document that is a 50% match for each of two topics will be positioned halfway between them. Documents that have similar topic profiles will be located near each other, reflecting semantic similarity directly in the spatial layout. In an alternative view (Figure 4), the documents are pinned and the topic nodes float between them. In both cases, the user can drag around any set of documents or topics to learn more about the strength of their connections to other nodes. The remainder of this section presents two scenarios which showcase these views.

Scenario 1: summarizing the research literature

Consider the task of searching an unfamiliar research literature, with the goal of identifying the maturity of existing technology towards a commercial problem. In our scenario, the user must determine whether it is possible to automatically identify personal names on foreign language websites, using a collection of 15,032 research papers on

computational linguistics [11].

The first step is to devise a query; with current tools like Google Scholar, the response to the query would be an ordered list of results. Only some of the resulting documents will be relevant, and almost surely there will be relevant documents that do not match the query. The user may then vary the search terms or navigate the citation links to try to get a complete sense of the research literature in this unfamiliar area.

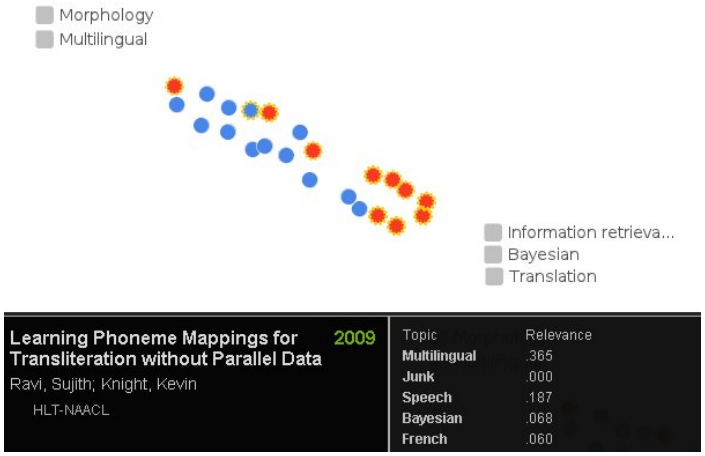

Figure 3: By arranging the topic centers into two points, the documents are shown linearly by relevance. The node colors indicate the results of two different queries.

With TopicViz, the first step is the same: the user supplies a search query. The results are shown in a list (the right panel of Figure 1). The user then drags as many documents as desired into the main area, which is called the Topic Field: each document is displayed as a node, and these nodes are surrounded by a ring of topic centers. The topic centers are "pinned," while the position of each document is set by the force-directed layout. Only the most relevant topics are shown — that is, the topics with the highest total values of θ_d, summed across all documents in the set — and the number of topics to show is a user-defined parameter. The panel on the lower-left shows the most relevant words for each topic, selected by mouseover. Document-topic relevance is shown both statically (by the document's position) and dynamically (by dragging the topic center around to see how the document nodes are affected). The dynamic view is key for overcoming the inherent limitations of 2D projection. If only two topics are shown, a document that is 50% relevant to each will have the same static position as a document which is only 5% relevant; however, the effect of dragging the topic centers will immediately reveal the difference in attraction.

In our scenario, the user recognizes the topic **multilingual** as especially relevant to the search – but other topics like **morphology** are not familiar.[1] Some of the terms associated with **morphology** are unknown (e.g., "morpheme" and "inflection"), but "affix" and "suffix" are recognized as parts of individual words (see Figure 2). Based on this insight, the user renames the topic from **morphology** to **subwords**. While this name is not typically used in the research literature, it helps the user relate the topic model to her pre-existing understanding.

Having identified **morphology** and **multilingual** as key topics of interest, the user again rearranges the topics, placing the relevant topics in one corner of the screen and the others in another corner. This causes the document nodes to form a line, with location governed by relevance

[1] The topic names are specified in advance, either by a domain expert or automatically [10]; the user is free to rename topics with more familiar terms.

to the topics of the interest (Figure 3). The user now removes documents that are not close to the desired topics by selecting and deleting their nodes.

The original list of query hits has now been culled to a set of documents that are closely related to multiple topics of interest. But the coverage of this document set depends on the quality of the original query. To make sure that important documents have not been missed, the user selects a subset of particularly promising documents and adds documents that cite them. The new documents do not exactly match the search query, but may still be relevant. They are laid out based on their relevance to the existing topics in the field, allowing the user to investigate their topical chacteristics and further refine the search.

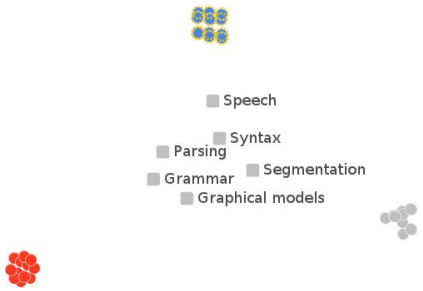

Figure 4: To compare topical emphasis of different authors, the user creates and pins sets of documents for each author; the unpinned topic centers float between them.

Ultimately, the user arrives at a set of documents that reflect the underlying semantics of the information search. By investigating the topic structure, the user has pruned away "false positives" that match the query but are in fact irrelevant; by exploring the citation graph, the user has identified "false negatives" that are relevant but did not match the original query. By interactively exploring the documents, topics, and terms that relate to the initial query, the user acquires a deeper understanding of the relevant area of the document collection.

Scenario 2: distinguishing author topics
Now consider a more experienced user who wants to identify the topical interests of several authors – perhaps to distinguish the specific contributions of multiple authors on a single paper. To do this, the user searches for papers by each author and drags them into the field. Unlike the previous view, the *documents* are pinned in place, and the topics float between them (this non-default behavior can be obtained using a toolbar at the top of the screen). The edges in the force-directed layout are bidirectional, and they work identically in this setting; the user need only pin sets of documents for each author, and then add relevant topics to the view. Figure 4 shows such a view for the relationship between the three authors of a well-known paper [7]; this view reveals that the author on top has focused more on the **speech** topic; the author to the lower left has focused more on **grammar** and **graphical models**; and the author on the lower right has focused more on the **segmentation** topic. Similar visualizations can be constructed to reveal the change in topical interests of the collection over time (by creating document groups for different years) or venue (by creating groups for specific conferences or journals).

Summary

Topic models can give powerful insights on document collections, but only if used in combination with a comprehensible presentation and an interaction design built around the sensemaking process. TopicViz presents an interactive visualization that places topic models in the context of a search interface, filling the same role currently played by keyword search. The key advantage of

TopicViz comes from coupling a model of latent document semantics with an interactive spatial visualization that allows the user to rapidly focus in on key areas of interest.

A clear goal for future work is empirical validation. We believe that TopicViz can serve as a platform for *in situ* studies of how topic models can best support document set exploration and sensemaking. From a visualization standpoint, we see several other directions for further development. Of particular interest is how to convey the meaning of individual topics: we plan to explore more spatial visualizations as well as alternative approaches such as DocuBurst [4]. An integrated presentation of document metadata such as time [8], authorship, and venue may also improve the practical usability of the system, while raising new questions about how to visualize such metadata jointly with the induced topics.

Acknowledgments This work was supported by AFOSR FA9550010247, ONR N0001140910758, NSF OCI-0943148, NSF IIS-0968484, NSF IIS-0713379, NSF CAREER DBI-0546594, and an Alfred P. Sloan Fellowship. Funding was also provided by the U.S. Army Research Office (ARO) and DARPA contract number W911NF-11-C-0088.

References

[1] M. Baldonado and T. Winograd. SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of CHI*, pages 11–18, 1997.

[2] D. M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, To appear, 2011.

[3] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans intepret topic models. In *NIPS*, 2009.

[4] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009.

[5] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR*, pages 318–329, 1992.

[6] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: Topic model for visualizing documents. In *Proceedings of KDD*, 2008.

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289, 2001.

[8] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *Proceedings of CIKM*, pages 543–552, 2009.

[9] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):46, 2006.

[10] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of KDD*, pages 490–499, 2007.

[11] D. R. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL anthology network corpus. In *Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61, 2009.

[12] H. Small. Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813, 1999.

[13] E. M. Talley, D. Newman, D. Mimno, B. W. Herr, H. M. Wallach, G. A. P. C. Burns, A. G. M. Leenders, and A. McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6):443–444, May 2011.

[14] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4:239–256, 2005.