Aaron Zinman
azinman@media.mit.edu

Doug Fritz
doug@media.mit.edu

TOPIC MODELS AND DATA PORTRAITURE

ABSTRACT

The art world uses the tradition of portraiture to semantically compress relevant information about their subjects into a single art work. We believe this tradition can be extended into the digital realm by the use of topic modeling to compress individuals' information into generative data portraits. Data portraiture has strong implications for navigating large social spaces, collaborative systems, and self reflection. We briefly showcase our works thus far to highlight potential directions in this emerging field.

INTRODUCTION

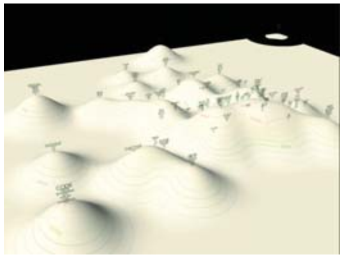
In cyberspace, we are bodiless. Despite the obvious and long desired advantages of removing race, gender, age, and other non-mental attributes from online interactions, the physical body remains a powerful force in face-to-face interactions. Stereotyping allows society to function as a whole, and minute physical gestures are important for efficient communication, trustworthiness, and expression of identity. In the art world, portraiture has been long-standing and rich tradition that exploits our ability to recognize these physical properties to obtain a multi-dimensional gestalt of character, form, and function. We believe that transferring this tradition into the digital realm can help individuals not only make better sense of strangers in the online spaces they inhabit, it can help organizations to understand the information flow within, facilitate better collaboration, and function ego-centrically as a digital mirror to better understand ourselves.

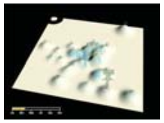
Data-driven portraiture, or simply data portraiture, is the generative process of creating visual representations of the self based on abstract data ather than facial features. These representations can function as a body for where we are bodiless, facilitating navigation of large quantities of human-centered information. Topic modeling plays a special role in data portraiture for its ability to aggregate large amounts of unorganized information into high-level categories. These categories and associated weighted vectors allows data portraiture to significantly advance in utility and power.

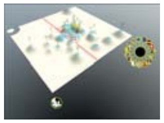
In our paper, and in this poster, we briefly outline our contributions of data portraiture using topic modeling.

Landscape of Words

Landscape of Words is a visualization created by Aaron Zinman and Alex Dragulescu that centered around applying Latent Dirichlet Allocation to – at the time – all of Twitter. It was constructed when the service had not quite reached the mainstream recognition it has today, and was meant to allow newcomers to become familiar with the service. Inspired by earlier work with landscape metaphors and Latent Semantic Analysis, we represent each topic as a "mountain" whose height is proportional to the total number of tweets that fall under that topic. Multi-dimensional scaling of Kullback-Leibler divergences grouped related topics on the flattened two-dimensional map. The number of topics was artificially reduced to not overload the landscape. Trade-offs in model completeness versus navigation are an inherent problem in interactive uses of topic models. The landscape is annotated using heat maps to display topic popularity across time, compare individuals, and social networks. Using a common model across all of Twitter keeps the location and contents of the topics static, thereby making it easier to compare like entities. We reserve a small disconnected area in the corner for Tweets that have low-probability assignments to the model, an important feature when presenting data via topic models.

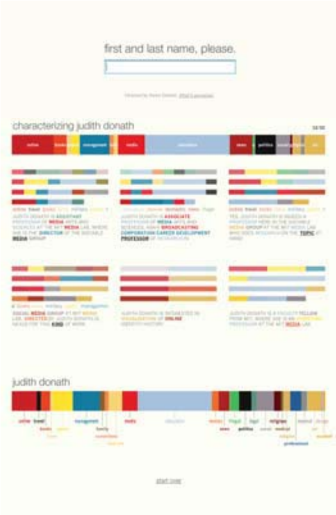






Personas

Personas is a piece originally designed for the Metropath(ologies) exhibit recently on display at the MIT Museum. Museum-goers approach the terminal and enter their name. Yahoo!'s BOSS service is then used to to find characterizing statements of the name on the web. Inference using Latent Dirichlet Allocation and Gibbs sampling is performed on the sentences, saving the results of each iteration of its word-topic assignments. Each iteration is then animated for the user, allowing them to cogitate on the information given, the oscillations inherent in the machine trying to make sense of the data, and the quality of the topic-word assignments. Meanwhile, an aggregated DNA-like strip of the document-topic vectors is continuously adjusted as the results change; it is the final data portrait. Personas is meant for participants to reflect on issues of privacy, identity, and their faith in data mining. Since being ported to the web, Personas has generated several million portraits.



first and last name, please.

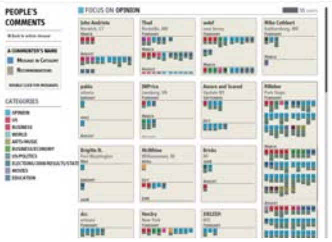
characterizing judith donath

judith donath

start over

Defuse

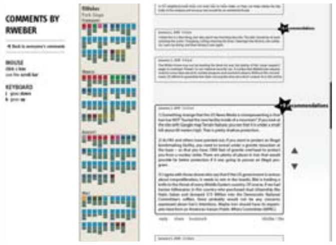
Defuse is a project currently under development that redefines navigating and participating in online discussions using a web-based commenting platform. Online commenting in large sites is a space that is ripe for data portraiture, as condensing the rich history of posters helps contextualize their messages for unfamiliar users, in addition to providing a quality feedback loop to the original posters themselves. Currently topic modeling is being used to generate clusters of discussion topics and conversational style, which are then used to showcase users by their past topical history as a high-level gestalt for their interest patterns over time. We are also experimenting with using topic modeling as a summarization tool for building navigable models of very activite participants and conversations by artifically setting the number of topics to a low value. Similarly, we are also experimenting with cross-cultural topic models (ccLDA) to separate comments based upon vocabulary usage as a proxy for social prototypes.



PEOPLE'S COMMENTS

FOCUS ON OPINION

CATEGORIES



COMMENTS BY RWEBER

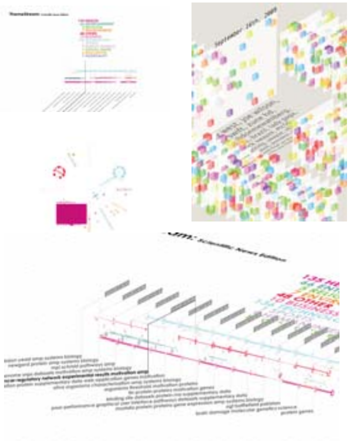
MOUSE

KEYBOARD

ThemeStream

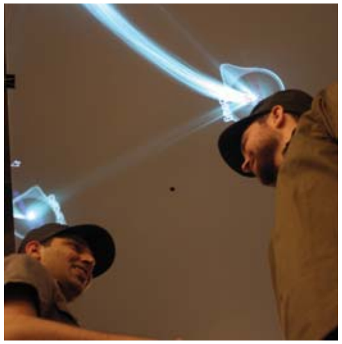
ThemeStream is a visualization of the major themes of one's personal RSS feeds. Zooming around a 3D interface, it provides differing levels of detail depending upon the camera view, alternating between high-level top-down views, and focused contextual views. Much in the way an artist paints the same model from different perspectives, creating varying views of data, attribution, and detail reveals more than any single view could provide. It uses topic modeling to compress individual time slices and OpenCalais for theme extractions.

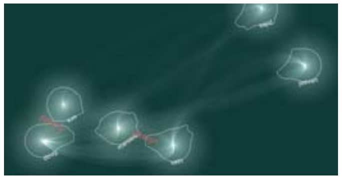
Each view is a transformation of the same consistent layout. The top view exposes major fluctuations in time and topic frequency changes as they flow from one time slice to the next. The three-quarters view reveals a localized meta-view, displaying the topics popular within each time slice. Finally, the front view accesses each of the individual items, which can be opened and viewed in full. Smoothly animating between these different models of the data gives an ability to drill down while maintaining a consistency, and thereby understanding, of the navigation context.



ConnectUs

ConnectUs is a visualization to better incorporate our digital personas into physical interactions. The project optically tracks and identifies participants moving around in a space. A priori, participants bind their physical identity to their online social networking accounts. Available data is used to create a common topic model. An ego-centric data portrait follows participants as they physically move, deforming as they come into contact with others in the room to show similarity. Topics unique to those in conversation are highlighted to function as a digital impetus for social interaction.





Fluid Interfaces Group · MIT Media Lab

MAS.714J / STS.445J Technologies for Creative Learning
Fall 2009