

The NIH Visual Browser: An Interactive Visualization of Biomedical Research

Bruce W. Herr II, Edmund M. Talley, Gully A.P.C. Burns, David Newman, Gavin LaRowe
ChalkLabs, National Institute of Neurological Disorders and Stroke, Information Sciences
Institute, University of California Irvine/NICTA Australia, ChalkLabs
{bh2@chalklabs.net, talleye@ninds.nih.gov, gully@usc.edu, newman@uci.edu, gl@chalklabs.net}

Abstract

This paper presents a technical description of the methods used to generate an interactive, two-dimensional visualization of 60,568 grants funded by the National Institutes of Health in 2007. The visualization is made intelligible by providing interactive features for assessing the data in a web-based visual browser, see <http://www.nihmaps.org>. The key features include deep zooming, selection, full-text querying, overlays, color-coding schemes, and multi-level labeling. Major insights, broader applicability, and future directions are discussed.

Keywords--- information visualization, scientific funding, topic modeling, information navigation.

1. Introduction

Within successful biomedical research, there are two elements that accompany every project: proposals and publications. Finding ways to analyze these elements in a framework that is intuitive and convenient for biologists is a very difficult undertaking. The scope and range of on-going projects and related data sets are not easily identifiable, let alone computable within large-scale analyses.

By providing a gestalt-like frame of reference and an expeditious visual entry point for large information spaces that is not available via textual reports, spreadsheets, or database results, visualization is now well established as both an academic discipline and a tool for improving scientific research. It is already well entrenched in the biomedical community where it is used to facilitate, enhance, and drive analytical and decision-making processes [1].

In this case study, we present a technical description of an interactive mapping tool based on a comprehensive collection of abstracts from proposals funded by the National Institutes of Health (NIH). The data provides a significant view of federally funded biomedical research extant within the United States. The primary purpose of this tool is to provide NIH administrators, portfolio

analysts, program officers, and researchers an intuitive, easy-to-use tool for navigating and analyzing large-scale funding data.

Although still in development, this tool currently provides:

- an orthogonal view of the entire collection
- a zoom-able component for viewing iteratively smaller sub-sets of the collection
- an interactive interface for selecting groups of documents for analysis
- a system of layers, overlays, color-coding schemes, and labels
- selection components for various fields such as institute, topic, or keyword

2. Data

The data used for textual analysis were titles and abstracts from NIH grants funded in 2007. In addition, we extracted information on Principal Investigator and funding Institute or Center (from among 27 distinct NIH Institutes and Centers) for each grant. The data is publically available on the NIH website [2] (specific project data available on request). After removing duplicate abstracts (~1% of the corpus) and grants with no abstracts (~10%), we analyzed a total of 60,568 abstracts associated with 25 separate Institutes.

3. Technique

3.1. Topic Modeling

We used topic modeling [3] to analyze the NIH abstracts, and to provide a robust means for computing similarities between abstracts. The topic model is a state-of-the-art Bayesian statistical machine learning technique for modeling collections of text documents. The intuition behind the topic model is that individual documents exhibit multiple concepts or topics. The topic model simultaneously learns a set of topics that together describe a collection, and the mixture of topics within each document in the collection. These topics – which are learned automatically from the data with no prior

knowledge – are usually semantically meaningful, and highly interpretable.

While the topics themselves produce a useful taxonomy or ontology to categorize individual abstracts, they also provide a reliable basis for computing similarity between any two abstracts. We ran a total of six independent topic models, initially learning $T=200$, $T=400$ and $T=600$ topics (each topic setting was run twice from different random initial conditions). The topic models were learned using Gibbs sampling, a standard technique for this model [5]. For each of the six runs, the Gibbs sampler was run for 500 iterations. At the end of 500 iterations, very similar topics in each run were manually merged using hierarchical agglomerative clustering, resulting in models having $T = 183, 186, 343, 349, 529$ and 535 topics. These six topic models were run for a further 500 iterations of the Gibbs sampler to reach stationarity.

3.2. Document Similarity

The layout and visualization of the 60,568 grants is completely based on pair-wise similarities computed for each pair of grants. While we did compute all 2 billion similarities, we only kept similarities that were above a threshold. Since there is no gold standard for measuring similarity between two text documents, we took a multi-resolution approach of computing a similarity score based on general and specific similarity. This was achieved by computing the similarity based on all six topic model runs, with topic resolutions ranging from 183 to 535 topics. Averaging over six different topic model runs also improves the similarity estimate.

The input to the graph layout algorithm is a list of pairs of nodes with an associated weight, given by the similarity. We computed the similarity between grant abstracts d_i and d_j using the average symmetric Kullback-Leibler divergence function [6], computed using the respective mixtures of topics $Pr(t | d_i)$, and $Pr(t | d_j)$:

$$KL^*(d_i, d_j) = \sum_t (\Pr(t | d_i) - \Pr(t | d_j)) \log \frac{\Pr(t | d_i)}{\Pr(t | d_j)}$$

This symmetric version (denoted by $*$) of KL-divergence was converted to a $[0,1]$ similarity for input into the graph layout algorithm, which is described in the next section.

3.3. Graph Layout - DrL

Layout is a critical factor in all visualization applications, especially for large-scale data sets. For laying out the data, the Distributed Recursive Graph Layout (DrL) algorithm developed by Shawn Martin was used for the data [7]. At its core, DrL is a force-directed layout algorithm based on VxOrd [8] and optimized for large graphs. It offers both a parallel and sequential version of the above force-directed algorithm, a recursive multi-level version for very large graphs, and the ability to add new nodes to a previously drawn graph.

DrL is ideally suited for large graphs; meaningful results have been obtained on graphs with up to 2 million nodes [23]. The input for DrL is a similarity network where the weight of each edge in the network represents how similar the two connected nodes are. When DrL is run, it produces a 2d layout where nodes that are similar are placed close to one another. This typically results in local clustering of related nodes and a global structure of related clusters.

For this project, quite a bit of time was spent on getting a good layout. This is due to the fact that there are so many variables in topic generation, grant-grant distance computation (0 is identical, infinity is very dissimilar), distance to similarity conversion (0 is dissimilar, 1 is very similar), and layout parameters. A change in parameters at any of these stages could drastically alter the layout. Further, we wanted the layout to come out where major research areas are strongly clustered with strands of grants between clusters representing ‘bridges’ which consist of projects laid out in a gradient between two topically defined research areas. These project bridges have the potential to add extra dimensionality to the layout, since their topical content provides concrete information on the relationships between clusters. The layout we eventually chose for this project is shown in Figure 1.

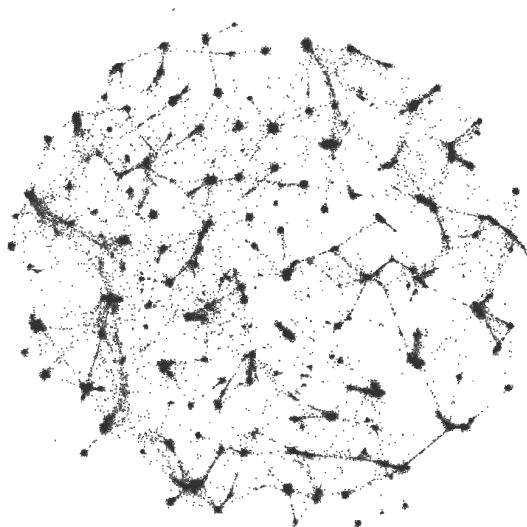


Figure 1. Final layout of the NIH grants

3.4. Visual Encoding

Making sense of the network layout without any labeling, coloring, or sizing is nearly impossible. So, to help in understanding we chose to color the nodes according to the institute which funded the represented grant. Unfortunately, there are 25 different institutes in the dataset, so there are many colors which are too similar to one another to tell apart. To deal with this we had to do several things. First, we chose colors such that similar colors were as far apart as possible and near colors used complimentary colors. While not perfect, this

helped to at least be able to notice different institutes intermingling. Second, when drawing the nodes, we sorted by institute first, and then drew from largest institute to smallest institute. Finally, in the user interface (discussed below), we allowed overlaying of institutes in order to bring all of their respective nodes to the top.

3.5. User Interface

To best understand and explore this visualization, one must interact with the user interface (UI) (accessible from <http://www.nihmaps.org/>, see Figure 2). The major features of the UI include a zooming and panning interface to the visualization, toggled overlays, a selection mechanism for selecting regions of the visualization, and a free text querying mechanism.

The zooming and panning interface is set up such that as the user zooms in, the grants stay essentially the same size, but the relative distances increase. This way at the top level one can see the major clusters, but with each successive level, the nodes become more and more distinct. At the lowest level, there are almost no overlapping grants. Also, when text overlays are enabled, the granularity of text increases with each level, gaining more labels and going from general (NIH institutes) to specific (the most popular topic in each region).

A toggling overlay control allows the user to turn on an overlay for each institute to see exactly where each of the institute’s grants fall. Further, more than one institute can be toggled on at a time allowing the user to see their overlap.

When, a user selects a region on the visualization, they are presented with a list of the individual grants in that region. In addition, a histogram is displayed showing the institutes represented in that region, see Figure 3.

The free text querying mechanism allows the user to search for grants which match a query. Matched grants are displayed exactly as grants in selected regions are. In addition, they show up on the visualization as individual markers allowing the user to see where on the map the grants lie, see Figure 4.

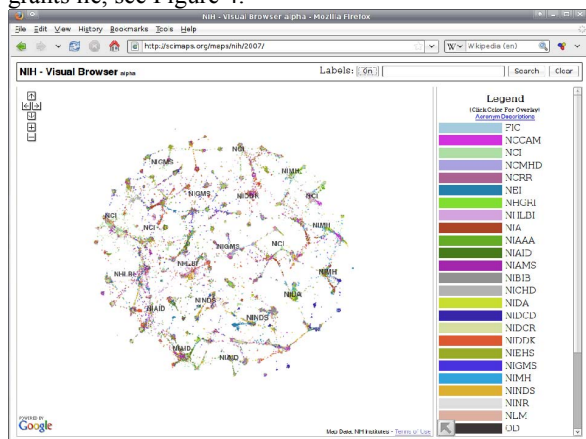


Figure 2. NIH Visual Browser

4. Discussion

The tool provides an intuitive spatial framework for domain experts to view and explore the topic-based-landscape of *any* document corpus. Given the importance of understanding the trends in scientific funding for working scientists, we specifically identified NIH grants as a high-value data set. We actively engaged with grant administrators at NIH to improve the design of the tool and to add new features. These include (a) analytics based on graphical selections on the map (Figure 3) and (b) searching capabilities (Figure 4). By presenting the data in a familiar map-based format (q.v. ‘Google Maps’), the visualization became far more practically useful and understandable for end-users.

The development of these tools relies on visual elements (i.e., dot color) that represent different partitions of the data based on attribute values. We are also able to provide alternative color schemes based on different attributes (for example, NIH institute vs. Grant Coordinator) to provide valuable new views of the data (implemented only in a previous, non-public implementation). The labels placed onto the map provide a mechanism to embed aggregated data from all points in a cluster onto the figure at several different levels. Thus, the wider application of this tool involves the careful choice of (a) which attributes to use to partition the data within the color scheme and (b) which attributes to use as labels placed over the map. These relatively simple choices have a large impact on the usefulness of the tool.

Performing LDA over large corpora is possible [18] and the pipeline underlying this approach scales well to the size of this analysis. Members of this team have been involved in the development of similar visualizations involving over 1 million documents [23].

5. Related Work

The original impetus of trying to derive semantic associations from a ‘bag-of-words’ comes from work done by Landauer and Dumais in Information Retrieval (IR) in regard to Latent Semantic Analysis (LSA) [10]. Moving beyond relatively coarse sets of vectors of words generated by LSA, subsequent probabilistic models were applied to this problem. Work done by Hofmann on probabilistic approaches to find semantically related sets of words, and thus documents, ultimately resulted in Probabilistic Latent Semantic Indexing (PLSI) [11].

Blei, Ng and Jordan derived a generative version of PLSI and called it Latent Dirichlet Allocation (LDA) [13], where inference was done using variational methods. Steyvers and Griffiths presented the now popular Gibbs sampling inference method for LDA [5]. Newman, Chemudugunta, Steyvers and Smyth developed a topic model for words and named entities [14].

Research regarding the display of large graph layouts has been going on for many decades now. Some of the other popular layout techniques that exist for large graphs include Large Graph Layout by Adia, Date,

Weiland, and Marcott [15], Cytoscape [16], and Tulip [17], et. Al.

'Maps of Science' (<http://mapofscience.com/>) is a website presenting visualization tools and maps of sets of journals and documents [20]. Inter-journal similarity scores were calculated with standard methods (cosine,

Jaccard *etc.*) from attributes of journals (including inter-citation and co-reference) rather than a lexical approach like LDA.

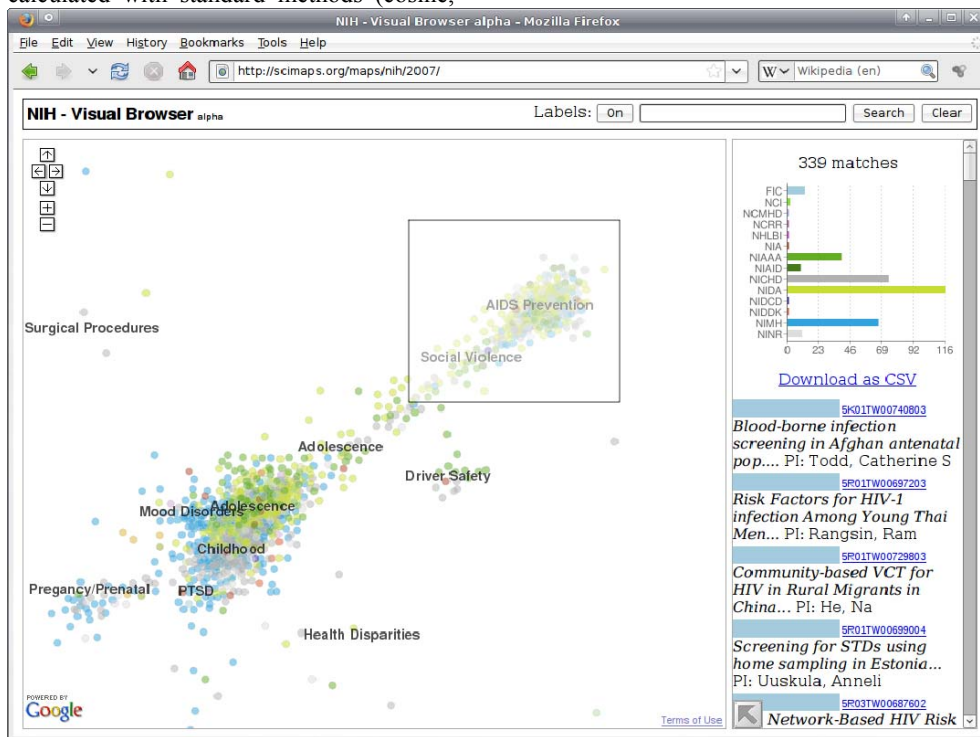


Figure 3. Cluster selection with results shown in the right-hand column.

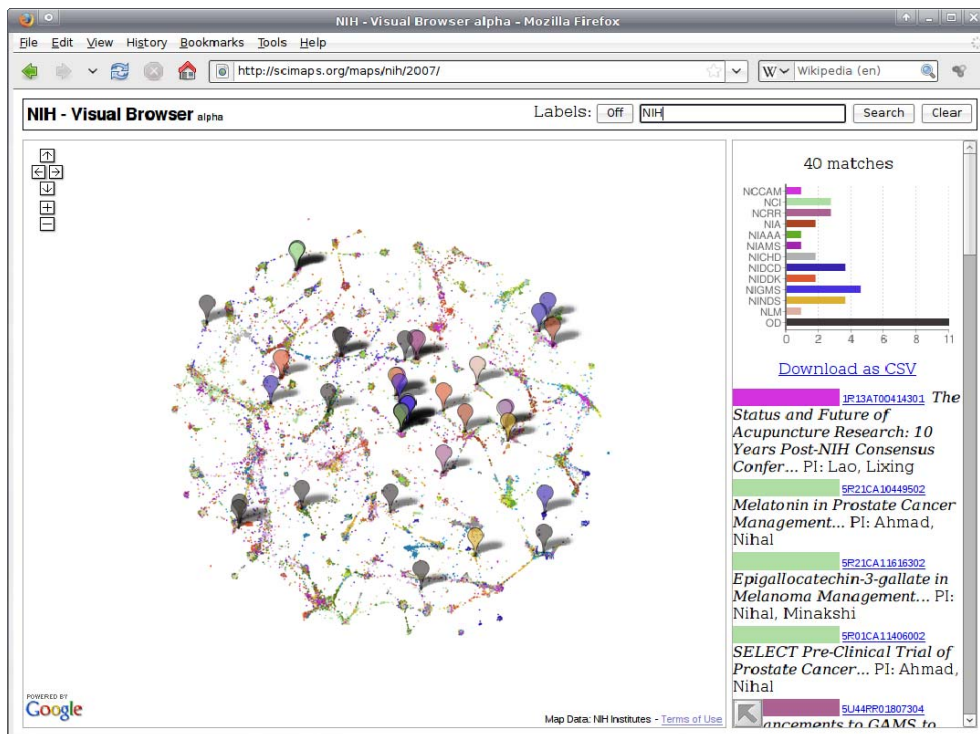


Figure 4. Query for 'NIH' in the title. Results are shown on the map and in the right-hand column.

6. Further Work

The NIH Visual Browser has received positive feedback from within and without the NIH. We are investigating how to perform combined analysis over multiple corpora (such as NIH grants and Medline) [19]. This line of inquiry requires investigating the relationships between topic-models from overlapping and or fully enclosed corpora. Constructing the system in a parallelizable format to allow the creation of smaller maps in real time is also of particular interest. This could provide interfaces to other search engines or external databases to provide near real-time modeling and visualization of related external data. Perhaps most significant, performing quantitative historical assessments to evaluate the role of previous NIH funding mechanisms and philosophies in the context of specific scientific fronts is another area we will explore.

A better understanding of the many interrelated variables involved in TOPICS+DrL style visualizations is essential. Currently much of our visualization efforts have been through trial and error. Our next major research hurdle will be in becoming more exact about what it takes to create the type of layouts we are looking for: strong clusters representing topics and thinner ‘bridges’ of inter-topic clusters connecting the topics.

7. Acknowledgements¹

This work is funded by the National Institute of Neurological Diseases and Stroke (NINDS) and the National Institute of General Medical Science (NIGMS). Early Support was provided by the National Science Foundation and unrestricted funds from the Information Sciences Institute at University of Southern California. Additionally, we gratefully acknowledge the early inspiration, vision, and support of Katy Börner and Padhraic Smyth. We are also deeply indebted to the contributions of Kevin Boyack, a leader in this field.

References

- [1] Tempany, C. M. C., and McNeil, B. J. 2001. Advances in biomedical imaging. *The Journal of the American Medical Association* 285, 5 (Feb.), 556-561.
- [2] report.nih.gov/crisp/. 20 Feb. 2009. CRISP - NIH Research Portfolio Online Reporting Tool (RePORT) <http://report.nih.gov/crisp/>.
- [3] Steyvers, M. & Griffiths, T. (2007). [Probabilistic topic models](#). In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum
- [4] Ibid.
- [5] Griffiths, T., & Steyvers, M. (2004). [Finding Scientific Topics](#). *Proceedings of the National Academy of Sciences*, 101 (suppl. 1), 5228-5235.
- [6] Kullback, S., and Leibler, R. A., 1951, On information and sufficiency, *Annals of Mathematical Statistics* 22: 79-86.

- [7] Martin, S., Brown, W.M., Klavans, R., & Boyack, K.W. (2008). DrL: Distributed recursive (graph) layout, SAND2008-2936J: Sandia National Laboratories.
- [8] Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using vxinsight for science and technology management. *Journal of the American Society for Information Science and Technology*, 53(9), 764-774.
- [9] nwb.slis.indiana.edu. 02 Feb. 2008. Network Workbench: A Workbench for Network Scientists. <https://nwb.slis.indiana.edu/community/?n=VisualizeData.DrL>
- [10] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211-240.
- [11] Hofmann, T. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.
- [12] Steyvers, M. & Griffiths, T. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.
- [13] Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, 2002.
- [14] Newman, D., Chemudugunta, C., and Smyth, P. 2006. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 20 - 23, 2006)*. KDD '06. ACM, New York, NY, 680-686.
- [15] Adai AT, Date SV, Wieland S, Marcotte EM. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *J Mol Biol*. 2004 Jun 25;340(1):179-90.
- [16] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13, 2498–2504.
- [17] Auber, D. Tulip. In P. Mutzel, M. Jünger, and S. Leipert, editors, *9th Symp. Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, pages 335-337. Springer-Verlag, 2001.
- [18] Newman, D., A. Asuncion, M. Welling and P. Smyth. “Distributed Inference for Latent Dirichlet Allocatino”, *Neural Information Processing Systems* 2007.
- [19] Burns, G.A., D. Newman, B. Herr, A. Holloway, P. Smyth, and K. Borner (2008), "Analyses over multiple scientific text corpora: combining MEDLINE, CRISP and the proceedings of the annual meetings of the Society for Neuroscience" Annual Meeting for the Society for Neuroscience.
- [20] Boyack, K.W., R. Klavans, and K. Börner (2005), "Mapping the backbone of science." *Scientometrics*, 64(3): p. 351-74
- [21] Boyack, K.W. (2004), "Mapping knowledge domains: characterizing PNAS". *Proc Natl Acad Sci U S A*, 101 Suppl 1: p. 5192-9
- [22] scimaps.org/maps/wikipedia. 24 Feb. 2008. Wikipedia Visualizations. <https://scimaps.org/maps/wikipedia/>.

¹ NSF (IIS-0513650)