Zoss, Angela M., Michael Conover and Katy Borner. 2010. Where are the Academic Jobs? Interactive Exploration of Job Advertisements in Geospatial and Topical Space. Sun-Ki Chai, John Salerno and Patricia L. Mabry, Eds. Advances in Social Computing: Third International Conference on Social Computing, Behavioral Modeling and Prediction, SBP10: Bethesda, MD March 30-31, Springer, pp. 238-247.

Where are the Academic Jobs? Interactive Exploration of Job Advertisements in Geospatial and Topical Space

Angela M. Zoss[1], Michael Conover[2] and Katy Börner[1]

[1] Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, IN 47405
[2] School of Informatics and Computing, Indiana University, Bloomington, IN 47405
{amzoss, midconov, katy}@indiana.edu

Abstract. This paper details a methodology for capturing, analyzing, and communicating one specific type of real time data: advertisements of currently available academic jobs. The work was inspired by the American Recovery and Reinvestment Act of 2009 (ARRA) [2] that provides approximately $100 billion for education, creating a historic opportunity to create and save hundreds of thousands of jobs. Here, we discuss methodological challenges and practical problems when developing interactive visual interfaces to real time data streams such as job advertisements. Related work is discussed, preliminary solutions are presented, and future work is outlined. The presented approach should be valuable to deal with the enormous volume and complexity of social and behavioral data that evolve continuously in real time, and analyses of them need to be communicated to a broad audience of researchers, practitioners, clients, educators, and interested policymakers, as originally suggested by Hemmings and Wilkinson [1].

Keywords: RSS feeds, data mining, data visualization, science map, visual interfaces, job market

1 Introduction

According to the U.S. Bureau of Labor Statistics (`http://www.bls.gov`) the U.S. unemployment rate rose to 9.8% in September 2009 from 9.7% in August, 2009. More than 214,000 people lost their jobs within one month. This is the highest unemployment rate since June 1983 when it was 10.1%. Academia, industry, and government are all affected. Many universities cut staff lines, reduced salaries by up to 20%, or have hiring freezes. Students that graduate this year or postdocs that are interested in moving on will face major competition for few jobs. Understanding the job market is an essential element of both informed career choices and scientific policy making.

The work presented here aims to capture and visually communicate exactly what academic job opportunities currently exist. Data from major job advertisement services was captured, processed and analyzed. The geospatial and topical distribution of available jobs is communicated to a broad audience using two different base maps: a map of the world and a map of all sciences. Methodological challenges comprise the robustness of real-time data analysis including stop-wording and matching algorithms and the legibility of visualizations. Practical challenges relate to the different temporal distributions and formats of diverse RSS feeds, the automatic identification of geolocations and topics, map labeling, search, and interactivity.

Solutions are expected to be valuable for other projects that aim to use multi-faceted, linked visualizations to help understand, model, and predict complex social and behavioral data in real-time. The remainder of the paper is organized as follows: Section 2 discusses related work, section 3 details the datasets used, section 4 presents data preparation and analysis, section 5 introduces the interactive visualizations, and section 6 discusses the strengths and limitations of the presented work together with an outlook to future work.

2 Related Work

The data analysis and visualization work presented here draws inspiration from projects in many fields, including those that deal with real-time data analysis and interactive visualizations and those that focus specifically on job market data.

2.1 Real-Time Data Analysis and Interactive Visualizations

There are few tools and services that support real-time data analysis. Among them is Google Trends [3], which is a service offered by Google that provides longitudinal data about Google searches performed on specified terms and topics. Users enter one or more search terms, and Google Trends produces a report with a plot of the usage frequency over time, information about the geographic distribution of the searches, and related Google News stories. Data can be exported as a CSV file, and some restrictions can be made on the report, i.e., limiting the report to a specific geographic region or time period.

Visualizations such as the Map of the Market by SmartMoney [4] provide up-to-date information on the size and trends of more than 500 stocks using a tree map visualization. The maps are updated every 15 minutes (with a 20 minute delay) based on stock data provided by ComStock Partners, Inc.; historical prices and fundamental data by Hemscott, Inc.; earnings estimates by Zacks Investment Research; and insider trading data provided by the financial division of Thomson Reuters. Stocks are grouped by industry. The size of a rectangle (an individual company) represents its market capitalization. Color gradation depicts the level of losses (bright red is -6 percent) or gains (bright green is +6 percent). Hovering the mouse over a rectangle brings up the company's name and advises whether its stock price is going up or down. Clicking on a rectangle provides more detailed information. Newsmap by Marcos Weskamp takes groupings from the Google News aggregator and displays it as a tree map in real-time [5]. Here, size is used to indicate the number of articles dealing with a particular topic. Color codes show what larger news category (e.g., business, entertainment) each topic belongs to.

The systems discussed so far visualize small to medium size datasets. However, there is an urgent need to make sense of larger amounts of data to understand their topic coverage and context. For example, the *Science Related Wikipedian Activity* map [6, 7] uses a base map of all Wikipedia articles. Overlaid are 3,599 math, 6,474 science, and 3,164 technology relevant articles. Four smaller maps show articles size coded according to article edit activity, number of major edits from January 1st, 2007 to April 6th, 2007, number of bursts in edit activity, and indegree, e.g., the number of times other articles link to an article. These visualizations serve to highlight current trends and predict future editing activity and growth in science, technology, and mathematics related Wikipedia articles. Similarly, *A Topic Map of NIH Grants 2007* shows all 60,000 grants awarded by the National Institutes of Health (NIH) in 2007 [8, 9]. It supports search, zoom and pan, color coding, and differential labeling for the different scales. By exploring this map, one can see what topics of research are being heavily pursued, how the topics relate to one another, and what research topics each institute is funding.

2.2 Job Market Data Analysis and Visualizations

There are a number of online sites that visualize employment (or unemployment) data. Among them are the Flowing Data Bleeding Country maps [10], the Slate interactive map of employment data [11], Recovery.gov maps of recovery funding and unemployment [12], indeed.com [13], coolworks.com [14], mapyourjob.com [15], and jobmaps.us [16].

Most sites focus specifically on the geospatial visualization of job data. Indeed.com aggregates jobs from many major job sites and displays them as circles of varying sizes that have been normalized by the population of the location. That is, a circle indicates that, e.g., 49 jobs have been posted for every 1000 people living in the city. The Slate interactive map also used circles of varying sizes but does not normalize by population density. The Bleeding Country and Recovery.gov maps color-code geographic regions (states or counties) by unemployment rate. Other jobs maps, such as those at coolworks.com, mapyourjob.com, and jobmaps.us, use a single flag on the map for each job, making it difficult to evaluate the strength of the job market in a particular location from a distant zoom level. Some sites list very brief snippets of the jobs that are being displayed on the map, but mapyourjob.com in particular includes a detailed table for the jobs listed. Each incorporates some sort of search or filter by topic except for Indeed.com, which is more of a static visualization of quarterly data than a way to browse individual jobs. Thus, many of the available maps incorporate both geospatial data and topical filters. To our knowledge there exists no site that serves topic maps of jobs.

3 Data Sets

Since December 2008, we have been collecting job postings from 380 RSS feeds, representing 36 distinct career sites (Figure 1). The feeds were selected from the BestCollegesOnline.com list of the 100 Best RSS Feeds for Recent College Grads [17]. Among these are Monster.com, CareerBuilder, Indeed, and other specialized job sites, many of which publish multiple RSS feeds. Closer examination of the data, however, revealed that there is considerable variety in the type and quantity of text included in the RSS feed items for each site. The feed items are often abridged descriptions of the jobs with links to a more complete job posting. The RSS feed itself may only contain the name of the job and the first sentence or two of the job description. Because of the limitations of using so short of a description for text-based analysis, a more complete sample data set was collected to prototype the system. The sample data set includes over 3,500

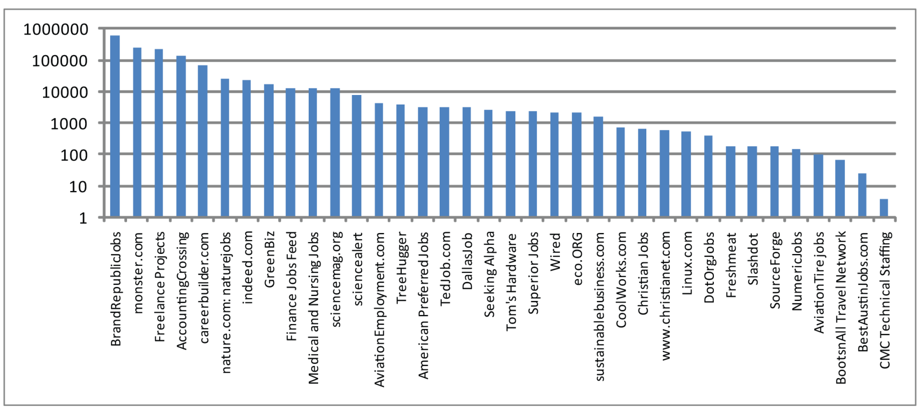

Fig. 1. Number of posts collected from each job site between December 2008 and October 2009. Each job site may have multiple RSS feeds, dividing jobs by topic or geographic area.

full-text, location-specific, time stamped job postings from Nature Jobs [18]. These posts were parsed and stored in a relational PostgreSQL database. The HTML files from these sites have clearly-delineated fields for important information, such as post title, location, employer, etc. These were harvested using screen scraping techniques. Ideally, the data collected would have high coverage (U.S. or world-wide) and high quality of location, topic, and salary data with few missing or unidentifiable values and added flags for those jobs that are funded by ARRA. This level of detail and delineation, however, is not yet available from major job posting sites.

4 Data Preparation and Analysis

For the sample data set, we extracted a timestamp for the posting, the title, the source (company or agency) responsible for the post, the full HTML-formatted text of the job description, and a URL linking to the site on which the post originally appeared.

In order to geo-locate jobs, the Google Maps geocoding API [19] was used to transform plaintext location strings into rich geographic information, including latitude and longitude.

In order to science-locate jobs, the job descriptions were processed as follows. First, the full text of each job description was stopword-filtered and tokenized into 1-, 2- and 3-grams. These n-grams were then scored for relevance using TFIDF (term frequency-inverse document frequency) term weighting, and these weights were summed to create a total strength of association between the job posting and the node or nodes to which the used keywords belong.

5 Visualization

Many of the sites discussed in the related work section use circle size coding to effectively denote the number of jobs, and we adopted this in our visualizations. Similarly, we adopted zoom, pan, search, and request of detail functionality following Shneiderman's Visual Information-Seeking Mantra [20].

Contrary to the other sites mentioned, our visualization provides two complementary views of the data: a geospatial view and a topical view. The geospatial view helps answer, "Where are the jobs?" The topical view helps answer, "What jobs exist?" Common visual metaphors are leveraged to give the user a sense of consistency. Specifically, because both visualizations are maps, users transfer an understanding of spatial relationships between the two. At the code level, both maps use the same Javascript library and server-side web service such that interaction mechanisms from one visualization are readily available to the other. The primary interaction affordances shared by the two visualizations include the circular markers (or icons) that are size-coded to represent job density, higher resolution (more markers) at lower zoom levels, a common search interface, and identical detail-on-demand behavior. Moreover, both of these visualizations were created with the Google Maps API, and user tests have demonstrated that this familiar interaction framework affords users an immediate understanding of the basic functionality of the interface, allowing them to begin exploring its features more easily.

When users click on an icon on either map, an Information Window pops up to show a list of the jobs that have been associated with that position (that is, with a location for the geographic visualization or with a scientific domain in the Map of Science visualization). When a user clicks on one of the job titles in the Information Window, the secondary window on the right of the web page displays more detailed information about the job.

5.1 Geospatial Visualization

The geospatial visualization behaves much like a traditional Google Map, with individual circle markers representing clusters of posts in a given geospatial area. Semantic zoom [21] is employed using MarkerManager [22] to ensure equal information density at different zoom levels. For example, at high zoom levels, where much of the U.S. is visible at a given time, fewer markers are displayed, associating jobs only with states and not with individual cities (Figure 2, left).

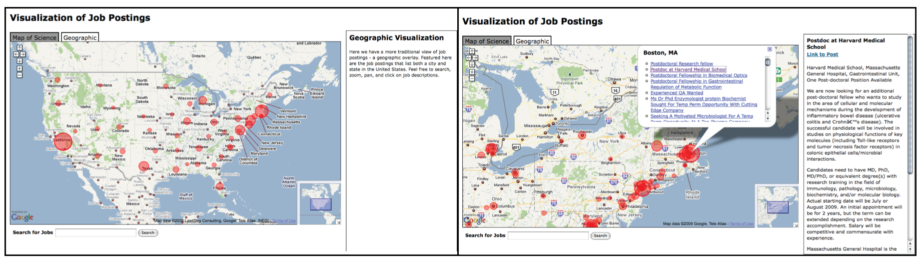

Fig. 2. U.S. level view of job postings on the geographic visualization, clustered by state (left). Lower zoom level of job postings on the geographic visualization – state clusters have broken apart into individual geographic locations at this level (right).

When users zoom in, however, the visualization displays a finer-grained representation of the geographic area, and markers corresponding to individual localities become visible (Figure 2, right).

Further research is warranted to investigate the best scaling technique for the geospatial visualization, whether it is using a density measure (number of jobs per unit population) or scaling linearly or by a power law to approximate real-world populations.

5.2 Map of Science Visualization

The Google Maps API allows users to create a custom map with custom tile sets, thus appropriating the standard pan and zoom actions from the map metaphor to explore other types of images. Here, we have created a custom Google Map with the UCSD Map of Science as a basemap. The UCSD Map of Science is the product of a large study by Klavans and Boyack supported by the University of California San Diego [23–25]. It uses 7.2 million papers and over 16,000 separate journals, proceedings, and series from the Web of Science by Thomson Scientific and Scopus by Elsevier over the five year period from 2001 to 2005. Bibliographic coupling using both highly cited references and keywords was applied to determine the similarity of journals. Using a hierarchical, multi-step clustering procedure, journals were grouped into 554 clusters, represented by 554 individual nodes in the network. Links denote strong bibliographic coupling relations. In its traditional format, the UCSD map has different sizes for the nodes to indicate the volume of publications from the 2001-2005 data set. Here we only use the structure of the network; see Figure 3.

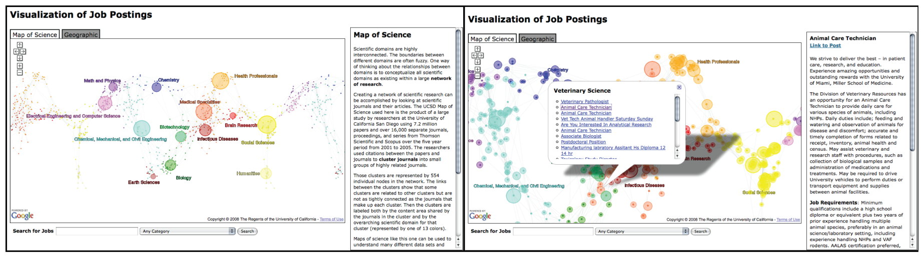

Fig. 3. High (left) and low (right) zoom views of the Map of Science visualization. The map is circular, so areas of the map are repeated side to side as users scroll back and forth. Postings are clustered by the 13 main scientific domains at the high zoom level and the 554 subdisciplines at the lower zoom level.

The clusters are grouped into the 13 overarching scientific domains identified by the analysis (e.g., "Math and Physics", "Humanities"). Each cluster also has its own name (a descriptive name of the subdomain, like "plant physiology") and a set of keywords (an average of about 130 keywords per cluster). Keywords located within a job posting can then be used to "science-locate" a job positing. However, if a posting contains keywords from several Map of Science clusters, that posting will appear in multiple nodes in the visualization. As can be seen in Figure 4, most job posts are associated with multiple nodes, though only a few have more than 15 associations. A larger data set with job posts from a variety of different job sites may produce a different distribution of associations between posts and nodes.

The process of creating a Google Map with custom tiles is partially outlined on Google's Map Overlays [26] page. More detailed instructions at Mapki [27] and a Photoshop script for creating custom tiles [28] were also very helpful. Finally, because the online documentation is a bit sparse and often uses different versions of the Google Maps API, we heavily relied on and modeled after one particular guide by Matthew Muro [29] and several examples, like the previously mentioned NIH map [8] and Google examples like the Tile Detector [30] and the LabeledMarker Marker Hider [31]. The tiles were created from a PostScript file of the base map that includes colored nodes of a uniform size.

6 Discussion and Outlook

The presented work makes several contributions. We have created geospatial and topical visualizations of job opportunities throughout the United States and

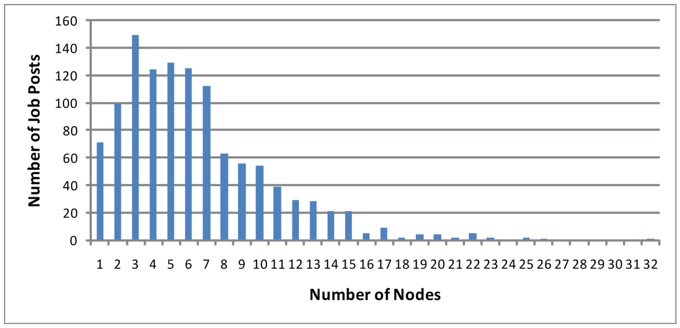

Fig. 4. Number of associations between Nature job posts and nodes. The y-axis is a count of the job posts that are associated with the number of nodes on the x-axis.

all sciences. These complementary visualizations use similar visual metaphors to afford the user unique insights into the continuously evolving scientific job market. Moreover, the approach enables decision makers and job seekers to get a high level overview of the relative distribution of employment opportunities in different domains, while at the same time providing a more detailed perspective of data from a prominent employment site that is both easy to use and insightful.

Future work will improve the online service by using higher quality job data sets, improve geolocation and the cleaning of job descriptions for keyword matching, and optimize the visual display of larger amounts of jobs. A significant challenge of this project involves the mapping of job data to nodes in the Map of Science. The Map of Science term data have been automatically extracted from the text in journal publications. Language use in scientific publications naturally differs from language use in job postings. Overly-broad words such as "chemists" or "economy" are uniquely associated with a single cluster in the Map of Science, whereas these words may appear more commonly in job postings. Analyses of the growing job posting data set could help establish baseline information about the differences in language use and suggest modifications to the keyword set that would produce more meaningful matches.

The connections between disciplines and the presence of multiple nodes per discipline may also cause conceptual problems for users. Because the Map of Science was created from bibliometric data and the journals were clustered by computational analysis, disciplines have multiple nodes, nodes have multiple journals and keywords, and one job might map to multiple nodes. On the other hand, some nodes may not have any job data associated with them. Without documentation, the connections between nodes, the size of "empty" nodes, and the differences between nodes of the same discipline may not be clear. The positions and shapes of continents and countries are extensively taught and used in school. Similar training might be required to fully utilize maps of science.

Additionally, the correct encoding mechanism to communicate the density of job postings in the geographic visualization warrants further evaluation. Important considerations include whether to use size-coding or color-coding and whether or not to scale raw values linearly, by a power law, or in relation to the size of a scientific domain or geographic location.

The next phase in the progression of the project is to conduct a usability study to establish standards and explore competencies of navigation for both of the visualizations. Showing the increase or decrease of jobs over time is a major challenge that will be addressed in the usability study. The final online service will contribute meaningful data and trend analyses to labor market research, especially when grounded by and compared with other data sets from agencies like the Bureau of Labor Statistics.

7 Acknowledgements

This project received a great deal of support from members of the Cyberinfrastructure for Network Science Center at the School of Library and Information Science at Indiana University. Consultations with Bruce Herr II and Russell J. Duhon were helpful in the conceptualization and implementation of this project.

This work is funded in part by the National Science Foundation under grant IIS-0715303 and the National Institutes of Health under grants RM-07-004 and 1U24RR029822-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Hemmings, J., Wilkinson, J.: What is a public health observatory? Journal of Epidemiology and Community Health 57 (2003) 324–326
2. The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education. http://www.ed.gov/policy/gen/leg/recovery/implementation.html.
3. Google Trends. http://www.google.com/trends.
4. Map of the Market at SmartMoney.com. http://www.smartmoney.com/map-of-the-market/.
5. Newsmap. http://newsmap.jp.
6. Herr II, B.W., Holloway, T., Hardy, E.F., Boyack, K.W., Börner, K.: Science Related Wikipedian Activity. 3rd iteration (2007): The power of forecasts, places and spaces: Mapping science edn. Places and Spaces: Mapping Science, Bloomington IN and Albuquerque, NM (2007) http://scimaps.org.
7. Math, Science, & Technology Articles in Wikipedia Visualization. http://www.gigapan.org/viewGigapan.php?id=4305.
8. Herr II, B.W., Burns, G., Newman, D., Talley, E.: A Topic Map of NIH Grants 2007. 5th iteration (2009): Science maps for science policy makers, places and spaces: Mapping science edn. Places and Spaces: Mapping Science, Bloomington, IN (2007) http://scimaps.org.
9. Herr II, B.W., Talley, E.M., Burns, G.A., Newman, D., La Rowe, G.: The nih visual browser: An interactive visualization of biomedical research. In: Proceedings of the 13th International Conference on Information Visualization (IV09), Barcelona, Spain, July 14-17, IEEE Computer Scoiety (2009) 505–509

10. Unemployment in the United States, 2004 to Present. http://projects.flowingdata.com/america/unemployment/.
11. An interactive map of vanishing employment across the country. http://www.slate.com/id/2216238/.
12. Where is the Money Going? http://www.recovery.gov/transparency/pages/home.aspx.
13. Where are the Jobs? — Indeed.com. http://www.indeed.com/jobtrends.jsp.
14. Job Map - CoolWorks.com. http://www.coolworks.com/job-map/.
15. MapYourJob.com : Find or Post your job on our map! http://www.mapyourjob.com.
16. JobMaps = Indeed Job Search + Google Maps. http://jobmaps.us.
17. 100 Best RSS Feeds for Recent College Grads. http://www.bestcollegesonline.com/blog/2008/08/28/100-best-rss-feeds-for-job-seekers/.
18. Science Jobs : Scientist Recruitment & Vacancies : Nature Jobs. http://www.nature.com/naturejobs/index.html.
19. Services - Google Maps API - Google Code. http://code.google.com/apis/maps/documentation/services.html#Geocoding.
20. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, Washington, DC, IEEE Computer Society (1996) 336–343
21. Furnas, G.W., Bederson, B.B.: Space-scale diagrams: Understanding multiscale interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Denver, Co., ACM Press (1995) 234–241
22. MarkerManager v1.0 Reference. http://gmaps-utility-library-dev.googlecode.com/svn/tags/markermanager/1.1/docs/reference.html.
23. Boyack, K.W., Klavans, R.: Map of Scientific Paradigms. 2nd iteration (2006): The power of reference systems, places and spaces: Mapping science edn. Places and Spaces: Mapping Science, Albuquerque, NM and Berwyn, PA (2006) http://scimaps.org.
24. Klavans, R., Boyack, K.W.: Is there a convergent structure to science? In Torres-Salinas, D., Moed, H., eds.: Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, CSIC (2007) 437–448
25. Maps of Science. http://www.mapofscience.com/.
26. Map Overlays - Google Maps API - Google Code. http://code.google.com/apis/maps/documentation/overlays.html.
27. Add Your Own Custom Map - Google Mapki. http://mapki.com/index.php?title=Add_Your_Own_Custom_Map.
28. Automatic Tile Cutter - Google Mapki. http://mapki.com/index.php?title=Automatic_Tile_Cutter.
29. Custom Google Maps. http://webtide.wordpress.com/2008/08/27/custom-google-maps/.
30. Google Maps JavaScript API Example: Tile Detector. http://code.google.com/apis/maps/documentation/examples/tile-detector.html.
31. LabeledMarker v1.3 Reference. http://gmaps-utility-library-dev.googlecode.com/svn/tags/labeledmarker/1.3/docs/reference.html.