How Can Geographic Information Retrieval Benefit from Geovisualization Principles?

Olga Buchel[1]
[1] INSIGHT Lab, Faculty of Information & Media Studies, University of Western Ontario

Abstract
This paper proposes to enhance geographic information retrieval interfaces with geovisualization techniques such as multiple representations and rich interactions. It supports its proposition by examples and illustrations of GeoPubMed prototype which was designed to enhance geospatial access to medical citations in PubMed database, the premier database of health publications. Such an enhancement results in expansion of user tasks which become more analytical than simple reference question-answering tasks.
Keywords: geovisualization, geographic information retrieval, multiple representations, analytical tasks, exploratory search.
Citation: Buchel, O. (2014). How Can Geographic Information Retrieval Benefit from Geovisualization Principles? In *iConference 2014 Proceedings* (p. 1066–1071). doi:10.9776/14383

Contact: obuchel@uwo.ca

1 Introduction

Geographic information retrieval interfaces are becoming common enhancements to search engines, social networking systems, bibliographic databases, digital libraries, image retrieval systems, and many other information systems. The majority of such interfaces, however, answer simple questions such as Where?, Who?, What?, and When? (Jones, 2007). Social networking systems go beyond these questions and add information about places, and opinions about them by means of linking user-contributed reviews to maps (Bilandzic et al., 2008). Questions Where?, Who?, When?, and What? along with reviews could be characterized as reference information. While answering reference questions is important for exploration of geographic localities, it is not the only type of information that geographic retrieval systems can and should inform users about. In contrast, interactive geovisualizations support a much wider range of exploratory tasks. These tasks could be beneficial to information seekers as well. Specifically, geovisualizations are concerned with analytical questions and tasks. For example, they focus on facilitating comparisons, associations, identification of similarities and differences, and other tasks (Andrienko & Andrienko, 2006). Such analytical tasks help researchers think, notice patterns and outliers, synthesize information, derive insights from massive data, and complete many other higher order cognitive activities.

These tasks could be beneficial to information seekers as well, especially when we take into account that many users of retrieval systems do not only come to find answers to reference questions but also to analyze published information. There is substantial evidence in research literature that researchers engage in analytical tasks as they analyze spatial diffusion of publications, spatio-temporal patterns, social networks, and research other aspects of documents (see, for example, recent studies by Merel et al., 2013; Pan et al., 2011; and Groneberg-Kloft et al., 2009). A search for "citation analysis" in PubMed retrieves around 3000 publications.

This paper outlines how blending analytical tasks with geographic retrieval tasks could facilitate detection of the expected and discovery of the unexpected phenomena in document collections. We do this on the example of GeoPubMed prototype that provides access to PudMed citations, one the most comprehensive collections of medical literature.

2 Background

Geographic information retrieval is an interdisciplinary concept that combines principles of information retrieval and digital maps. Early research efforts in geographic retrieval were focused on conceptualization of retrieval algorithms, geospatial relationships, automated indexing, development of gazetteers, modeling of metadata and ontologies, and conceptualization of basic interactions such as navigating, browsing, and searching (Hill, 2006; Larson, 1996).

Meanwhile, geovisualization researchers have been preoccupied with a different research agenda. They were concerned with exploratory tasks (Andrienko & Andrienko, 2006), knowledge discovery, visual analysis (MacEachren et al., 2010), and other higher level cognitive activities. Such tasks and activities required not only highly interactive maps, but also a range of additional representations that can inform users about changes in objects' attributes in space and time. Geovisualization tools help users make discoveries, generate hypotheses, reason, and understand. This paper demonstrates how geovisualization principles developed by geovisualization researchers can complement and improve representations and interactions with geographic information retrieval results.

3 GeoPubMed Prototype

GeoPubMed prototype is a search interface for PubMed that groups search results by locations. It retrieves names of countries, states/provinces from authors' affiliation fields in PubMed. Geographic names are retrieved with the help of the gazetteer (a dictionary of geographic place names that binds names with coordinates) that is used for query expansion. When users specify search statements, their queries are expanded with geographic place names; the number of results for each location and location coordinates are recorded in a KML file which is later used for visualization. The gazetteer contains only place names of locations that are present in PubMed; it does not have place names that do not have any matching documents in PubMed. The gazetteer is stored in MySQL database, the visualization is based on Google Maps interface with some additional representations and interactions developed in PHP, JavaScript, and D3.js. The prototype uses a number of parallel searches to expedite the retrieval.

3.1 Representations

The template works with versions of LibreOffice and OpenOffice.Visualization of results in GeoPubMed has multiple representations. We use a heatmap of all publications in PubMed as a base layer; a layer with graduated circles to show search results. Density of the heatmap represents density of results for each location. Currently the heatmap coverage is limited to the US. To generate this layer we queried PubMed with BGN gazetteer that has 159,608 populated places for the US. This helped us identify 19,538 placenames mentioned in PubMed. Other placenames were not found in medical citations and for this reason are not used for retrieval.

The results are distributed to multiple zoom levels that show countries, states/provinces, and smaller locations respectively. Each representation of results has an index of placenames, tightly coupled with the map. Both layers are shown in Figure 1.

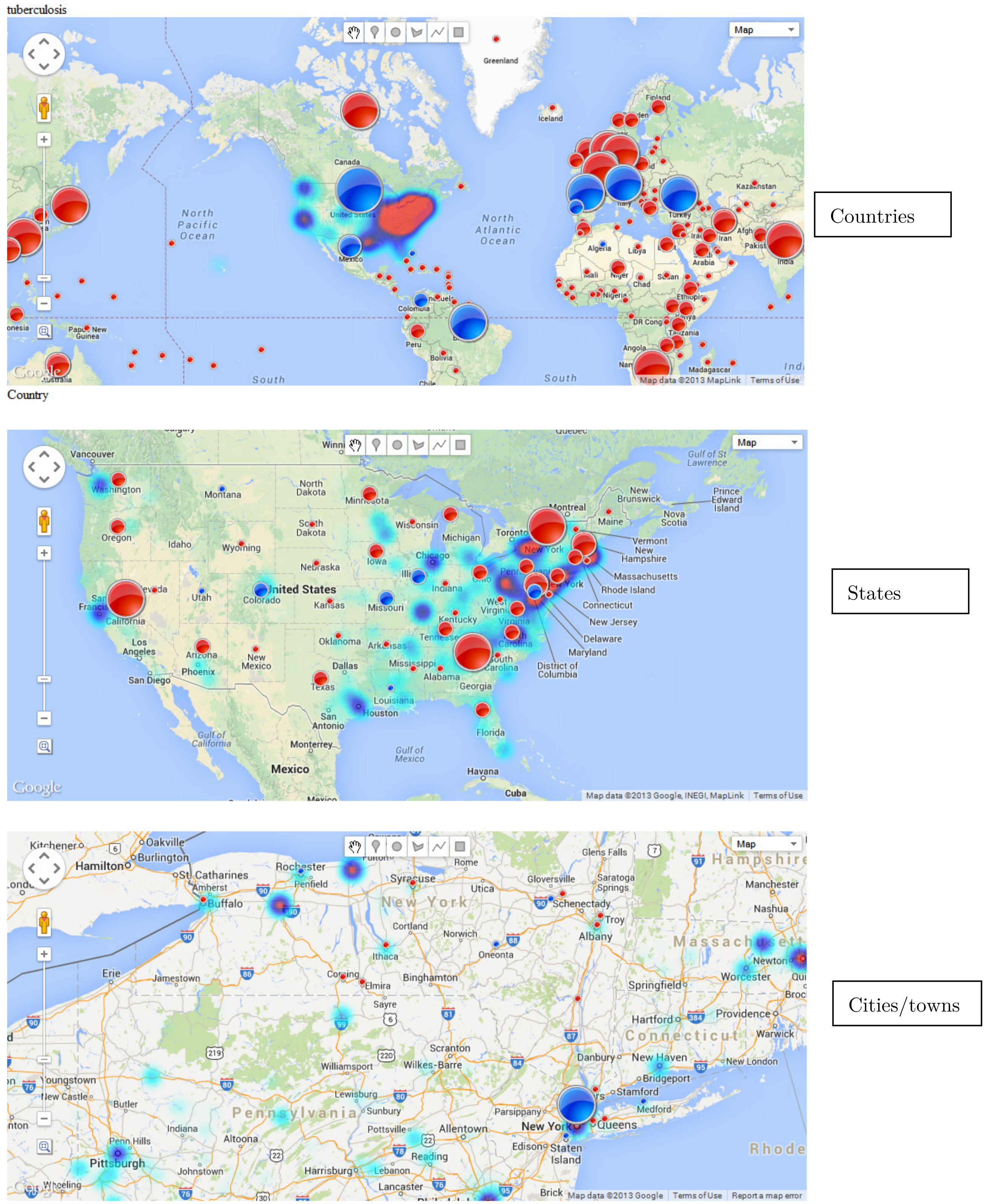

Figure 1: Search results for “tuberculosis”. Results are distributed to multiple zoom levels and are shown against the heatmap.

Graduated symbols represent central points of locations (countries, regions, and smaller locations). Symbols are linked to a series of representations that show additional aspects of documents such as clinical facets (Fig. 2.a), ages of patients (Fig. 2.b), and years of publication (Fig. 2.c).

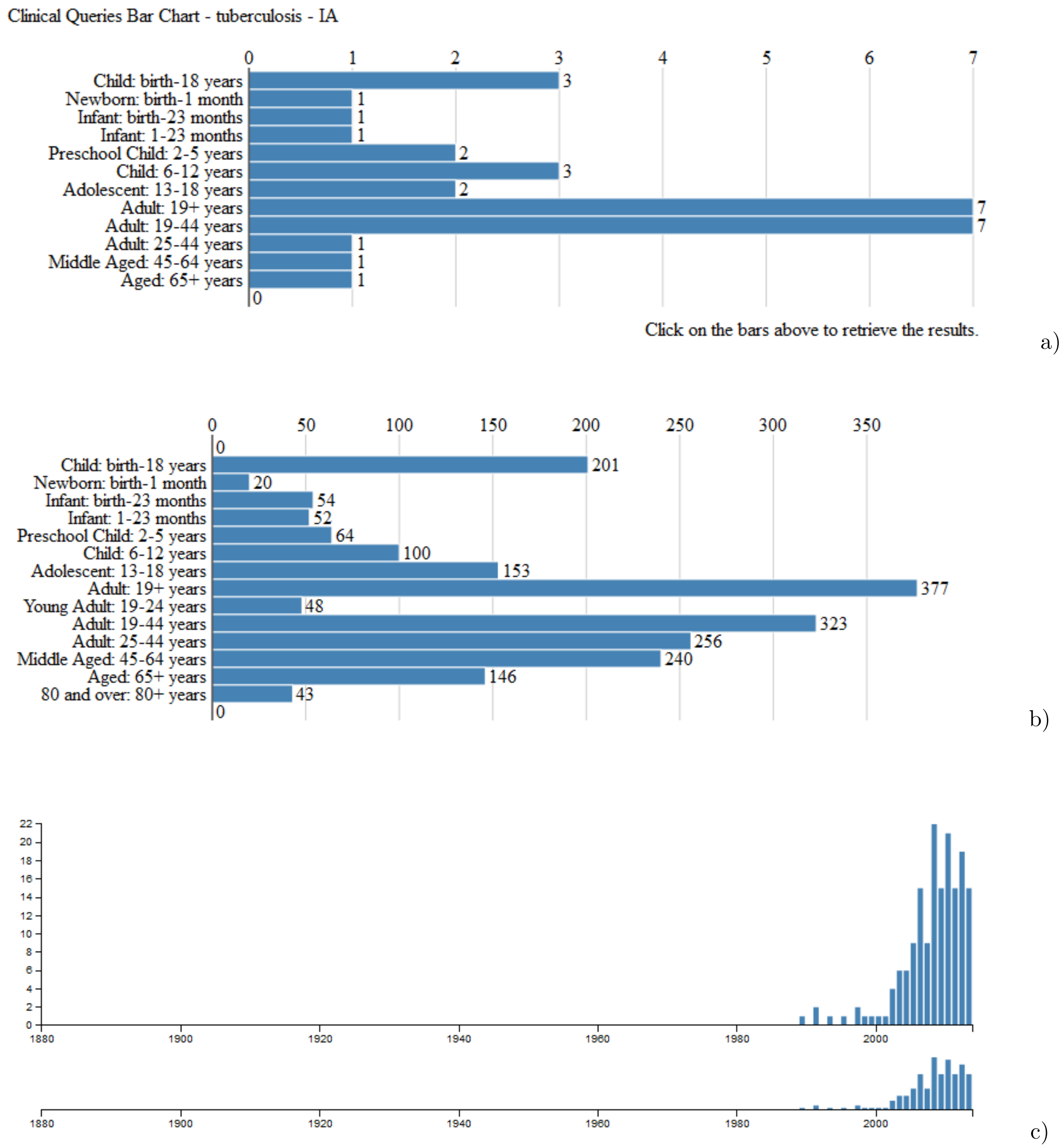

Figure 2: a) a bar chart showing clinical facets; b) a bar chart showing ages of patients described in medical citations; c) a column chart of years of publication.

3.2 Interactions

GeoPubMed has a number of interactions that increase the chance of discovery in the visualization. Besides searching, it supports different kinds of sorting in the location index, visual annotation of markers, interactive tours, linking, and zoom dependent selection. Map index can be sorted alphabetically and according to the number of search results. Markers on the map and locations in the index change their colors as users open them up. This helps users to keep track of their previous actions. In addition, markers and locations in the index can be deleted or highlighted as special to aid users annotate relevant and less relevant documents. Selection at the level of countries allows users to select multiple countries and view them through the lens of additional representations similar to the ones linked to markers. This facilitates answers about geographic location which are not explicitly represented in an information system. For example, with this selection, the results from European countries can be grouped together and users could get access to treatments in Europe in the representation of clinical facets. Selection affects ranking of search results: ranking for individual countries is different from ranking for multiple countries where results from different countries are mixed.

4 Knowledge Patterns

The approach described in this poster offers a number of benefits to information seekers. It allows them to complete a number of analytical tasks and discover interesting patterns about medical knowledge while they search for information. First of all, distributing the results to multiple zoom levels shows cognitively plausible patterns of diseases, treatments, social networks, and diffusion of authors publishing in the same journal which allows users generate research hypotheses about etiology of diseases, medical practices, and awareness of diseases. Users can easily identify hot spots or deserts of medical research at various levels of geographic specificity. They can explore distribution of treatments and draw inferences about how some treatments are more popular in some countries than in others. Researchers can study international collaborations and their publishing productivity. Together all these knowledge patterns may lead to new scientific discoveries and better understanding of public health issues.

Currently, a user study is underway evaluating the usability of knowledge patterns. The goal of the study is to assess merit of these patterns in analytical tasks and activities such as hypothesis generation, reasoning, spatio-temporal analyses and other activitie.

5 Conclusion

This paper demonstrates how merging geovisualization principles of multiple representations and rich interactions enhance tasks of information seekers. Specifically, their tasks become more analytical. These new tasks enable researchers not only to ask questions Where?, When?, Who?, and What? but also to discover and explore interesting knowledge patterns at various levels of geographic granularity (at the level of countries, provinces/states, and cities). These patterns may lead to new research questions, hypotheses, and new ways of thinking.

6 References

Andrienko, N., Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data*. Berlin, Germany : Springer.

Bilandzic, M., Foth, M., De Luca, A. (2008). CityFlocks: designing social navigation for urban mobile information systems. *Proceedings of the 7th ACM conference on Designing interactive systems*. ACM.

Groneberg-Kloft, B., Scutaru, C., Fischer, A., Welte, T., Kreiter, C., & Quarcoo, D. (2009). Analysis of research output parameters: density equalizing mapping and citation trend analysis. *BMC health services research*, *9*(1), 16.

Hill, L.L. (2006). Georeferencing : The Geographic Associations of Information. Cambridge, MA: MIT Press.

Jones, M. (2007). Google's geospatial organizing principle. *Visualization Viewpoints* , 8-13.

Larson, R.R. (1996). Geographic Information Retrieval and Spatial Browsing. In L. C. Smith, & M. Gluck (Eds.), *Geographic Information systems and Libraries: Patrons, Maps, and Spatial Information* (pp. 81-123). Urbana-Champaign: GSLIS, University of Illinois at Urbana-Champaign.

MacEachren, A.M., Stryker, M.S., Turton, I.J., Pezanowski, S. (2010). HEALTH GeoJunction: place-time-concept browsing of health publications *International Journal of Health Geographics* 2010, **9**:23 (18 May 2010) .

Merel, S., Villarín, M. C., Chung, K., & Snyder, S. (2013). Spatial and thematic distribution of research on cyanotoxins. *Toxicon*, *76*, 118-131.

Pan, R. K., Kaski, K., & Fortunato, S. (2011). World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, *2*, 902-902.

7 Table of Figures