UrbanForest: Seeing the data forest despite the trees

Preprint · October 2020

DOI: 10.13140/RG.2.2.22300.72325

2 authors:

Akm Zahirul Islam
New Jersey Institute of Technology
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Aritra Dasgupta
New Jersey Institute of Technology
51 PUBLICATIONS **816** CITATIONS
SEE PROFILE

To appear in IEEE VIS Posters 2020.

UrbanForest: Seeing the *data forest* despite the trees

Akm Islam*and Aritra Dasgupta†
New Jersey Institute of Technology

ABSTRACT

Data is generally considered as a starting point in analytical processes. But difficulty in finding relevant data often precedes downstream processes and can become a bottleneck for the intended analytical outcomes. For example, open data repositories from many major cities contain thousands of data sets that domain experts can potentially use for deriving insights; but they are confronted with the challenge of *too many* alternatives, either not knowing where to start or having to manually download each data set for further investigation. We argue that that *data discovery* can be conceived as a visual analytic task, where, by using guided analysis, domain experts can interactively discover associations across many data sets. We realize the data discovery task in *UrbanForest*, an interactive visual interface, that lets analysts find and link openly available data sets from a number of data portals using visual representations of attribute co-occurrence and similarities across data sets at the meta-data level. We envision *UrbanForest* being especially useful for domain experts, like urban planners and social scientists, who can formulate and test alternative data-driven hypotheses without having to download individual data sets.

1 INTRODUCTION

Visual analytic techniques generally work on the premise that analysts have data sets at their disposal for wrangling, modeling, learning, visualization, etc. However, in many real-world scenarios, the problem of *data discovery* precedes all downstream analytical tasks: analysts spend significant time and effort for finding useful data even before they can start analyzing them. For example, social scientists often need to work with open data portals with the goal of finding relevant data that help them formulate or address emerging hypotheses about urban societies, related to education, economics, crime, etc. The transformative potential of big data in shaping new theories and paradigms in social science has been widely acknowledged, while at the same time, it has been emphasized that experts' knowledge need to be tightly coupled with computational methods for maximizing technological outcome.

Open data portals, such as the ones from New York City and Chicago [1, 5], while characterized by the volume and potential value of the data sets, offer very little support for social scientists to systematically navigate the data landscape. In theory, a treasure trove of open data exists, which if semantically linked and analyzed, can shed light on social systems ina much more holistic way than possible before [7]. For example, healthcare policy-makers can connect data that contain economic indicators with hospital data about disease and treatment patterns for better targeting policies based on people's needs, requirements, and affordability [6]. School and college administrations can work towards more inclusive education policies by better understanding racial and gender disparities using data related to rankings, admission policies, and student demographics and preferences [4]. Researchers studying social behavior can use openly available healthcare data and compare actual disease statistics to disease perception and opinions as shaped by the media [3, 9] for flagging unreliable reporting trends.

However, for performing such nuanced and deep analyses, availability of open data is only a necessary first step. Visual analytics can potentially lead to a greater return on investment of exterts' time from their interaction with the open data portals. The high-level needs from analytical solutions is to facilitate sense-making, whereby experts can formulate new hypotheses about urbal planning, social science, etc. Appropriate levels of transparency needed so that experts can incoporate their domain knowledge while benefitting from analytical solutions [2]. Transparency is needed to both seamlessly access relevant or linked data and also get guidance from the system that can lead to surprise findings, in other words, discovering the unexpected [8].

To address these needs, we present UrbanForest, with the vision that people with a non-technical background can seamlessly access open data sets and quickly discover relevant data for their own analysis in an integrated manner, without having to manually download and analyze individual data sets.

2 DATA DISCOVERY USING URBANFOREST

We address the challenge of scale and complexity that analysts face while working with open source data repositories. There are too many datasets and various categories that could be of interest, and they are available, yet, there is little guidance on how they could be more accessible for answering specific analysis questions (e.g., what is the connection among health, crime and education patterns in across neighborhoods with particular income levels?). Understandably, answering such questions require semantically integrating diverse attributes. Interfaces like the NYC open data and Boston open data serve as collections with immense value for advancing social science research or evaluating alternative urban policies. However, to tap into the value of these collections, currently, analysts spend a lot of manual effort for reconciling information from many data sets. This process is not only time-consuming but could also be counter-productive with many explorations leading to analytical dead ends. It is impossible to quickly find hidden associations among data sets without manually downloading data sets and post-processing them using analytical tools.

UrbanForest aims to minimize the human effort in data discovery tasks and drastically increase analysts' productivity and return on investment of time for data-driven analysis. We envision the main users of UrbanForest to be researchers in social science or humanities and policy-makers who are interested in developing new hypothesis and models about urban environments, specifically on the cross-cutting issues of smart cities, sustainable development of cities, energy and environmental implications of smart cities, issues of social equity in cities and their causes, and improving access to healthcare and education facilities. UrbanForest is also geared towards the goal of democratization of urban data. We envision reaching thousands of users, many of whom might not be domain experts, but laymen trying to better understand their neighbourhood characteristics from the data they can access.

UrbanForest saves the time and effort spent by analysts in seeking data by providing them with a holistic picture of "what is out there","how they are related", and "what could be interesting". We currently have a prototype implementation of the first two aspects,

*e-mail: azi3@njit.edu
†e-mail: aritra.dasgupta@njit.edu

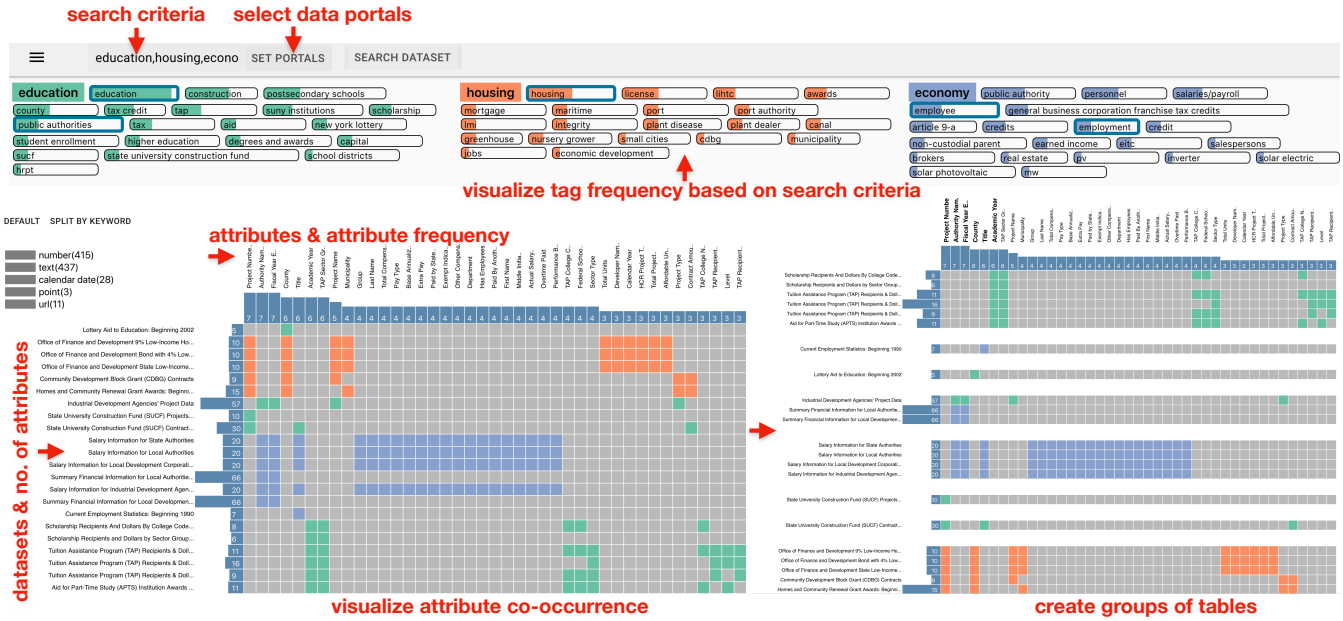

Figure 1: **The UrbanForest interface accessible at datadiscovery.njitvis.com** comprises: a tag frequency view, that shows the number of tables described by particular keywords that are associated with users' search tags (education, housing, economy), a matrix view, where each row represents a data table and each column represents an attribute and the colors represent the individual search tags; and a group view: where, a user selects a number of attributes and finds the groups formed by different combinations of the chosen attributes.

while the last aspect, based on suggesting interesting data sets is work in progress. We illustrate the functionalities of our prototype (accessible at *datadiscovery.njitvis.com*) with respect to the following three key tasks (Figure 1).

Describe search tags: Each data set in a data portal is described by a set of tags. In our tool, we automatically extract these tags for each data set and are a part of the meta-data specification for the data set. Tag frequency indicates the number of times a specific tag is used to describe the retrieved data sets, independent of the search keywords. For example, let us say we search by keywords, health and education, and 3 data sets, h1, h2, h3 are returned for the health keyword and 2 data sets, e1 and e2 are returned for the education keyword. Further, the tags for data set h1 are mental health and child care and those for e2 are mental health and education. In this example, the tag frequency of mental health is 2. Although h1 and e2 are retrieved for satisfying two different query keywords, they have the same tag. Hence we count their net occurrence as the tag frequency. For visualizing this, we use the tag frequency view on the top, as shown in Figure 1. The colors indicate the different keywords and the filled area indicates the relative frequency of the tags. We normalize the frequency values on a scale of 0 to 1 by using the maximum frequency value across the tags.

Visualize attribute co-occurrence: User can select multiple tags and dynamically understand co-occurrence patterns of attributes in a matrix view (Figure 1). The matrix is optimized with respect to information density such that the top-left cells mostly indicate attribute presence (color is indicative of a search keyword) instead of absence (gray color). The height of a bar over each column indicates the frequency of that attribute across all datasets in the matrix and the width of a bar by the side if each row indicates the total number of attributes for each dataset.

Create groups of tables: From all data sets that are associated with the selected tags, user can create groups or data subsets with shared attributes. As shown in Figure 1, on the right, user can select a number of attributes and groups get created based on which combinations of attributes are present in each group. We check, for every possible subset from the chosen set of attributes, whether that attribute is present or absent in a dataset. Then we create groups based on shared attributes. If a group has only one dataset, that means only that dataset contains the specific combination of attributes from the chosen set.

These tasks ultimately result in a merged dataset which a user can download as a csv file and continue the analysis offline. We plan to integrate more implicit and explicit suggestion mechanisms for guiding the analysts in creating these semantically linked datasets based on their interests and preferences.

References

[1] L. Barbosa, K. Pham, C. Silva, M. R. Vieira, and J. Freire. Structured open urban data: understanding the landscape. *Big data*, 2(3):144–154, 2014.

[2] R. Barcellos, J. Viterbo, L. Miranda, F. Bernardini, C. Maciel, and D. Trevisan. Transparency in practice: using visualization to enhance the interpretability of open data. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, pp. 139–148. ACM, 2017.

[3] B. Combs and P. Slovic. Newspaper coverage of causes of death. *Journalism Quarterly*, 56(4):837–849, 1979.

[4] K. N. Gulson and S. Sellar. Emerging data infrastructures and the new topologies of education policy. *Environment and Planning D: Society and Space*, 37(2):350–366, 2019.

[5] M. Kassen. A promising phenomenon of open data: A case study of the chicago open data project. *Government Information Quarterly*, 30(4):508–513, 2013.

[6] P. Kostkova, H. Brewer, S. de Lusignan, E. Fottrell, B. Goldacre, G. Hart, P. Koczan, P. Knight, C. Marsolier, R. A. McKendry, et al. Who owns the data? open data for healthcare. *Frontiers in public health*, 4:7, 2016.

[7] D. A. McFarland, K. Lewis, and A. Goldberg. Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, 47(1):12–35, 2016.

[8] J. J. Thomas and K. A. Cook. A visual analytics agenda. *Computer Graphics and Applications, IEEE*, 26(1):10–13, 2006.

[9] M. E. Young, G. R. Norman, and K. R. Humphreys. Medicine in the popular press: the influence of the media on perceptions of disease. *PLoS One*, 3(10):e3552, 2008.