Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A brute force method for spatially-enhanced multivariate facet analysis

Anthony C. Robinson[a,*], Sterling D. Quinn[b]

[a] GeoVISTA Center, Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802, USA
[b] Department of Geography, Central Washington University, 301 Dean Hall, Ellensburg, WA 98926, USA

ARTICLE INFO

Keywords:
Spatial analysis
Geovisual analytics
Faceted search
Computational methods

ABSTRACT

Faceted search is a common approach for helping users query multivariate data. While the method is found widely in contemporary tools, so far there has been little exploration of its potential to incorporate a spatial perspective. In this article we extend multivariate faceted search through the application of a brute force computational process to reveal facet combinations that have spatially-interesting results. We explore the potential utility of spatially-enhanced facet combinations in case study analyses of multivariate spatial data from learners in a massive open online course and multivariate spatial data from restaurant inspections. Spatially-enhanced facet combinations improve on ordinary faceted search by helping analysts understand which combinations have significant spatial footprints. We also show how this method can be integrated into a geovisual analytics system through a simple user interface. Finally, we draw on our case study analyses to highlight important challenges and opportunities for future research.

1. Introduction

Approaches for exploring high-dimensional spatial data often invite an analyst to forage through geographic features and their attributes with the goal of constructing insights and knowledge about particular phenomena (Andrienko et al., 2007; Pirolli & Card, 2005; Sacha et al., 2014). This usually involves manipulating user interface (UI) components such as range sliders and dropdown menus to perform dynamic queries (Shneiderman, 1994). In contemporary website design these combinations of query parameters are often symbolized on the interface using a technique called faceted search. As a user makes selections to narrow items of interest across different categories, each categorical choice is represented as a facet that can be toggled on or off (Fig. 1).

Faceted search alone is a simple and powerful way for users to develop and refine queries for simple tasks like finding products to buy, but in a multivariate analysis context with spatial data where the goal is to identify and evaluate interesting patterns, users are likely to miss interesting combinations as the method does not support an exhaustive approach. For example, in a census data analysis context a user may string together a query that retrieves all counties that have population above a certain median age, an average income above the poverty level, at least one college or university, and more than two hospitals. In this case, the faceted search method by design leads analysts down increasingly narrow paths to an outcome, and it relies on users to self-recognize when they have narrowed criteria too far to make a meaningful discovery. In the present work we draw upon prior work to develop so-called *multivariate faceted search*, which does not assume the independent categorical hierarchy that is found in traditional faceted search (Ben-Yitzhak et al., 2008). Multivariate faceted search supports exploratory analysis of variables that may be correlated with one another. The approach we present here attempts to help users avoid the problem of narrowing queries into empty sets by using computational methods to automatically extract patterns, and then employing interface cues to direct user attention toward those patterns. In this paper we describe a new approach that combines the simplicity of faceted queries with computational analysis to identify facet combinations that have significant spatial clusters. Specifically, we outline a method for batch computing faceted queries and feeding the most fruitful results into a LISA (local indicators of spatial association) analysis to determine which facet combinations result in interesting spatial patterns. Based on our review of the literature, we have found no prior work proposing similar approaches for spatially-enhancing facet analysis. The use of faceted search is widespread in spatial analysis applications, and our work provides a novel approach that allows analysts to conduct multivariate exploration based not only on attribute combinations, but also based on whether or not that pattern has any spatial significance. In case studies examining student engagement with a massive open online course (MOOC), and high-dimensional data on restaurant food safety violations in New York City, we show how our method can be used to help analysts explore facet combinations that have significant spatial patterns.

In addition to extending the basic method of faceted search to incorporate the spatial dimension explicitly, we believe this approach has

* Corresponding author.
E-mail addresses: arobinson@psu.edu (A.C. Robinson), sterling.quinn@cwu.edu (S.D. Quinn).

https://doi.org/10.1016/j.compenvurbsys.2017.12.003
Received 14 June 2017; Received in revised form 18 December 2017; Accepted 24 December 2017
0198-9715/

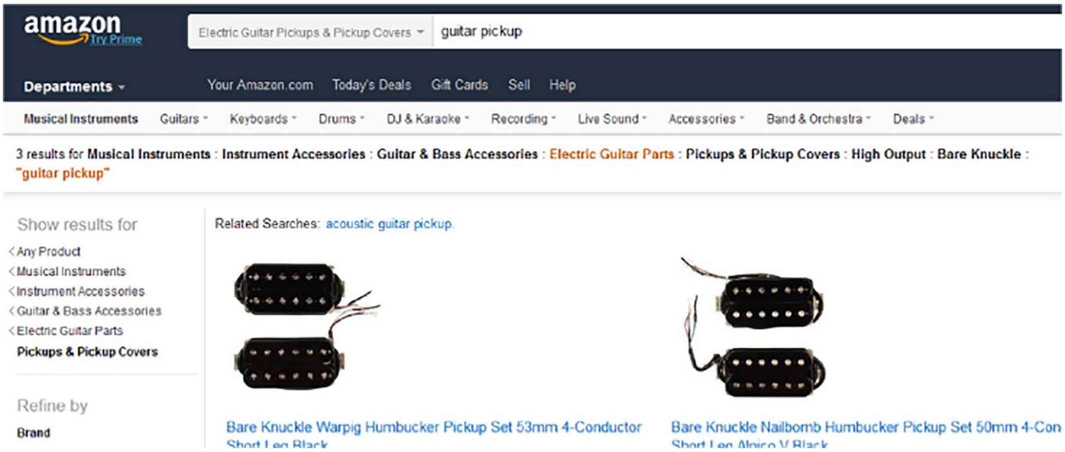

Fig. 1. Faceted search as implemented on Amazon.com.

potential advantages over black box spatial data mining methods in that its results may be more easily explained to non-expert users and can be implemented into a simple visual analytics system interface. Our experience working with this method and exploring its utility in two case studies suggests that the method can help reveal interesting patterns in multivariate spatial data by calculating and highlighting the spatial significance of a faceted search, and in doing so can help save analysts the trouble of manually exploring a very large number of combinations on their own, which is impractical for most realistic data analysis contexts. We also show that the computational burden for calculating facet combinations is not a major impediment to utility in the context of our two case studies. Furthermore, the spatial facet combination approach is possible to integrate with geovisual analytics systems, as we show in our case study analysis of learner data from a massive open online class.

In the sections that follow we explore related research on computational approaches for pattern analysis with spatial data, with a focus on brute force methods and their potential connection to faceted search. We then provide a methodological design for the brute force calculation of spatially-enhanced facet combinations and highlight its potential utility through its application to two case study datasets. Next we reflect on what we have learned from implementing spatially-enhanced facet combinations, including potential limitations of this approach associated with the scale and complexity of a given dataset. Finally, we propose future research efforts that could tackle these concerns as well as other new opportunities that are prompted by our progress so far.

2. Related work

This research extends past work on computational approaches for supporting pattern analysis of spatial information, the use of faceted search as a common means for information retrieval, and the incorporation of spatial components into linked data frameworks.

The use of data mining and spatial computation techniques as a "first pass" to guide users of visualization tools was suggested by Shneiderman (2002), who hypothesized that metrics for the degree of clustering could help users decide what to explore. Beale (2007) noted that today's data deluge makes it difficult to identify what is truly interesting or noteworthy. Beale describes the calculation of a "surprise factor" to guide which data should be prioritized as interesting for users of an information visualization system. In GIScience, Guo, Peuquet, and Gahegan (2002) offered methods for interactive exploration of spatial clusters, and incorporated an entropy-based metric evaluating the "interestingness" of subspaces in the dataset. A very wide range of approaches are available now for conducting spatial data mining (Shekhar, Evans, Kang, & Mohan, 2011). A popular approach today is to fuse computational methods for pattern discovery in spatial data with visual interfaces to support analytical reasoning – a subdomain of GIScience known as geovisual analytics (Andrienko et al., 2010; Andrienko et al., 2016). Recent work in geovisual analytics has sought to understand mobility patterns in large spatio-temporal datasets (Landesberger et al., 2016), spatial aspects of social media in crisis situations (MacEachren et al., 2011), and periodicity in spatiotemporal event sources (Swedberg & Peuquet, 2017). Ongoing geovisual analytics research involves many additional domains.

We complement these efforts through the development of a batch computation of queries that identifies spatially-clustered facet combinations and suggests these patterns for exploration to the end user in a simple way. In the next two sections we explore progress on developing brute force approaches and their applications in spatial analysis, as well as the development of faceted search as a popular paradigm for information retrieval.

2.1. Brute force analysis

Computational data analysis approaches can include the development of models that abstract data and relationships, or sampling to use data subsets in order to extract and extrapolate patterns. They can also include so-called brute force approaches, also known as exhaustive search methods. Brute force methods offer the advantage of exploring all possible outcomes (Gonzalez, Diaz-Herrera, & Tucker, 2014), but can see increased computational processing times with very large or complex datasets. In spite of that disadvantage, brute force approaches are not irrelevant to contemporary spatial analysis contexts. While much attention today is placed on massive, streaming data sources of increasing dimensional complexity in the era of big data (Gandomi & Haider, 2015) and spatial big data (Li et al., 2016; Robinson et al., 2017; Tsou, 2015), there remain plenty of scenarios in which geographical analysis concerns less daunting data sources, for example, in the contexts of demographic analysis with census information, or public health analysis of patient records. In these cases it is often not essential to avoid any pre-processing steps, as the context for analysis concerns retrospective investigation or long-term analysis. In addition, the rise of parallel computing and scalable cloud computing approaches open the door a bit wider to spatial analysis methods that are exhaustive in nature, as it becomes increasingly possible to scale computer power to a problem on demand.

There are a number of precedents in GIScience that apply brute force approaches to spatial analysis problems. For example, Openshaw's Geographical Analysis Machine (GAM) exhaustively searches geographic space at many scales to search for clusters (Openshaw, Charlton, Wymer, & Craft, 1987). Successors to GAM include the widely utilized SaTScan space-time cluster detection method, which also employs an exhaustive approach (Kulldorff, Rand, Gherman, Williams, &