A Visualization-Analytics-Interaction Workflow framework for Exploratory and Explanatory Search on Geo-Located Search Data using the Meme Media Digital Dashboard

Jonas Sjöbergh, Xingkai Li, Randy Goebel, Yuzuru Tanaka
Meme Media Lab, Hokkaido University, Sapporo, Japan
Alberta Innovates Centre for Machine Learning, University of Alberta, Edmonton, Canada
{js,tanaka}@meme.hokudai.ac.jp, {xingkai, rgoebel}@ualberta.ca

Abstract

Modern geo-position system (GPS) enabled smart phones are generating an increasing volume of information about their users, including geo-located search, movement, and transaction data. While this kind of data is increasingly rich and offers many grand opportunities to identify patterns and predict behaviour of groups and individuals, it is not immediately obvious how to develop a framework for extracting plausible inferences from these data. In our case, we have access to a large volume (more than half a billion individual records) of real user data from the Poynt smart phone application, and we have developed a generic and layered system architecture to incrementally find aggregate items of interest within that data. "Interest" is based on the semantics of the data, so include time and space correlations, e.g., are people searching for dinner and a movie; distributions of usage patterns and platforms, e.g., geographic distribution of Android, Apple, and BlackBerry users; and clustering to identify interesting and relatively complex search and movement patterns, e.g., consumer trajectories from key word searches.

Our integration of visualization tools is thus guided top-down, by semantic concepts in the application domain, rather than by bottom-up tool development. Our presentation here is preliminary in that we provide sketches of case-studies that demonstrate an application specific integration of the three major components of modern visual analytics: visualization, analytics, and interaction (VAI).

Our case-study sketches show how an interactive system for visual data exploration can be used to alternate between exploratory search – looking for ideas and new hypothesis in data – and explanatory search – looking for evidence to support a hypothesis. While we have not yet formulated experiments to directly measure the cognitive efficacy of our experimental system, we believe that our semantically-driven VAI workflows and the integration of visual methods and interaction provides some useful ideas about how to extend current frameworks for visual analytics systems.

Keywords— **Visualization, Analytics, Interaction, Geo-Located Search**

1 Introduction

An increasing number of sources of large data are now publicly available (*e.g.*, [8]). But in many cases, despite the intuition that these data hold valuable inferences, one may not have a clear idea of how to model those data, or exactly what kinds of valued information could be extracted. In such cases, being able to freely explore the data is useful, and because of data volumes, and like many current visual analytics researchers (*e.g.*, [21], [13], [20]), we believe visual exploration and interaction with visualizations of data to be important — even necessary.

A good example of large publicly available data is a variety of geo-coded data from smart phones (*e.g.*, [5]), which provided the basis for a spatial visualization of data, time, location, and activities associated with those time and geo-spatial coordinates. In our case, we have access to a significant volume of search records from the popular smart phone application Poynt[1], which provides about 20 million gps-enabled smart phone users with the ability to access a variety of data, including business and private phone numbers, restaurants, events, movies, and gas stations, all indexed by geo-location of the handset user. For example, a Poynt user can search for the cheapest gas within a specified distance from a specific location.

Each individual use of one of these geo-located searches creates a search record (described below), which provides that user's location, time of search, and category of search (*e.g.*, gas station, movie, restaurant, and a variety of others). In our case we have over 531 million individual records of Poynt user search records, which include

[1]See www.poynt.com

1550-6037/15 $31.00 © 2015 IEEE
DOI 10.1109/iV.2015.60

both searches across several categories (*e.g.*, gas, movies, restaurants) as well as Yelp[2] classified keyword searches.

Our general motivation is to investigate a potential framework for deploying analytics on the user search records, to find potential business value. Broadly speaking, the potential business value lying behind the rapidly accumulating search records (about 20,000,000 records per day) is that associated with a variety of user profiling initiatives, *e.g.*, the suggestions of Amazon. The difference here is that, in addition to preferences for products (*e.g.*, books, movies), there is extra information in terms of time and place of request across the spectrum of business places, personal phone numbers, events, movies, restaurants, *etc*. Monetizing the potential value of these data is similar to the challenge of online advertising placement, but with the added complexity of geo-location, including individual user movement.

In the general analytics research community that considers geo-located data and events, the focus of value has been in identifying geographic trajectories based on large volumes of geo-coded data (*e.g.*, [7]).

Here we are interested in a more general framework for understanding such data, including those motivated by identifying patterns that may provide business value for the application provider. The Poynt data can be used for improving search services and for serving advertisements that are likely to be relevant to the user (*e.g.*, avoid showing advertisements for things that are not available in the city the user resides); but perhaps there are many other interesting things hiding in the data?

Our overall methodology is to use use the Poynt data to help guide the coupling of a variety of visualization, analytical, and interactive (VAI) tools embedded within our general purpose *Digital Dashboard* visual exploration system. Our system allows the federation of many data sources ("data mash-up") and it supports interaction with all visualization results. As we will show in some mini-case studies, it is easy to do *exploratory searches*, which are searches where general visual inferences can be applied to expose a next "step" or cycle of visualizations guided by semantic domain constraints. In those cases, a user wants to explore data to find interesting artifacts which lead to ideas for new hypotheses about that data. For example, one might ask to visualize the distribution of iPhone, Android, and Blackberry users within some specific region, in order to observe any interesting emerging patterns. A user can also switch to *explanatory searches*, where a particular hypothesis might be confirmed or denied, depending on the available data, *e.g.*, to confirm the hypothesis that New York iPhone users search for movies more than New York Blackberry users. And then, perhaps most importantly, a user can alternate between the two types of searches. We hope to help articulate these kinds of VAI workflows by making them explicit, which we believe will help identify classes of VAI workflow relevant to specific visual analytics problem solving challenges.

The remainder of our presentation is organized as follows. The next section introduces the structure and content of the Poynt data, both to expose its richness, but also to provide some intuition for the complexity of the potential inferences that may emerge from the semantics of such data. This is followed by a brief summary of the general themes of research on visual analytics, where the major components of producing visualizations from data, applying analytics techniques, and interacting with those visualizations provides the basis for making visual inferences. Then follows the more detailed description of the background of our digital dashboard, and its configuration for our particular framework for VAI workflow. Included here are case study segments intended to show how the dynamic switch between exploratory and explanatory VAI workflow can help expose inferences otherwise left unrevealed.

We conclude with a summary, and some discussion about what priorities might be to exploit this preliminary investigation in a more formal framework.

2 Data Description

The data we use in our examples is collected through the smart phone application Poynt[3]. The application has about 20 million users and it allows access to and searching of a variety of data, including business and private phone numbers, restaurants, events, movies, and gas stations. Search results can be ordered by geo-location, for instance by shortest distance to the location of the user when searching (using the GPS of the smart phone).

Each search generates a search record that contains the **user ID** of the user doing the search, the **location**, the **time**, the **device type** (iPhone, Android, etc.) and the **search category** (*movie*, *restaurant*, etc.). Some records also have a search query string, the number of results returned by the search, and the action taken by the user upon seeing the results (for example calling the phone number of one of the search results, asking for a route on the map to the location of the result, or clicking on a Web link).

About 20 million records per day are generated. Our data includes over 531 million records collected over three different five week periods, but here we concentrate only on about one third of that data, taken from five weeks from the summer of 2011. In addition, our descriptions here constrain the geography of our data to a rectangular area containing the southern part of Ontario in Canada and

[2]Classified directory search, for example see www.yelp.ca

[3]http://www.poynt.com

Pennsylvania and New York in the USA.

3 A brief summary of visual analytics

Current research in exploratory visual analytics informs our development of a framework for visual analysis of commercial geo-position data (*e.g.*, [22]). In addition, the semantics driven top down approach to the choice of tools and technique integration is similar to that of [9], where New York taxi data is used to drive a variety of VAI workflow structures. Note that, while we are still developing a perspective on how to turn application independent notions of visual analytics in logics of visualization, we find these specific applications and the semantics of their domains more informative than those generalizations derived from a more abstract history of visual analytics and the VAI workflow framework (e.g., [1, Ch.5]).

Like us, other research that distinguishes explanatory and exploratory search have noted the need for exploratory search, typically as a preliminary step to understand large volumes of data (*e.g.*, [12], [21]). From our viewpoint, exploratory visualization is a good first step in becoming familiar with large data, especially with the integration of visualization and machine learning (*e.g.*, clustering) techniques. But our experience with our digital dashboard and the complexity of the Poynt data is that the visual analytics workflow must provide for rapid and frequent alternation between exploratory and explanatory interaction, much like that related to those techniques provided within text and knowledge domains of interactive abductive diagnostic reasoning (*e.g.*, [14]).

Our observation is that there is acknowledged value in distinguishing "exploratory" visual search from "explanatory" visualization search, but that the nature of the human interaction is such that there is currently little precision in the distinction. For example, Wood [21] speaks to exploratory visual analysis, and emphasizes exploration but with a complete pipeline of transforming data into pictures, where interaction is used to make selections on a variety of dimensions. But as stated in the conclusion "Some understanding was achieved ... " which is positive, but never explicitly crosses the exploratory boundary to consider confirmatory visual evidence or explanation for a hypotheses. The point is not that their system fails to provide methods to identify visual evidence for a hypothesis on that domain's data; but rather that the boundaries are not clear. This is also the case for two other significant systems ([4], [9], [2]), where the general term "exploration" provides a basis for exposing a variety of interesting visualization rendering and interaction techniques, but with no clear boundaries on switching from exploratory to explanatory visual search.

We believe this is a symptom of the state of the art, rather than a fundamental deficit of the cited work. We note that there are a variety of existing systems and techniques that provide components suitable for use in each of the visual analytics workflow segments. For example, visualization itself is rarely about merely redrawing base data in a picture space, but often uses a pipeline of both syntactic and semantic transformations before actually rendering a picture (*e.g.*, [3]). The syntactic components of a visual analytics framework have a long history (e.g.,[16], the evolution of visual analytics clearly requires more emphasis on the semantics of visual manipulation (e.g., [1]. Similarly, the analytics phase of visual analytics often works directly on alternative visualizations (*e.g.*, [6], [10]). And finally, the wide range of visual interactions can be both syntactic and semantic (depending on the domain of the base data), and beg the issue of things as basic as Norman's interaction cycle (*e.g.*, see [19, Ch. 5]).

To clarify our own mini-case studies and the demonstration of the value of switching between explanatory and exploratory work, we first use a simple definition of visual analytics (taken from [22, P. 174]) and use it to discuss a simple visualization workflow, within which we can distinguish a variety of both exploratory and explanatory visualization and interaction.

The definition we adopt is simple:

Visual Analytics consists of three aspects of transforming and manipulating data to support humans drawing inferences from that data:

visualization
: which is the transformation of base data or any of its abstractions to some kind of rendering as pictures or visualizations in 2, 3, or 4D space,

analytics
: which is the transformation of base data or any of its abstractions by machine learning or analytics methods, to expose relationships otherwise left implicit, and

interaction
: which provides a human interactor with a repertoire of tools and methods to adjust, change or otherwise interrogate the visualizations arising directly from visualizations or from their analytical transformations.

We abbreviate this definition of visual analytics as VAI, so that we can refer to the idea of VAI workflow. Each component of visual analytics can provide a repertoire of tools and techniques for each of V, A, and I, which can be deployed in a variety of workflows, to provide a human interactor the basis for doing visual analytics on the base data of interest.

4 Components of VAI workflow

The potential repertoire of visualization, analytics, and interaction tools (VAI) is vast. Just the number and variety of papers, systems, and ideas on how to transform base data into pictures is overwhelming; the same is true for analytics tools, which are as broad as the whole spectrum of machine learning techniques. Perhaps the most constrained category is interaction, only because the spectrum is at least constrained by the capabilities of human interactors and the technologies to support interaction with data (*e.g.*, touch screens, 3D motion sensors).

Because of this broad variety of potentially useful methods and tools, we believe that the selection of a small set of VAI methods should be constrained by the semantics of the domains that are under investigation. So in the case of visualization tools for the Poynt data, an initial focus is on linking space, time, and search term data. And a guiding abstraction model for that data is to consider individual and aggregate behaviour of individuals generating that data, *e.g.*, to try and identify instances of consumer trajectories.

In the case of visualization tools, we first consider how an important foundation of such semantically guided top down design first requires a foundation arising from the presentation of multiple linked views (*e.g.*, [18], [15]), as well as interaction and direct manipulation (*e.g.*, [11, 18, 15, 16]) which facilitate the process of visual exploration of data. These techniques provide the bottom up basis for any VAI workflow. In our case here, the analytics tools emerge when the multiple linked views are dynamically adjusted by human interactors, who can control time series, spatial and position linked views, to reveal a variety of clusters that can be hypothesized as components of the abstract idea of consumer trajectories.

Roberts [15] surveyed the area of Coordinated Multiple Views (CMV), namely Multiple Linked Views. That work introduced Coordinated Multiple Views as a specific exploratory visualization technique, from which users may find insightful relationships and features from target data.

Interaction with quick visual feedback is indispensable to visual exploration techniques. As described in [15], a large variety of interaction strategies are integrated by CMV systems, in which users can interact with data in various ways, including both indirect direct manipulation. Indirect manipulation includes the idea of *dynamic queries*, which allows a user to interact with sliders, menus and buttons. This can be used to filter data and constrain how the information is displayed. Direct manipulation with grab, pinch, and stretch actions can be combined to provide semantic interaction (*e.g.*, to filter or select elements from the visualization). The principle approach of direct manipulation is so-called *brushing*, where selecting (and highlighting) elements in one display concurrently propagates to other linked displays. Shneiderman et al. [16] presented a task by data type taxonomy as design guidelines for visual information exploration tasks. The taxonomy connects each of seven identified data types (1-, 2-, 3-dimensional data, temporal and multi-dimensional data, and tree and network data) with the appropriate tasks. This provides the repertoire of direct interactions to explore data of this type (i.e, overview, zoom, filter, details-on-demand, relate, history, and extract).

These basic VAI techniques form the basis of the domain specific configuration of the Digital Dashboard, described in the next section.

5 The Digital Dashboard System

The *Digital Dashboard* [18] is an experimental framework for rapidly prototyping systems for visual exploration of data. It provides general methods for direct manipulation which makes interaction intuitive and easy, even for novice users, and all visualization results allow further interaction. The framework is implemented as a Web-top application and runs in a Web browser on a user's local machine.

Because the framework is component based, new data sources can easily be combined to federate data from separate sources ("data mash-up"). It is also easy to incrementally add new types of VAI methods.

The *Digital Dashboard* supports multiple linked views of a single collection of base data. When selections or groupings are done in one view, all linked views are immediately updated. Components that not only provide visualization but also do advanced analyses (*e.g.*, data mining, clustering, time-series) can also be linked. Such components will automatically recalculate whatever analysis they do, as necessary, because of changes in any linked components. All linked visualization components will be automatically updated as analysis results are completed.

As described above, our informal characterization of visual analytics as VAI workflow provides a simple framework in which to select a number of visualization, analytics, and interaction tools appropriate to the semantics of the domain.

For the scope of this paper and the properties of the Poynt data described above, we have focused on the following:

V (visualization)

With the abstract goal of making visual inferences about Poynt user behaviour, as either individuals or in aggregate, we have configured two kinds of basic visualization techniques to provide two alternative views of the geo-located data. In addition to the standard Cartesian geo-location map mashup, we add an alternative geo-location visualization tech-

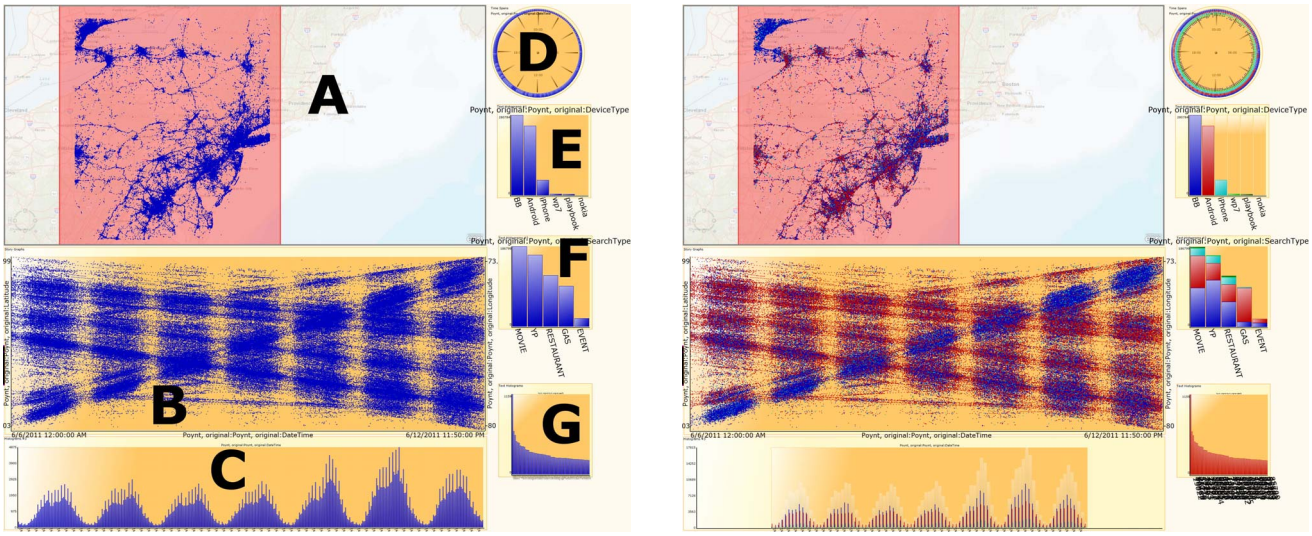

Figure 1: Left: The *Digital Dashboard* displaying some Poynt data. Right: Searches grouped by device type.

nique called Storygraph [17], which provides an alternative to the Cartesian map display and retains geo-coordinates for each individual Poynt record, but also clearly shows a time dimension of each Poynt request. This is augmented with auxiliary visual tools that can be used to select subsets or views on the Poynt data, including selections in Cartesian space, temporal-Cartesion space (Storyline), and subsets of both time (*e.g.*, "Every Monday") and search attributes (*e.g.*, "movie and restaurant search on Fridays.").

A (analytics)

It is often difficult to separate the transformations of base data to visual representations and many of the possible analytics methods that might be deployed to provide insight into those data turned into pictures. That is only because the process of visualization is itself an abstraction process, just as are all of machine learning methods. In our VAI framework, the desire is to be clear about what kinds of analytics methods are potential useful in exposing visual inferences on the application domain. In this case, the analytics methods largely arise from the visual aggregation of Poynt data records, arising from the selection of subsets of the data (typically with geographic constraints and Poynt record attribute space selections), which is then rendered to create an aggregate representation of those selections in a small number of visualization variations (as described in the "V" segment above).

I (interaction)

The natural semantics of the Poynt domain suggests that a human interactor should be able to visually perform selections on the data, both in terms of spatial extend and attribute extend (including time). In this case, our reconfigurable Digital Dashboard provides for direct touch manipulation of spatial extent, attribute selection, and temporal extent.

A selection of VAI methods suggested by the properties of the Poynt data are shown in Figure 1. We mostly focus on four types of visualization components: A) a Cartesian map, to show geospatial data on a map and B) StoryGraph [17], to show both geospatial and time distribution; C, E, F, G). These foundational components for visualizing space and temporal extend are augment here with 2D histograms, to show data distributions in a variety of selected Poynt record attributes, together with a clock D), to show distribution over the time of day. Note that, because the process of visualization is itself a dimensionality or abstraction step, it is not always easy to distinguish implicit "Analytics" from explicit "Analytics" as suggested in the definition we have adopted. For example, the Storygraph visualization in component B of Figure 1 already creates a visual clustering of Poynt events in time, just by the nature of how events are aggregated.

Similarly, for example, the direct manipulation of both the Cartesian visualization (by doing multiple region selections) and the temporal Storyline visualization (but choosing time extent, and linking the selections from previous Cartesian selections) provides the basis to directly explore things like "What is the temporal clustering of Poynt users in a certain geographic area search for a restaurant and a movie."

6 A Framework for VAI workflow

Here we explain a few mini-case studies in terms of how the selected VAI tools are exploited to do exploratory or explanatory search, and then considered any refinements of visualizations, analytics, and interaction to help improve our sense of the effectiveness of the obtained visual inferences.

We use the expressions *Exploratory Search* and *Explanatory Search* to distinguish two different ways of investigating our data. *Explanatory Search* is when we already have an idea or hypothesis in mind, and we seek to see if the data provides confirming support (or not). An example is the hypothesis that "People are more likely to search for a movie on Fridays than on Mondays." We open the data, select data from only Mondays and Fridays, and then we visually inspect the number of searches in the movie category to support our hypothesis (or to discard the hypothesis if the data indicates that it is wrong).

By *Exploratory Search* we mean using tools to look at a variety of the data to find "something interesting," especially when we lack any precise idea of what we are looking for. An example of an *Exploratory Search* could be that we want to look at the data and ask ourselves "Are there any differences in behaviour between Android users and iPhone users?" We could then visualize the data in a variety of ways and display search records from Android devices and iPhone devices separately, to see if something stands out.

The idea of exploratory versus explanatory search is not a mode or distinguished operation in the *Digital Dashboard*. But it is a cognitive style of user by human users. In many cases, it is common to switch back and forth between *Exploratory Search* and *Explanatory Search*. Starting with a rough idea, we might do some *Exploratory Search* and see something in the data that prompts us to formulate a clearer hypothesis. We can then switch to *Explanatory Search* to find support, or discard that hypothesis it if it turns out to be wrong. In the case that lack of support or disconfirming evidence is found, we might return to *Exploratory Search*, to look for a similar but refined hypothesis for further *Explanatory Search*, or we can do further *Exploratory Search* on a new subset of the data based on the previous hypothesis. This kind of distinguished cognitive activity bears more investigation, but here we simply provide more detailed scenarios that further articulate the distinction.

We acknowledge that several have noted the important role of distinguishing exploratory and explanatory visualization interaction (*e.g.*, [13],[21], [4], [9]). We also note that no visualization systems we know of provide any explicit record of posting hypotheses to be confirmed or extracting potential hypotheses from exploration, thus the ongoing debate about the difference (*e.g.*, [19]). However, with our simple VAI framework, we can at least consider simple visualization analytics interaction workflows that capture some of the intent of changing from exploratory and explanatory mode. We even expect that, in future visualization systems, such workflows will become part of a standard library of abstract methods that provide a starting point for the kinds of visualize inference required by any particular application domain, as is the case in the taxi data of Ferreira [9], the geo-spatial mashups of Wood et al. [21], or the visualization of web-based search of Dörk et al. [4].

7 VAI example scenarios

Here we show some scenarios using the data described in Section 2 with the system described in Section 5. In Figure 1 the system provides a variety of methods to display selected data. For example, there is a typical Cartesian map (**A**) showing the locations of all the searches. There is also a 24-hour clock (**D**), with midnight at the top and noon at the bottom, showing the time of the day that the searches were done. We also provide histogram charts showing the number of searches at different times (**C**), the number of searches per device type (**E**), the number of searches per search category (**F**), and the number of searches per user (**G**), for the most frequent users.

There is also a StoryGraph [17] display (**B**). A StoryGraph makes it easier to see both temporal distribution and geographical distribution. The horizontal axis is time, and we can see that there are regular vertical bands with very few data. These are the nights; there is very little search activity during the nights. The left hand vertical axis is the latitude of the location of the search and the right hand vertical axis is the longitude of the location of the search. So geographic location, otherwise appearing as an x,y coordinate in a Cartesian map view, now becomes a straight line from the left to the right. A data point is drawn on the line from its latitude value on the left axis to its longitude value on the right axis, on the location on the line corresponding to the time of this search query on the horizontal (time) axis.

7.1 Explanatory Search of Geographical Differences

A first simple example scenario is an explanatory search to see if the data supports the hypothesis: "There will be many queries that are common but that occur only in Canada, and similarly queries that occur only in the USA, and there will also be common queries that occur in both regions." To confirm this hypothesis, we simply group the data into searches from locations in the USA and searches from locations in Canada by interacting with the Cartesian map visualization and then look at the frequent search strings.

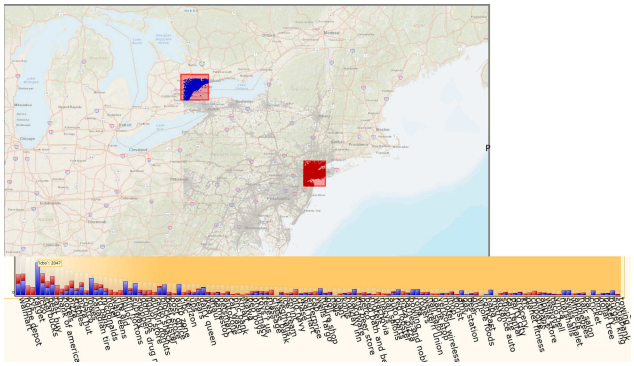

Figure 2: Canada contrasted with the USA

The resulting visualization in Figure 2 shows that frequent strings such as "CIBC" (a Canadian bank), "LCBO" (the Liquor Control Board of Ontario), "Canadian Tire," occur only in the Canadian group. There are also searches such as "Bank of America," "CVS," and "Target," that occur only in the USA, and common queries that occur both in Canada and the USA, such as "UPS," "Costco," or "Starbucks," thus confirming the hypothesis.

7.2 Exploratory Search of Device Types

One exploratory investigation of interest might be to determine if users of different devices behave in different ways. We start an exploratory search to find any interesting differences between users of different device types, by interacting with the histogram showing the device types, grouping the data by the type of device used. The resulting visualization is shown in Figure 1 (right image). We can see that the most frequent type is BlackBerry, followed fairly closely by Android. There is a small iPhone category, and the remainder are extremely low in volume.

Severals things emerge in the visualization of the different device types. For example, in the StoryGraph component there is a thick band going from the lower left to the upper right that consists mainly of BlackBerry searches. Geographically, this is the Toronto area, and on the map we can also see that Toronto has mainly BlackBerry searches. For other areas, there is no such BlackBerry domination. BlackBerry seems to be strongly correlated with the Canadian areas, which might be because BlackBerry is a Canadian company with headquarters near Toronto.

Another thing that stands out is that, while BlackBerry is the largest group for most search categories, for *Gas* searches the BlackBerry group is very small. Also, while BlackBerry is the most commonly used device, none of the users with the most searches use BlackBerry.

Interacting with the frequent user visualization and selecting only data from these users shows that they all have very large numbers of *Gas* searches. Since there seems to be something that makes the *Gas* searches different from the other categories, we explore this category of search a little more.

We interact with the histogram of search types and select only the *Gas* searches. We also reset the device type histogram to show all the data as one group. The resulting visualization is shown in Figure 3. The BlackBerry bar in the device type histogram is now very small, despite being the largest when all search categories were shown. Something that also stands out very clearly is that on the Cartesian map, there are no longer any searches at all from Canada; all searches in the *Gas* category come from the USA. The Poynt search application does not seem to support *Gas* searches in the Canadian version.

Since the BlackBerry device type was strongly correlated with Canada, the explanation for the under representation of BlackBerry devices in the *Gas* category is simply that this category is only available in the USA, where BlackBerry users are fewer.

The most frequent users in the frequent user histogram have an unusually high volume of searches when showing only the *Gas* searches. In one case, one user has over 2,000 searches during five weeks.

In the right image of Figure 3, four users with very many searches in the *Gas* category are shown. In the StoryGraph visualization we can see that there are straight lines of regularly spaced dots. This means that the users are in more or less the same location and search for gas at regular intervals. Sometimes this goes on for more than 24 hours without interruption. This suggests that these are automatically generated searches. Perhaps these very frequent users have a navigation application that continuously refreshes a map with the nearby gas stations and their prices?

There are also instances where there is a gap in a long line with a short line nearby filling this gap, and then the long line continues again. This shows a user searching for gas in one location many times, then moving to a new location where more gas searches are generated for a time, and then the user moves back to the first location again.

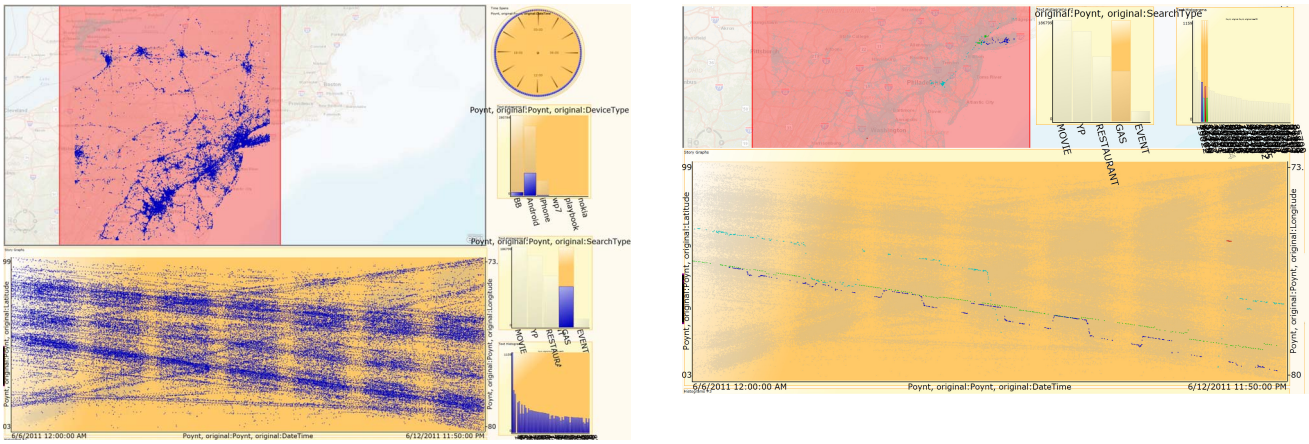

Figure 3: Left: Restricting the data to only searches in the *Gas* category. Right: Further restricting the visualization to only four users with abnormally high numbers of searches in this category.

7.3 Explanatory and Exploratory Search of Time

What users search for is likely to be different at different times of the day and at different days of the week. We start with a simple hypothesis: "There are more searches for 'taxis' at late hours during the weekends than at other times of the day and other days of the week". We will also explore the data to see if there are any other differences that stand out while we check our hypothesis.

We first interact with the clock component to restrict the times of the searches and to group the data into groups that represent "morning," "noon," "night." The resulting visualization is shown in Figure 4 (left image). The bottom bar chart shows the most frequent search strings.

We find many query strings that appear in only one or two of the groups. Some examples include: "pizza." "ice cream," "red lobster," and "liquor," which are never searched for in the morning. As expected, "bar" and "strip-club," are generally searched for at night. On the contrary, "bank" or "Bank of America" are not searched for at night. At night there are many searches for "cab" and "taxi," though these do occur at other times too. Common noon searches include "Home Depot," "Best Buy," "Walmart," and "Costco."

Next we interact with a histogram showing the weekdays of the searches, grouping them into searches that happened during Monday to Thursday ("weekdays") and searches from Friday to Sunday ("weekends"). We also set the StoryGraph to overlay all five weeks of data on top of each other. This way, differences between weekends and weekdays may become more apparent than if all weeks are shown separately. The resulting visualization is shown in the right image in Figure 4.

Searches for "Walmart" are much more common on weekdays than on the weekends, as is for instance "Home Depot." There are also more searches for "bowling" or "gym" on weekdays, while "bar" is more common on weekends. This is also reflected in the Yelp clustering of the search strings, For example, the "Active life" category (which includes gyms, bowling, etc.) has more weekday searches and the category "Nightlife" (which includes bars and restaurants) has more weekend searches.

Our hypothesis regarding "taxi" searches can now be explained. The "taxi" searches are the fifth bar from the left in the frequent search queries visualization. The part of the bar that corresponds to weekend nights makes up more than half the total searches for taxis, confirming our hypothesis.

8 Conclusions

We have presented an application instance of the *Digital Dashboard* framework, and described its use for interactive data exploration of Poynt geolocated search data. In our case study of that data, we have distinguished the concepts of *explanatory search*, where a user can search a visualized subset of data to support or reject a hypothesis, and that of *exploratory search*, where one looks at a variety of visualized views of the data, in order to get ideas for new hypotheses. The alternation of these styles of visualization and interaction have proven very useful when dealing with large multi-dimensional data sets.

The case study examples we have described only scratch the surface of what is possible, much remains to be explored. For example, one could image building and saving multiple hypotheses on a variety of data, and manage a large number of machine learning techniques, perhaps simultaneously, to investigate a variety of exploratory and explanatory structures emerging from the data.

We have not yet considered an evaluation the simple VAI workflows from a cognitive effectiveness viewpoint,

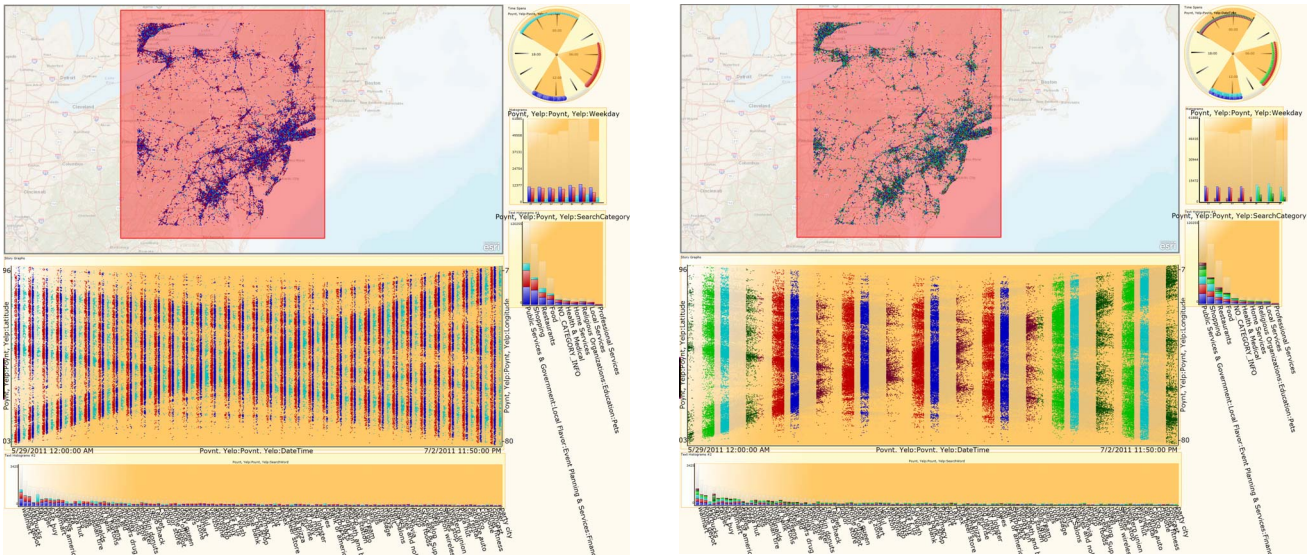

Figure 4: Left: Searches grouped by the time of day. Right: Also grouping by weekend or weekday, while folding all weeks into one.

in order to try and accurately measure the effectiveness of VAI tool selection alternatives. However, we believe that this VAI framework is a prerequisite to the design of such experiments, because it provides a basis for considering how the selected toolsets can be used to improve the efficacy of visual inference for this particular domain.

In addition, there is no theoretical basis to guide the choice and development of an efficient and semantically useful repertoire of user interactions on the visualizations within the *Digital Dashboard*. While we have exposed some of the interactions that seem of value (*e.g.*, the interactive adjustment of the time clock to help see time series dependencies), much remains to be done in terms of user evaluation of preferred visual inferences and valued interaction.

In this case, these kinds of future investigations are well-supported by our system, and analysis of other kinds of data (*e.g.*, snow accumulation and removal data in large cities, traffic flow and accident patterns, trends in health and epidemiology) are underway.

Acknowledgements

We would like to thank Poynt for providing access to a sample of their data, and to the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates Technology Future (AITF) and Japan Science and Technology Agency (JST) for their generous support of this research.

References

[1] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. *Visual Analytics of Movement*. Springer, Heidelberg, Germany, 2013.

[2] Natalia V. Andrienko, Gennady L. Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *J. Vis. Lang. Comput.*, 14(6):503–541, 2003.

[3] Ed Huai-hsin Chi. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization 2000 (INFOVIS'00), Salt Lake City, Utah, USA, October 9-10, 2000.*, pages 69–75, 2000.

[4] M. Dörk, S. Carbondale, C. Collins, and C. Williamson. VisGets: coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008.

[5] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proc. Nat. Acad. Sci.*, 106(36):15274–15278, September 2009.

[6] Justin Fagnan, Osmar R. Zaïane, and Randy Goebel. Visualizing community centric network layouts. In *16th International Conference on Information Visualisation, IV 2012, Montpellier, France, July 11-13, 2012*, pages 321–330, 2012.

[7] Katayoun Farrahi and Daniel Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27, January 2011.

[8] A. Ferlitsch, K. Wischnofske, and W. Mayall. Americas Open Geocode (AOG) Database. http://opengeocode.org/about.php.

[9] Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, December 2013.

[10] Peter Gatalsky, Natalia Andrienko, and Gennady Andrienko. Interactive analysis of event data using space-time cube. In *Proceedings of the Eighth International Conference on Information Visualisation*, IV '04, pages 145–152, 2004.

[11] Randy Goebel, Wei Shi, and Yuzuru Tanaka. The role of direct manipulation of visualizations in the development and use of multi-level knowledge models. In *Proceedings of the 17th International Conference on Information Visualisation*, IV '13, pages 325–332, London, UK, 2013.

[12] D. Guo, J. Chen, A.M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006.

[13] Daniel A. Keim, Leishi Zhang, Milos Krstajic, and Svenja Simon. Solving problems with visual analytics: Challenges and applications. *Journal of Multimedia Processing and Technologies*, 3(1):1–11, 2012.

[14] Lorenzo Magnani and Ping Li, editors. *Model-Based Reasoning in Science, Technology, and Medicine*, volume 64 of *Studies in Computational Intelligence*. Springer, 2007.

[15] Jonathan C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, CMV '07, pages 61–71, Washington, DC, USA, 2007.

[16] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *In IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[17] Ayush Shrestha, Ying Zhu, Ben Miller, and Yi Zhao. Storygraphs: Extracting patterns from spatio-temporal data. In *Proceedings of IDEA'13*, pages 96–104, Chicago, IL, USA, 2013.

[18] Jonas Sjöbergh and Yuzuru Tanaka. From multiple linked views to multiple linked analyses: The Meme Media Digital Dashboard. In *Proceedings of the 18th International Conference on Information Visualisation*, IV '14, pages 170–175, Paris, France, 2014.

[19] R. Spence. *Information Visualization: Design for Interaction (3rd Edition)*. Springer, 2014.

[20] G-D. Sun, Y-C. Wu, R-H. Liang, and S-X. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 13(5):852–867, 2013.

[21] J. Wood, J. Dykes, A. Slingsby, and K. Clarke. Interactive visual exploration of a large patio-temporal dataset: reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183, 2007.

[22] Leishi Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim. Visual analytics for the big data era - a comparative review of state-of-the-art commercial systems. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 173–182, Oct 2012.