The Development of a Hypothesis-driven Framework for Commercial Geo-position Data Visual Analytics

by

Xingkai Li

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

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 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. This includes 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 relatively complex search and movement patterns we call "consumer trajectories."

Our pursuit of these kinds of patterns has helped guide our development of conceptual tools and visualization tools in aid of investigating the geo-located data, and finding both interesting and useful patterns in that data, in a hypothesis-driven process. Included in our system architecture is the ability to consider the difference between exploratory and explanatory searches on data patterns, as well as the deployment of multiple visualization methods that can provide alternatives to help expose patterns. Here we provide examples of formulating hypotheses on geo-located behaviour, and how visual analytics can help formulate hypotheses, and confirm or deny the value of such hypotheses as they emerge.

Acknowledgements

First of all, my sincere gratitude goes to my supervisor Randy Goebel, for his patient guidance on this research, the persistent support to my professional activities and personal development, as well as the encourage and care given to my life as a friend, during my not so short master's studies.

I would also like to thank my colleagues Xiaoxiao Li and Mi-young Kim for the discussions and help on Yellow Pages keyword semantic labeling in this work, and Jonas Sjöbergh from Meme Media Lab, Hokkaido University for the collaboration work on the Webble Dashboard and the kind company during my visit to the lab, and Poynt Corporation for providing the raw data.

In the end, to my beloved friends and family, who are my motivation to move forward and be better, I can never say thank you enough.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The popular smart phone application, "Poynt"[1], 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.

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.*, movie, restaurant, and a variety of others). Our task 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.

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.*, [5]). What is more interesting in the Poynt data context is the development of analytic methods to identify what we refer to as

[1] see http://www.poynt.com

"consumer trajectories," which are a combination of geo-location trajectories and consumer interest patterns, as evident in the classes of search that a Poynt user can conduct.

In general, the identification of consumer trajectories is now the whole focus of our Poynt analytics framework. We will complete our scientific research by clarifying the framework, and providing initial solutions to the challenges of identifying consumer trajectories.

We will focus on refining the identification of possibly valuable business hypotheses based on identifying general patterns of consumer trajectories. Specifically, by incrementally varying constrained geo-extent and time extent, we conduct:

1. explanatory search to confirm visual understanding of expected hypotheses (*e.g.*, movie nights on Friday and Saturday);

2. exploratory search to find "interesting" phenomena, then constraint adjustments to find supporting evidence for possible hypotheses.

1.1 Statements of Contributions

The hypothesis of our research is that a data artifact called consumer trajectories can be identified within the Poynt data, by the development and use of a variety of data filtering and visualization tools. Our search for consumer trajectories within the Poynt data has driven our development of a variety of tools, and the contribution of our research is the incremental development and testing of those tools, and the development of a method of their use to expose what we have conceived as the basic components of consumer trajectories.

The tools include the development of filtering and selection methods on the Poynt data, to select geographic regions and temporal regions, visualization methods (Cartesian, Storyline), and the use of interactive visual analytics using Webbles.

The methods of use include the distinction between a user's exploratory and explanatory use of the tools, as exposed in the case studies on finding a variety of correlations on Poynt user data, beginning with the simplest hypotheses on where and when people search for movies, all the way to semantic classification of keyword searches, all displayed in both Cartesian and Storyline pictures.

1.2 Thesis Outline

This dissertation is organized in four parts. The first part is entirely contained in Chapter 2 which gives the description of related work, especially most important to our research. In the second part, we present the methodology adopted in our research of commercial geo-position data visual analytics, with a framework for hypothesis-driven analysis introduced in Chapter 3 and the visualization approaches utilized in the framework, Cartesian Geography Map and Storyline, described in Chapter 4. Part three explains the case studies and discusses the corresponding results. Chapter 5 showcases a typical example of the application of exploratory and explanatory searches, as well as a case study on the comparison of different source devices. In Chapter 6 we explore the investigation of Poynt Data with the Webble Dashboard, a software component based system which uses multiple linked views and allows interactions via direct manipulation. Finally, Chapter 7 summarizes our work and contributions, and suggests possible future directions.

Part I

Related Work

Chapter 2

A Survey of Work on Geo-position Data Visual Analytics

In this chapter, we give the description of the related work in several areas that are associated with our research of geo-position data visual analytics.

2.1 Analysis of Geo-located Data

2.1.1 Identifying Geographic Trajectories

There are researches focusing on the identification of geographic trajectories with large volumes of geo-coded data [5, 23].

Farrahi and Gatica-Perez [5] proposed an unsupervised methodology to discover users' location routines, from the location dataset of the Reality Mining project [4], with the help of probabilistic topic models. In the raw location data, a user's location is represented by its nearest cell tower. The authors assigned "location labels" (*home*, *work*, *out*) to users' locations, and applied a novel bag representation for location sequences to capture both fine-grain and coarse-grain time factors, as well as transitions of location labels. Location routines (patterns of location label transitions over time, *e.g.*, "going from *home* to *work* between 9-11 am," "going from *work* to *out* in the evening," and "at *work* early in the morning,") were mined as latent topics by the topic models. Specifically, Latent Dirichlet Allocation (LDA) identifies daily location routines dominating the entire user group's activities, while Author Topic Model (ATM), taking into account both user and topic, detects routines shown by certain subgroups of users.

Zheng and Xie [23] develop models for generic and personalized travel recommendations by mining a large volume of user-generated GPS location traces. Specifically, the TBHG (Tree-Based Hierarchical Graph) is introduced to learn histories of user travel sequences, based on which a HITS (Hypertext Induced Topic Search)-based inference model is used to infer top popular locations and travel sequences in a given geo-region for the generic travel recommendation. On the other hand, the personalized travel recommendation for a user is achieved by first calculating correlation between locations with the travel experiences of the user and the visited location traces of others, and then building an item-based CF (Collaborative Filtering) model with the calculated correlation incorporated, to predict the user's interests in unvisited locations, and finally making the personalized recommendation for that user.

The data for analysis in [5, 23] differ from normal commercial geo-position data created by users of smartphone mobile applications (*e.g.*, the Poynt data), in that user locations in [5, 23] are periodically and densely collected with fine time granularity (in seconds), thus are massive and nearly of equal size per user, but the total number of users is small (around 100); on the other hand, the Poynt data is the opposite: with millions of users using the application, the record set of one user represents all his/her search behaviours, hence it varies from user to user in set size and the distribution of associated records over time. In addition, the data in [5, 23] are simple logs of location traces with no extra attributes contained, while the Poynt dataset is rich in attributes to describe consumer search behaviours in the Poynt app.

2.1.2 Analyzing Geo-tagged Social Media Data

Twitter, a popular micro-blogging and social networking application, has accumulated enormous amount of geo-tagged data with texts (status updates, or "tweets"), which provides valuable data sources for researches in many areas, *e.g.*, Natural Language Processing, Social Network Analysis. Many researches have been conducted which apply Twitter data for geo-located data analysis [2, 11, 9].

Brennan et al. [2] studied how individuals contribute to the spread of flu-like illness based on interpersonal interactions, with the help of status updates created by travelling Twitter users. Specifically, by focusing on those who tweeted from two or more airports and are identified by a binary Support Vector Machine (SVM) as sick passengers, estimated are a number of latent variables indicating flu signals, *e.g.*, the volume of sick passengers, and the number of people they physically encountered. Based on the inferred latent variables, a regularized regression model is learned for the prediction of flu spread, which outperforms the baseline model ignoring people's health states. The paper shows that the latent features learned on the basis of Twitter data can help improve the predictions of flu prevalence in a given area.

Kamath et al. [11] analyzed the spatio-temporal dynamics of Twitter hashtags based on a sample of 2 billion geo-tagged tweets, for understanding meme diffusion and information propagation. Specifically, the authors investigated how location, time, and distance impact hashtag adoption. Also examined was the spatial propagation of hashtags in terms of their focus, entropy, and spread. As a result, this study shows that the physical distance between locations strongly constrain the adoption of hashtags, and hashtags are mostly a local phenomenon with long-tailed life spans. The authors also discovered that how fast a hashtag will reach its peak is determined by its purpose and global awareness, and hashtags exhibit spatial and temporal locality since they normally spread over small geographical areas but at high speeds.

Hong et al. [9] proposed a method which models topical diversity, geographical diversity, and user interest distribution, in order to uncover geographical topical patterns in different languages and users' common topics of interest from geo-tagged Twitter messages. Statistical topic models are utilized to combine tweet content and geographic locations, and sparse coding techniques are adopted for an efficient and effective implementation. The model can be applied to applications including user profiling, content recommendation, and by outperforming several state-of-the-art algorithms, has demonstrated its effectiveness in the task of location predictions of new messages.

There are also researches based on geo-located data created by Youtube users. For example, Brodersen et al. [3] conducted an analysis of the properties of geographical popularity of YouTube videos on a corpus of over 20 millions geo-coded

YouTube videos. New measures (e.g., view focus, view entropy) were applied to quantify video popularity distribution across different geographic regions. The authors investigated how social sharing impacts the spatial popularity of a video and how the popularity of a video geographically evolves over its lifetime, and found that Youtube video consumption displays strong geographic locality of interest.

2.2 Exploratory Visual Analytics

Researches in exploratory visual analytics correlate closely with our development of the framework for visual analysis of commercial geo-position data. While the research area of exploratory visual analytics extensively covers a variety of topics, in this section, we focus on multiple linked views [20, 15] as well as interaction and direct manipulation [7, 20, 15, 16], which facilitate the process of visual exploration of data.

Multiple linked views can serve as the base visualization framework of exploratory visual analytics. Sjöbergh and Tanaka [20] introduced a software component based system, the *Digital Dashboard*, which utilizes multiple linked views for visual exploration of data. In the system, each view is a pluggable visualization component, and all views are interactive and connected. When visually exploring data, the adoption of multiple linked views enables that interactions in one view, *e.g.*, selections or groupings of visualized content, are automatically reflected in other views, since all views are “linked”. Roberts [15] surveyed the area of Coordinated Multiple Views (CMV), namely Multiple Linked Views. This paper introduced Coordinated Multiple Views as a specific exploratory visualization technique, benefiting from which users may find insightful relationships and features from target data.

Interaction 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. Indirect manipulation and direct manipulation are the two styles of interaction. Indirect manipulation includes *dynamic queries* which enables users to interact with sliders, menus and buttons to filter data and constrain how the information is displayed, while direct manipulation techniques allow users to manipulate with visualization displays directly (*e.g.*, filter or select elements from the visualization). The principle approach of direct manipulation is *brushing*, where selecting (and highlighting) elements in one display concurrently makes the corresponding information in other linked displays highlighted. Shneiderman [16] presented a task by data type taxonomy as design guidelines for visual information exploration tasks. The taxonomy connects each of the seven data types (1-, 2-, 3-dimensional data, temporal and multi-dimensional data, and tree and network data) with the appropriate tasks to explore data of this type, selected from a repertoire of seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extract) which are supported by direct manipulation. Sjöbergh and Tanaka [20] introduced the *Digital Dashboard*, which allows interaction through direct manipulation in all view components. Goebel et al. [7] focused on semantically-coupled direct manipulation under complex natural and unnatural knowledge model contexts. The paper argues that, due to the growing volume and increasing complexity of data, the role of direct manipulation guided by the semantics of multi-scale models becomes increasingly critical in providing users with tools to build and debug complex scientific models.

In our development of a hypothesis-driven framework for commercial geo-position data visual analytics, we distinguish the roles played by "Exploratory search" and "Explanatory search". In this regard, not only can multiple linked views and interaction techniques help with the exploratory process searching for insightful features of data to propose new hypotheses ("Exploratory search"), but also facilitate the collection of evidence to confirm or deny proposed hypotheses ("Explanatory search"). We also believe that the *Digital Dashboard* by Sjöbergh and Tanaka [20] holds the most promise for supporting more sophisticated interactive visual analytics.

2.3 Spatio-temporal Data Visualization

Visualization is a key part in visual exploration of data. Many researches about spatio-temporal data visualization have been conducted to integrate time and location (*i.e.*, latitude and longitude) in a two-dimensional (2D) or three-dimensional (3D) visualization space [17, 18, 13, 6, 12], in addition to rendering them separately (which we do not cover in this section).

In a 3D integrated visualization of spatio-temporal data, latitude, longitude and time are plotted as three independent dimensions in a 3D chart. For instance, the space-time cube method (STC) [13, 6] represents time as an additional dimension to the 2D geo-spatial map plane. Kapler and Wright [12] presented a prototype called GeoTime for displaying and tracking events, objects and activities in a combined spatial and temporal 3D view, and demonstrated its effectiveness when applied to the analysis of complex past and future events within a geographic context. The advantage of 3D integrated visualization is that the time graph does not occlude the 2D map, since time and location are plotted in separate dimensions. However, inherently in a 3D view of spatio-temporal data, time is difficult to be aligned with location, and the perception of depth is a problem. Also, space-time paths (line segments connecting events) plotted in a 3D view are often ambiguous. As more data are visualized in a 3D view, its inherent problems bring more confusion hindering the understanding and analysis of the visualization result.

On the other hand, attention has been drawn to designing innovative 2D integrated views of spatio-temporal data, so that the inherent problems of 3D visualization could be avoided. Shrestha et al. [17, 18] proposed a novel approach called *Storygraph*, which integrates time and location in a 2D visualization space based on parallel coordinates [10]. Compared to 3D methods, *Storygraph* reduces cluttering and occlusion, and helps track time in case of clustered events. Based on *Storygraph*, a storyline is built by connecting all the events of an individual sequentially, in order to track the movement of characters and the interactions between them over time. Storyline helps discover relationships and patterns among spatially and temporally scattered events.

Storygraph-based storyline is suitable for visualizing spatio-temporal consumer trajectories in our framework, since this 2D approach avoids the inherent problems of 3D integrated visualization and is easy to implement. Although it still suffers occlusion and cluttering to some extent when data gets dense, we consider the adoption of the strategy *Frequency Plots* proposed by Artero et al. [1], suggested by the measures taken by Shrestha et al. [18] to alleviate the issue of cluttering and overplotting in *Storygraph*.

Part II

Methodology

Chapter 3

A Framework for Hypothesis-driven Visual Data Analysis

The volume of individual search records created by the Poynt application is understandably very high: even with one sample of just five weeks of historical data, we have over 178 million records. From the technical viewpoint, analytics researchers could, without business guidance, produce at least 178 million analytic outcomes, which would render the analytics process useless. Of course the alternative is to guide the development of analytics for Poynt by ensuring there is a top down business model that guides the search for analytics consequences of interest. In this chapter, we will explain how such a framework is developed.

The basic idea is to create an analytics framework which is guided by business-relevant hypotheses, *e.g.*, "People who search for movies are most active on the afternoons of weekends." In the case of this hypothesis, we would want to deploy an analytics process that could aggregate the dataset to either confirm or refute the hypothesis. In this case, if we find a trend that movie searches are clustered around, say Friday, Saturday or Sunday afternoons, then we have provided the basis for the business model to exploit that hypothesis (*e.g.*, by increasing the price of movie advertising within that time frame).

Overall, we believe that a Poynt analytics framework should be to develop analytics tools that help confirm or refute potentially valuable business hypotheses. In our initial development of such a framework, we have focused on two categories of analytics hypotheses:

1. hypotheses about the behaviour of individual users in one single category of search (*e.g.*, time distribution of all users in movie searches);

2. hypotheses about correlations amongst multiple search categories by individual users (*e.g.*, how many times does an individual searching for a movie also search for a nearby restaurant).

We believe that we can develop such categories of business model driven hypotheses, expanding these as appropriate, and then construct an analytic tool kit that helps confirm or deny such hypotheses in data. In this way, business decisions can be decoupled from the need for analytics knowledge.

3.1 Poynt Geo-search Data

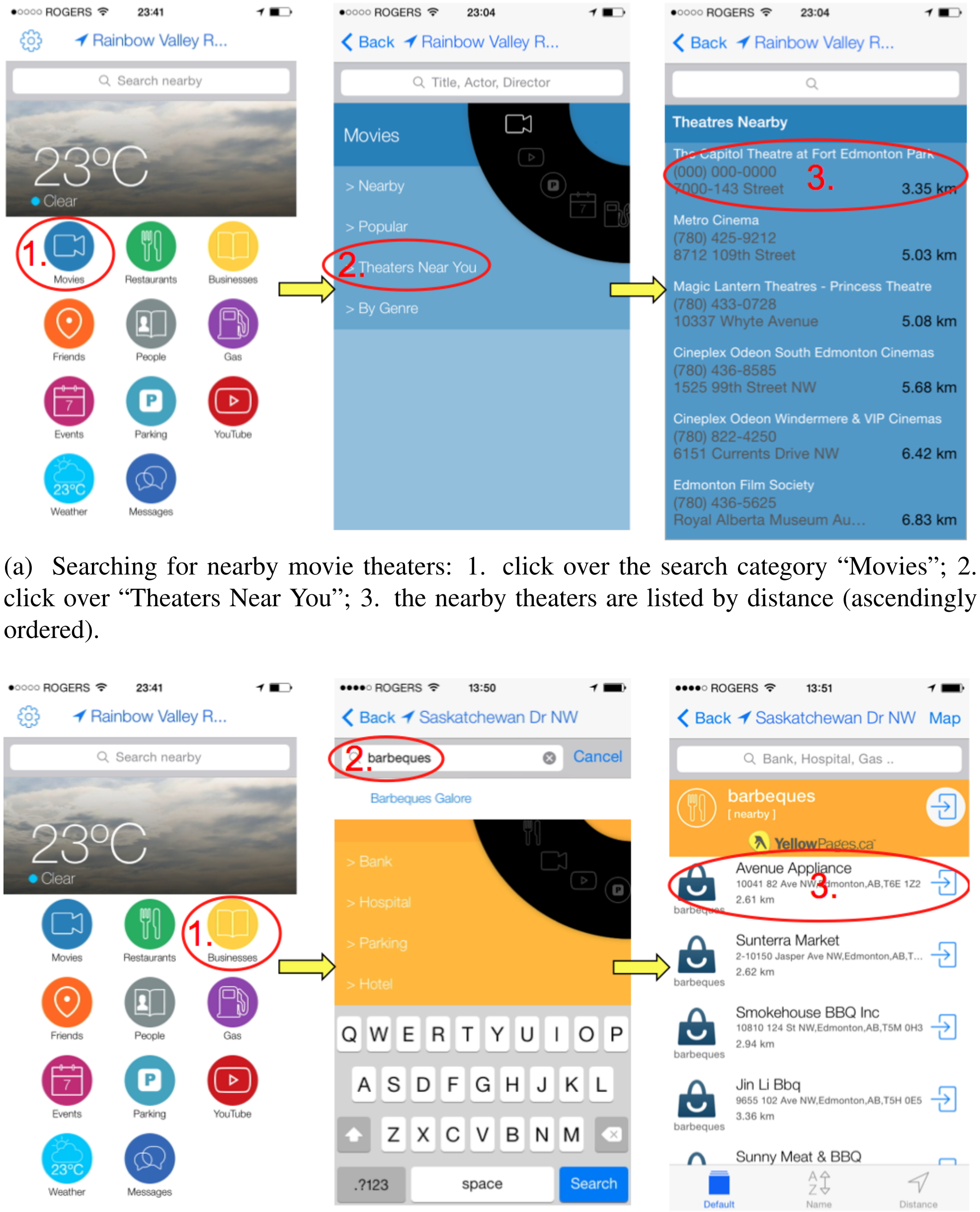
As briefly mentioned in Chapter 1, the local mobile search application, Poynt, provides a location-based local search service for users to query about nearby movies, restaurants, businesses, gas prices, events and many other varieties. For example, Figure 3.1 showcases several typical use cases of the application.

All user search behaviours are collected by the Poynt corporation in the form of search records, in order to keep track of what was searched for, where and when, by whom. The volume of individual search records created by the approximately 20 million Poynt users is understandably very high. Three datasets are granted to us by the Poynt corporation for analysis purpose, which, numbering around 531 million search records in total, respectively cover the three periods of five consecutive weeks from May 29/2011, October 30/2011, and April 29/2012.

3.2 Exploratory and Explanatory Searches

In proposing new hypotheses and collecting evidence to confirm or deny proposed ones, the framework should take into account both explanatory and exploratory searches.

Explanatory search is to confirm/refute our understanding of expected hypotheses (*e.g.*, movie nights are typically on Friday and Saturday). Exploratory search



The Capitol Theatre at Fort Edmonton Park
(000) 000-0000
7000-143 Street 3.35 km

Metro Cinema
(780) 425-9212
8712 109th Street 5.03 km

Magic Lantern Theatres - Princess Theatre
(780) 433-0728
10337 Whyte Avenue 5.08 km

Cineplex Odeon South Edmonton Cinemas
(780) 436-8585
1525 99th Street NW 5.68 km

Cineplex Odeon Windermere & VIP Cinemas
(780) 822-4250
6151 Currents Drive NW 6.42 km

Edmonton Film Society
(780) 436-5625
Royal Alberta Museum Au... 6.83 km

(a) Searching for nearby movie theaters: 1. click over the search category "Movies"; 2. click over "Theaters Near You"; 3. the nearby theaters are listed by distance (ascendingly ordered).

Avenue Appliance
10041 82 Ave NW,Edmonton,AB,T6E 1Z2
2.61 km

Sunterra Market
2-10150 Jasper Ave NW,Edmonton,AB,T...
2.62 km

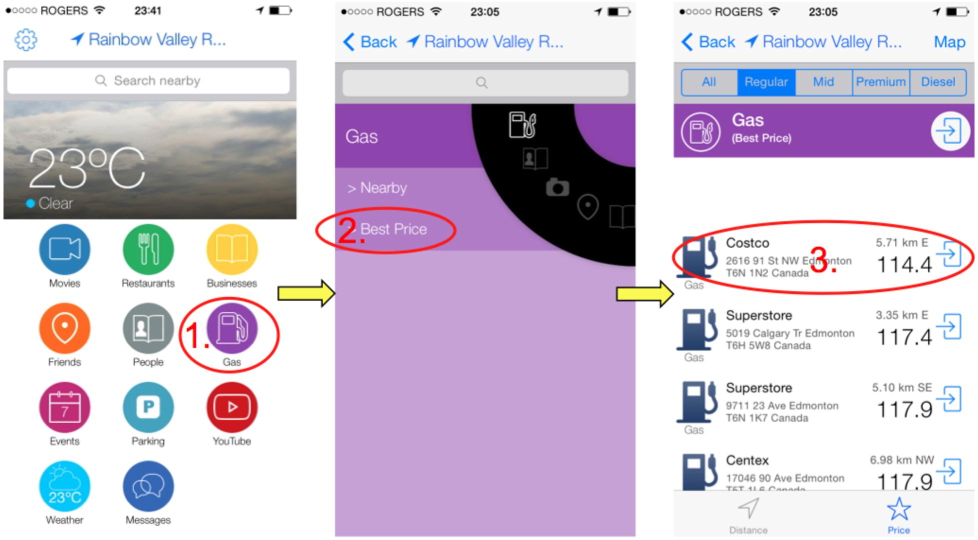
Smokehouse BBQ Inc
10810 124 St NW,Edmonton,AB,T5M 0H3
2.94 km

Jin Li Bbq
9655 102 Ave NW,Edmonton,AB,T5H 0E5
3.36 km

Sunny Meat & BBQ

(b) Searching for nearby "barbeques" businesses: 1. click over the search category "Businesses"; 2. type in the query keyword "barbeques"; 3. the nearby businesses related to "barbeques" are listed by distance (ascendingly ordered).

Figure 3.1: Typical use cases of the Poynt application. *(cont.)*



Costco
2616 91 St NW Edmonton
T6N 1N2 Canada
5.71 km E
114.4
Gas

Superstore
5019 Calgary Tr Edmonton
T6H 5W8 Canada
3.35 km E
117.4
Gas

Superstore
9711 23 Ave Edmonton
T6N 1K7 Canada
5.10 km SE
117.9
Gas

Centex
17046 90 Ave Edmonton
T5T 1L6 Canada
6.98 km NW
117.9

(c) Searching for nearby best gas price: 1. click over the search category "Gas"; 2. click over "Best Price"; 3. the nearby gas stations are listed by price (ascendingly ordered).

Figure 3.1: Typical use cases of the Poynt application.

is to find "interesting" phenomena, after which potential hypotheses can be formulated and pursued.

3.3 Filtering of Poynt Data to Focus Evidence Compilation

Any analytics framework requires us to identify appropriate subsets of the raw data, so that our analytics work can be focused on categories of hypotheses. This provides a more efficient and effective analysis, especially considering the computational complexity of analyzing the large volume of spatio-temporal Poynt search records.

Based on our framework development to date, we have adopted the following filtering constraints, for the purpose of targeting subsets of search records for experimental usage:

- **Time Span**: temporal scope restricting when target records are generated.
- **Region**: constraint controlling the two spatial attributes of target records, latitude (LAT) and longitude (LON). A target region is described in the form of a bounding rectangle, represented by two corner points, SW (South West corner) and NE (North East corner).

- **Search Category**: categories of search records, *i.e.*, Movie, Restaurant, Gas, Yellow Pages and Event.
- **User Group**: constraint to target the group of users. We currently use Search Density (SD) as the measure to select target users. It is defined as the number of search records by an individual user within a certain time span. By setting a range of record numbers and a time span, we obtain a scope of search density, *e.g.*, users who search between one to ten times in the same category in time span (week=5, year=2011), and can thus target those users whose corresponding search density falls within that scope. In the definition of search density, records by an individual user could also be further specified by search type and region if necessary. Any record whose user is in the target user group belongs to our filtered sub-dataset.
- **Source Device**: types of mobile devices held by users, *i.e.*, Android, Blackberry, and iPhone.

Generally speaking, we could gather a subset of search records for certain experimental usage (*e.g.*, geo-related hypotheses on gas search), by using a combination of filtering constraints (Figure 3.2). The selection of filtering constraints is not independent of the hypothesis to be investigated: searching for common patterns supporting the idea of "dinner and a movie" could well involve different filtering constraints than something like "how far will some one drive for cheap gas?" In the current framework, the selection of these constraints is an approximation to a business hypothesis, in the sense that the subset of data selected by these constraints is that data from which a hypothesis is supported or refuted.

3.4 Examples of Simple Hypothesis Investigation

3.4.1 Hypothesis One

Hypothesis One is in the first category of hypotheses about the distribution of individual searches in one category. For example if we hypothesize that movie searches are most active on weekends. This example investigates how records of one search

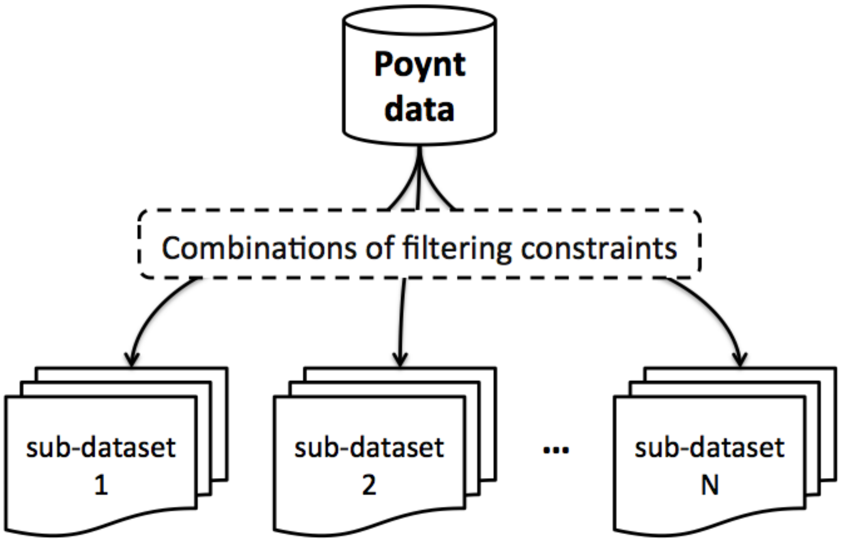


Figure 3.2: Choosing sub-datasets with combinations of filtering constraints.

type are distributed across the hours within one week, to show the corresponding peak searching periods during that week. Note that the graph version of the movie search distribution (Figure 3.3) confirms a hypothesis about movie searches most-active in the afternoon of a day and peaking on Friday and Saturday afternoons (as marked by red circle in the diagram).

3.4.2 Hypothesis Two

Hypothesis Two is in the second category which covers hypotheses on correlations amongst different search categories. For instance, what an individual might search for, movies or events, in 30 minutes after looking up nearby restaurants? The example discusses the possible correlation of one search category (Movie or Event) to the other (Restaurant). We simply suppose that a search α is "correlated" to a search β (both α and β are made by the same individual) if α is conducted within certain time window t after β. The "correlation rate" of search category *A* to category *B* ($CR_{A\text{-}B}$) in certain time span s is defined as the percentage of searches in category *A* that are conducted within s and correlated to some search in category *B*. Note that *A*-*B* differs from *B*-*A* in correlation direction. Therefore, the correlations between *A* and *B* can be measured bidirectionally by $CR_{A\text{-}B}$ and $CR_{B\text{-}A}$.

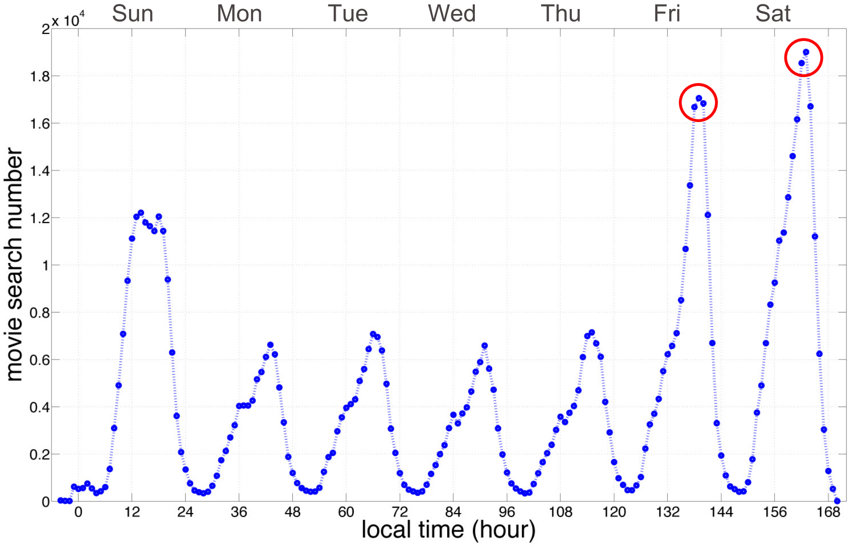


Figure 3.3: Hourly number distribution of movie searches within one week.

In the preliminary experiment, we choose two target pairs of search categories, [Movie, Restaurant] and [Event, Restaurant], and examine only the unidirectional correlations of Movie (M) or Event (E) to Restaurant (R), *i.e*, M-R and E-R. Time window t and time span s are set to twenty minutes and one hour respectively. Therefore correlation rate is computed for every hour across the entire target week. In Figure 3.4, the hourly number distributions of searches in total and searches correlated to any R search are visualized respectively in blue and red curves for the investigation of the correlations M-R and E-R. It indicates that the search category Movie is more closely correlated to Restaurant than Event in the afternoons, especially on Friday and Saturday. But the graph presentation in Figure 3.4 is immediately a challenge to interpret; it is clear we need to identify appropriate visualization methods to clearly and quickly interpret for each type of hypotheses.

3.5 Consumer Trajectories

Under the context of spatio-temporal search data analysis, a consumer trajectory is a chronologically ordered sequence of search records generated by one individual user. Given a subset of search records filtered by a combination of constraints (discussed in Section 3.3), a set of consumer trajectories can be identified by group-

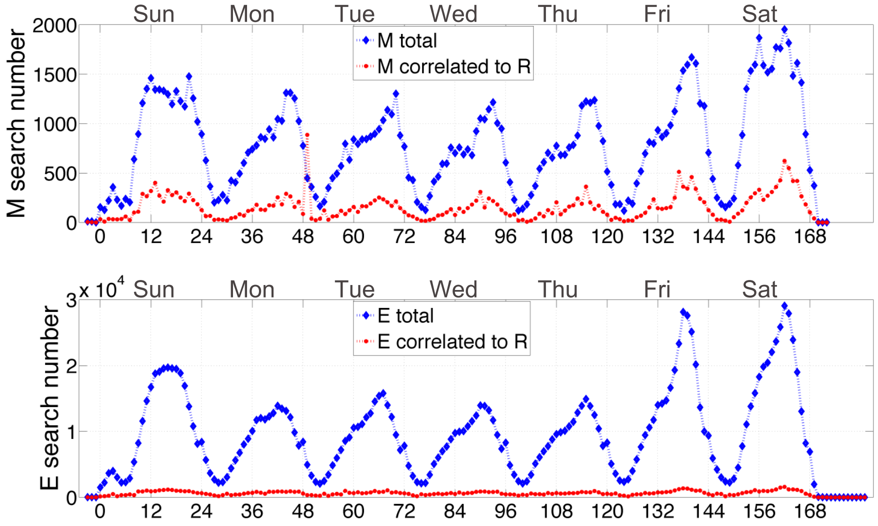


Figure 3.4: Investigate the correlations Movie-Restaurant and Event-Restaurant: the distribution of searches in total (blue) and the distribution of searches correlated to any Restaurant search (red) over the hours in one week.

ing records by individual user and ordering each group of records into sequence chronologically. Therefore we can vary the geo-extent, time frame, search type, *etc.*, to look for identifying features relevant to the target hypothesis among the corresponding consumer trajectories. Practically, we would normally restrict that one consumer trajectory involves at least two records under the specified constraints. Of course the challenge is to find appropriate selection, clustering, and visualization techniques to support a human user's confirmation of consumer trajectories.

Chapter 4

Visualization Approaches

The Poynt data in our case studies provide the basis for identifying consumer trajectories built on various subsets of search records, filtered by combinations of constraints representing the intuitive semantics of consumer trajectory components (as described in Section 3.5). Three major attributes, record transaction time, region (latitude and longitude), and node label type, are involved with the target data to be visualized.

Generally speaking, we are interested in visual evidence that exhibit spatial and/or temporal patterns for the exploratory and explanatory investigations of semantic hypotheses. In the framework of hypothesis-driven visual data analysis, we need multiple visualization schemes to depict the target data from different perspectives, to reveal unexpected features, and to aid exploratory searches for proposing new hypotheses. Exploiting the data with multiple visualizations throughout the constraint space could also focus and amplify evidence that help confirm or refute proposed hypotheses during explanatory searches.

Specifically, we use two fundamentally different visualization approaches: the Cartesian Geography Map and the Storyline Visualization.

4.1 Cartesian Geography Map

A Cartesian Geography Map visualizes a target set of consumer trajectories under the context of a conventional geography map. Each search record is marked by its normalized latitude and longitude in the first quadrant of a rectangular coordinate system where the horizontal and vertical axes represent longitude and latitude respectively. Both axes are bounded by the minimum and maximum values of the corresponding attributes (LON_{min}/LAT_{min} and LON_{max}/LAT_{max}) from the target dataset. Therefore, the origin of the coordinate system in a Cartesian Geography Map represents (LON_{min}, LAT_{min}) and the top-right corner denotes (LON_{max}, LAT_{max}). Node label types of the involved search records are distinguished in the map by their predefined node colors. Search record nodes in one consumer trajectory are connected by grey trajectory lines. An example of Cartesian Geography Map is shown in Figure 4.1a which visualizes consumer trajectories under the target region enclosed by the blue rectangle in Figure 4.1b.

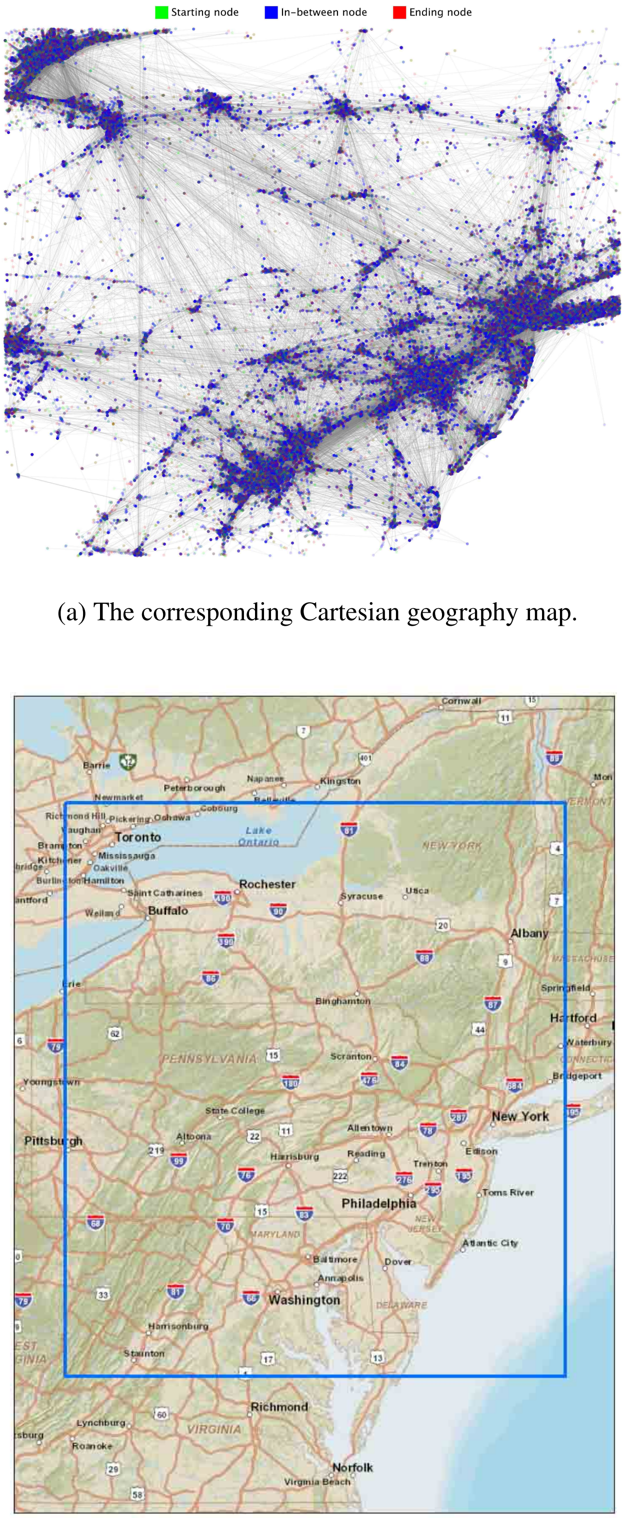
Obviously, it is straightforward to observe the spatial distribution of target consumer trajectories in a Cartesian geography map. The drawback, however, is the lack of temporal traces in one single map. To investigate how the target consumer trajectories distribute over time, we have to first divide the time span into sub-partitions with certain time granularity (*e.g.*, by day or six-hour interval), then apply multiple Cartesian geography maps to visualize respectively the subsets of consumer trajectories on the divided time span sub-partitions. This makes the approach inappropriate for temporal pattern related hypothesis analysis.

4.2 Storyline Visualization

To incorporate in one visualizing diagram how consumer trajectories distribute over time, we introduce Storyline, a visualization approach which is based on a method named "Storygraph" proposed by Shrestha et al. [17, 18].

4.2.1 Storygraph

Storygraph is a 2D diagram consisting of two parallel vertical axes $V_\alpha \subset \Re$ (on the left) and $V_\beta \subset \Re$ (on the right), and an orthogonal horizontal axis $H \subset \Re$ [18]. In our case, V_α and V_β represent, respectively, the latitude and longitude coordinates of a point on a Storygraph plane, while H represents time. All three of the axes are bounded at both ends by the minimum and maximum values of the corresponding



(a) The corresponding Cartesian geography map.

(b) The geography map of the target region (in the blue rectangle).

Figure 4.1: An example of Cartesian Geography Map.

attributes in the dataset. Values in the axes are in ascending order from left to right horizontally and bottom to top vertically.

Obviously Storygraph is suitable for visualizing spatio-temporal search records. We normally refer to a node representing a search record plotted in Storygraph as an *event* because it indicates where and when a consumer search was conducted. Figure 4.2 illustrates how two events in a regular 2D Cartesian map, both coded with the same location coordinates (lon, lat) but with different timestamps t and $t + 1$, are presented in Storygraph. The line segment connecting the two points on the vertical axes, $lat \in V_\alpha$ and $lon \in V_\beta$, indicates the location coordinates (lon, lat) of the corresponding Storygraph event nodes on it. We refer to this type of lines in Storygraph as *location lines*. Location lines are important to Storygraph in that any event node plotted without it in Storygraph loses the location information and represents some event occurred at certain time with unknown location coordinates.

As shown in Figure 4.2, the coordinates of an event (lon, lat, t) on a Storygraph plane are determined by the intersection of its location line and the vertical line $x = t$ which indicates the event time t. The function $f(lon, lat, t) \rightarrow (x, y)$ which maps an event (lon, lat, t) to the coordinate (x, y) on a 2D Storygraph plane can be formally written as follows:

$$x = t \tag{4.1}$$

$$y = \frac{(lon - lat)(x - T_{min})}{T_{max} - T_{min}} + lat \tag{4.2}$$

where T_{min} and T_{max} are respectively the minimum and maximum timestamps of the dataset.

Figure 4.2 also shows the advantage of Storygraph over Cartesian Geography Map when visualizing data with the temporal attribute.

4.2.2 Generating Storygraph-based Storylines

As described in Section 3.5, a consumer trajectory is a chronologically ordered sequence of search records (events) by one individual user. Based on Storygraph, the Storyline visualization of a target consumer trajectory is constructed by sequentially connecting all involved events visualized in Storygraph with trajectory lines,

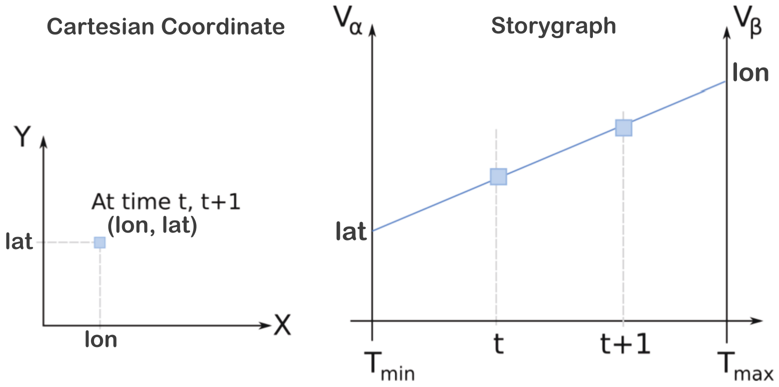


Figure 4.2: Cartesian versus Storygraph [18].

thus telling a story about the spatio-temporal events of the user. Therefore the main components of a consumer trajectory visualized in Storyline include the Storygraph nodes and the corresponding location lines for all involved events, as well as the trajectory lines connecting adjacent nodes. Algorithm 1 describes how to construct a Storygraph-based Storyline visualization given a set of target search records. In this algorithm, transparency level is used to help indicate the relative densities of components that overlap in the visualization.

In the original scenario of Storyline based on Storygrah by Shrestha et al. [17, 18], event nodes and connecting lines are distinctly colored by individual user to help visualize the movements of characters and the interactions amongst them on events. We introduce a node label, by which Storygraph nodes of consumer trajectories are colored, to Storyline as an extra dimension (in addition to location and time) created for the attribute semantically matching the hypothesis under analysis (*e.g.*, search type, source device, and semantic cluster). Unlike the original authors, we are more interested in the application of Storyline to discover general visual patterns shown spatially and/or temporally under assorted contexts of node labels over large numbers of users, instead of single user's behaviors.

Figure 4.3a showcases an example of Storygraph-based Storyline Visualization where location lines are colored in light blue and trajectory lines in grey. We can see that, as with the increase of the number of target consumer trajectories to be visualized, the problem of occlusion, cluttering and color mixture turns visual con-

Algorithm 1 Construct Storygraph-based Storyline visualization.

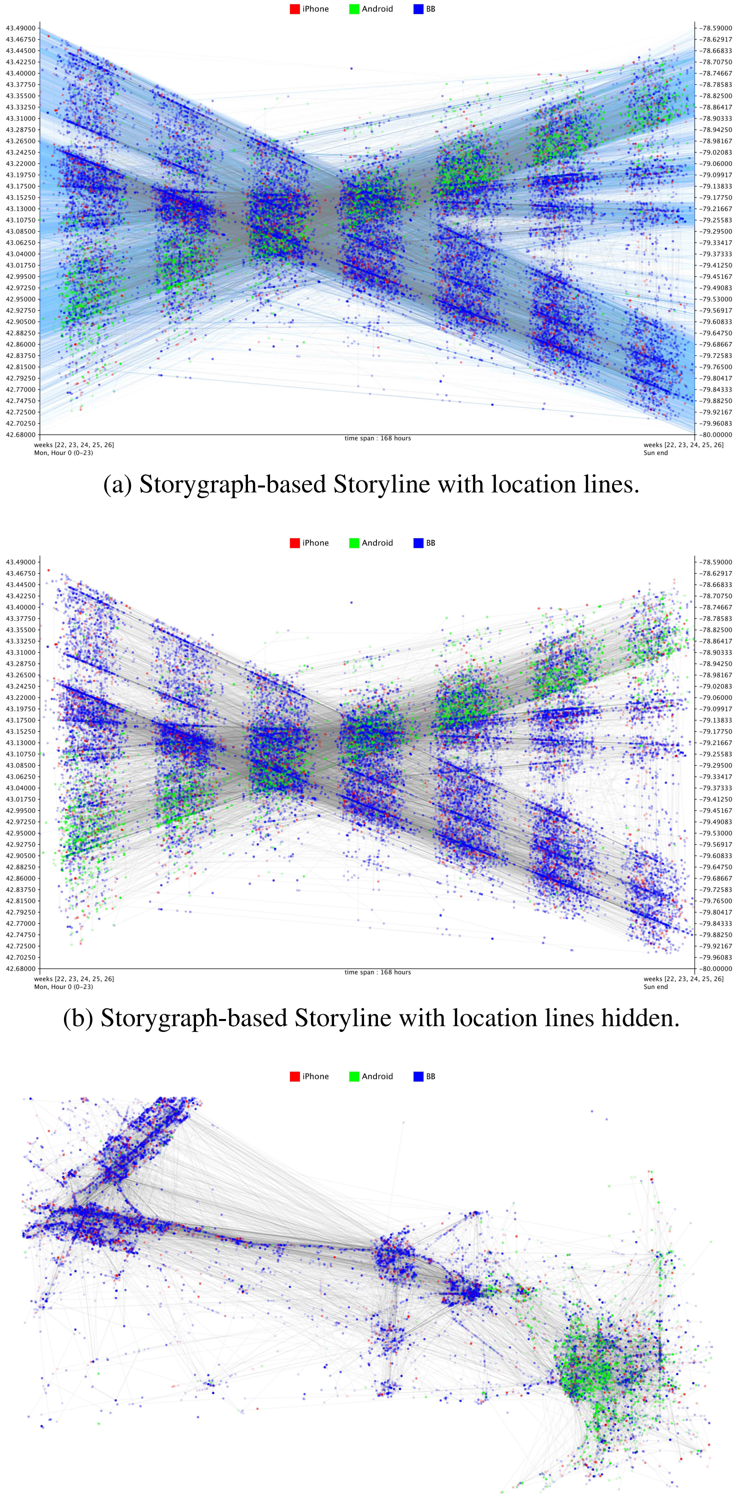
```
**Given**: a set of target search records
Preprocess the original set of search records:
group records by individual user and order each group of records chronologically, thus constructing $S$, the set of target consumer trajectories to be visualized (each group of time-ordered records is a consumer trajectory of the individual user);
retrieve the minimum and maximum values of the attributes latitude, longitude and transaction time of all records in $S$: $LAT_{min}/LAT_{max}$, $LON_{min}/LON_{max}$, $T_{min}/T_{max}$.

$prev_x \leftarrow 0, prev_y \leftarrow 0$
**for** each consumer trajectory $ct$ in $S$ **do**
    **for** each search record $r$ $(lat, lon, t, label)$ in $ct$ **do**
        Normalize $lat$, $lon$, and $t$ respectively such that $lat$ and $lon$ are normalized to the interval $[1, L]$, and $t$ is normalized to the interval $[1, W]$ (all rounded to the nearest integer):
            $lat_{norm} \leftarrow \left[\frac{lat-LAT_{min}}{LAT_{max}-LAT_{min}}(L-1)\right]+1$
            $lon_{norm} \leftarrow \left[\frac{lon-LON_{min}}{LON_{max}-LON_{min}}(L-1)\right]+1$
            $t_{norm} \leftarrow \left[\frac{t-T_{min}}{T_{max}-T_{min}}(W-1)\right]+1$

        Compute the pixel coordinate $(curr_x, curr_y)$ of the current trajectory node $r$ (according to Equations 4.1, 4.2):
            $curr_x \leftarrow t_{norm}$
            $curr_y \leftarrow \left[\frac{curr_x-T_{min}}{T_{max}-T_{min}}(lon_{norm}-lat_{norm})\right]+lat_{norm}$

        **if** $prev_x \neq 0$ **and** $prev_x \neq curr_x$ **and** $prev_y \neq curr_y$ **then**
            Draw trajectory line segment joining pixels $(prev_x, prev_y)$ and $(curr_x, curr_y)$ with the RGB color and transparency level pre-defined for trajectory line;
        **end if**
        $prev_x \leftarrow curr_x,\ prev_y \leftarrow curr_y$

        Draw Storygraph location line segment joining pixels $(1, lat_{norm})$ and $(W, lon_{norm})$ with the RGB color and transparency level pre-defined for Storygraph location line;
        Draw trajectory node centered at pixel $(curr_x, curr_y)$ with the RGB color and transparency level pre-defined for trajectory node of the type $label$;
    **end for**
**end for**
```

fusion in the visualization. Storygraph location lines in particular contribute the most. Under such circumstances, hiding location lines helps alleviate the issue without losing much of the involved events' location information (see Figure 4.3b). This is useful, since events in one consumer trajectory normally align with each other around the hidden location lines and the trajectory lines connecting them give a close sense of the event location (if the user rarely moves) or the user movement, especially with the property of the Poynt data whose consumers exhibit geographical locality to some extent while conducing searches. Therefore we normally hide Storygraph location lines in Storyline visualization when the target data is dense in the 2D visualizing space.

4.3 Augmenting Storyline with Frequency Plots

When the target dataset scales up drastically, hiding location lines in a normal Storygraph-based Storyline visualization (Section 4.2) is not enough to ensure a visually satisfactory result for hypothesis analysis. Since Algorithm 1 simply draws consumer trajectory components in Storyline overlapping each other under certain transparency levels, the issues of occlusion, cluttering and color mixture created by the overlapping of trajectory lines and Storygraph event nodes (without location lines) cannot be ignored. Figure 4.4a exhibits such a problematic example.

Based on the above motivation, a strategy named *Frequency Plots* [1] is introduced to augment normal Storygraph-based Storyline (described in Section 4.2) when confronting the cluttering and occlusion issues caused by scaled-up data. The idea of *Frequency Plots* is to associate the content to be drawn, in our case assorted consumer trajectory components in Storyline, with frequency information in a pixel-wise manner in order to reflect the relative density of involved components in the visualization space. Specifically, for each pixel, the color indicates the type of the most-frequent component on the spot (if any), and the lightness is set proportionally to the corresponding frequency value so that in general components associated with higher frequency superimpose those with lower frequency in terms of pixel lightness.



(a) Storygraph-based Storyline with location lines.

(b) Storygraph-based Storyline with location lines hidden.

(c) Corresponding Cartesian Geography Map.

Figure 4.3: Examples of Storygraph-based Storyline with or without location lines.

Considering that the method of Frequency Plots is applied to large dense data, location lines of Storygraph event nodes are normally hidden in a Frequency Plots augmented Storyline visualization. Therefore the consumer trajectory components to be visualized in a Storyline augmented by Frequency Plots are Storygraph event nodes of different label types and trajectory line, and the frequency information is cumulatively gathered pixel-wise by component type (note that event nodes of different labels are considered as components of different types, thus their frequency information are computed separately). In addition, it is inappropriate to mix up the frequency information of event node and trajectory line when applying Frequency Plots since they are components of two different general classes and the reflected "relative density" should be in terms of other components in the same class. Therefore, in the actual implementation of Frequency Plots augmented Storyline, we handle event node and trajectory line separately with Frequency Plots in two independent layers under the same discreet screen system. To render the final Storyline output, we have to make a decision on which layer of the two shows on top. Practically speaking, event node is more important than trajectory line since the latter is for auxiliary purpose to indicate transition or connectivity, and general patterns involved with event node labels (*e.g.*, search categories, source devices) are of most interest. Therefore, a Frequency Plots augmented Storyline visualization is constructed with the layer of event nodes superimposed on that of trajectory lines. Figure 4.4b shows an example of a Frequency Plots augmented Storyline, which, compared to its corresponding normal Storyline visualization (Figure 4.4a), reduces visual occlusion and cluttering, and exhibits relatively clearer clusters of consumer trajectories.

Besides, the individual layer of trajectory lines in Figure 4.4b is shown in Figure 4.4c. When the data is extremely dense, the layer of trajectory lines could be completely covered by the layer of event nodes in a Frequency Plots augmented Storyline. Under such circumstances, we could check the layer of trajectory lines individually if necessary since it helps provide a general sense about where strong patterns of node labels reside in the visualization space.

In summary, the strategy of Frequency Plots helps avoid the issue of color mixture in a normal Storyline, which relies on simple transparency levels to reflect relative density, and alleviates occlusion by selectively displaying the "most important" content (in terms of component pixel-wise frequency). In addition to the above, we enable interactive zooming in Storyline visualization, which allows the analysis of selected subareas with finer granularity and helps alleviate the problems of cluttering and occlusion. Note it is clear that general methods for removing visual clutter will provide motivation for the development of dynamic interactive rendering (*e.g.*, [21, 7, 8]).

Details about building a Storyline visualization augmented by Frequency Plots is described in Algorithm 2. An example augmenting Figure 4.4a is shown in Figure 4.4b. Our contribution to the original usage of Frequency Plots by Artero et al. [1] is applying Frequency Plots for visualization augmentation under the context of multiple types of components to be visualized (*e.g.*, event nodes of different label types, trajectory line) in ARGB color space.

4.4 Questions Answered by the Visualizations

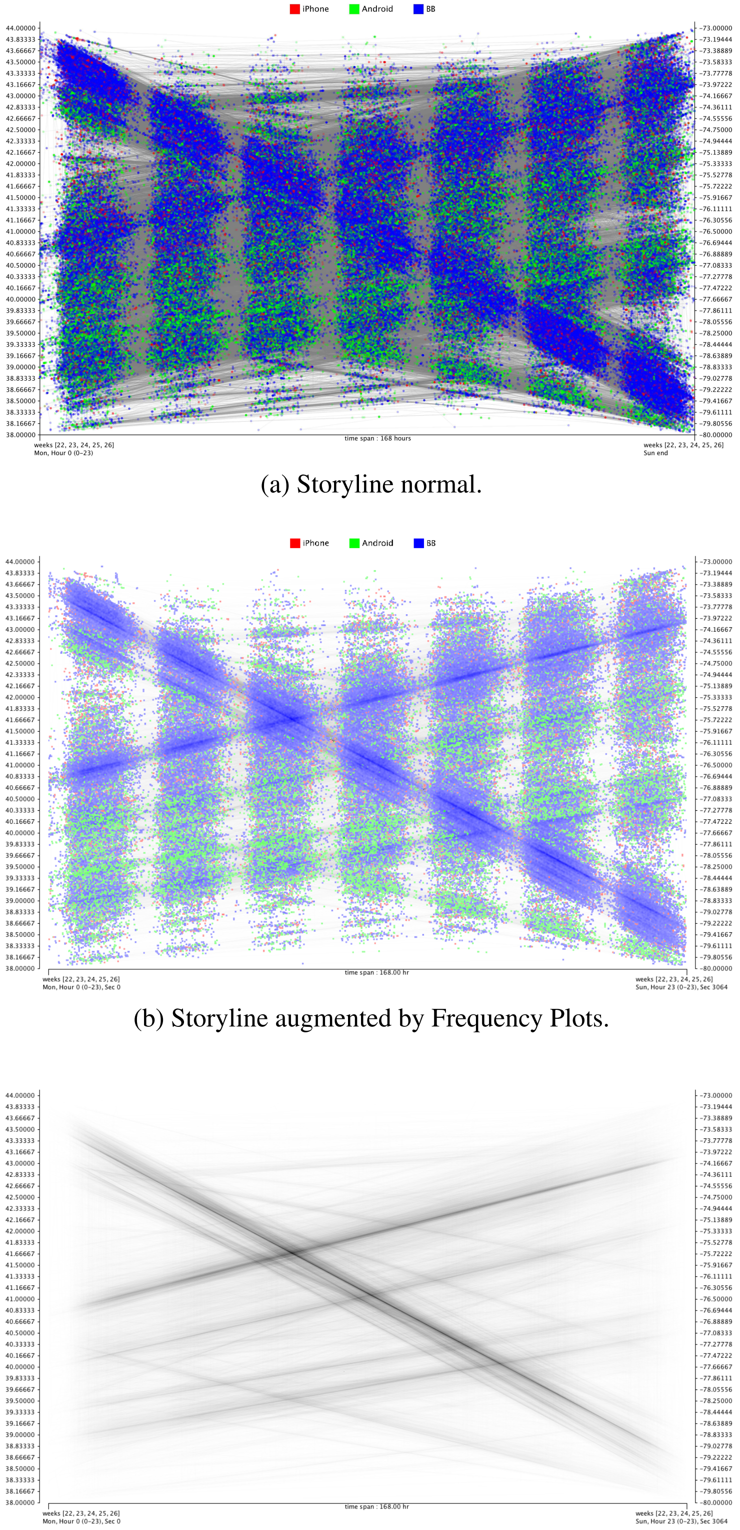
This section summarizes the kinds of questions that the selection, filtering and visualization tools can address.

4.4.1 Search Category Selection and Comparison

From the original relational database, the search category selection tool chooses target Poynt search categories (*e.g.*, Figure 4.5). By default, consumer trajectories built on the data of the selected search categories are visualized in a Cartesian geography map over the entire time span and spatial region of the dataset.

4.4.2 Spatial Abstraction and Comparison

After target search categories are selected, the geographic selection tool defines a region that can be mapped to a Cartesian visualization, so as to show the spatial distribution of the consumer trajectories built on the Poynt user queries of the target search categories in the chosen region (*e.g.*, Figure 4.6).



(a) Storyline normal.

(b) Storyline augmented by Frequency Plots.

(c) individual Storyline layer of trajectory lines.

Figure 4.4: An example of Storyline augmented by Frequency Plots.

Algorithm 2 Construct Storyline visualization augmented by Frequency Plots.

REQUIRE:
S: the set of target consumer trajectories to be visualized;
LAT_{min}/LAT_{max}, LON_{min}/LON_{max}, T_{min}/T_{max}: the minimum and maximum values of the attributes latitude, longitude and transaction time of all records in S.

[Step 1] Initialize pixel color matrix $G_{L \times W}$, whose dimension L and W are defined by the plot's pixel resolution (L is determined by the vertical resolution, W is determined by the horizontal resolution), with zeros:

Let $G_{[p][q]} = 0$, for $p = 1, \ldots, L$ and $q = 1, \ldots, W$

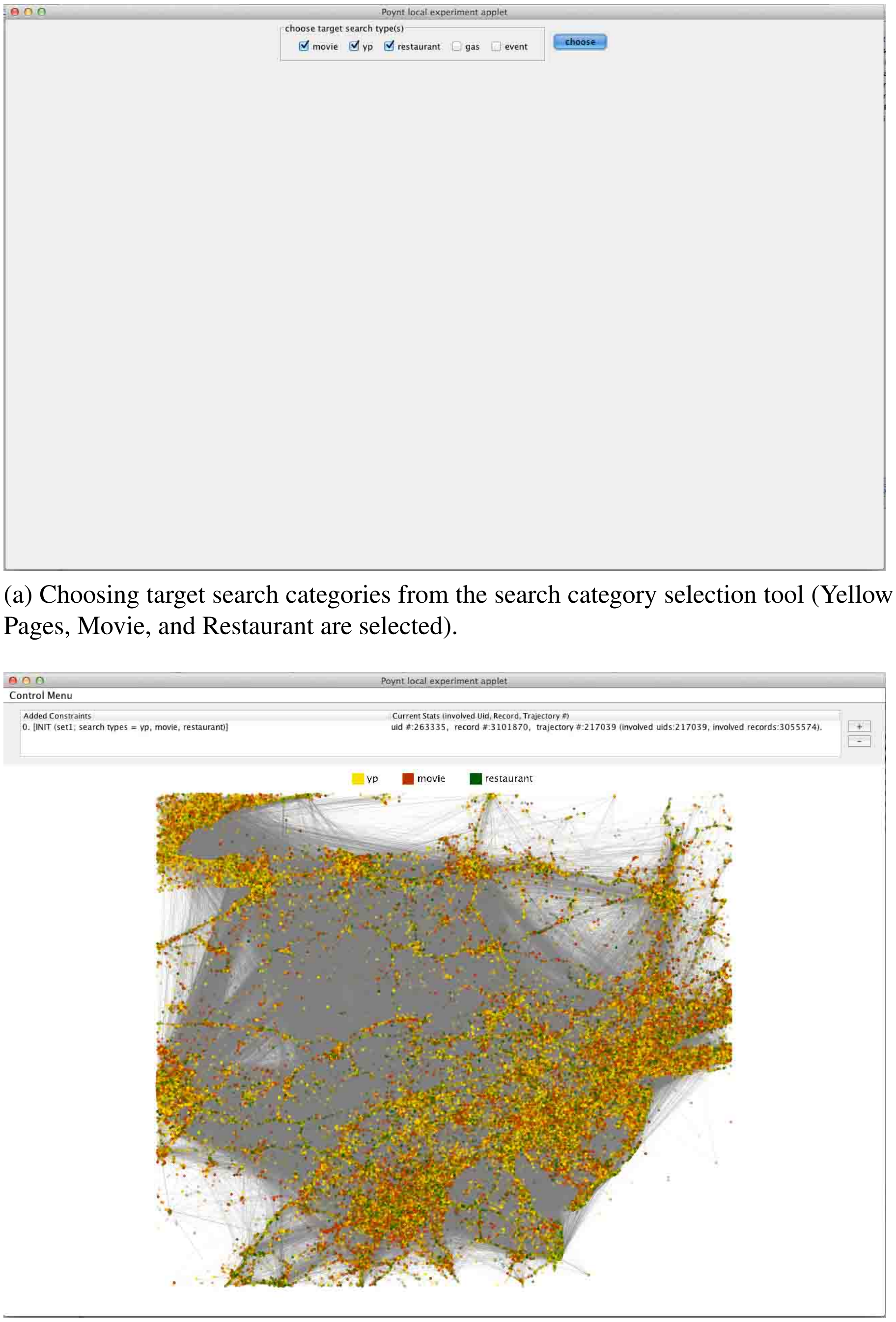
[Step 2] Normalize each record r $(lat, lon, t, label)$ in S to $(lat_{norm}, lon_{norm}, t_{norm}, label)$ such that lat and lon are normalized to the interval $[1, L]$, and t is normalized to the interval $[1, W]$ (all rounded to the nearest integer):

$$lat_{norm} \leftarrow \left[\frac{lat - LAT_{min}}{LAT_{max} - LAT_{min}}(L-1)\right] + 1$$
$$lon_{norm} \leftarrow \left[\frac{lon - LON_{min}}{LON_{max} - LON_{min}}(L-1)\right] + 1$$
$$t_{norm} \leftarrow \left[\frac{t - T_{min}}{T_{max} - T_{min}}(W-1)\right] + 1$$

[Step 3] Visualize S with normalized records: draw consumer trajectory nodes (Storygraph event nodes) and trajectory lines

3.1 compute pixel frequency matrix of trajectory lines $P\_tjln_{L \times W}$ and pixel frequency matrix by node label type $P\_type_{L \times W \times K}$ (K is the number of record node label types)
Let $P\_tjln_{[i][j]} = 0$, for $i = 1, \ldots, L$ and $j = 1, \ldots, W$
Let $P\_type_{[i][j][k]} = 0$, for $i = 1, \ldots, L$, $j = 1, \ldots, W$ and $k = 1, \ldots, K$
$prev_x \leftarrow 0, prev_y \leftarrow 0$

```
**for** each consumer trajectory $ct$ in $S$ **do**
    **for** each search record $r$ $(lat_{norm}, lon_{norm}, t_{norm}, label)$ in $ct$ **do**
        compute the pixel coordinate $(curr_x, curr_y)$ of the current node $r$ (according to Equations 4.1, 4.2):
            $curr_x \leftarrow t_{norm}$
            $curr_y \leftarrow \left[\frac{curr_x - T_{min}}{T_{max} - T_{min}}(lon_{norm} - lat_{norm})\right] + lat_{norm}$
        $P\_type_{[L+1-curr_y][curr_x][label]} \leftarrow P\_type_{[L+1-curr_y][curr_x][label]} + 1$
        **if** $prev_x \neq 0$ **then**
            Use the Bresenham algorithm to compute the coordinates $(t_x, t_y)$ of all pixels in the line segment joining pixels $prev : (prev_x, prev_y)$ and $curr : (curr_x, curr_y)$; for each position $(t_x, t_y)$ computed, increment the corresponding pixel frequency in matrix $P\_tjln$:
                $P\_tjln_{[L+1-t_y][t_x]} = P\_tjln_{[L+1-t_y][t_x]} + 1$
        **end if**
        $prev_x \leftarrow curr_x, \ prev_y \leftarrow curr_y$
    **end for**
**end for**

**3.2** update $G$ with $P\_tjln$ and $P\_type$
**for** each $P\_tjln_{[i][j]} \neq 0$, with $i = 1, \ldots, L$ and $j = 1, \ldots, W$ **do**
    compute $alpha$ (color lightness of pixel $(j, L + 1 - i)$) proportionally to $P\_tjln_{[i][j]}$;
    set $G_{[i][j]}$ with ARGB color value determined by $alpha$ and the pre-defined RGB color value of trajectory line;
**end for**
**for** each matrix index pair $(i, j)$ satisfying $\exists k' \in [1, K]$ such that $P\_type_{[i][j][k']} \neq 0$, with $i = 1, \ldots, L$ and $j = 1, \ldots, W$ **do**
    $label\_type \leftarrow \{k \mid \forall k' \in [1, K] : P\_type_{[i][j][k']} \leq P\_type_{[i][j][k]}\}$
    compute $alpha$ (color lightness of pixel $(j, L + 1 - i)$) proportionally to $P\_type_{[i][j][label\_type]}$;
    set $G_{[i][j]}$ with ARGB color value determined by $alpha$ and the RGB color value pre-defined for trajectory node of the type $label\_type$;
**end for**

**[Step 4]** Display the resulting Storyline visualization augmented by Frequency Plots stored in matrix $G$
```



choose target search type(s)

☑ movie ☑ yp ☑ restaurant □ gas □ event

(a) Choosing target search categories from the search category selection tool (Yellow Pages, Movie, and Restaurant are selected).

Added Constraints

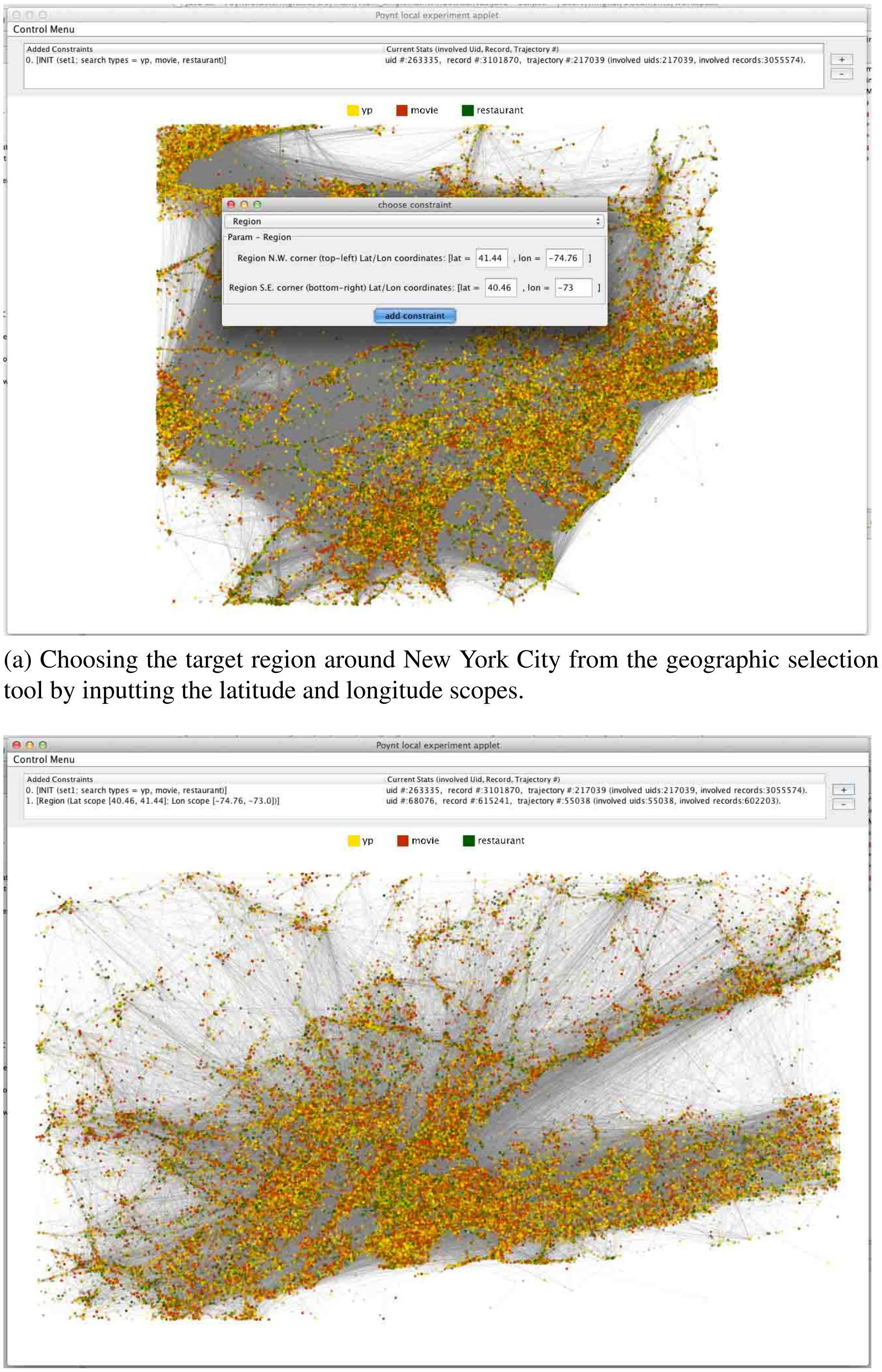
0. [INIT (set1; search types = yp, movie, restaurant)]

Current Stats (involved Uid, Record, Trajectory #)

uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).

(b) The consumer trajectories built on the data of the selected search categories (over the entire time span and spatial region of the dataset) are visualized in a Cartesian geography map by default.

Figure 4.5: An example of Poynt search category selection.



Added Constraints
0. [INIT (set1; search types = yp, movie, restaurant)]

Current Stats (involved Uid, Record, Trajectory #)
uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).

Region
Param – Region
Region N.W. corner (top-left) Lat/Lon coordinates: [lat = 41.44 , lon = -74.76]
Region S.E. corner (bottom-right) Lat/Lon coordinates: [lat = 40.46 , lon = -73]

(a) Choosing the target region around New York City from the geographic selection tool by inputting the latitude and longitude scopes.

Added Constraints
0. [INIT (set1; search types = yp, movie, restaurant)]
1. [Region (Lat scope [40.46, 41.44]; Lon scope [-74.76, -73.0])]

Current Stats (involved Uid, Record, Trajectory #)
uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).
uid #:68076, record #:615241, trajectory #:55038 (involved uids:55038, involved records:602203).

(b) The consumer trajectories built on the data of the selected search categories (Yellow Pages, Movie, and Restaurant) in the chosen region are visualized in a Cartesian geography map.

Figure 4.6: An example of region selection (under Yellow Pages, Movie and Restaurant search categories).

This spatial abstraction can be further refined by restricting the kinds of target source devices with the source device selection tool (*e.g.*, Figure 4.7).

4.4.3 Temporal Abstraction and Comparison

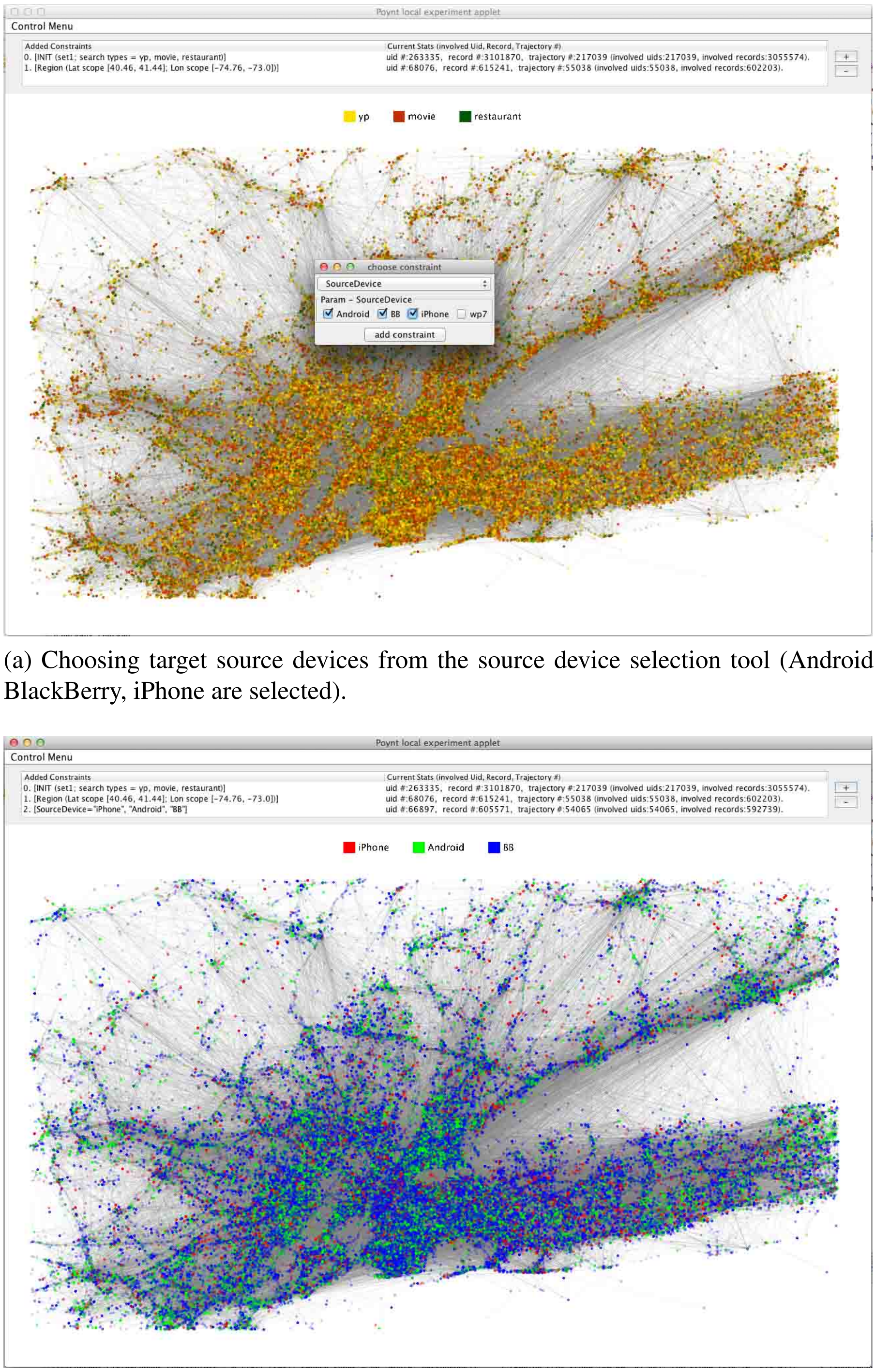
After target search categories are selected, the temporal selection tool can be used to set a time span that can be mapped to a Storyline visualization, so as to show the temporal distribution of the consumer trajectories built on the Poynt user queries of the target search categories within the selected time span (*e.g.*, Figure 4.8).

Furthermore, like the spatial distribution tool, alternative selections of other attributes (*e.g.*, source device) can be further selected, to show the temporal distribution of the consumer trajectories built on the selected data.

4.4.4 Yellow Pages Keyword Cluster Selection and Comparison

The semantic distribution of keywords used in Yellow Pages search can be investigated by using a variety of clustering methods on keyword data, to visually investigate how different keyword clusters impact both spatial and temporal distribution of Poynt user key word search behaviour. In our cases studies, we did not pursue many alternative key word classifications, but have created Yellow Pages keyword clusters by hand (see Section 6.2.1 for details).

When the category Yellow Pages is selected as the only search category, we can further choose target Yellow Pages keyword clusters from the Yellow Pages keyword cluster selection tool, and show the spatial and temporal distributions of the consumer trajectories built on the Poynt user data of the selected keyword clusters (by default over the entire time span and spatial region of the dataset) (*e.g.*, Figure 4.9).



Added Constraints	Current Stats (involved Uid, Record, Trajectory #)
0. [INIT (set1; search types = yp, movie, restaurant)]	uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).
1. [Region (Lat scope [40.46, 41.44]; Lon scope [-74.76, -73.0])]	uid #:68076, record #:615241, trajectory #:55038 (involved uids:55038, involved records:602203).

SourceDevice

Param – SourceDevice

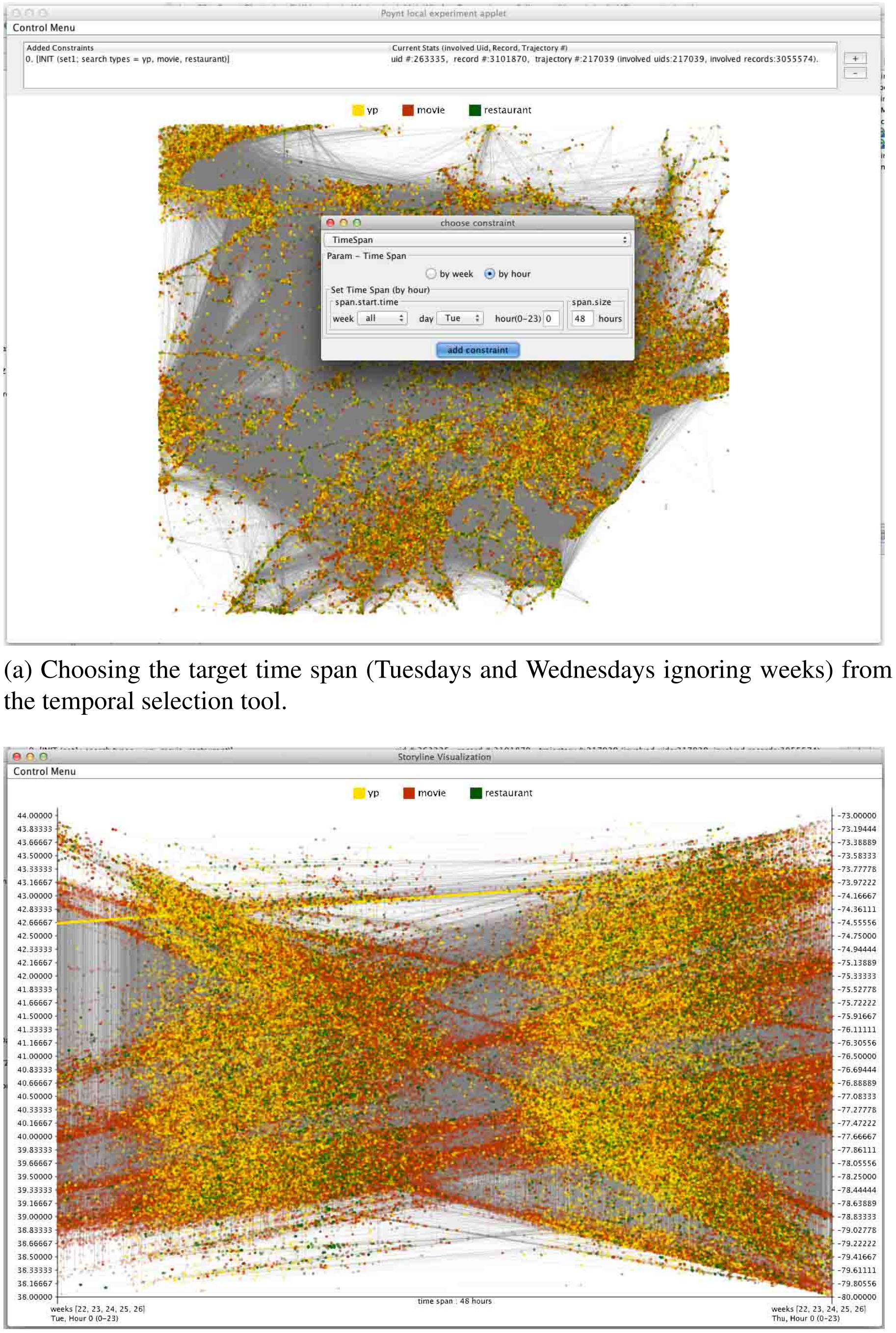
☑ Android ☑ BB ☑ iPhone □ wp7

(a) Choosing target source devices from the source device selection tool (Android, BlackBerry, iPhone are selected).

Added Constraints	Current Stats (involved Uid, Record, Trajectory #)
0. [INIT (set1; search types = yp, movie, restaurant)]	uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).
1. [Region (Lat scope [40.46, 41.44]; Lon scope [-74.76, -73.0])]	uid #:68076, record #:615241, trajectory #:55038 (involved uids:55038, involved records:602203).
2. [SourceDevice="iPhone", "Android", "BB"]	uid #:66897, record #:605571, trajectory #:54065 (involved uids:54065, involved records:592739).

(b) The consumer trajectories built on the data of the selected search categories (Yellow Pages, Movie and Restaurant) in the chosen region and further refined by the chosen source devices (Android, BlackBerry, iPhone) are visualized in a Cartesian geography map.

Figure 4.7: An example of spatial abstraction further refined by restricting the kinds of target source devices.



Added Constraints
0. [INIT (set1; search types = yp, movie, restaurant)]

Current Stats (involved Uid, Record, Trajectory #)
uid #:263335, record #:3101870, trajectory #:217039 (involved uids:217039, involved records:3055574).

TimeSpan

Param – Time Span

○ by week ● by hour

Set Time Span (by hour)

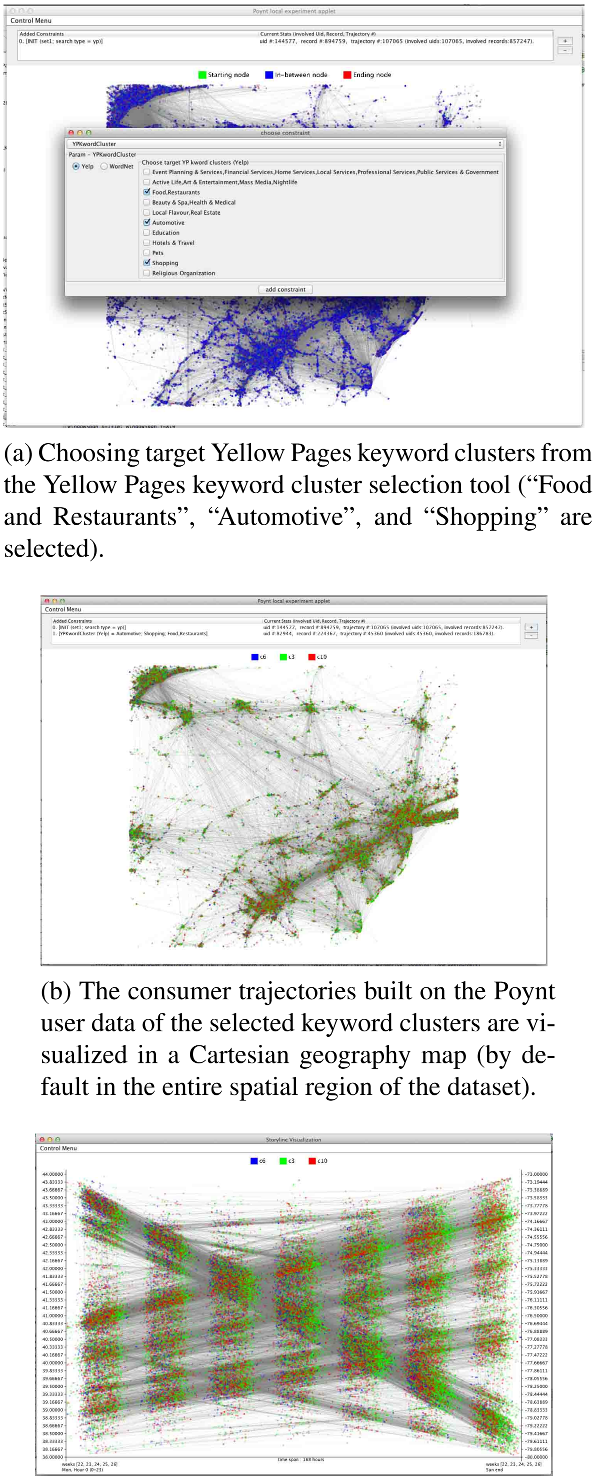
span.start.time: week all, day Tue, hour(0–23) 0

span.size: 48 hours

(a) Choosing the target time span (Tuesdays and Wednesdays ignoring weeks) from the temporal selection tool.

(b) The consumer trajectories built on the data of the selected search categories (Yellow Pages, Movie and Restaurant) within the chosen time span are visualized in a Storyline visualization.

Figure 4.8: An example of time span selection (under Yellow Pages, Movie and Restaurant search categories).



Added Constraints
0. [INIT (set1; search type = yp)]

Current Stats (involved Uid, Record, Trajectory #)
uid #:144577, record #:894759, trajectory #:107065 (involved uids:107065, involved records:857247).

YPKwordCluster

● Yelp ○ WordNet

Choose target YP kword clusters (Yelp)

- □ Event Planning & Services,Financial Services,Home Services,Local Services,Professional Services,Public Services & Government
- □ Active Life,Art & Entertainment,Mass Media,Nightlife
- ☑ Food,Restaurants
- □ Beauty & Spa,Health & Medical
- □ Local Flavour,Real Estate
- ☑ Automotive
- □ Education
- □ Hotels & Travel
- □ Pets
- ☑ Shopping
- □ Religious Organization

(a) Choosing target Yellow Pages keyword clusters from the Yellow Pages keyword cluster selection tool ("Food and Restaurants", "Automotive", and "Shopping" are selected).

Added Constraints
0. [INIT (set1; search type = yp)]
1. [YPKwordCluster (Yelp) = Automotive; Shopping; Food,Restaurants]

Current Stats (involved Uid, Record, Trajectory #)
uid #:144577, record #:894759, trajectory #:107065 (involved uids:107065, involved records:857247).
uid #:82944, record #:224367, trajectory #:45360 (involved uids:45360, involved records:186783).

(b) The consumer trajectories built on the Poynt user data of the selected keyword clusters are visualized in a Cartesian geography map (by default in the entire spatial region of the dataset).

(c) The consumer trajectories built on the Poynt user data of the selected keyword clusters are visualized in a Storyline visualization (by default over the entire time span of the dataset).

Figure 4.9: An example of Yellow Pages keyword cluster selection.

Part III

Case Studies and Discussions

Chapter 5

Case Studies to Confirm Evidence for Consumer Trajectories

In this chapter, we experiment on target data in the form of consumer trajectories and visualize them in Cartesian Geography Map and Storygraph-based Storyline, in order to tell spatio-temporal stories and analyze more complex hypotheses about the data with exploratory and explanatory searches.

5.1 A Glimpse of the Experimental Data

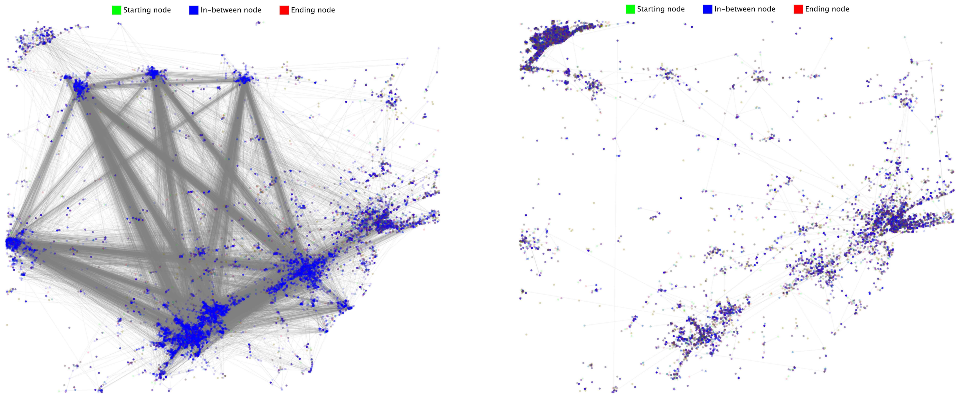
The dataset of Poynt search records based on which target consumer trajectories are constructed for experiment temporally spans five consecutive weeks, Week of May 29 to Week of June 26, in the year 2011. The general geographic region in our case studies (the area enclosed by the blue rectangle in Figure 4.1b, $LAT[38, 44]$ and $LON[-80, -73]$) mainly involves two states in the northeast US, Pennsylvania and New York, and southern Ontario in Canada. Covered are several metropolitan areas including New York City, Toronto, Washington, and Philadelphia.

5.2 Exploratory and Explanatory Searches

Exploratory and explanatory searches (Section 3.2) are important to the framework of hypothesis-driven visual data analysis. In this section, we show with an example how exploratory and explanatory searches are alternated and interleaved during the processes of proposing new hypotheses and confirming/refuting proposed ones.

First we start with a simple hypothesis: "People who search for movies are most active on the afternoons of the weekend." It is explanatory search since we have an expected answer in mind ("the afternoons of the weekend") before the investigation. This question has already been answered in Section 3.4.1 by checking the hourly distribution of movie searches within one week. From the histogram in Figure 3.3, we conclude that movie searches are most-active in the afternoon of a day and peak on Friday and Saturday afternoons. Therefore the explanatory search ends with our proposed hypothesis refuted.

Next, we would like to compare how users with different source devices (*e.g.*, Android and BlackBerry) behave on movie search during one of its peak periods, Friday afternoons. This is an exploratory search triggered on the previous explanatory search. Among the visualization results, we first notice an evident difference from the Cartesian geography maps that Android users (Figure 5.1a) exhibit strong trajectory connectivity between metropolitan centers. This may indicate that Android users tend to move more actively in long distances than BlackBerry users during the peak period of Movie search, which we then propose as a new hypothesis and try to confirm with more evidence. This is a typical example that shows how interesting phenomena appearing unexpectedly during exploration may help invoke explanatory searches with new hypothesis inspired by the exploratory findings.

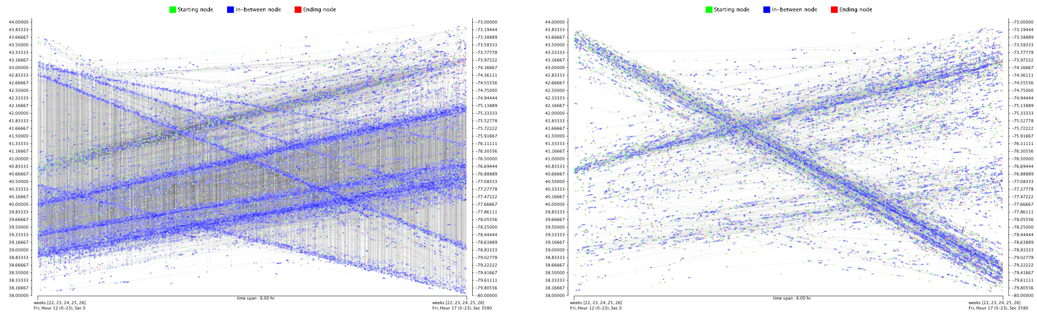


(a) Android.

(b) BlackBerry.

Figure 5.1: Cartesian Geography Map visualizations of movie searches by Android/BlackBerry users on Friday afternoons.

On the other hand, there are large numbers of vertical trajectory lines existing only in the Storyline visualization of Android users (Figure 5.2a), which indicates that those involved movie searches are abnormally made within extremely short time interval by the same individual users at multiple locations that are distant from each other geographically. Further investigation shows that these abnormal consumer trajectories belong to only a few Android users who make an abnormally high number of searches periodically (*e.g.*, about 12,000 movie searches per week), and the involved searches are normally conducted by one individual user within short time intervals at different locations far away from each other geographically (*e.g.*, searches conducted around New York City and Philadelphia, respectively, in just two minutes). The strong trajectory connectivity shown exclusively in the Cartesian map of Android users also results from these abnormal users. Apparently we should ignore these abnormal users when analyzing consumer behaviors. In all the following experiment sections, the records associated with abnormal Android users are filtered out from target data whenever applicable, in order to avoid the corresponding noise that they create in the analysis of consumer behaviors.

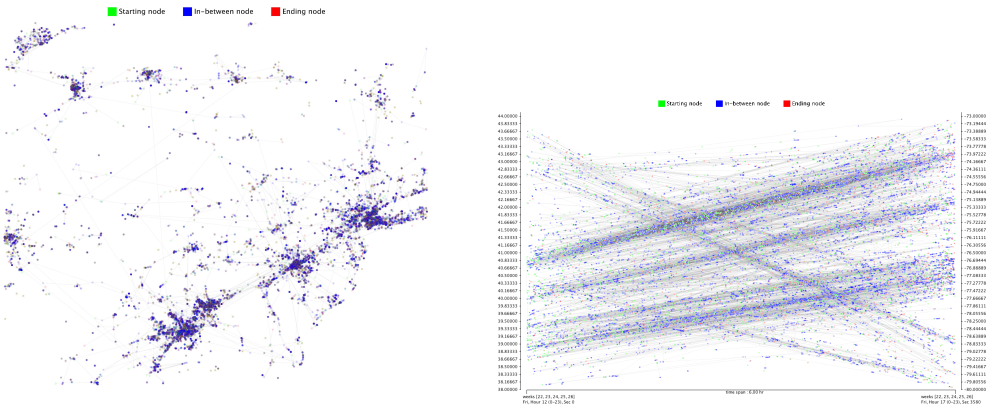


(a) Android.

(b) BlackBerry.

Figure 5.2: Storyline visualizations of movie searches by Android/BlackBerry users on Friday afternoons.

Figure 5.3 shows the visualizations after abnormal users are removed from Android users. As we can see in Figure 5.3a and Figure 5.3b, the strong trajectory connectivity in Figure 5.1a and the prevailing vertical trajectory lines in Figure 5.2a disappear accordingly. Besides, after the removal of abnormal users, there is no significant difference between Android and BlackBerry users in terms of active long-distance movements in both visualizations. Therefore our second explanatory search terminates by denying the proposed hypothesis. Meanwhile, supported to some extent is the exploratory search for the comparison of Android and BlackBerry users on movie search during the peak period, by comparing Figure 5.1b with Figure 5.3a and Figure 5.2b with Figure 5.3b. The observation is that Android users are mostly active across the metropolitan areas from the US side, while BlackBerry users mainly gather around Toronto, then New York City.



(a) Cartesian geography map.

(b) Storyline.

Figure 5.3: Visualizations of movie searches by Android users (with abnormal users removed) on Friday afternoons.

As you can see in the above example, the co-display of both a Cartesian and Storyline visualization provides two different perspectives on the same selection of data and helps reveal unexpected features. We have implemented a graphical tool assisting with the interactive investigation of Poynt data in exploratory and explanatory searches. It enables dynamic selection of filtering constraints and interactive zooming over the rendered Cartesian and Storyline visualizations. As we further develop our tools for managing multiple visualization methods, we will work to extend the repertoire of dynamic interactions to support the idea of exploratory and explanatory visual analytics. In this regard, the visualization architecture of Sjöbergh and Tanaka, Webbles [19], holds the most promise for supporting more sophisticated interactive visual analytics.

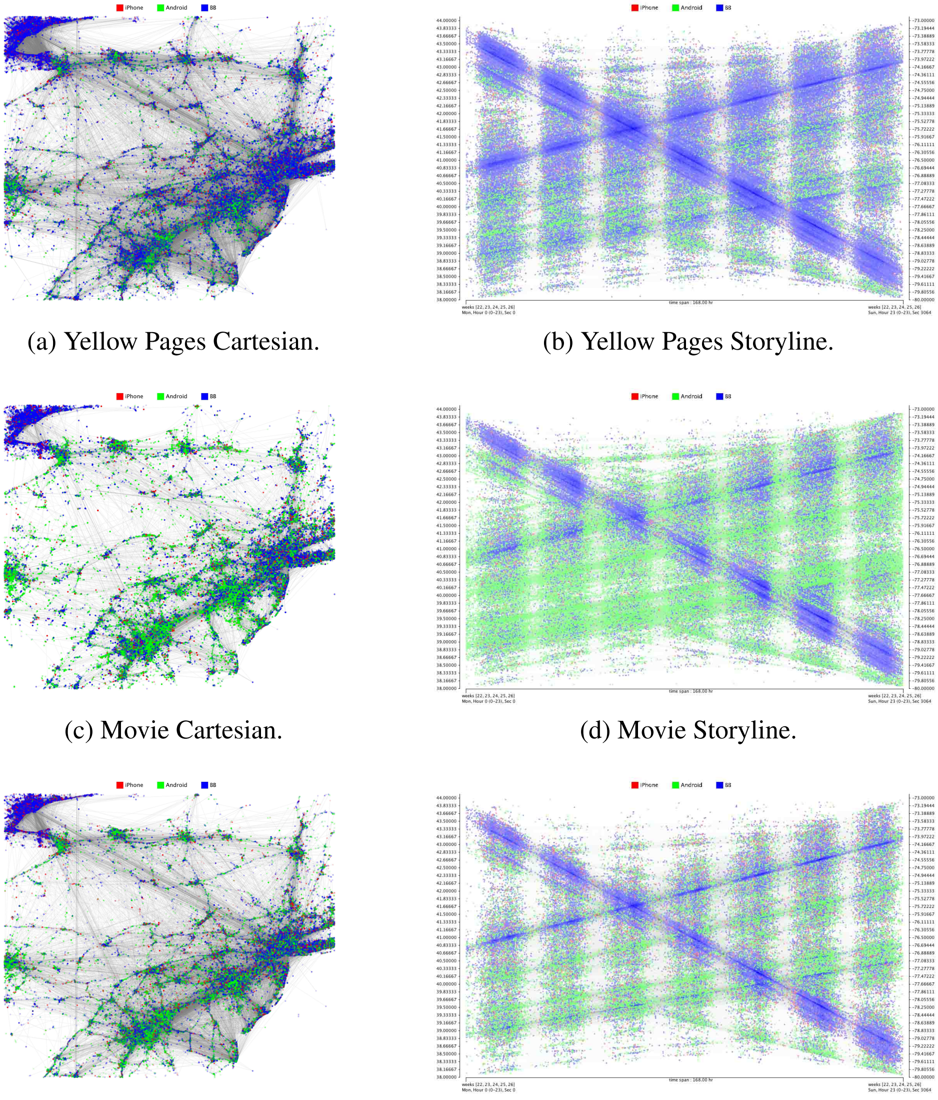
5.3 Comparison of Different Source Devices

The idea of comparing the behaviours of users with different source devices (*e.g.*, Android versus BlackBerry) has been briefly investigated as an example of exploratory searches in Section 5.2, which helps reveal the existence of abnormal Android users. In this section, we make the comparison and investigate the behaviours of users with different source devices in a more comprehensive manner. The records associated with abnormal Android users are filtered out from the data for the experiment, in order to avoid the corresponding noise created in the analysis of consumer behaviors.

5.3.1 Exploratory Search

We first explore the features exhibited by different source device groups under a single target search category (Yellow Pages, Movie or Restaurant). Figure 5.4 shows the results of both Cartesian Geography Map and Frequency Plots augmented Storyline visualizations exploring how Android (green), BlackBerry (blue) and iPhone (red) users behave differently under each single search category in the entire target region. First, it is obvious that iPhone users only account for a very small proportion of the target data over the three search categories, hence the comparison is mainly made between Android and BlackBerry devices. Apparently BlackBerry users prevail in southern Ontario from the Canadian side for all three search categories, while Movie and Restaurant searches in the US, except certain areas around New York City, are more prevalent with Android users, and BlackBerry users in the US only act much more actively in YP searches.

The dominance of BlackBerry users in southern Ontario from the Canadian side matches well the fact that the company BlackBerry, formerly known as Research In Motion (RIM), is headquartered in Waterloo, Ontario, Canada. Besides, the observation that BlackBerry users are more active in Yellow Pages searches in both the US and the Canadian side may indicate that BlackBerry devices are more popular for business or work-related use.



(a) Yellow Pages Cartesian.

(b) Yellow Pages Storyline.

(c) Movie Cartesian.

(d) Movie Storyline.

(e) Restaurant Cartesian.

(f) Restaurant Storyline.

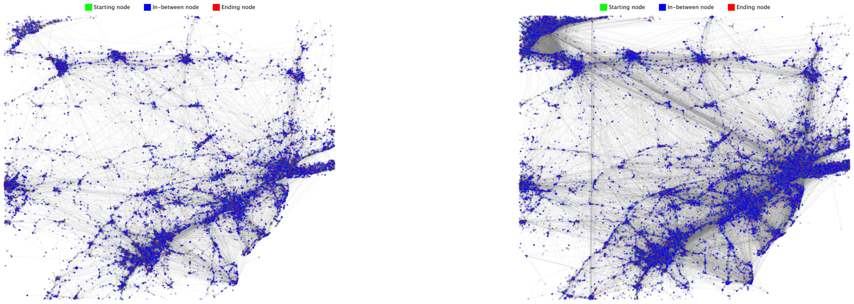
Figure 5.4: The comparison of different source devices under single search type over the entire target region by the results of Cartesian Geography Map and Frequency Plots augmented Storyline visualizations (Android (green), BlackBerry (blue), iPhone (red)).

5.3.2 Explanatory Search

Previously in Section 5.2, visual inference refuted the hypothesis that “Android users tend to move more actively in long distances than BlackBerry users during the peak period of Movie search,” as the result of the explanatory search example, which fails to differentiate Android from BlackBerry in terms of the consumer behaviors of geographical movement activeness in Movie searches. However, from the exploratory results in this section (shown in Figure 5.4), there are many more grey consumer trajectory lines shown in the Cartesian geography map of YP search (Figure 5.4a) than that of Movie search (Figure 5.4c), which indicates stronger spatial transitions among consumers’ YP search behaviors. Also considering the exploratory indication noted above (that BlackBerry devices are more popular for business or work-related use since they are more active in YP searches), we are interested in distinguishing the characteristics of Android and BlackBerry users exhibited in YP searches from the perspective of geographical transitions, thus proposing the following hypothesis and starting the corresponding explanatory search: “BlackBerry users are geographically more active than Android users when searching for Yellow Pages businesses.”

Figure 5.5 shows the Cartesian Geography Map visualizations of YP searches with single brand of source device targeted (Android or BlackBerry) over the entire target region, for comparing the difference in geographical transition activeness between Android and BlackBerry users in the explanatory search. From the results, BlackBerry (Figure 5.5b) exhibits stronger trajectory connectivity between metropolitan areas than Android (Figure 5.5a), which supports the proposed hypothesis.

Digging further, we continue the explanatory search with a zoomed-in focus on the region from the city Toronto in southern Ontario in Canada, all the way south to the city Buffalo in the US, as shown in Figure 5.6a. The small region is specifically selected to span the border area between Canada and the US, with the Canadian side dominated by BlackBerry users and the US part mostly covered by Android users. The distinguishing difference of geographical distribution between Android (green) and BlackBerry (blue) devices is shown in both the Cartesian Geography



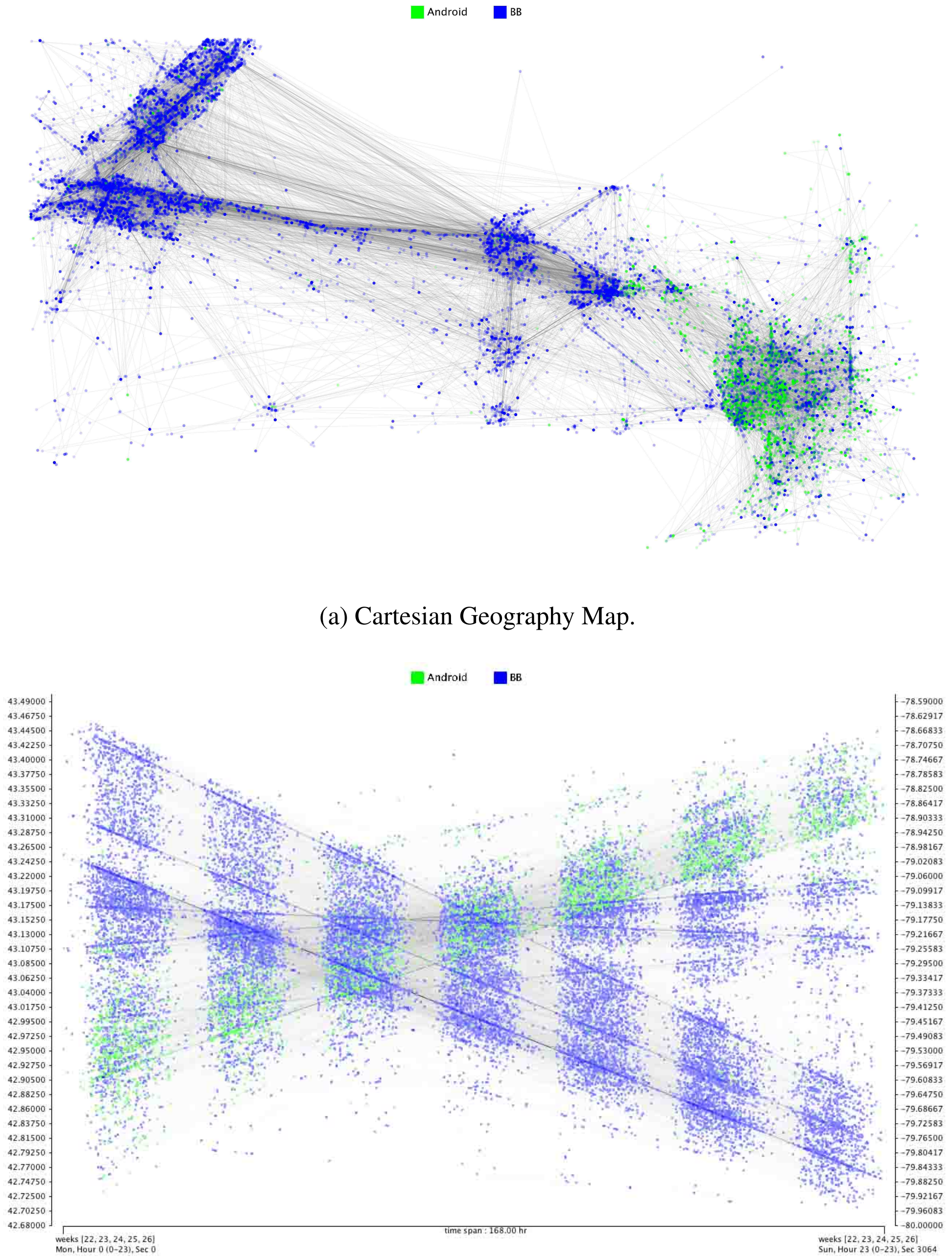
(a) Android.

(b) BlackBerry.

Figure 5.5: Cartesian Geography Map visualizations of YP searches with Android and BlackBerry respectively targeted over the entire target region for the comparison of geographical transition activeness between Android and BlackBerry users.

Map (Figure 5.6a) and Storyline (Figure 5.6b) visualizations.

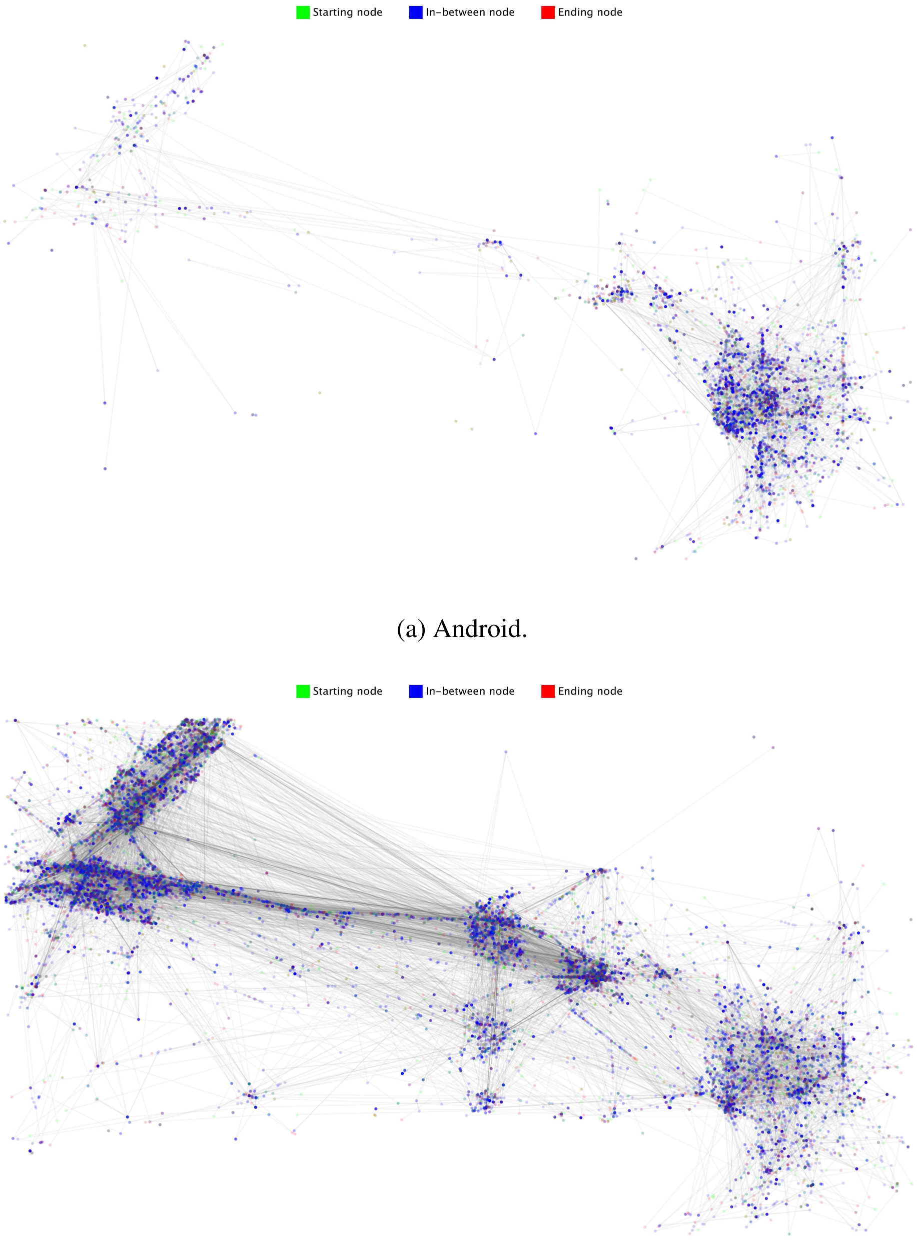
Figure 5.7 shows the results of a similar comparison between Android and BlackBerry users for the explanatory search as Figure 5.5, with a zoomed-in focus on the selected region spanning the border area between Canada and the US (specifically from the city Toronto in southern Ontario in Canada, all the way south to the city Buffalo in the US). We can see from Figure 5.7 that in YP searches, Android users (Figure 5.7a) have much fewer grey trajectory lines indicating weaker geographical transition activeness than BlackBerry users (Figure 5.7b), considered as the supportive evidence to the target hypothesis.



(a) Cartesian Geography Map.

(b) Storyline (augmented by Frequency Plots).

Figure 5.6: YP searches of Android (green) and BlackBerry (blue) users in the selected small region spanning the border area between Canada and the US, from the city Toronto in southern Ontario in Canada, all the way south to the city Buffalo in the US.



(a) Android.

(b) BlackBerry.

Figure 5.7: Cartesian Geography Map visualizations of YP searches with Android and BlackBerry respectively targeted over the selected region covering the border area between Canada and the US, for the comparison of geographical transition activeness between Android and BlackBerry users.

Chapter 6

Utilizing the Webble Dashboard for Poynt Data

6.1 The Webble Dashboard and Searches on Spatio-temporal Data

6.1.1 Webble World and the Digital Dashboard

Webble World[1] [14] is a framework built for users to share, configure and combine pluggable software modules, *i.e.*, Webbles, which encapsulate both content and functionalities as web resources. The purpose of Webble World is to practice and study the evolution of human knowledge, with Webble objects serving as smart containers of digital knowledge based on the philosophy of memes [22]. In general, Webbles are reusable, editable, compoundable software components which wrap content and functionalities to be plugged in at runtime. A Webble object evolves every time users of Webble World reuse/edit and republish it.

The Digital Dashboard [20, 19] is an application of Webble World in exploratory visual analytics, which uses multiple linked views of data. Each view is a plugged-in Webble visualization component. All views in the Digital Dashboard are interactive with their respective visualization result via direct manipulation [16, 15], and are connected in the sense that interactions in one view, *e.g.*, selections or groupings of visualized content, are automatically reflected in other linked views.

We generally refer to the Digital Dashboard with plugged-in Webble visualiza-

[1] see http://www.meme.hokudai.ac.jp/WebbleWorldPortal

tion components as the “Webble Dashboard.” The Webble Dashboard is suitable for visual exploration tasks in that the framework based on pluggable software components makes it easy to incorporate Webbles for new visualization methods and to prototype new interfaces. It also allows interactions with multiple linked views which is convenient for exploratory searches. Directly interacting with the visualization results is very useful for exploring interesting properties of data that emerge in a visualization. For problems that deal with sparse, low quality and high dimensional data (and are thus difficult to model), interactive visual exploration of the data with the Webble Dashboard helps in hypothesis generation and exposing ideas on how to model or analyze the data, and works more effectively when done by domain experts to exploit their domain knowledge and intuitions.

6.1.2 Exploring Spatio-temporal Data with the Webble Dashboard

The Webble Dashboard with components specifically built for visual exploration of spatio-temporal data (*e.g.*, Poynt data) is shown in Figure 6.1. The visualization components, based on content and functionality, can be generally categorized into the following groups:

- **spatial**: the Cartesian Geography Map visualization
- **temporal**: the 24-hour clock view and the histogram of record transaction time
- **spatio-temporal**: the Storygraph visualization
- **other**: mainly histograms indicating data distribution over certain attributes, *e.g.*, search category, source device, and YP query keyword in Poynt data

There are procedures to follow when we conduct visual exploration on target spatio-temporal data; in other words, we interact in a rational sequence of interactions by view category (*e.g.*, spatial or temporal) with each view component to select multiple data groups for comparison. This helps the user to identify possible

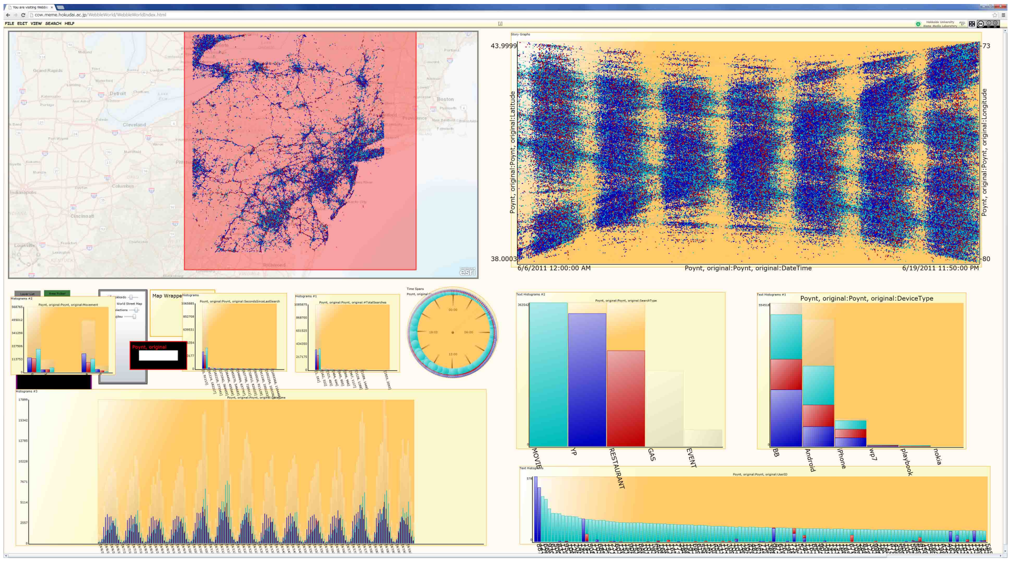

FILE EDIT VIEW SEARCH HELP
Story Graphs
43.9999
38.0003
73
80
Poynt, original:Poynt, original:Latitude
Poynt, original:Poynt, original:Longitude
6/6/2011 12:00:00 AM
Poynt, original:Poynt, original:DateTime
6/19/2011 11:50:00 PM
Map Wrapper
Time Space
Text Histograms #2
Poynt, original:Poynt, original:SearchType
363542
MOVIE
YP
RESTAURANT
GAS
EVENT
Text Histograms #1
Poynt, original:Poynt, original:DeviceType
554518
BB
Android
iPhone
wp7
playbook
nokia
Text Histograms
Poynt, original:Poynt, original:UserID
578
esri

Figure 6.1: The Webble Dashboard with view components specifically built for the visual exploration of Poynt spatio-temporal data.

distinguishing differences exhibited in other linked views. In our case, here are the repertoire of possible interactions:

1. interact with the temporal or spatio-temporal views to select different temporal intervals for comparison, for instance, different days of the week (*e.g.*, weekdays against weekends, Mondays and Tuesdays versus Fridays and Saturdays) and various time slots in a day (*e.g.*, morning versus noon versus evening), and then observe the spatial view and other plugged-in histograms.

2. interact with the spatial view to select different regions for comparison, for instance, different countries (*e.g.*, Canada versus the US) or cities (*e.g.*, Toronto versus New York City) and downtown areas against corresponding outskirts in selected cities, and then observe the temporal/spatio-temporal views and other plugged-in histograms.

3. interact with each of the other histogram views to select different value groups of the corresponding attribute for comparison, for instance, different search categories (*e.g.*, Movie versus YP) and different source devices (*e.g.*, Android versus BlackBerry) in Poynt data, and then observe the temporal/spatio-temporal/spatial views.

6.2 Investigating Poynt Data with the Webble Dashboard

Since the Poynt data is typically spatio-temporal data with extra search category dependent attributes, it is interesting to investigate it for distinguishing features and patterns with the Webble Dashboard. Specifically, we concentrate on hypotheses associated with the semantics of Yellow Pages (YP) query keywords, and features involving user movement distances.

6.2.1 Yellow Pages Keyword Semantics

During the preliminary exploratory search on the semantics of YP user query keywords, we notice that certain query keywords are sensitive to certain times or locations. Temporally, for instance, "hotel," "bar," "liquor store" and "taxi" are mostly queried during the evening, while higher percentages of "Starbucks," "Tim Hortons" and "McDonalds" occur in the morning. Some interesting yet unexpected findings include that gym and fitness related queries, *e.g.*, "ymca" and "planet fitness," are more actively made during weekdays rather than weekends. Spatially, for example, "Sears," "Home Depot" and "Walmart" are searched more frequently around city outskirts or suburbs, while "parking" and "LCBO" are mostly searched in downtown areas. We also find it not surprising that some query keywords exhibit geographical locality in their geo-spatial distributions, *e.g.*, "CIBC" and "LCBO" are queried only in Canada, while "Bank of America" is only from the US side. But there still are interesting discoveries beyond our expectations, *e.g.*, "Dairy Queen" and "Ikea" are searched more in Canada than in the US.

Motivated by the above preliminary explorations of the spatio-temporal characteristics shown among YP query keywords, we are interested in further clustering YP query keywords semantically, and observe potential patterns amongst keyword groups with different semantics.

Creating Yellow Pages Keyword Semantic Groups

When a user searches for the information of some business entity in the nearby neighborhood, if the query is well-formed and has mappings of nearby business entities returned by the app, the user can click over any entry in the returned listings to obtain its corresponding information (phone call, map route, web link, *etc.*). Normally, YP query keywords are nouns (*e.g.*, “pizza,” “post office,” “hotels”) or business named entities (*e.g.*, “pizza hut,” “cibc”). Categorized as follows are the problematic instances in the raw YP query keywords which have no mappings of nearby business entities returned by the Poynt app:

1. the keyword contains ill-formed units with typos, symbols, number signs. *E.g.*, “papa jogbs,” “cfgj$$$,” “drn0i0.”

2. the keyword consists of well-formed units, but no nearby business entities are found, as either all the units combined together do not make much sense semantically and could not be mapped to any valid business entity by the app, or there are business entities mapped, but none of them are located in the nearby neighborhood of where the user types in the query.

An important circumstance not covered by the above categorization is that even if what a user types in has nearby business entities mapped by the app, the returned listings may not contain any entry relevant enough to what the user originally intends to search for, suggested by no subsequent click action, which implies the user’s negative feedback to the YP search results.

The target data that we expect for the experiments associated with YP keyword semantic groups are well-formed YP search keywords with “valid” semantics in the sense that not only there are mappings of nearby business entities returned for each keyword input, but the listings do actually contain what the user originally intends to search for, which could be confirmed by the user’s click behaviors over the listed YP business entries. It is difficult to tell the quality of the returned YP search results without knowing the user’s click actions afterwards as an implicit source of positive feedback. Fortunately, we have a special category of Poynt data collected, Clickthru, recording users’s click actions in certain sections of the app (mainly in

YP and Restaurant). Based on Clickthru data, we could "verify" YP search data by focusing only on those chronologically immediately followed by a click action of the same user. For example, Table 6.1 lists several YP search records followed by a Clickthru record of the type map, all conducted chronologically within a short time frame by the same individual. The last YP search record with the query keyword "*wilmington trust bank*" is therefore verified by the user's map click behavior and selected as a target for the experiments. Normally, the YP search records selected by the verification of Clickthru data have query keywords with more precise identification of nouns or business entities.

*** several YP search records by one individual**				
transaction date	**longitude**	**latitude**	**query keyword**	**results**
11-06-24 13:07:38	-75.7165	39.4496	"*wimimgton ytrust babk*"	0
11-06-24 13:08:16	-75.7165	39.4496	"*wimimgton trust*"	0
11-06-24 13:13:10	-75.7165	39.4496	"*wilmington trust bak*"	0
11-06-24 13:20:35	-75.7165	39.4496	"***wilmington trust bank***"	**20**
*** a Clickthru record indicating the same user's map click action**				
transaction date	**longitude**	**latitude**	**click section**	**type**
11-06-24 13:21:25	-75.7165	39.4496	YP	map

Table 6.1: An example of YP search records verified by Clickthru data.

Based on the Clickthru-verified YP search records, we create mappings from the query keywords to their semantic categories with the help of the Yelp Business Search API[2]. The Yelp Business Search API recognizes business named entities with a hierarchical semantic category structure and we choose its topmost level as our general categories for building keyword semantic groups. Also considered by the Yelp API when mapping a query keyword to a semantic category is where the keyword was searched for, which means that one same keyword queried at different locations may be mapped to different semantic categories due to assorted local business distributions in their respective surroundings. Given access to the Yelp API, we send each keyword with its corresponding location coordinates in the target YP search data and get the category information of the first business returned by the API (if any). If there is no category information obtained, we then normalize the

[2] see http://www.yelp.com/developers/documentation/v2/search_api

keyword using Google "didyoumean,"[3] and resend the normalized candidate to the API. If there is still no category information retrieved, we remove the first word in the keyword and send it to the API as the final trial. After the above procedures, all keywords without category information in the end are assigned with the group label "NO_CATEGORY_INFO." Table 6.2 lists all top-level semantic categories of the Yelp Business Search API that are utilized for creating YP keyword semantic groups and the distributions of the target Clickthru-verified YP search records among them.

Yelp top-level category	**Category number**	**Percentage**
Shopping	c1	25.20%
Restaurants	c2	18.64%
Food	c3	9.46%
Automotive	c4	7.62%
Hotels and Travel	c5	6.49%
Active Life	c6	4.16%
Financial Services	c7	3.83%
Beauty and Spas	c8	3.73%
Health and Medical	c9	2.99%
Home Services	c10	2.76%
Local Services	c11	2.19%
Nightlife	c12	1.90%
Pets	c13	1.56%
Arts and Entertainment	c14	1.37%
Public Services and Government	c15	1.23%
Education	c16	0.76%
Event Planning and Services	c17	0.58%
Professional Services	c18	0.46%
Real Estate	c19	0.39%
Religious Organizations	c20	0.29%
Mass Media	c21	0.10%
Local Flavor	c22	0.05%
NO_CATEGORY_INFO	c23	4.23%

Table 6.2: The Yelp top-level semantic categories and the distribution of the target Clickthru-verified YP search records among them.

As the final step to the creation of YP keyword semantic groups, the Yelp top-level categories are manually grouped in terms of semantic similarity. All the group-

[3] see https://developers.google.com/custom-search/json-api/v1/overview

ing details are shown in Table 6.3, including the mappings of the created semantic groups to their corresponding Yelp top-level categories, and the distributions of the target Clickthru-verified YP search records among the created semantic groups (based on the category-wise percentage figures in Table 6.2).

Keyword semantic group	**Yelp category number(s)**	**Percentage**
Food and Restaurants	c2, c3	28.10%
Shopping	c1	25.20%
Services	c17, c7, c10, c11, c18, c15	11.05%
Automotive	c4	7.62%
Life and Entertainment	c6, c14, c21, c12	7.53%
Beauty and Health	c8, c9	6.72%
Hotels and Travel	c5	6.49%
Pets	c13	1.56%
Education	c16	0.76%
Local Flavour and Real Estate	c19, c22	0.44%
Religious Organization	c20	0.29%
NO_CATEGORY_INFO	c23	4.23%

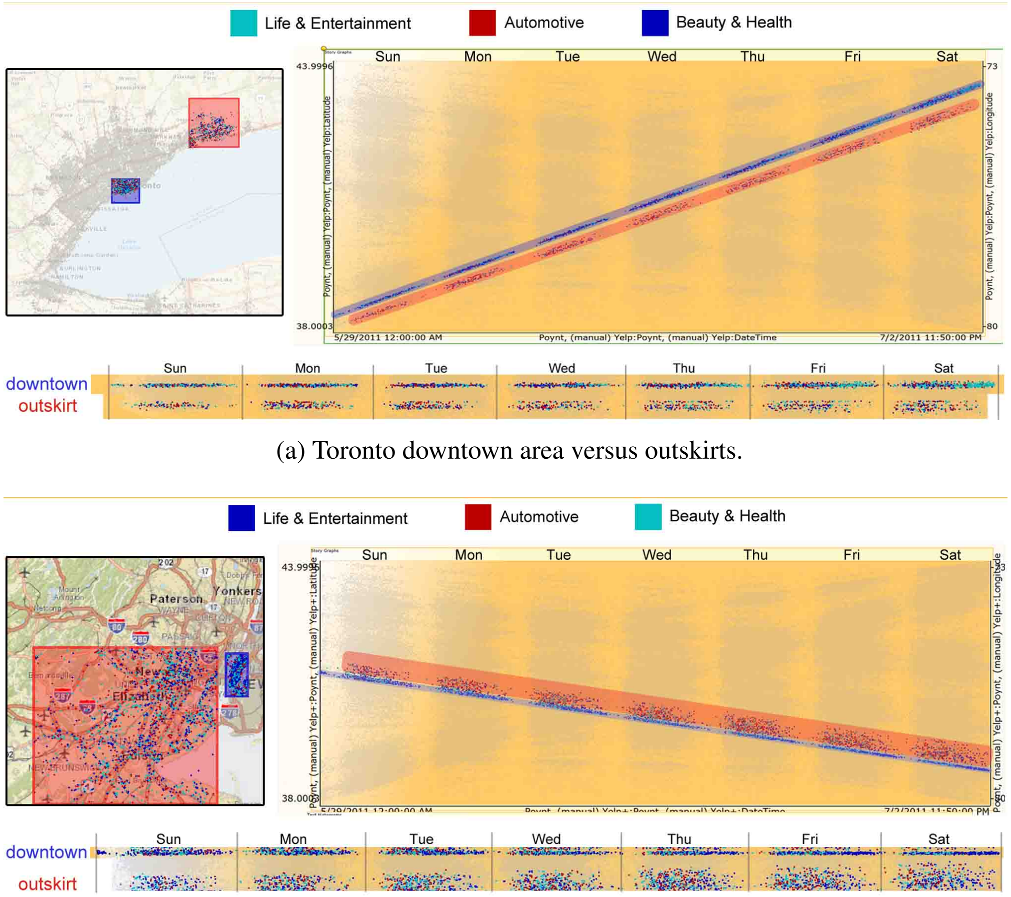
Table 6.3: The mappings of the manually created keyword semantic groups to their corresponding Yelp top-level categories, and the distribution of the target Clickthru-verified YP search records among the manually created semantic groups.

Experiments

We first start with an exploratory search which focuses on singular keyword semantic group from Table 6.3. Revealed are several simple temporal patterns that make sense and are consistent with the semantics of the group. For example, the semantic group “Services” distributes evenly over weekdays and exhibits minor drops on weekends, while “Hotels and Travel” behaving similarly to “Services” on weekdays shows an increase on weekends. In addition, “Automotive” and “Services” have relatively larger ratios during weekdays than other groups.

As we search further with multiple keyword groups involved, we choose three targets from the created keyword semantic groups (see Table 6.3), “Automotive,” “Life and Entertainment,” and “Beauty and Health,” which have similar sizes (all around 6% to 8% of the target data) yet cover totally different semantic topics. We separately compare the selected groups’ temporal distributions around downtown areas and corresponding outskirts in terms of days of the week in the Storygraph component. Figure 6.2 displays the visualization results of the exploratory search on the relevant Webble components, which compare downtown areas (marked by a blue shadow) and corresponding outskirts (marked by a red shadow) in the two cities Toronto and New York City respectively. It shows that activities of the three target groups in both cities have longer temporal segments in the evening till midnight in downtown areas than around outskirts. Also, downtown areas on Sundays exhibit distinguishing early-morning activities right after midnight, which are continuing extensions of activities from late night of previous Saturdays. Among the selected groups, "Life and Entertainment" contributes the most to all the above differentiations between downtown areas and corresponding outskirts. It is most active in downtown areas during the second half of each day of the week till late, especially on Fridays and Saturdays, which matches well the semantics of the group. Besides, geographically speaking, "Life and Entertainment" in downtown New York City is relatively more evenly distributed during the second half of each day of the week, while downtown Toronto exhibits more distinguishing differences between Fridays/Saturdays and other days of the week in terms of the group's peaking activities.

Normally we expect some interesting phenomenon to appear unexpectedly during an exploratory search, based on which we could conduct corresponding explanatory searches with certain hypotheses inspired by those exploratory observations. The processes of the explanatory searches to reach the confirmation or refutation of the proposed hypotheses, in return, provide valuable assistance to the continuance and conclusion of the exploratory search. For example, previously in the experiment section "Exploratory and Explanatory Searches" (Section 5.2), the exploratory search comparing how Android and BlackBerry users behave during one of movie search's peak periods reveals an unexpected difference in the Cartesian geography maps that Android users exhibit strong trajectory connectivity between metropolitan centers. Accordingly we hypothesize that Android users tend to move more actively in long distances than BlackBerry users during the peak period of movie search. In the explanatory search for the confirmative or refutative evidence,



(a) Toronto downtown area versus outskirts.

(b) New York City downtown area versus outskirts.

Figure 6.2: The exploration of the temporal features of the selected keyword semantic groups (“Automotive,” “Life and Entertainment,” “Beauty and Health”) in the Storygraph component, comparing downtown areas (in the blue shadow) and corresponding outskirts (in the red shadow) in the two cities Toronto and New York City respectively.

we notice abnormal Android users from the Storyline visualizations, who contribute to the strong trajectory connectivity shown exclusively in the Cartesian map of Android users, and by comparing the visualization results after the removal of abnormal Android users, the proposed hypothesis is refuted. The exploratory search also benefits from the removal of abnormal Android users after which a preliminary conclusion about the geographic difference between Android and BlackBerry users is reached.

However, often nothing beyond our expectations stands out during exploratory searches. Under such circumstances, the observed characteristics and patterns are the conclusion of the exploration, with no explanatory searches initiated. Most likely the observations are self-explanatory and consistent with the semantics of the target data artifacts. The exploratory investigation regarding multiple YP keyword semantic groups in this section is a typical example which triggers no explanatory searches and is concluded with the observation that the keyword semantic group “Life and Entertainment” contributes the most to the differentiations between downtown areas and corresponding outskirts. What the keyword group “Life and Entertainment” covers semantically, *e.g.*, “night life,” “bar,” “night club,” explains the observed temporal features that actitivies in downtown areas span longer temporally in the evening till midnight and exhibit distinguishing early-morning activities right after midnight on Sundays. And when compared with the other two target groups, “Life and Entertainment” is most active in downtown areas during the second half of each day of the week till late, especially on Fridays and Saturdays.

6.2.2 User Movement Distances

We are also interested in investigating the characteristics of geographic movements across search behaviors in consumer trajectories, by individual user. Since the Poynt search data is discrete, the actual movement trace between two search records cannot be determined for calculating the exact movement distance. Due to this uncertainty, we simplify the movement distance between two search records to be the geographic length of the line segment connecting the corresponding nodes in the Cartesian geography map.

Labeling Search Records by User Movement Status

We label target search records in terms of user movement status to help indicate the state of movement that the corresponding user was in when conducting a search, *e.g.*, querying about hotels in a travelling trip (under the state of long-distance movements), or looking for a restaurant and movie theater for movie night in an everyday neighborhood around home (under the state of short-distance movements). Specifically, for each search record we calculate the average movement distance ($avgmd$), the average of the movement distance from the record appearing immediately before it in the consumer trajectory (if any) to this record and the distance from this record to the one immediately after it in the consumer trajectory (if any). According to the corresponding value of $avgmd$, each search record is categorized into one of the following four groups of user movement status:

- **NO-MOVE**: no movement ($avgmd = 0$)
- **ONE-BLOCK**: movement within one street block ($avgmd$ is in the distance scope of one street block)
- **INTRA-CITY**: movement within a city ($avgmd$ exceeds one-block distance and is in the distance scope of multiple street blocks in a city)
- **INTER-CITY**: movement between cities ($avgmd$ exceeds multiple-block distance)

Experiments Exploring Features of User Movements

We are interested in exploring and comparing the spatio-temporal features exhibited when users are searching under the status of either short-distance or long-distance movements. Preliminarily, we choose the two movement status groups, ONE-BLOCK and INTER-CITY, to represent the user states of short-distance movements and long-distance movements respectively.

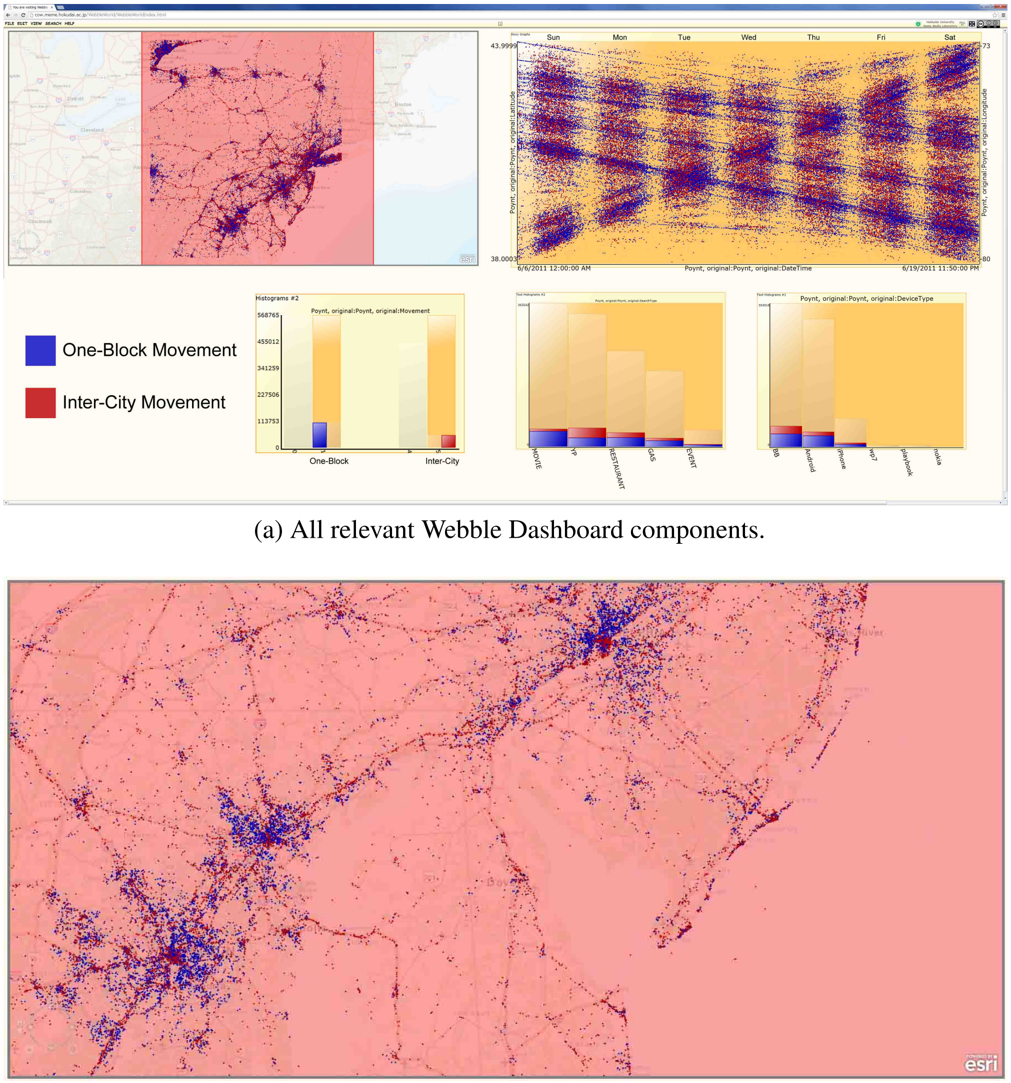
Figure 6.3a displays the visualization results of the exploratory search comparing the differences in the spatio-temporal features shown between ONE-BLOCK (marked in blue) and INTER-CITY (marked in red) on the related Webble components.

From the zoomed-in Cartesian geography map component (Figure 6.3b), it shows that the routes connecting towns/cities are mainly covered by red INTER-CITY nodes, and they also extend to the central parts of the major metropolitan areas, which are dense gatherings of nodes mostly covered by blue ONE-BLOCK nodes. Also note that the majority of ONE-BLOCK nodes are centralized around urban regions, while INTER-CITY nodes are more scattered across the entire map.

On the other hand, temporally speaking, the Storygraph view (Figure 6.3c) suggests that INTER-CITY events distribute mainly during the daytime (around 6 a.m. to 6 p.m. typically) and few occur late at night or early in the morning around midnight, while ONE-BLOCK generally has a longer temporal span for each day of the week which normally starts early in the morning from midnight or lasts late in the evening till midnight, and it even remains active for the whole night in certain regions. The temporal features about INTER-CITY and ONE-BLOCK are interesting observations based on which some explanatory search is suggested.

Extended Experiments on Explanatory Searches

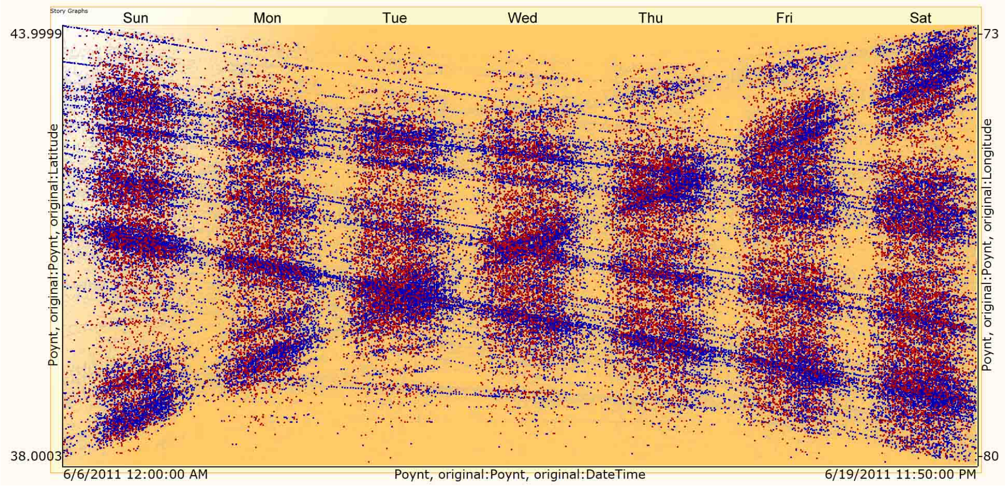
In addition to the Cartesian geography map and the Storygraph components based on which the geographical and temporal features are revealed as the results of the previous exploratory search about user movement statuses, the Webble Dashboard (see Figure 6.3a) also contains a histogram component which indicates the distribution of records categorized as ONE-BLOCK or INTER-CITY movement status over different search categories (see Figure 6.4 enlarged for details). It shows that Movie, Gas and Event searches are much more frequent under the status of ONE-BLOCK (short-distance movements) than INTER-CITY (long-distance movements), while YP reaches an approximately equal division of shares by ONE-BLOCK and INTER-CITY. The difference in the distribution of movement status groups ONE-BLOCK and INTER-CITY over different search categories sheds an intriguing clue which inspires us to start explanatory searches about the characteristics of target search categories in terms of user movements.



(a) All relevant Webble Dashboard components.

(b) The Cartesian geography map zoomed in for details.

Figure 6.3: The visualization results of the exploratory search on the relevant Webble Dashboard views regarding the spatio-temporal features of user movements: ONE-BLOCK movement (blue) and INTER-CITY movement (red) represent the user states of short-distance movements and long-distance movements respectively. *(cont.)*



(c) The Storygraph component.

Figure 6.3: The visualization results of the exploratory search on the relevant Webble Dashboard views regarding the spatio-temporal features of user movements: ONE-BLOCK movement (blue) and INTER-CITY movement (red) represent the user states of short-distance movements and long-distance movements respectively.

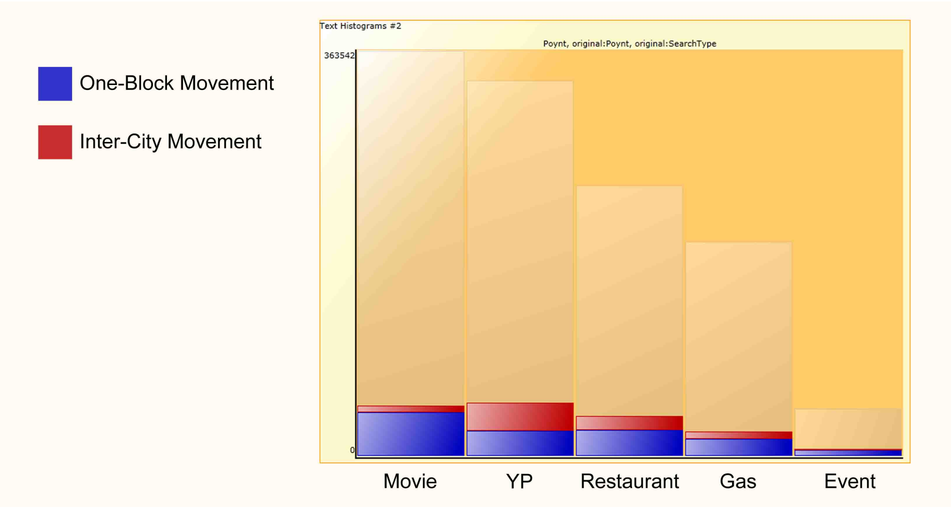


Figure 6.4: The histogram component indicating the distribution of records categorized as ONE-BLOCK or INTER-CITY movement status over different search categories.

In the extended experiments on corresponding explanatory searches, we include all four groups of user movement status, with NO-MOVE/ONE-BLOCK representing short-distance movements and INTRA-CITY/INTER-CITY representing long-distance movements, and select two typical search categories with distinguishingly different ONE-BLOCK and INTER-CITY distributions in Figure 6.4, Movie and YP, as the target search categories.

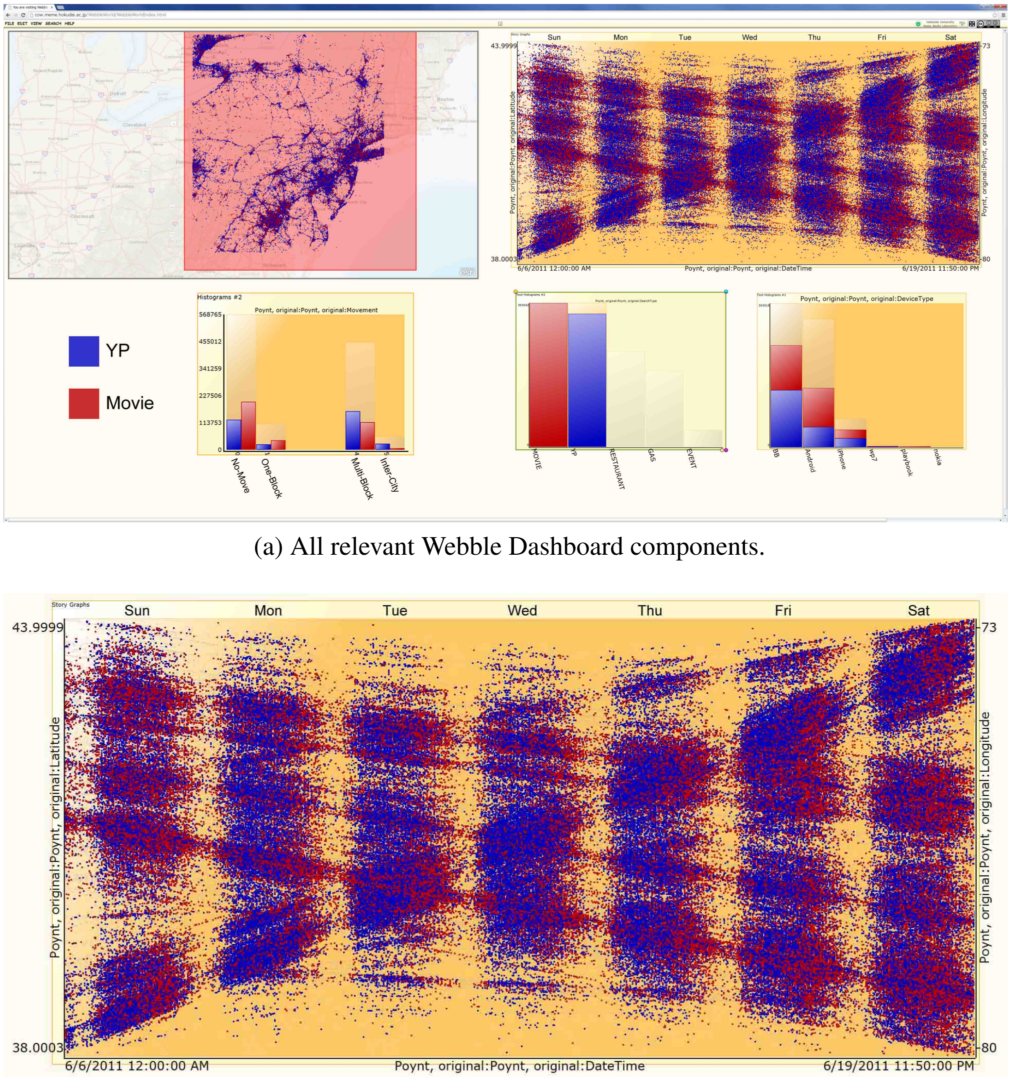
We hypothesize that users tend to search for movies around the second half of the day and lasting late till the midnight in short-distance movements, while YP businesses are more typical of daytime queries (around 6 a.m. to 6 p.m. typically), when users are moving in relatively longer distances.

Figure 6.5a shows the visualization results of the explanatory search for the proposed hypothesis on the relevant Webble Dashboard components. The temporal distribution of YP and Movie searches shown in the Storygraph component (Figure 6.5b), together with the distribution of YP and Movie searches over different user movement status groups shown in the histogram component (Figure 6.5c, with NO-MOVE/ONE-BLOCK representing short-distance movements and INTRA-CITY/INTER-CITY representing long-distance movements in the experiment), are the confirmative evidence which support the proposed hypothesis. In addition, there exhibits a resemblance of the spatial features shown in the Cartesian geography maps between YP/Movie search (in Figure 6.5d) and the long-distance/short-distance movement status (in Figure 6.3b).

The characteristics of YP and Movie searches in terms of user movements (discussed in the above explanatory search) conforms with the real-life cases that people search for movies before a movie night in a familiar everyday neighborhood around home (under the state of short-distance movements); and people query about business entities in a relatively new environment on the go (under the state of long-distance movements).

6.2.3 Limitation of Investigation with the Webble Dashboard

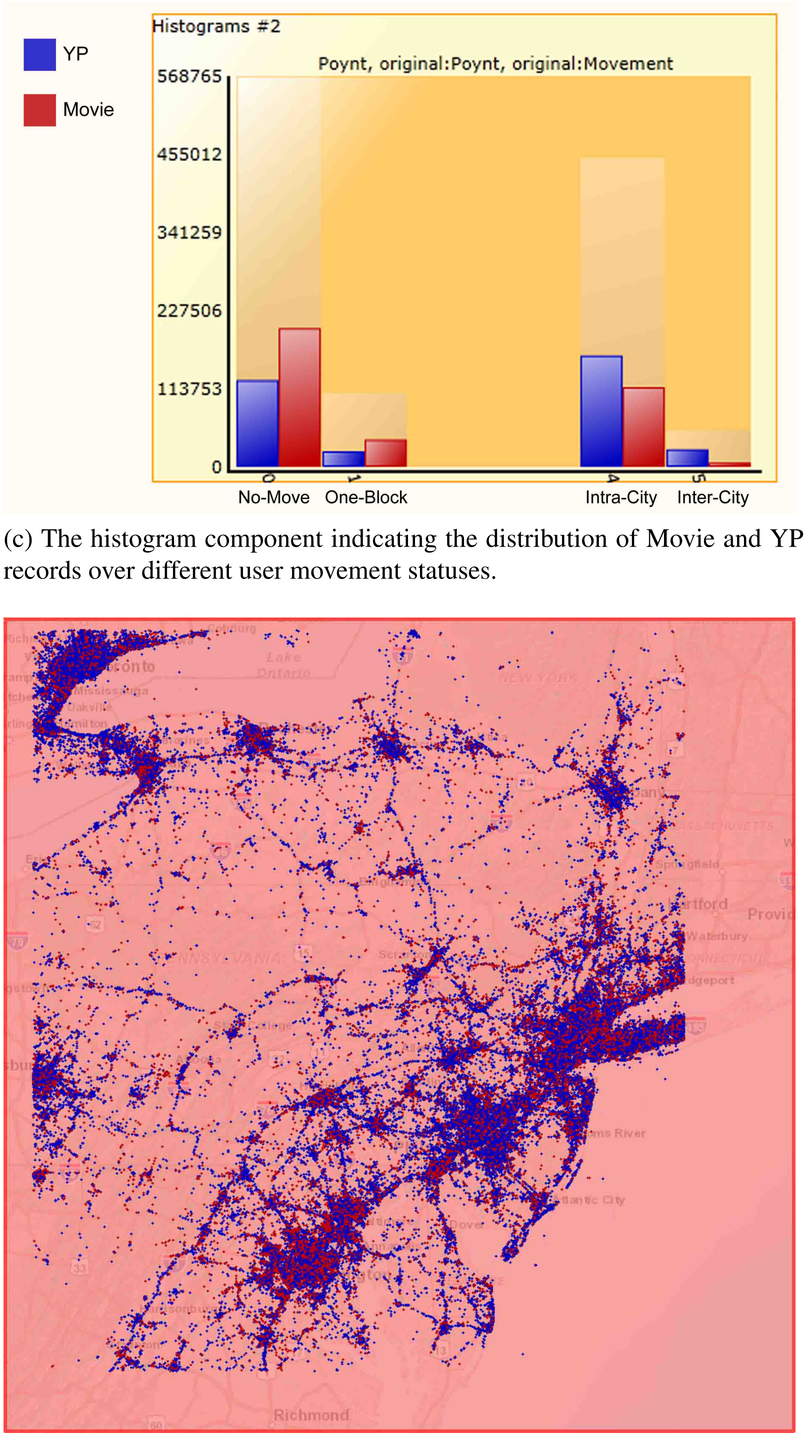
While the Webble Dashboard enables easy deployment of visualization components and rapid interactivity, the total volume of data allowed for analysis is restricted



(a) All relevant Webble Dashboard components.

(b) The Storygraph component.

Figure 6.5: The visualization results of the explanatory search on the relevant Webble Dashboard views about the characteristics of the search categories Movie (red) and YP (blue) in terms of user movements. *(cont.)*



(c) The histogram component indicating the distribution of Movie and YP records over different user movement statuses.

(d) The Cartesian geography map.

Figure 6.5: The visualization results of the explanatory search on the relevant Webble Dashboard views about the characteristics of the search categories Movie (red) and YP (blue) in terms of user movements.

since the system runs on the web browser end. The available memory space is therefore much smaller than what is provided for an application running locally in the RAM or remotely on a server. This limitation could be alleviated by the filtering of data to focus on a relatively smaller portion that can be loaded integrally by the system.

Part IV

Conclusions

Chapter 7

Conclusions and Future Work

7.1 Summary and Conclusions

Our approach to the visual analysis of a large volume of geo-located individual search records requires the development of a conceptual framework that provides a systematic method of filtering and selection, in order to focus our search for semantically relevant data artifacts. In our case, our overall goal is to identify time sequence components we call "consumer trajectories," which we hypothesize as clustered time-series events of individual users correlated with some category of search (*e.g.*, a sequence of search records by an individual user searching for inexpensive fuel).

This kind of framework requires a variety of filtering and visualization techniques, organized in a system that supports a kind of hypothesis-driven process of visually identifying interesting data artifacts within selected data, and then using a variety of data selection and visualization techniques to adjust the parameters of those artifacts, in order to further support them, or to dismiss them as semantically unsupported. This can typically be done by changing the selection of data related to any particular artifact hypothesis, for example, extending the geographic region in which it is contained, or viewing the same data across a number of different time segments.

Our framework further acknowledges that no one visualization method will suffice to provide alternative views for the same data artifacts, so we compare two fundamentally different visualization methods, based on a Cartesian coordinate display, or a Storyline display. In this way, two quite different views of the same data provide a human user with a broader view to confirm visual inferences about interesting data artifacts. In addition, the case studies demonstrate the effectiveness of combining "exploratory" and "explanatory" searches in our hypothesis-driven framework for visual analysis tasks.

Furthermore, our adoption of the Webble Dashboard implementation creates a highly interactive environment for visual exploration of data, in which the shift and movement between "explanatory" and "exploratory" searches have a broader repertoire of visual actions which are rapid and more informative.

7.2 Review of Contributions

We have developed a variety of tools for selecting and filtering the original Poynt data, and demonstrated their use in case study scenarios for identifying increasingly complex individual and aggregate Poynt user activities, culminating in the identification of consumer trajectories, *i.e.*, the visual display of focused Poynt user activities, localized in space and time, and heuristically alleged to represent specific consumer goal activities.

In Section 3.3, we noted how selection of spatial regions and time duration can provide the basis for visualizing both spatial and temporal properties of the Poynt data, for drawing exploratory inferences about spatial distribution and temporal distribution. We noted that the tools provide the flexibility to explore the Poynt data space, and are even more useful when combined with the direct interaction of the Webble Dashboard system (Section 6.1).

In Part III, we demonstrated how the interleaving of exploratory and explanatory uses of the tools, while still localized subjectively in the mind of the user, can reveal data artifacts that would otherwise be unexposed, *e.g.*, as in the detection of artificial movement of Android users (Section 5.2). The incremental stories of the development of the case studies (Sections 5.3, 6.2) show how one can formulate and then test hypotheses on components of consumer trajectories (*e.g.*, classify Poynt user movements in small spatial clusters, like NO-MOVE, ONE-BLOCK, INTER-

CITY, *etc.*, in Section 6.2.2), and how variation in Yellow Pages search terms can change the categories in which Poynt users search, and so help determine another component of consumer trajectories (Section 6.2.1).

Overall, the contributions are about how the tools can be used, within an emerging methodology of "explore" and "explain", to identify consumer trajectories, using multiple visualizations, side by side, with interaction as a mechanism to focus attention on data artifacts of interest. These data artifacts would otherwise remain unidentified by ordinary relational query languages on the half a billion Poynt records.

7.3 Future Work

As is obvious, while we have a complete prototype that provides access to our 531 million geo-located search records, we have only begun to investigate the possible emergent relationships amongst that volume of data, beginning with the idea of finding semantic data artifacts we call "consumer trajectories." We are continuing our work with simultaneous refinement and improvement of the framework, as well as deeper exploration of more sophisticated semantic artifacts, hypothesized as consumer behaviour.

For future work, one promising direction is designing interactions for complex filtering constraints (*e.g.*, Search Density and User Group), in order to facilitate exploratory and explanatory visual analytics for complex hypotheses. On the other hand, we are interested in automating some interactive visual exploration processes, *e.g.*, building a program which explores the constraint space to generate all visualization results and automatically focuses on "interesting" differences exhibited amongst them, with the help of image processing techniques.

We also plan to evaluate the process of hypothesis-driven visual analysis in our framework, based on accuracy and speed of hypothesis confirmation. Formal evaluation requires more precise formulation of hypothesis spaces, and measures of coverage achieved with analytics tools.

Bibliography

[1] Almir Olivette Artero, Maria Cristina Ferreira de Oliveira, and Haim Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 81–88, 2004.

[2] Sean Brennan, Adam Sadilek, and Henry Kautz. Towards understanding global spread of disease from everyday interpersonal interactions. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2783–2789, 2013.

[3] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 241–250, 2012.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Randy Goebel, Wei Shi, and Yuzuru Tanaka. The challenge of semantic symmetry in visualization. In *Proceedings of the 18th International Conference on Information Visualisation*, IV '14, pages 27–33, Paris, France, 2014.

[9] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 769–778, 2012.

[10] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer (Special Issue on Computational Geometry)*, 1(2):69–91, 1985.

[11] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, 2013.

[12] Thomas Kapler and William Wright. Geotime information visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 25–32, 2004.

[13] M. Kraak. The space-time cube revisited from a geovisualization perspective. In *Proceedings of the 21st International Cartographic Conference (ICC) "Cartographic Renaissance"*, pages 1988–1996, 2003.

[14] Micke Kuwahara and Yuzuru Tanaka. Webble world – a Web-based knowledge federation framework for programmable and customizable Meme Media objects. In *The IET International Conference on Frontier Computing 2010*, pages 372–377, Taichung, Taiwan, 2010.

[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 *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–343, Washington, DC, USA, 1996.

[17] Ayush Shrestha, Ying Zhu, Ben Miller, and Yi Zhao. Storygraph: Telling stories from spatio-temporal data. In *Proceedings of the 9th International Symposium on Visual Computing*, ISVC '13, July 2013.

[18] Ayush Shrestha, Ying Zhu, Ben Miller, and Yi Zhao. Storygraphs: Extracting patterns from spatio-temporal data. In *Proceedings of the KDD 2013 Workshop on Interactive Data Exploration and Analysis*, IDEA '13, August 2013.

[19] Jonas Sjöbergh and Yuzuru Tanaka. Visual data exploration using Webbles. In *Proceedings of the Webble World Summit 2013*, volume 372 of *Springer CCIS*, pages 119–128, 2013.

[20] 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.

[21] Robert Spence. *Information Visualization: Design for Interaction (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[22] Yuzuru Tanaka. *Meme Media and Meme Market Architecture*. IEEE Press, Piscataway; NJ; USA, 2003.

[23] Yu Zheng and Xing Xie. Learning travel recommendations from user-generated gps traces. *ACM Trans. Intell. Syst. Technol.*, 2(1):2:1–2:29, January 2011.