2014 18th International Conference on Information Visualisation

Towards the identification of consumer trajectories in geo-located search data

Xingkai Li, Randy Goebel
Alberta Innovates Centre for Machine Learning
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: xingkai@ualberta.ca, rgoebel@ualberta.ca

Jonas Sjöbergh
Meme Media Laboratory Hokkaido University
Hokkaido University
Sapporo 060-8628, Japan
Email: js@meme.hokudai.ac.jp

***Abstract*—Modern geo-positioning 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 information extraction, machine learning, and visualization methods that provide systematic tools for investigating the geo-located data, and for the development of both conceptual tools and visualization tools in aid of finding both interesting and useful patterns in that data. Included in our system architecture is the ability to consider the difference between exploratory and explanatory hypotheses on data patterns, as well as the deployment of multiple visualization methods that can provide alternatives to help expose interesting patterns.

In our introduction to our framework here, we provide examples of formulating hypotheses on geo-located behaviour, and how a variety of methods including those from machine learning and visualization, can help confirm or deny the value of such hypotheses as they emerge. In this particular case, we provide an initial basis for identifying semantically motivated data artifacts we call geo-located consumer trajectories. We investigate their plausibility with a variety of time and space series clustering and visualization models.

***Index Terms*—geo-located search, clustering, visualization, geo-trajectories.**

I. Introduction

The popular smart phone application "Poynt"[1] provides about 10 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 in 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 value is similar to 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.*, [1]). What is more interesting in the Poynt data context is the development of analytic methods to identify what we refer to as "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 : explanatory search to confirm visual understanding of expected hypotheses (*e.g.*, movie nights on Friday and Saturday); exploratory search to find "interesting" phenomenon, then constraint adjustments to find supporting evidence for possible hypotheses.

II. A Framework for Hypothesis-driven Analysis

The volume of individual search records is understandably very high: even with one sample of just five weeks of historical

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

1550-6037/14 $31.00 © 2014 IEEE
DOI 10.1109/IV.2014.39

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 data set 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:

(i).

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, certainly 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.

A. *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 is to find “interesting” phenomenon, then constraint adjustments to find supporting evidence for possible hypotheses.

B. *Filtering of Poynt Data to Focus Evidence Compilation*

Any analytics framework will require us to identify appropriate subsets of the raw data, so that our analytics work can be focused on categories of hypotheses. This will provide 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 Type**: 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.*, [1,10] 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 1). 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.

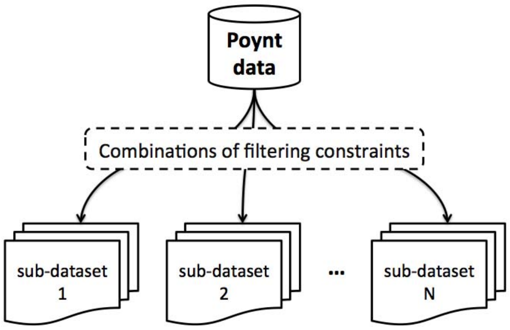


Fig. 1: Choosing sub-datasets with combinations of filtering constraints.

C. Examples of Simple Hypothesis Investigation

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 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 2) 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).

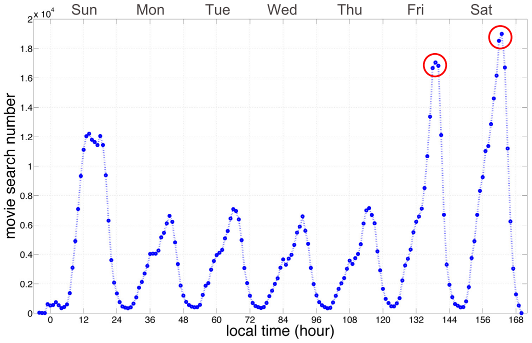


Fig. 2: Hourly number distribution of movie searches within one week.

2) Hypothesis Two: Hypothesis Two is in category (ii) 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 $CR_{B\text{-}A}$.

In the preliminary experiment, we choose two 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, 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 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.

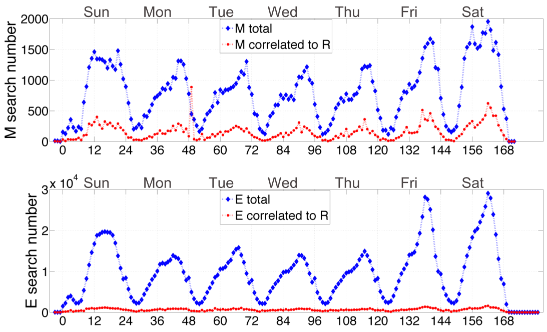


Fig. 3: Investigate the correlations Movie-Restaurant and Event-Restaurant: the distribution of searches in total (blue) and the distribution of searches correlated to any Restaurant search (red) over the hours in one week.

D. Consumer Trajectory

Under the context of spatiotemporal 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 II-B), a set of consumer trajectories is built by grouping 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 normally restrict that one consumer trajectory involves with 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.

III. Visualization Approaches

The Poynt data in experiment are consumer trajectories built on various subsets of search records that are filtered by combinations of constraints representing assorted semantics (II-D). Three major attributes of search records, 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 evidences that exhibit spatial and/or temporal patterns for the exploratory/explanatory investigation of semantic hypotheses. In the framework of hypothesis-driven analysis, we need multiple visualization schemes to depict the target from different perspectives and reveal unexpected features, for the purpose of exploratory searches for proposing new hypotheses. Exploiting the data with multiple visualizations throughout the constraint space could also help target and amplify evidences that confirm or refute proposed hypotheses in explanatory searches.

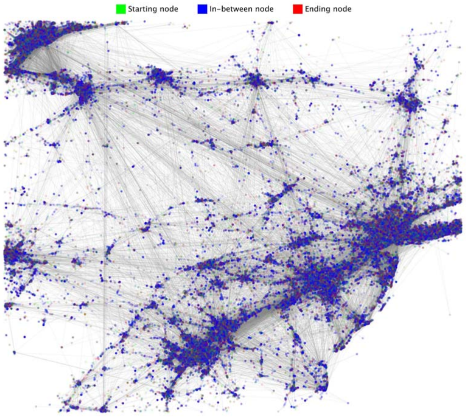


Fig. 4: An example of Cartesian Geography Map visualization

More specifically, a Cartesian Geographic Map (III-A) and Storyline Visualization (III-B) are the two main visualization approaches used in our framework.

A. Cartesian Geography Map

Cartesian Geography Map visualizes a target set of consumer trajectories under the context of a conventional geography map. Each involved 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 set. Therefore the origin of the coordinate system represents (LON_{min}, LAT_{min}) and the top-right corner denotes (LON_{max}, LAT_{max}). Node label types of search records are indicated in the map by their predefined node colors. Search record nodes in one consumer trajectory are connected by grey trajectory line. An example of Cartesian Geography Map is shown in Figure 4 which visualizes the target region enclosed by the blue rectangle in Figure 5.

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) and apply multiple Cartesian geography maps to visualize respectively subsets of consumer trajectories on time span sub-partitions. This makes the approach inappropriate for temporal pattern related hypothesis analysis.

B. Storyline Visualization

To incorporate 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.* [2], [3].

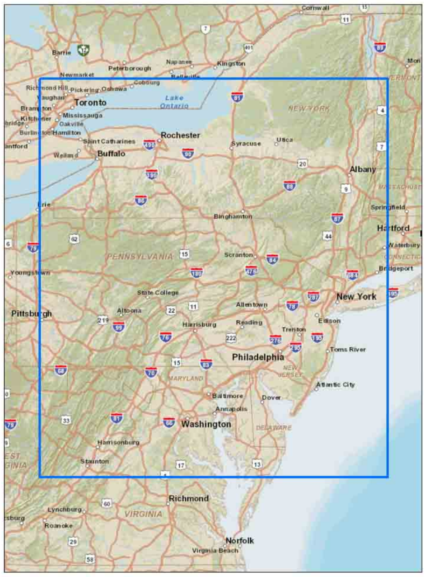


Fig. 5: The geographic map of the target region

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 $H \subset \Re$ [3]. 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 attributes in the data set. Values in the axes are in ascending order from left to right horizontally and bottom to top vertically.

Obviously Storygraph is suitable to visualize spatiotemporal search records. We normally refer to a node of search record plotted in Storygraph as an *event* because it indicates where and when a consumer search was conducted. Figure 6 illustrates how two events on 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 line is important to Storygraph in that any event node plotted without it in Storygraph loses the location information and represents some event occurred at that time with unknown location coordinates.

As shown in Figure 6, 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:

$$x = t \tag{1}$$

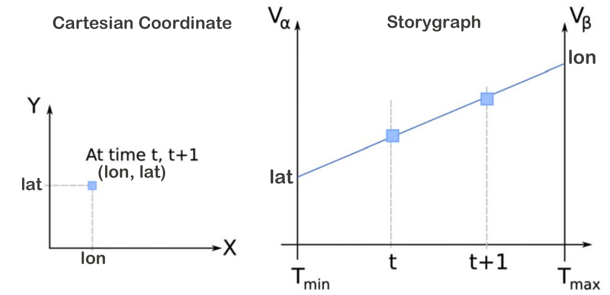


Fig. 6: Storygraph versus Cartesian [3]

$$y = \frac{(lon - lat)(x - T_{min})}{T_{max} - T_{min}} + lat \tag{2}$$

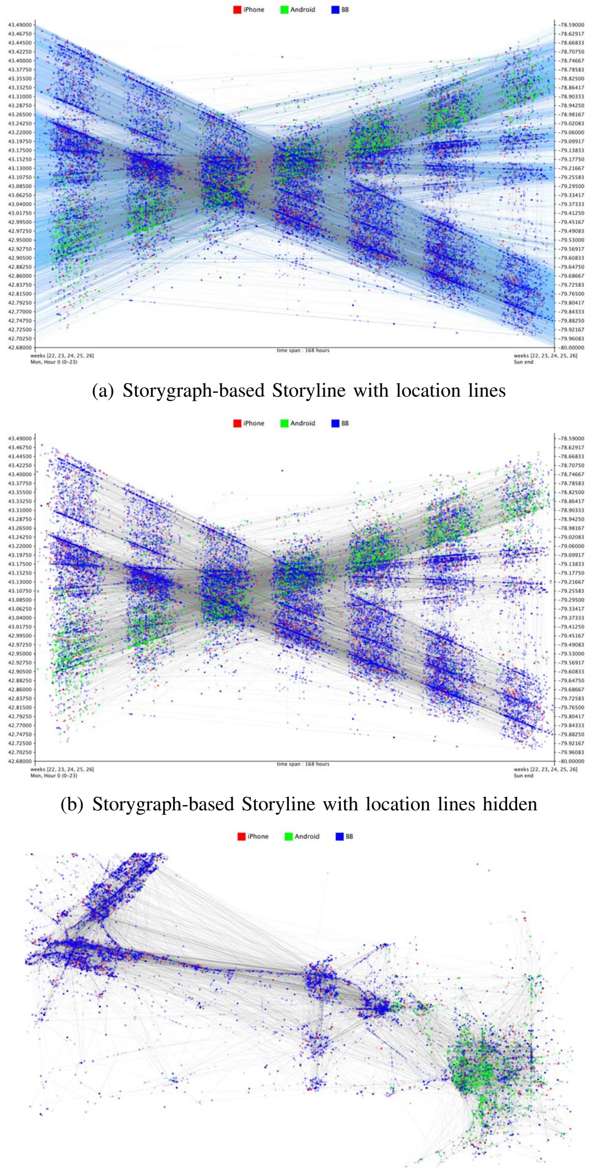
where T_{min} and T_{max} are the minimum and maximum timestamps of the data set.

Figure 6 also shows the advantage of Storygraph over Cartesian Geography Map when visualizing data with the temporal attribute.

2) Generate Storygraph-based Storyline: As described in II-D, 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 connecting sequentially all involved events visualized in Storygraph with trajectory lines, thus telling a story about the spatiotemporal 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. Besides, transparency level is used to help indicate the relative density of consumer trajectory components that are drawn overlappingly.

In the original scenario of Storyline based on Storygrah by Shrestha *et al.* [2], [3], consumer trajectories are distinctly colored by user to help visualize the movement of individual characters and the interactions between them on events. We introduce 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 behaviours.

Figure 7(a) 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 worse and brings more confusions to 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 7(b)). 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 event location (if the user rarely moves) or user movement, especially with the property of the Poynt data whose consumers exhibit geographical locality while conducing searches. Therefore we normally hide Storygraph location lines in Storyline visualization when the target data is dense in the 2D visualizing space.



(a) Storygraph-based Storyline with location lines

(b) Storygraph-based Storyline with location lines hidden

(c) Corresponding Cartesian Geography Map

Fig. 7: Examples of Storygraph-based Storyline visualization with and without location lines

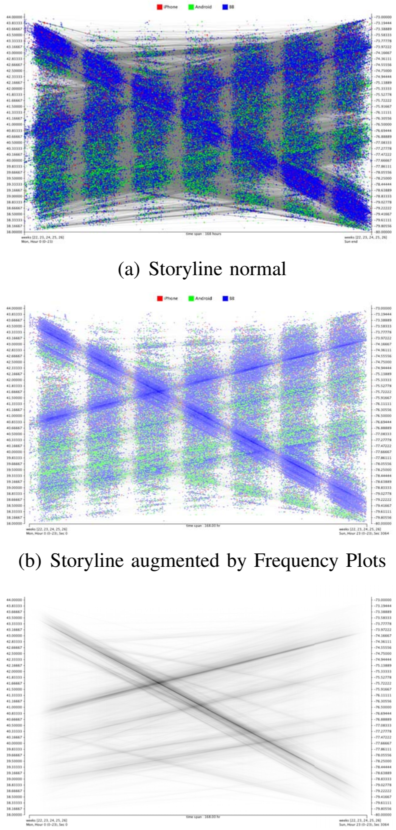
C. Augment Storyline with Frequency Plots

When the target dataset scales up drastically, hiding location lines in the normal Storyline visualization (III-B) is not enough to generate a visually satisfactory result for hypothesis analysis. The issue of occlusion, cluttering and color mixture brought by the overlapping of trajectory lines and Storygraph event nodes (without location lines) remains at an unnegligible level. Figure 8(a) exhibits such an example.

Based on the above motivation, a strategy named *Frequency Plots* [4] is introduced to augment normal Storygraph-based Storyline (described in III-B) 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 components associated with higher frequency superimpose those with lower frequency in terms of pixel lightness values.

Considering that 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 mainly involved are Storygraph event nodes of different label types and trajectory line, and the frequency information is cumulatively gathered pixel-wise by component type (event nodes of different labels are considered as components of different types).

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 classes and the reflected "relative density" should be in terms of other components in the same class. 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 obtain the final Storyline, 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 geographcial transition or connectivity, and general patterns involved with event nodes labels (*e.g.*, search categories, source devices) is what interest us the most. Therefore, a Frequency Plots augmented Storyline visualization is constructed with the layer of event nodes superimposed on that of trajectory lines. Figure 8(b) and Figure 8(c) show respectively an example of Storyline augmented by Frequency Plots and its individual layer of trajectory lines. When the data is extremely dense, the layer of trajectory lines could be completely covered by the layer of event nodes. Under such circumstances, we could check the layer of trajectory lines individually If necessary. Besides, the individual layer of trajectory lines helps provide a general sense about where strong patterns with node labels reside.



(a) Storyline normal

(b) Storyline augmented by Frequency Plots

(c) individual Storyline layer of trajectory lines

Fig. 8: An example of Storyline augmented with Frequency Plots

To sum up, the strategy of Frequency Plots helps avoid the issue of color mixture in the normal Storyline which relies on simple transparency levels to reflect relative density, and alleviates occlusion by selectively displaying the most important (in terms of component frequency) pixel-wise content. 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 problem of cluttering and occlusion.

An example augmenting Figure 8(a) is shown in Figure 8(b). Our contribution to the original usage of Frequency Plots introduced by Artero *et al.* [4] 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.

IV. Experiments to confirm evidence for consumer trajectories

In this section, 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 spatiotemporal stories and analyze more complex hypotheses with exploratory and explanatory searches.

A. A Glimpse of the Experimental Data

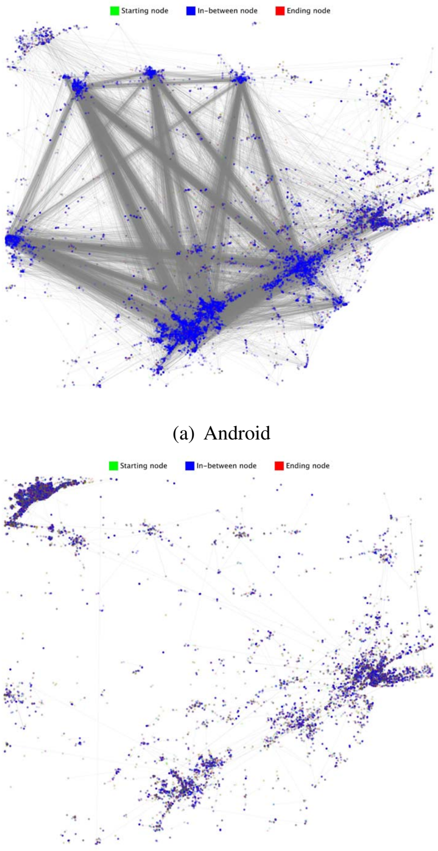
The dataset of Poynt search records based on which target consumer trajectories are constructed for experiment temporally spans five consecutive weeks, Week of May 29 to Week of June 26, in the year 2011. The general geographic region in experiment (the area enclosed by the blue rectangle in Figure 5, $LAT[38, 44]$ and $LON[-80, -73]$) mainly involves with two states in the northeast US, Pennsylvania and New York, and southern Ontario in Canada. Covered are several metropolitan areas including New York, Toronto, Washington, and Philadelphia.

B. Exploratory and Explanatory Searches

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

First start with a simple hypothesis, as considered above: “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 II-C1 by checking the hourly distribution of movie searches within one week. From the histogram in Figure 2, 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.

Then 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 9(a)) 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 propose as a new hypothesis and try to confirm with more evidences gathered. This is a typical example that interesting phenomena appeared unexpectedly during exploration may help invoke explanatory searches with new hypothesis inspired by the exploratory findings.



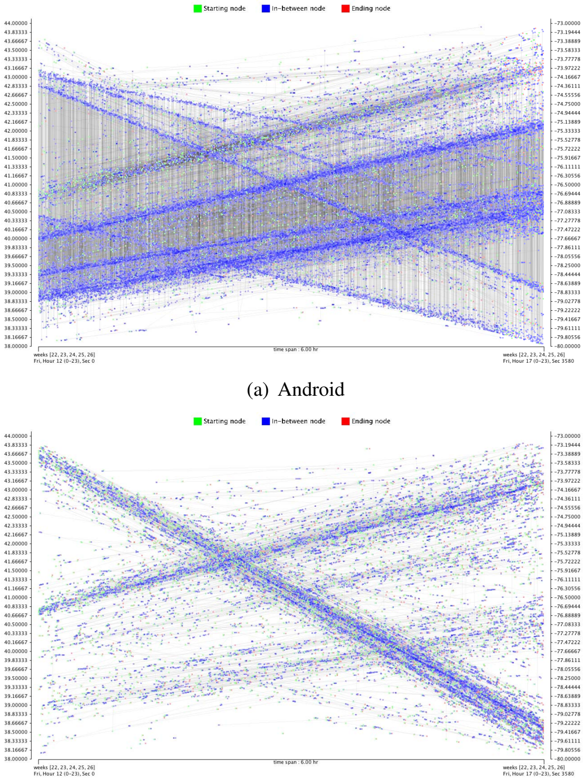
(a) Android

(b) BlackBerry

Fig. 9: Cartesian geography maps 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 10(a)), which indicates that those involved movie searches are abnormally made within extremely short time internal 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 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 filter out abnormal users in the analysis of consumer behaviours.

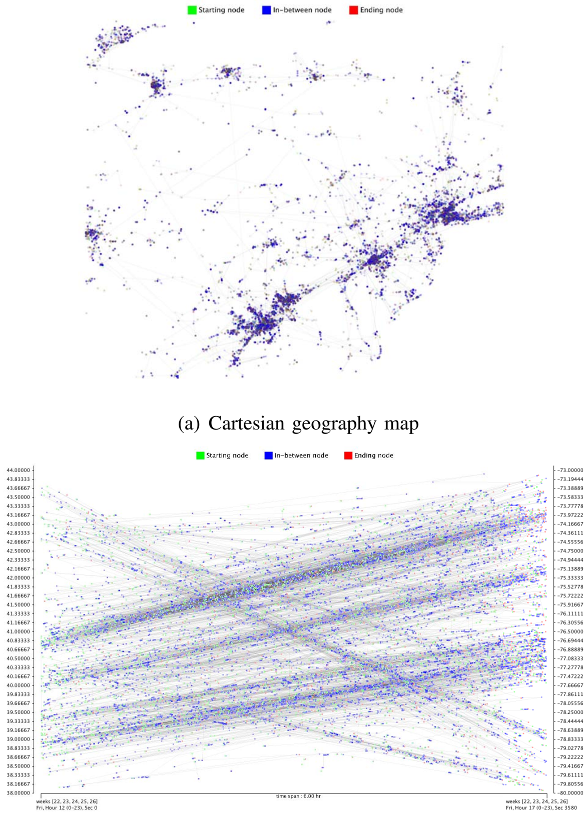
Figure 11 shows the visualizations after abnormal users are removed from Android users. As we can see in Figure 11(a) and Figure 11(b), the strong trajectory connectivity in Figure 9(a) and the prevailing vertical trajectory lines in Figure 10(a) disappear accordingly. Besides, after the removal of abnormal 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 hypoth-



(a) Android

(b) BlackBerry

Fig. 10: Storyline visualizations of movie searches by Android/BlackBerry users on Friday afternoons.



(a) Cartesian geography map

(b) Storyline

Fig. 11: Visualizations of movie searches by Android users (with abnormal users removed) on Friday afternoons.

esis. Meanwhile, answered 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 9(b) with Figure 11(a) and Figure 10(b) with Figure 11(b). 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.

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 both 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 [5], holds the most promise for supporting more sophisticated interactive visual analytics.

V. Summary and Future Work

Our approach to the 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 extents.

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 of interesting data artifacts. For future work, we plan to adopt the Webbles implementation to create 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.

As is obvious, while we have a complete prototype that provides access to our 178 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 to work with simultaneous refinement and improvement of the framework, as well as deeper exploration of more sophisticated semantic artifacts, hypothesized as consumer behaviour that will emerge.

Acknowledgments

This research is supported in part by the Alberta Innovates Centre for Machine Learning (AICML), the Alberta Innovates Technology Futures innovates Centre Of Research Excellence (iCORE), the Canadian Natural Sciences and Engineering Research Council (NSERC), Alberta Innovates Technology Future (AITF), and the Meme Media Lab of Hokkaido University.

References

[1] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 3:1–3:27, Jan. 2011. [Online]. Available: http://doi.acm.org/10.1145/1889681.1889684

[2] A. Shrestha, Y. Zhu, B. Miller, and Y. Zhao, "Storygraph: Telling stories from spatio-temporal data," in *Proceedings of the 9th International Symposium on Visual Computing*, ser. ISVC'13, Rethymnon, Crete, Greece, Jul. 2013.

[3] ——, "Storygraphs: Extracting patterns from spatio-temporal data," in *Proceedings of the KDD 2013 Workshop on Interactive Data Exploration and Analysis*, ser. IDEA'13, Chicago, IL, USA, Aug. 2013.

[4] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualizations," in *Proceedings of the IEEE Symposium on Information Visualization*, ser. INFOVIS'04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 81–88. [Online]. Available: http://dx.doi.org/10.1109/INFOVIS.2004.68

[5] J. Sjöbergh and Y. Tanaka, "Visual data exploration using Webbles," in *Proceedings of the Webble World Summit 2013*, ser. Springer CCIS, vol. 372, Erfurt, Germany, 2013, pp. 119–128. [Online]. Available: http://dr-hato.se/research/digitaldashboard2013.pdf