Tiramisu: Information from Live Data Streams

FINAL RESEARCH REPORT

Anthony Tomasic (PI), Joseph Giampapa (principle author), Steven Gardiner, Sophia Deng, Aaron Steinfeld

Contract No. DTRT12GUTG11

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

The Problem

The primary source of information for rider safety with respect to dynamic events such as cancelled buses, detours, traffic conditions and other factors is the transit system website. Although technological enhancements, such as real-time tracking, rider alert RSS feeds, and Twitter feeds, are available for transit users, such information sources do not always report updates reliably. For example, some real- time tracking systems stop transmitting updates about a bus if the bus takes a detour from its scheduled route. Rider alerts might temporarily suspend updates due to holidays or employee PTO. Notifications through social media frequently offer information of relevance, but are often difficult to understand, or for one to comprehend the impact of the message on their trip. Paradoxically, users are still faced with a lot of effort at navigating the appropriate information sources, finding and understanding messages and updates of relevance to their trip.

Our project goal is to access transit service live update data feeds, identify the routes and stops on which their updates will have an impact, and provide an integrated display of that information in the user's Tiramisu smart phone app. The extracted information can be used in multiple ways to improve the interactive experience of the user and keep them better informed. For example, if bus stops will be discontinued between certain hours of a day for, say, water main repair, and other stops will be established during that period, a search of nearby stops will reflect such temporary changes. This, combined with vehicle fullness data from the Tiramisu system, will allow riders to identify alternate transit trip options that are not too full to board.

Approach

The primary source of information for rider safety with respect to dynamic events such as cancelled buses, detours, traffic conditions and other factors is the transit system website. Although technological enhancements, such as real-time tracking, rider alert RSS feeds, and Twitter feeds, are available for transit users to access, such information sources may be difficult to access by transit users, or do not always report updates reliably.

To illustrate this problem, consider the live services that the Port Authority of Allegheny County offers. One of them is called, "Rider Alerts", an RSS feed, which requires "Live Bookmarks", "Outlook", or "My Yahoo!" to access its alerts. At one point, the latest alert was nearly a month old with the simple message, "There are no major service alerts at this time." Assuming that a transit rider has an RSS reader app on their mobile smart phone, it is easy to imagine how they would overlook consulting this site, and miss any unexpected updates.

A second source of updates is the real-time tracking system. However, many of these systems drop data streams for buses that need to detour off the scheduled route, leaving riders in the dark on whether their bus will arrive or is serving their stop. Sole reliance on the real-time arrival data feed can condition a rider to believe that service on specific routes is unavailable, when indeed, it is only dealing with contingencies. Further, it does not update information on the transit service schedule or indicate if the bus is so full that it is not allowing any more riders.

The third source of rider updates is through social media messaging, such as through Twitter feeds. The Port Authority of Allegheny County offers a Twitter feed where its tweet notifications are prominently posted. Twitter catalogs rider "responses" to tweets as a "conversation" to the Port Authority tweet, and are therefore not immediately visible through the feed. Yet, sometimes, rider "responses" ask a question that is unrelated to the original post, and indicate other possible concerns. What follows is a recent post (personal name is anonymized):

> **Port Authority PGH @PGHtransit** 6h6 hours ago
> On Tuesday (11/11) in observance of Veterans Day, the Downtown Service Center will be closed and twitter will not be staffed.
> 0 replies 2 retweets 0 favorites
>
> **@g...user** 5h5 hours ago**@PGHtransit** Is Glenbury Street on the South Pat Way now open? 0 replies 0 retweets 0 favorites
>
> **Port Authority PGH @PGHtransit@g...user** Not sure if it's open to all traffic yet but our buses are getting through to/from the busway. 0 replies 0 retweets 0 favorites11:59 AM - 7 Nov 2014
>
> **@g...user** 4h4 hours ago **@PGHtransit** Thank you!

This exchange reflects a partial status update on an unrelated topic: it is clear that a route is now open, but there is still doubt about on-time status as well as which stops may be open, closed, or relocated.

Methodology

Tiramisu Transit feature upgrades included the ability allow transit users to communicate with social media – to send and receive messages – by way of their Tiramisu app. In addition to the human factors convenience, this facilitates the enrichment of data that is being transmitted, such as by providing access to a rider's GPS location along with the messages they originate or read. The focus of this project effort was on automatically incorporating information of relevance to a rider's planned or actual transit trip, fused from multiple live transit data sources, and integrated into the app user experience. For example, nobody wants to see detour alerts for a route they never use so it is important to only surface alerts for the correct routes and provide ways for users to surface timely information about the routes they care about.

Safety is enhanced in a variety of ways, both directly, and indirectly. A direct way is that messages of an emergency or distressed nature will be automatically geo-located and associated with a service, route, and/or stop. Indirectly, the safety benefits are to allow for the avoidance and decongestion of an area impacted by an event or accident. Not only is congestion on the public transit service reduced or avoided, but also by better up-to-date trip planning, congestion at the transit stops can be avoided, as well.

All members of this project are from the Rehabilitation Engineering Research Center on Accessible Public Transportation (RERC-APT), which is housed in the Robotics Institute. The RERC-APT is the main funding source for research on and with the Tiramisu Transit app. Tiramisu Transit, LLC is the lab's spinout and is responsible for app deployment. This technology effort was focused on back-end server components for enhancing rider

awareness of detours and alerts. It was designed to leverage the new messaging system within the Tiramisu app, which already has a very large deployed base of daily app users.

Findings

Attaching Transit Tweets to Transit Objects

As mentioned, a big challenge when working with tweets about transit is to attach specific tweets with specific transit routes and stops. This can be very difficult for a variety of reasons. Tweets specify service interruptions in terms of routes and landmarks: e.g. *Service interruption on buses 67 and 69 between Beeler and Wightman*. Actual stop information can only be recovered via reference to route information, a GIS database and a database of bus stops. Stops are removed from a route when there is a partial service interruption; sometimes alternative stops are added. Often tweets provide tiny URLs to full HTML web pages in which all affected and new stops are listed. But HTML pages provide graphical markup for information, not logical markup, and it is not always possible to parse HTML tags with a static grammar to extract the affected stops. There is too much variability in the HTML markup. This is an age-old web problem, and one that the W3C has attempted to address by developing the Web Ontology Language, OWL. Unfortunately, OWL has not enjoyed as much widespread use as HTML has.

For this work we examined the Port Authority of Allegheny County (PAAC) and Metropolitan Transportation Authority (MTA). To illustrate the scale of this problem, Port Authority has 207 routes and 7,107 stops while MTA has 323 bus routes and 15,382 stops (as of when this research was done). In an analysis of 834 twitter messages that spans 10.5 calendar months, PAAC tweets only during canonical business hours (one 8-hour shift), and generates from 0 to 30 tweets in a day, with a mean of 2.4 per weekday. Although PAAC does not use tags to identify message types, only messages of a service nature are originated directly by PAAC. Promotional messages, traffic messages, and transit police messages searching for missing children or persons of interest are all in the form of retweets from other organizations. Only one verb is used in the infinitive tense, and most PAAC tweets analyzed could be described by ten high-level grammar rules.

The MTA, on the other hand, tweets 24 hours a day, 7 days a week, with around 175 tweets per day. MTA tweets use hash tags to identify message type tags and route numbers. They use a consistent vocabulary of abbreviations, and sometimes add a generic URL to their web site for more information. While MTA messages showed a bit more variability in verb usage, their messages displayed a remarkable similarity to those of the PAAC in terms of specifications of the affected services, time expressions and durations, causes for the interruptions, and recommendations.

Fortunately, tweets from the two agencies we examined have similar content and structures. In the examples below (Table 1), green text corresponds to system identified routes, maroon text corresponds to route direction, and orange text corresponds to geographic regions. Each agency follows a different style and structure, but it is still possible for computer systems to extract and useful information for linking tweets to specific vehicle routes.

Table 1. Tweets with identified routes, directions, and geographic regions

Port Authority of Allegheny County (@PGHtransit)
Gearing St. closure to affect the 44 Sat. 8/22 from 2p-6p. Details: http://t.co/KLGhW1h6sF
Closure at Bloomfield Bridge to affect the 93 beginning Mon. 8/24 at 6a until the end of Sept. Info: http://t.co/PndOqKosP4
Sporting events in the north shore to affect buses this weekend: http://t.co/ILPlS494i6 & http://t.co/c9JgPqxauK
OB Single-tracking on Blue Line between King's School Crossover & Library Station, Sun. 8/23 from 7a-7p. Details: http://t.co/RZeVcVB5wj
Metropolitan Transportation Agency (@NYCTBus)
#ServAdv: s/b, #M15 & #M15-SBS buses expect delays, due to heavy traffic on 2 Av from 110 St to 67 St. Allow additional travel time.
#ServAdv: n/b, #M2 buses are detoured, due to construction on Audubon Av from 167 St to 168 St. See http://mta.info
#ServAdv: Due to the United Nations General Assembly expect delays in Midtown Manhattan on #Local and #Express bus service.
#ServAdv: Shore Rd bound #B64 & b/d #B70 buses are detoured due to paving on 8 Av b/t 72 St & Bay Ridge Av. See http://mta.info

Initial work by the team involved implementing the following three types of parsers in native Python3: (1) a regular expression parser for time expressions as generated by the transit agencies, (2) a chart parser for recognizing full and constituent parses, and (3) a fuzzy matcher to extract the pithy advisories that defy characterization via a static grammar, but can be identified and extracted by isolation. Transit route and stop databases were converted into Python3 data structures as a mockup for an eventual interface to the transit databases. Semantic frame schemas were also created for situating and interpreting the parsed text. Once working, the gathered information can be written to a database for use by Tiramisu and other transportation systems.

A large part of our effort was determining the best structures to draw "correct" extractions from each individual tweet. The final approach is likely to feed agency-specific information, like route short names and numbers (e.g., 69 Trafford), into automated manipulations designed to capture the most likely variations used by transit riders. Much of this data is likely to be obtainable from public agency data (e.g., GTFS, real-time arrival feeds, etc.). However, we feel additional information sources will probably be needed. For example, a tweet about a problem near "the Consol center" will not link since no stops have this name in their descriptions. Utilization of a GIS source will likely be needed. While individual rules can be written, the challenge is finding an approach that is scalable to multiple cities.

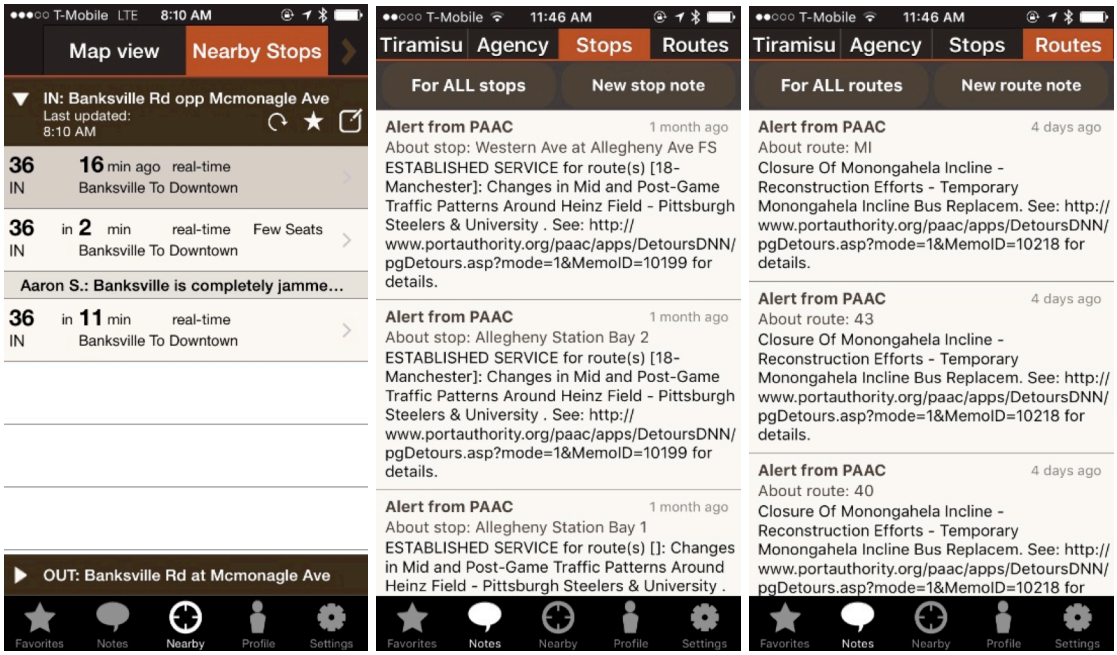

Figure 1. Examples of Tiramisu Notes: rider generated (left), system generated stop notes (middle), system generated route notes (right). Notes can be filtered by stop or route.

Attaching Detour Alerts to Transit Objects

In parallel with this effort, the team developed parsers specific to Port Authority alerts and integrated the output into Tiramisu Transit. The system extracted alert content from the Port Authority detour notices, identified the affected stops and routes, and loaded the information into the Tiramisu messages database. Within Tiramisu, riders can also enter their own messages if they have accounts.

Trip messages are those attached to specific route-time pairs (e.g., the 36 inbound at 8:10 am). *Route* messages are either old trip messages or messages that reflect all trips within a route (e.g., all 36 trips). Figure 1 shows a rider-generated trip message, system generated bus stop messages, and system-generated route messages. These are all screenshots from the publicly deployed Tiramisu system.

Other Achievements

The team was active in a variety of stakeholder activities, including regular professional service and meetings with interested industry visitors. For example, Aaron Steinfeld continues to serve on the National Academies of Science, Transportation Research Board, Standing Committee on Accessible Transportation and Mobility (ABE60)[1]. He is the Co-Chair of the Technology subcommittee. Joseph Giampapa participates as a Friend of TRB committees: ABE60 and ABJ70, the Standing Committee on Artificial Intelligence and

[1] https://www.mytrb.org/CommitteeDetails.aspx?CMTID=1164

Advanced Computing Applications. He was also a member of the ABJ70's program committee for the 2016 TRB annual conference and reviewed manuscripts for the conference and journal.

Conclusions

Our main finding from this effort is that current transit agency practices lead to semi-structured alerts and tweets. The higher degree of structure, as compared to regular natural language, increases the feasibility of properly linking messages with explicit routes, stops, and other transit elements.

However, there are still challenges with integrating multiple forms of reasoning and pattern matching, and with acquiring location-specific entities (e.g., Console Energy Center) that require hand labeling prior to system deployment. There is potential for linking other geographic data structures to bootstrap this process into more automated methods, but this is a larger research challenge.

We have also demonstrated that semi-structured alert messages can be properly connected to routes and stops for integrated information sharing to end users. For example, Tiramisu Transit is now displaying alerts from the Port Authority.

Recommendations drawn from the project:

1. While some agencies are starting to include alert messages into their real-time data streams in ways that eliminate the need for natural language processing, this is still very rare. More agencies should explore this option if their data systems support the feature.
2. Agencies should adopt and utilize structured language to facilitate high system accuracy. For example, including # before route names and being consistent in direction labels (OB, n/b, etc.).
3. Methods are needed for linking semantic labels of geographic data to explicit GPS coordinates (e.g., Console Energy Center, Shadyside, etc.).