Interactive Summarization of Social Media

Wen Li
Delft University of Technology
Delft, Netherlands
wen.li@tudelft.nl

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

Arjen P. de Vries
CWI
Amsterdam, Netherlands
arjen@acm.org

ABSTRACT

Data visualization and exploration tools are crucial for data scientists, especially during a pilot study. In this paper, we present an extensible open-source workbench for aggregating, summarizing and filtering social network profiles derived from tweets. We briefly demonstrate its range of basic features for two use cases: geo-spatial profile summarization based on check-in histories and social media based complaint discovery in water management.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data mining;
J.4 [**Computer Application**]: Social and Behavioral Science

Keywords

Twitter, Data Visualization, Data Summarization

1. INTRODUCTION

For more than a decade, social media services have been experiencing massive growth rates and cover critical shares of the developed world's population. More recently, they draw substantial attention from various disciplines of the research community. For example, Twitter, as one of the most popular online social media platforms boasts more than 200 million active users producing half a billion tweets (short messages, photos, web links, etc.) per day [8]. The service inspired a wealth of interesting studies, such as detecting earthquake-related events [7], investigating how people influence others on social media [1] and predicting users' mobility [3]. Many of these quantitative, data-driven studies are enabled by the high degree of diversity, coverage and scale at which information is available on Twitter. Qualitative approaches, however, may regard these exact properties as obstacles. They often involve the exploration of a set of carefully chosen, focused samples. Under this setting, how should one select the right test subjects among millions of users? How should the available data be adequately partitioned? And, finally, how should the research insights be abstracted and communicated to third parties? The numerous individuals and research groups around the globe that rely on samples from Twitter and YouTube as a scientific resource have come up with their very own, individual solutions to these questions. In this paper, we briefly describe an extensible open-source workbench to facilitate easy profiling of data samples in an interactive manner. The workbench provides researchers, data architects and users with an easy handle on understanding data collected from Twitter through private crawls; encouraging reuse of code within the research community.

IIiX '14 Aug 26-29 2014, Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637050

2. RELATED TOOLS

2.1 Social Media Visualization

The great popularity of online social media resources among researchers calls for a diverse arsenal of tools for data organisation and visualisation. Marcus et al. [5] proposed a system called *TwitInfo* for analysing and visualising the shifting sentiment of Twitter streams. It shows the distribution of tweets over time, their sentiment categories, a map for geo-tagged tweets and a list of relevant tweets given a query. *Tweet Sentiment Visualization App* (`http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app`) visualises tweets alongside a sentiment chart for classified tweets, maps of geographic spread and word clouds. *Sentiment140* (`http://www.sentiment140.com`) also shows sentiment classification of tweets expressed in a number of pie and bar charts. *Socialmention* (`http://www.socialmention.com`) offers sentiment visualization including social influence indicators. The commercial service *UberVU* (`https://www.ubervu.com`) provides a wide range of analysis tools for brand management and marketing efforts. Besides those sentiment-centric visualization tools, there are many other system for social media visualisation. *Livehoods* (`http://livehoods.net`) is an interactive map of user mobility based on check-ins from Foursquare in eight selected metropolitan areas. *Vizify* (`http://vizify.com`) is a personal profiling assistant that can automatically generate infographics showing aspects such as favourite topics, living locations, connectivity, etc. *TwiqsNL* (`http://twiqs.nl`) is a Hadoop-based search engine for Dutch tweets collected from Twitter's Streaming API, which depicts tweets matching query filters like keywords or hash-tags in distributions over time, location and user categories, as well as the textual content in a word cloud. Crisees [6] is proposed for monitoring realtime so-

cial events on Twitter, especially for crisis discovery. It is composed of a list view of tweets related to a queried event, a map showing the whereabouts of those tweets and a list of related media. Eddi [2] is a Twitter client optimized towards better tweet reading experience. It groups tweets into topics and showing them along with a trending timeline and a word cloud.

Each of these tools is tailored for their specific narrow purposes and they are neither open source nor customizable for other applications.

2.2 Online Chart Generation

There is a wide range of online chart generation solutions that accept arbitrary data streams in standardized formats. Some simple ones are Web-based versions of spreadsheet applications, e.g., *iCharts* (`http://www.icharts.net`), *Chart Maker* (`http://almaer.com/chartmaker`). The more sophisticated examples of this class offer greater control over design and layout and facilitate generation of posters and infographics. Popular examples of this category include *Infogra.me* (`http://infogra.me`), *Piktochart* (`http://piktochart.com`) and *VANNGAGE* (`http://vanngage.co`). These tools only provide limited means of importing data, and often lack the ability for automatic data aggregation. Users have to manually input data points in each chart and their interactivity is limited.

Many Eyes (`http://www-958.ibm.com/software/data/cognos/manyeyes`) and *StatPlanet* (`http://www.statsilk.com`) are two online applications that are similar to our social media workbench. They have both been designed for interactive data visualization. Many Eyes is hosted by IBM as a Java applet-based Web application. It supports 17 types of charts and endorses users for sharing their dataset and inferred charts. The main drawback of the application is that charts cannot be connected for interactive operation, so interactions on one chart do not propagate to other display elements. StatPlanet supports such propagation of interaction on a set of charts, but the Flash based interface is limited to geographical visualization and not customizable. Additionally, both of them require users to manually input data.

In contrast, our proposed workbench is designed and implemented towards a system that can help users compose comprehensive interfaces containing interactively connected charts of different types. It also allows for easy modular expansion, embedding and modification.

2.3 Comparison to Existing Libraries

To maximize the compatibility and flexibility of our workbench, it was implemented in JavaScript. Although there are several existing JavaScript libraries for chart generation, they are often not easy to use, such as *D3.js* (`http://d3js.org`). For a research or prototype system, using these libraries can cost significant time and effort. Our workbench, on the other hand, focuses on providing a simple domain specific language (DSL) for composing charts so that users can focus more on what types of charts are suitable for their tasks rather than how to program them. Table 1 shows a list of popular charting libraries for Web development and their individual strengths and limitations in 4 dimensions, i.e., the diversity of built-in charts, the support of DSL (Domain Specific Language) templating, the effort to program a prototype with them, and the support of interaction. As shown, *DC.js* (`https://github.com/dc-js/dc.js`) is the library that comes closest to meeting our requirements. There is however still a need for programming efforts before it can be used for general purposes. In addition, we also integrage a world cloud drawing library *wordcloud2.js* (`http://timdream.org/wordcloud2.js`) and a library *GMaps.js* (`http://hpneo.github.io/gmaps`) for easier accessing Google Maps. Consequently, we encapsulate all these charting modules into *directives* in the framework *AngularJS* (`http://angularjs.org`), which can then be used as a DSL for composing a workbench for exploring social media dataset.

3. SYSTEM DETAILS

We would like to stress that the proposed system is not "yet another library for online chart generation". It is designed for facilitating researchers to efficiently build a prototype interface for data aggregation and analysis. For this exact reason we ensured its easy extensibility.

3.1 Features

The key features of the proposed Social Media Workbench are:

- Easy composition of new chart layout with built-in directive tags.
- It can be used as a standalone application as well as embedded in another interface.
- It is able to load data from local storage or 3rd party APIs on the fly.
- Built-in pie, line, bar charts, word clouds and maps.
- Drill down operations for data via charts.
- APIs for extending and embedding the workbench into another Web application.
- Source code available under MIT License.

Figure 2 shows an example interface of the Social Media Workbench, summarizing a user's activity in terms of a number of pie charts, bar charts as time lines as well as categorical and geo-spatial spread on an interactive map. All charts in the interface are dynamic; interacting with any of the charts (e.g., clicking on a slice in a pie chart) triggers immediate updates of all other charts (e.g., focusing on the selected share of data). The chart layout can be easily customized through template editing. As an example, a tweet word cloud could be inserted by simply adding a single tag.

3.2 Architecture

The workbench is built as a Web application to make optimal use of state-of-the-art technology and allow for flexible platform-independent deployment either locally or on any popular *PaaS* (Platform as a Service). As shown in Figure 1, the workbench has two parts:

Computation of statistics and chart plotting are accomplished on the JavaScript front end. From the data stream provided by data APIs, the front end will extract properties of each item in the stream and send them to corresponding charts. When the user selects, for example, a particular category of places, a filter will be applied to the stream of items where only the matched items will account for the update and all other charts will be synchronized to the new selection criteria.

The back end is a dedicated server for hosting data APIs, providing easy handles on data from Google Datastore or

Table 1: Comparison of existing libraries

Name	Built-in Charts	DSL Templating	Programming effort	Interactivity
D3.js	–	No	JS + SVG + CSS	NA
DC.js	++	No	JS	Cross Charts, Drilling Down
Chart.js	++	No	JS	Zooming
amCharts	+++	No	JS	Zooming
Google Chart Tools	++	No	JS	Data Visibility
HiCharts	+++	No	JS	Data Visibility
YUI Charts	+++	No	JS	Data Visibility
plot.ly	+++	No	JS or Python or Matlab	Zooming, Data Visibility
Angular-Chartjs	++	Yes	TAG + JS	No

other data APIs online (e.g., Twitter). For example, users can feed a list of tweet IDs or user IDs or search queries for the workbench to automatically obtain all corresponding tweets.

The front end and back end are decoupled by simple REST APIs, through which JSON formatted data is communicated between them. So the workbench can either run as a standalone application (with data served by built-in back end running on Python) or be embedded in another application where the front end pulls JSON formatted stream of data from that application.

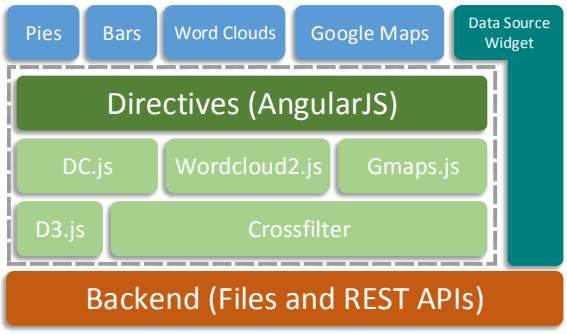

Figure 1: The Architecture of Social Media Workbench

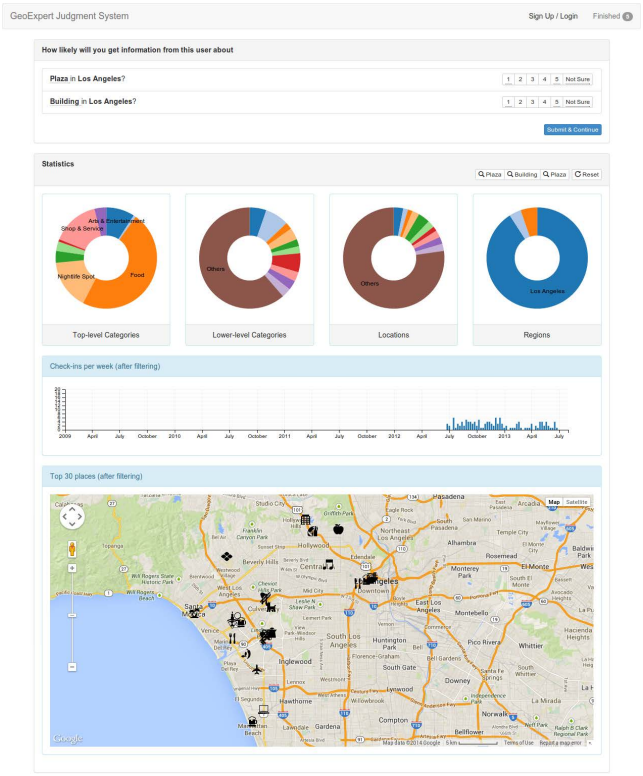

Figure 2: The geo-expertise annotation interface.

4. USE CASES

In this section we demonstrate our workbench in two different projects to show its flexibility and ease of customization.

4.1 Geo-Spatial Summaries

The demonstrated workbench was originally used for assessing users' expertise towards *points of interest* (POI), in order to improve ranking algorithms [4]. The assessment task is a complex process in which assessors are required to look deeply into users' check-in profiles, each of which can be composed of up to thousands of individual geo-tagged tweets. It would not be effective nor efficient to let assessors go through all available tweets to find out whether the user knows about a given place or a type of places. Instead, we created an interface based on the proposed workbench to show as many dimensions of a profile as possible, without exceeding or cluttering the confines of a single page. The aggregated diagram lets assessors focus on the actual check-ins rather than the content of tweets.

As shown in Figure 2, we include in the interface 4 pie charts for showing the distribution of check-ins across different types of locations and a chart of geo-tagged tweets distributed over time. Additionally, we provide a map to show the spatial distribution of check-ins. Our assessors can use this interface to comfortably inspect different dimensions of users' check-in profile. For example, they can click on any of the pie charts to select a type of place. Other charts including the map will immediately show only the check-ins at the selected type of places. After exploring a user's profile, assessors are required to evaluate the user's expertise for each given place or type of places, based on the visualized summarization and assign the scores accordingly at the top of the interface.

We offer a refined and comprehensive interface for the as-

sessors to avoid biased answers caused by lack of information or by obtrusive interfaces. For this project, we include charts and maps in the workbench to display the users' profiles in as many dimensions as possible. Moreover, the interactivity of our proposed workbench helps assessors to look into different dimensions and form their own ways of exploring profiles for the tasks. For example, when assessors are evaluating a users' expertise towards a place, they may also want to know users' check-ins at other similar types of places. The selection can be narrowed down to include only the desired type of places with a single click. The category distribution of check-ins and all other charts will be immediately updated to the new selection. In the original study, we encouraged assessors to explore new ways of using the workbench, for example, looking into the check-in distributions only in the past month by selecting check-ins on the bar chart, or investigating check-ins in different cities by clicking on the pie chart for check-in distribution over cities. By giving assessors freedom to explore the data, we hope to minimize the possibility of leading them towards false conclusions.

4.2 Complaints Discovery in Social Channel

We used the workbench for another project which leverages social media channels to collect complaints about damages caused by heavy rain in Rotterdam, Netherlands. Based on these complaints the municipalities may discover vulnerable locations in urban areas and then try to improve them.

As the first step in the pilot study, we need a tool for exploring social media for potential textual features characterizing complaint tweets. For this, we compose an interface based on the previously described workbench for summarizing the tweets relevant to a given search query. With this tool, we find various keywords to identify tweets regarding complaints about heavy rain damages so that we can monitor such keywords on the Twitter Streaming API.

In this adaptation, we include a bar chart, a word cloud and a structured data table. The bar chart shows the time distribution of relevant tweets, the word cloud contains frequent words in the selected tweets and the data table will show tweets sorted by their relevance towards the filter queries. A thin wrapper over a 3rd party full-text search library is programmed so that the front end can select and access relevant tweets though a REST API. Figure 3 shows an example query of *wateroverlast* which means *flood* in Dutch. As can be seen, the relevant tweets peak after the 11th of October 2013, the day on which an exceptionally heavy storm hit the city of Rotterdam. We can either zoom into a particular day to check the word distribution of tweets by select a specific range on the time line, or select another word as a new query by clicking on that word in the word cloud. With this tool, we refined our keyword set to monitor and collect more tweets about upcoming complaints.

5. CONCLUSION

In this paper, we described the first version of the *Social Media Workbench*, an open-source system that is designed to help researchers explore large social media data sets originating from platforms such as Twitter. The tool is an ongoing engineering effort, and in the future, we will integrate more content-related aggregation and summarization tools such as latent semantic analysis or unsupervised clustering functionality based on user defined fields. In this way, we hope to reduce the need for code replication throughout

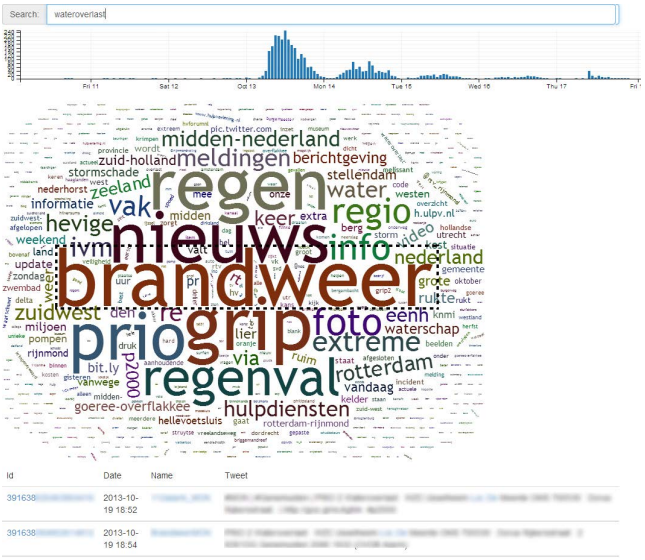

Figure 3: Water Complaints profiling tool

the research community by giving a flexible, easy-to-adapt environment for qualitative research on potentially large-scale data. The current version of the tool along with the two described use cases, focuses mainly on meta information such as geo-tags that associated to tweets. The Social Media Workbench is available for download at `https://github.com/spacelis/portraitist`.

6. REFERENCES

[1] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter Eytan. In *WSDM '11*, pages 65–74. ACM Press, 2011.

[2] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *UIST '10*, pages 303–312. ACM Press, October 2010.

[3] Wen Li, Carsten Eickhoff, and Arjen P. de Vries. Want a coffee? Predicting Users' Trails. In *SIGIR '12*, pages 1171–1172. ACM Press, August 2012.

[4] Wen Li, Carsten Eickhoff, and ArjenP. de Vries. Geo-spatial Domain Expertise in Microblogs. In *ECIR '14*, volume 8416, pages 487–492. 2014.

[5] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI '11*, pages 227–230, May 2011.

[6] David Maxwell, Stefan Raue, Leif Azzopardi, Chris Johnson, and Sarah Oates. Crisees : Real-Time Monitoring of Social Media Streams to Support Crisis Management. In *ECIR '12*, pages 573–575. 2012.

[7] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors Takeshi. In *WWW '10*, pages 851–860. ACM Press, April 2010.

[8] Twitter. FORM S-1 REGISTRATION STATEMENT, 2013.