CitizenHelper: A Streaming Analytics System to Mine Citizen and Web Data for Humanitarian Organizations

Prakruthi Karuna, Mohammad Rana, Hemant Purohit
Humanitarian & Social Informatics Lab, Department of Information Sciences & Technology
George Mason University
Fairfax, VA, USA
{pkaruna,mrana3,hpurohit}@gmu.edu

Abstract

Vast information available on social media through citizen sensing, complemented by diverse Web sources such as blogs, and news present an unprecedented opportunity for organizations to improve their work practices using such data. We will demo a novel, scalable and interactive streaming analytics system with a collaborative visual dashboard. The proposed system mines heterogeneous information streams from social and web platforms for analyzing real world events in humanitarian domain, along dimensions of information source demographics, time, location, and content summaries. Specific use-cases of gender-based violence and global displacements will be presented to demonstrate contrasting analyses of diverse data sources, unlike existing social computing systems with single data source, in humanitarian domain.

Introduction

Social media and open web data platforms have revolutionized information production and consumption in society. Social media has empowered information sourcing and sharing at an unprecedented scale, however, it has challenged small to large organizations for extracting relevant information at a rapid pace and scale. Within the humanitarian domain, non-profit organizations are often resource-scarce to develop sophisticated tools and techniques to improve their work practices by mining novel data sources. Mining behaviors from social media as well as intelligence from open Web sources including news, and linked data for real-world events present an opportunity to assimilate relevant knowledge for improving absorptive capacity (Cohen and Levinthal 1990) and work processes of complex humanitarian organizations (Kovács, Tatham, and Larson 2012; Meier 2015). Social computing systems for large-scale humanitarian event analyses are limited by a single data source analysis, such as UN Global Pulse's monitoring system[1] and CrisisTracker (Rogstadius et al. 2013), which may not provide diverse perspectives about narratives of events (e.g., social media versus local and global news). Furthermore, existing systems do not allow collaborative visual analytics, to help customize, save and share insights among team members. We propose a system *CitizenHelper*, which can mine citizen-generated data from diverse social platforms and open web data sources, to produce multidimensional analyses of real-world events along the facets of information source (*who*), and key information attributes—time (*when*), location (*where*), and topics (*what*). We first describe system architecture based on opensource technologies (Apache Kafka and Spark, ElasticSearch and Kibana).

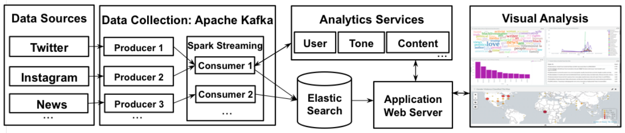


Figure 1: System Architecture for contrast stream analytics.

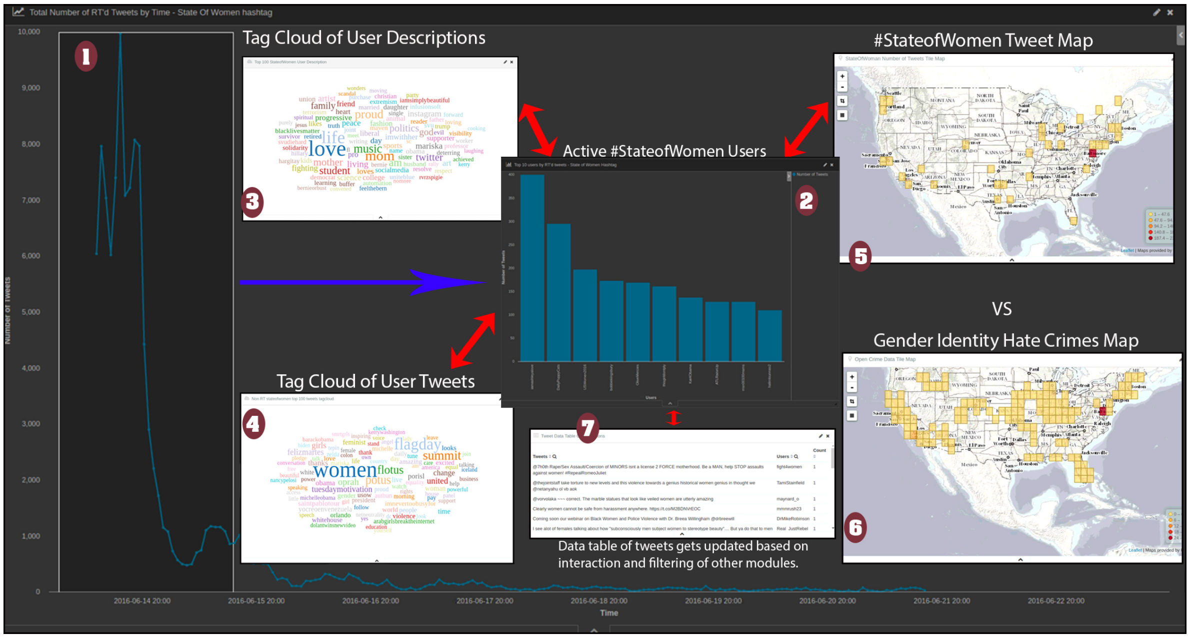
System Design

Data Collection. *CitizenHelper* uses an opensource distributed computing platform *Apache Kafka* to collect data (see Figure 1), which provides flexibility to scale producers (information sources), and consumers (information processors), in addition to a streaming data buffer—valuable for slow downstream processors when needed. System currently supports realtime data collection using Streaming and Location APIs of Twitter, as well as Instagram and Facebook (for public groups and pages). Additionally, the system supports collection of news (including GDELT (Leetaru and Schrodt 2013)) and blogs streams as well as data from Web knowledge bases including Wikipedia, and OpenGov Data.

Metadata Processing. The proposed system connects data collection components from *Apache Kafka* to processors in opensource stream computing framework *Apache Spark*. Different processors perform analytics on the streamed content by leveraging various analytics services, to extract and associate enriched metadata such as information provider classification (e.g., gender, user type such as organization), content classification for topics, intent, etc.

Data Storage and Visual Analytics. *CitizenHelper* stores raw data in a file system for long-term archiving, and processed data with extracted metadata in a *ElasticSearch* database, which supports a frontend visualization dashboard *Kibana* for streaming analytics. Our visual dashboard is

[1] `http://post2015.unglobalpulse.net/`



Tweet Data Table

Tweets	Users	Count
@7h0th Rape/Sex Assault/Coercion of MINORS isnt a license 2 FORCE motherhood. Be a MAN, help STOP assaults against women! #RepealRomeoJuliet	fight4women	1
@thejointstaff take torture to new levels and this violence towards a genius historical women genius in thought we @netanyahu ol vb aok	TamiStainfield	1
@vorvolaka ~~~ correct. The marble statues that look like veiled women are utterly amazing	maynard_o	1
Clearly women cannot be safe from harassment anywhere. https://t.co/M2BDNVrEOC	mmmrush23	1
Coming soon our webinar on Black Women and Police Violence with Dr. Breea Willingham @drbreewill	DrMikeRobinson	1
I see alot of females talking about how "subconsciously men subject women to stereotype beauty".... But ya do that to men	Real_JustRebel	1

Data table of tweets gets updated based on interaction and filtering of other modules.

Figure 2: Illustration of analysis using *CitizenHelper* system widgets (numbered for explanation) for #StateOfWomen anti-GBV campaign, which provided insights on participating demographics for explicit self representation of family relationships and gender identity. Furthermore, content analysis can be explored by user types, such as an organization versus an individual.

composed of different analytical widgets, such as volume trend graph of Twitter posts (tweets) over time (Fig.1, widget 1). These widgets have two unique features. First, when a user interacts with a widget and modifies an analysis unit on the widget (e.g., time slice on a trend graph, region of interest on the map, topical tag in the word cloud list), then all analytical widgets get updated corresponding to that change in the analysis unit. Second, the visual dashboard supports collaborative teamwork by allowing saving and sharing of a state of the dashboard by an end user, which in turn allows another collaborating team member study the same set of analyses from his/her colleague. Also, these widgets can be repositioned and deleted as needed to avoid visual information overload. System details with exemplary analyses and demos are available at: `http://ist.gmu.edu/~hpurohit/humanitarian-informatics-lab/icwsm17-citizenhelper.html`

Analysis Types.

- *Temporal diffusion*, to study engagement over time (e.g., via Retweet trends; Fig1. widgets 1 & 2).
- *Demographics*, to study engaged user types (e.g., via analysis of user profile summaries for self representation of gender, relationships, occupations, etc.; Fig1. widget 3),
- *Geographical engagement*, to study gaps between offline statistics about issues and corresponding online engagement in campaigns (e.g., via correlation by states using OpenGov Data; Fig1. widget 5 & 6), and
- *Content Practices*, to study diverse narratives of information sources (e.g., via summarized topics on pre-classified tweets and news for subjectivity, Fig1. widgets 4 & 7).

Event Scenarios and Conclusion

The proposed system has analyzed several humanitarian events in the year 2016, especially gender-based violence (GBV) and global displacements (GD), in collaboration with two non-profit organizations. For the live demo, we will present analyses of more than 30 million tweets, and news for GBV and GD with a focus on engagement by information source types for anti-GBV campaigns and displacement, as well as narratives across media.

Authors would like to thank the collaborators at Civic Nation and Internal Displacement Monitoring Centre for guiding the system design for the humanitarian domain.

References

Cohen, W. M., and Levinthal, D. A. 1990. Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly* 128–152.

Kovács, G.; Tatham, P.; and Larson, P. D. 2012. What skills are needed to be a humanitarian logistician? *Journal of Business Logistics* 33(3):245–258.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2. Citeseer.

Meier, P. 2015. *Digital humanitarians: how big data is changing the face of humanitarian response*. Crc Press.

Rogstadius, J.; Vukovic, M.; Teixeira, C.; Kostakos, V.; Karapanos, E.; and Laredo, J. A. 2013. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development* 57(5):4–1.