IBM System G SOCIAL MEDIA SOLUTION: ANALYZE MULTIMEDIA CONTENT, PEOPLE, AND NETWORK DYNAMICS IN CONTEXT

Ching-Yung Lin, Danny Yeh, Nan Cao, Jui-Hsin Lai, Chun-Fu (Richard) Chen,
Conglei Shi, Jie Lu, Jason Crawford, Keith Houck, Yinglong Xia, Sabrina Lin,
Richard B. Hull, Fenno F. Heath III, Piyawadee Sukaviriya, SweeFen Goh*

IBM T.J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598
{chingyung, dlyeh, nancao, larrylai, cfchen, conglei.shi,
jielu, ccjason, khouck, yxia, sabrinal,
hull, theath, noi, sweefen}@us.ibm.com

ABSTRACT

We present IBM System G Social Media Solution, which includes a suite of applications designed for in-context monitoring, exploration, and analysis of social multimedia content as well as related people and network dynamics. Each individual application focuses on a unique aspect of social media data analysis in relevant context; collectively, they provide a comprehensive set of tools for exploring and analyzing real-time and historical social media data at large scale. The solution is empowered by a unified data management platform, based on a property graph model, to efficiently handle a large variety of social media applications.

Index Terms— Social media, graph, network, analysis, multimedia, contextual

1. INTRODUCTION

Online social media have become prevalent platforms for the masses to interact and disseminate a large amount of information (e.g., memes, opinions, rumors, etc.). Information that may potentially induce adverse outcomes for organizations, including governments and corporations, can occur anywhere at any time and spread much faster than traditional media. Distilling useful information in a timely fashion from the overwhelming amount of social multimedia content is an important but challenging task. However, what is more important and challenging is the ability to understand relevant information in dynamic context, especially the interrelationships among information content, people, and network dynamics.

IBM System G Social Media Solution aims to help analysts monitor and explore dynamic social media data and conduct in-context analysis with a suite of applications that analyze and present data in various contexts. The solution offers a variety of analysis-rich information, including sentiment mined from both text and multimedia, emotions, personality traits, influence of individuals, impact of conversations, connections between tweets, users, hashtags, etc., as well as flows and spread of information. All such information greatly facilitates analysts in discovering, understanding, and tracking dynamic social media movements in rich contexts.

2. SYSTEM OVERVIEW

2.1. System Architecture

Social media data are inherently linked and form large heterogeneous graphs. Analysis of social media data often takes into consideration both different types of entities and the relationships between them. Therefore, it is not only natural but also beneficial to apply graph-based representations and technologies to social media data analysis. IBM System G Social Media Solution is built on top of *IBM System G Graph Computing Platform*, which provides a comprehensive software stack for Big Data Analytics. Fig. 1 illustrates the system architecture. The server side handles data collection, storage, retrieval, and analysis. Particularly, the *Database* layer organizes data of various types using the property graph model, consisting of a graph structure with vertices and edges, and the attributes associated with each vertex and edge, a.k.a. graph properties [1]. The *Middleware* layer includes several runtime libraries specifically designed for property graph computations to provide graph computing

This material is based upon work partially supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under Agreement Number W911NF-12-C-0028. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

*Chun-Fu Chen is affiliated with Department of Electrical Engineering, National Cheng Kung University. He performed the work while as a visiting scholar at IBM Research.

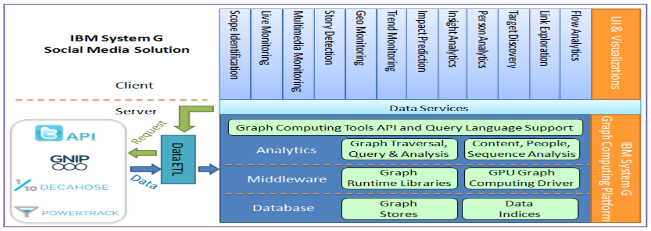
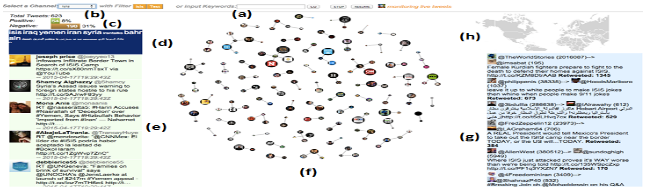


Fig. 1. System architecture diagram.



joseph price @joeyyeo13
Infowars Infiltrate Border Town in Search of ISIS Camp https://t.co/sX80nmTsxT via @YouTube
— 2015-04-17T19:29:43Z

Shamcy Alghazzy @Shamcy
Syria's Assad issues warning to foreign states hostile to his rule http://t.co/5AJrwF83yy
— 2015-04-17T19:29:42Z

Mona Anis @monaanis
RT @nasseratta5: #Hariri Accuses #Nasrallah of "Deception" over #Yemen, Says #Hizbullah Behavior "Imported from #Iran" — Naharnet http://t....
— 2015-04-17T19:29:42Z

#AbajoLaTirania. @TrancayHuye
RT @mendoszita: "@CNNMex: El líder de #ISIS podría haber aceptado la lealtad de #BokoHaram http://t.co/1ZgWvp7ZnC"
— 2015-04-17T19:29:42Z

debbierice55 @debbierice55
RT @UNGeneva: "Families on brink of survival" says @UNOCHA's @JensLaerke at launch of $247m #Yemen appeal - http://t.co/loz7mTH6e4 http://t...

@TheWorldStories (2016087)-->
@imsabat (195)
Female Kurdish fighters prepare to fight to the death to defend their homes against ISIS. http://t.co/KZM8DtrAAB **Retweeted: 1345**

@philippenis (38335)--> @HoodsMarlboro (1037)
leave it up to white people to make ISIS jokes then whine when people make 9/11 jokes **Retweeted: 673**

@3bdulla (286638)--> @lAlrawahy (612)
[illegible] Hobart Airport [illegible] http://t.co/t5dLHvq7cx **Retweeted: 529**

@FredZeppelin12 (23973)-->
@LAGraham64 (706)
A REAL President would tell Mexico's President to take out the ISIS camp near the border TODAY, or the US will....TODAY. **Retweeted: 384**

@AllenWest (380512)--> @sundoghigh (5949)
Where ISIS just attacked proves it's WAY worse than we're being told http://t.co/135WSpcZxp http://t.co/PF1q3YXZN7 **Retweeted: 170**

@4FreedominIran (3409)-->
@ShahnazP40 (532)
#Breaking Join ch.@Mohaddessin on his Q&A

Fig. 2. Live monitoring.

primitives. These primitives serve as building blocks for constructing various high-level graph analytics. The *Analytics* layer provides tools to traverse graph, retrieve relevant data, and conduct graph and data analytics. The *Data ETL* (extract, transform, load) module connects to the data sources to collect raw data, extract entities and relationships between entities, transform them into graph representations (vertices and edges with properties), and load into the graph-based data stores and indices, all in a dynamic and continuous fashion. The client side provides interactive Web-based user interfaces to gather information requests from users, interact with the backend through a *Data Services* module, and visualize the results returned by the module. Users can switch between different applications by using the uniform navigation bar at the top of all interfaces, or simply by following certain links provided in one application's interface for invoking other related applications. For example, the user profile images displayed in all applications are linked to the application that provides person analytics of the corresponding users.

2.2. Data Acquisition and Scope Identification

Currently IBM System G Social Media Solution focuses on Twitter data. The system can consume tweets through multiple means such as Twitter public API, and GNIP Decahose and PowerTrack subscriptions. The administrator sets up via an admin console channels for data collection. Each data channel corresponds to a topic of interest, which can be broad or narrow depending on the Twitter queries/filters/rules specified for the channel. Users can further define scopes of their interests within a data channel by creating filters through the ***Scope Identification*** application. Each filter consists of one or more terms ANDed or ORed together. An interactive visualization showing terms related via co-occurrences in tweets is provided to help users choose terms for defining filters. The application also includes an advanced setting which allows users to explore dynamically recommended related terms and create composite filters by grouping multiple filters. During monitoring, exploration and analysis, users can switch between data channels freely and apply any filter defined for the selected data channel to further constrain data in this channel.

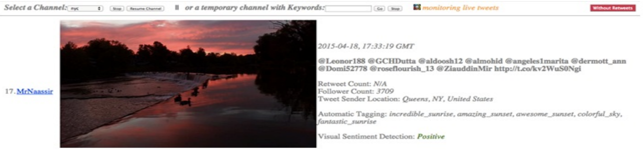
2.3. Applications

IBM System G Social Media Solution includes three sets of applications. ***Content Monitoring and Analysis*** applications focus on content and aim to answer questions such as what text/images/hashtags are tweeted, how popular they are, and what are the sentiment and impact of individual tweets or collections of tweets. ***People Analysis*** applications focus on analyzing emotions, personalities, trust and behaviors of people base on their tweeting activities. These applications help analysts to understand better the individuals in social movements and gain insights into the driving factors of various phenomenons. ***Network Exploration and Analysis*** applications support exploration and analysis of links and dynamic information flows in the networks, to address questions such as how tweets are propagated in the networks through people over time, and whether there are any anomalies during information dissemination (e.g. rumor spreading).

3. CONTENT MONITORING AND ANALYSIS

The ***Live Monitoring*** application provides real-time monitoring of tweets (including retweets) that are relevant to user interests. Fig. 2 displays a screenshot of the application's user interface, which contains multiple UI components. The input components (a) and (b) allow a user to input the keywords or select the data channel for which s/he wants to monitor. The statistics components (c) and (d) display aggregate statistics calculated from the tweets since the start time of the monitoring, including sentiment information based on tweet text and a word cloud. The list of current tweets (e) is dynamically updated as tweets come in, with the most recent tweets displayed at the top. The retweet graph (f) visualizes the retweeting relationships (edges) between users (nodes). The size of each node corresponds to the number of followers. Hovering over a node shows the ID and profile image of the corresponding user. Hovering over an edge shows the content of the associated retweet. Details of the retweets are also provided in (g). The map (h) indicates the location information of the tweets whenever available.

The ***Multimedia Monitoring*** application displays real-time tweets containing images along with automatically calculated visual sentiment to assist analysts in understanding social movements from a visual perspective. The visual sentiment of an image is determined by the sentiment of the Adjective and Noun Pairs (ANPs), such as happy dog, horizontal text, etc., that are automatically generated to describe the image. A prediction model, established using a deep learn-



17. MrNaassir

2015-04-18, 17:33:19 GMT

@Leonor188 @GCHDutta @aldoosh12 @almohid @angeles1marita @dermott_ann @Domi52778 @roseflourish_13 @ZiauddinMir http://t.co/kv2WuS0Ngi

Retweet Count: *N/A*
Follower Count: *3709*
Tweet Sender Location: *Queens, NY, United States*

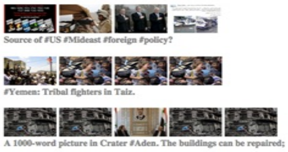
Automatic Tagging: *incredible_sunrise, amazing_sunset, awesome_sunset, colorful_sky, fantastic_sunrise*

Visual Sentiment Detection: *Positive*

Fig. 3. Multimedia monitoring.

ing framework that correlates thousands of ANPs with image features, is used to generate a list of ANPs given the features extracted from the image [2]. Fig. 3 displays a screenshot of the application's interface, which contains an image of beautiful sunset. The automatically generated ANPs are *incredible sunset*, *amazing sunset*, *awesome sunset*, *colorful sky*, *fantastic sunrise*, which are close to the human's descriptions of the image. Since the sentiment associated with these ANPs is positive, the visual sentiment of the image is set to positive.

The ***Story Detection*** application (Fig. 4) groups tweets containing images into "stories" to help analysts quickly get a sense of rapidly developing storylines in social media. Each group contains tweets with similar images. The similarities between the images are determined based on the similarities between the ANPs associated with these images. This approach enables images that are close to one another at the semantic level to be grouped together, even though they may not be similar in low-level image features.


Source of #US #Mideast #foreign #policy?

#Yemen: Tribal fighters in Taiz.

A 1000-word picture in Crater #Aden. The buildings can be repaired;

Fig. 4. Story detection

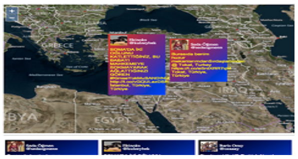
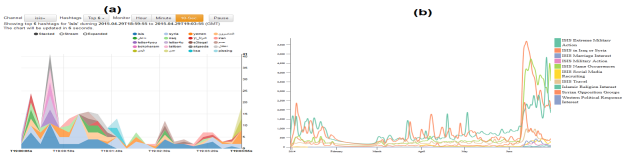


Fig. 5. Geo monitoring

The ***Geo Monitoring*** application visualizes tweets where they happen, as they happen on a world map. As shown in Fig. 5, users can pan and zoom in to get a high-resolution view of any location in the world and monitor real-time tweets from that location. These tweets are also displayed at the bottom of the interface, together with the profile images of the authors. Clicking on a profile image goes to the Person Analytics application (Sec. 4) for an in-depth analysis of the corresponding user.

The ***Trend Monitoring*** application (Fig. 6) provides timeline views of the popularity of hashtags (a) or topics (b) relevant to a given data channel. Users can interact with chart legends to hide/show particular hashtags or topics, or mouse over to get detailed count information at each time slice. The unit of the time dimension can be switched between 10 seconds, 1 minute and 1 hour for hashtag monitoring. Topic monitoring is provided over a longer time span of several months.



Showing top 6 hashtags for 'isis' during 2015-04-29T18:59:55 to 2015-04-29T19:03:55 (GMT)
The chart will be updated in 6 seconds.

Fig. 6. Trend monitoring.

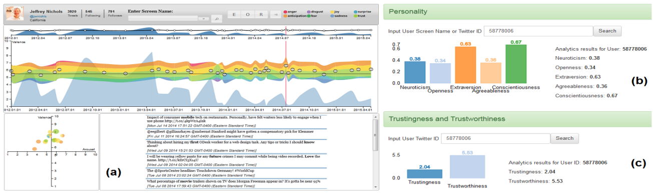
The ***Insight Analytics*** application applies analysis to collections of tweets defined based on time, location, or topic. The interface provides an aggregate view of several statistics and analysis results (e.g. tweet count, collective sentiment and emotion, etc.) for each tweet collection. This application can serve as a "social thermometer" to gauge reactions from the public during a particular time period, at a particular location, or about a particular topic.

Information disseminated via social media may have significant business impact. The ***Impact Prediction*** application aims to dynamically capture and analyze "virtual social conversations" formed around various topics, and predict their potential impact to the business that may be affected. Since Twitter users often use hashtags to participate in particular social conversations, the application extracts virtual social conversations by grouping tweets around common hashtags. All tweets within a single conversation are analyzed to extract a set of features based on tweet content (e.g. percentage of keyword coverage), author information (e.g. number of identified influencers), and other metadata (e.g. location, language). Then a regression-based prediction model is applied to these features to calculate an impact score of the conversation. The prediction model is created with the help of domain knowledge provided by subject matter experts, and can be dynamically updated given user feedback.

4. PEOPLE ANALYSIS

The ***Person Analytics*** application conducts multi-dimensional emotion analysis, personality analysis, and trust analysis of a given Twitter user based on his/her tweets. Emotion analysis detects the user's expressed emotions at different time points and summarize those emotions to reveal the user's emotional style [3]. Personality analysis focuses on the Big 5 personality traits of the user. Trust analysis calculates the user's trustingness and trustworthiness by looking at his/her interactions with others via tweets [4]. The interface (Fig. 7) provides an interactive visualization of the emotion analysis result (a) to create a visual emotional profile of the target user, and displays personality (b) and trust scores (c).

The ***Target Discovery*** application analyzes users and detects outliers based on the unsupervised Time-Adaptive Local Outlier Factor model [5] and visualizes them in context through novel visualizations and multiple coordinated contextual views (Fig. 8). Particularly, it uses ego-centric glyphs to visually summarize a user's behavior and effectively present the user's activities, features, and interactions in social media. The glyphs are placed on a triangle grid to capture similarities



Personality

Input User Screen Name or Twitter ID 58778006 Search

Analytics results for User: **58778006**
Neuroticism: **0.38**
Openness: **0.34**
Extraversion: **0.63**
Agreeableness: **0.36**
Conscientiousness: **0.67**

Trustingness and Trustworthiness

Input User Twitter ID 58778006 Search

Analytics results for User ID: **58778006**
Trustingness: **2.04**
Trustworthiness: **5.53**

Fig. 7. Person analytics.

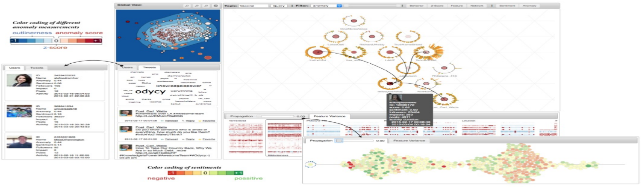

Color coding of different anomaly measurements
outlierness anomaly score
z-score
Color coding of sentiments
negative
positive

Fig. 8. Target discovery.

among users and facilitate comparisons of behaviors among different users. This application provides a powerful visual analytic tool for analysts to easily inspect behaviors of one or more users from different perspectives, which facilitates identification of target users for a variety of purposes such as marketing and detecting anomalous users or social bots.

5. NETWORK EXPLORATION AND ANALYSIS

The ***Link Exploration*** application allows analysts to easily explore and discover both direct and indirect connections between various entities (e.g. tweets, users, hashtags) in the heterogeneous graph representation of Twitter data. A standard node-link visualization is used to display a sub-graph retrieved based on a user query specified via the interactive query panel. For example, Fig. 9 shows the 2-hop ego network of a specific hashtag (a) and the 2-hop ego network of a specific user (b), both of which link together tweet, user, hashtag, image, and time nodes via create, retweet, mention, reply, or contain relationships. Hovering over a node displays more detailed information of this node, left-clicking on a node retrieves a new sub-graph using the selected node to query, and right-clicking on a user node invokes person analytics.

Analysis of retweeting activities can reveal not only how information spreads but also malicious social campaigns against certain entities. The ***Flow Analytics*** application focuses on analyzing retweet sequences to detect anomalous ones which may indicate rumors or other malicious actions.

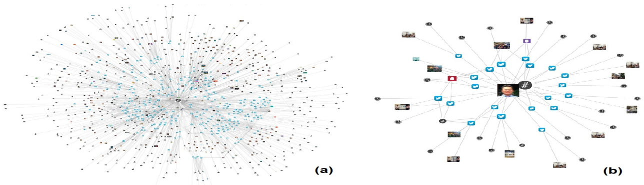


Fig. 9. Link exploration.

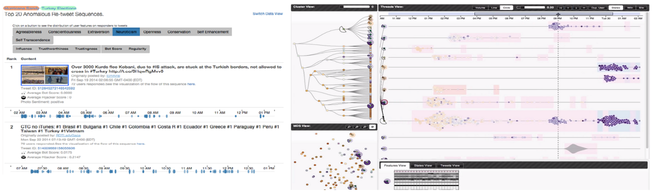


Fig. 10. Flow analytics.

Given a retweet sequence, the application evaluates how anomalous the sequence is using One-Class Conditional Random Fields [6]. The user interface provides two different views with interactive visualizations to present top anomalous retweet sequences in rich context, which allows analysts to easily explore, understand, and validate analysis results [7].

6. SUMMARY

IBM System G Social Media Solution offers a unified platform for conducting various social multimedia analyses in context in an efficient and productive way. This paper showcases the tools and interfaces included in the solution.

7. REFERENCES

[1] Y Xia, I Tanase, L Nai, W Tan, Y Liu, J Crawford, and CY Lin, "Explore efficient data organization for large scale graph analytics and storage," in *IEEE Big Data*, 2014, pp. 1–8.

[2] T Chen, D Borth, T Darrell, and SF Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *CoRR*, vol. abs/1410.8586, 2014.

[3] J Zhao, L Gou, F Wang, and M Zhou, "PEARL: An interactive visual analytic tool for understanding personal emotion style derived from social media," in *IEEE VAST*, 2014.

[4] H Borbora, M Ahmad, K Haigh, J Srivastava, and Z Wen, "Robust features of trust in social networks," *Social Netw. Analys. Mining*, vol. 3, no. 4, pp. 981–999, 2013.

[5] M Breunig, HP Kriegel, R Ng, and J Sander, "LOF: Identifying density-based local outliers," in *ACM SIGMOD*, 2000, vol. 29, pp. 93–104.

[6] Y Song, Z Wen, CY Lin, and R Davis, "One-class conditional random fields for sequential anomaly detection," in *IJCAI*, 2013, pp. 1685–1691.

[7] J Zhao, N Cao, Z Wen, Y Song, Y Lin, and C Collins, "FluxFlow:visual analysis of anomalous information spreading on social media," in *IEEE VAST*, 2014.