The Multimedia Challenges in Social Media Analytics

Tat-Seng Chua
School of Computing, National University of Singapore
chuats@comp.nus.edu.sg

ABSTRACT

With the popularity and wide acceptance of social networks, users are now sharing information on multiple aspects of their life and on a wide range of social networks. In the meantime, there is a huge amount of situational information generated by sensor devices, often as part of human activities. Thus for any given entity, we can now find a wide variety of social, device and structured information from multiple sources. The generation of reliable social media analytics with respect to any entity is hence a highly challenging (multimedia) task. The key challenges include the ability to: (a) gather "representative" data about an entity from multiple sources; (b) handle the increasing amount of non-textual media content; (c) detect and track sub-topics around the target entity, along with deep analysis tools such as named-entity extraction, visual concept detection and sentiment analysis; and (d) generate predictive and prescriptive analytics. This talk describes a live social observatory system that we have developed and our research efforts to tackle the above challenges. In particular, we outline our research to transform unstructured live social media streams into descriptive, predictive and prescriptive analytics.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - social media content analysis and fusion; H.3.4 [**Information Storage and Retrieval**]: Systems and Software - social observatory system; J.4 [**Computer Applications**]: Social and Behavioral Sciences - social media analytics

General Terms

Theory

Keywords

Social Media Analytics; Live Social Observatory

1. THE RICH SOCIAL MEDIA ENVIRONMENT

We now live in a rich social media environment. Many of us freely and spontaneously generate and share contents of various types as part of our daily activities, including making comments, sharing photos, checking-in to venues, asking and answering questions. This is done through a wide variety of social networks. As a result, a huge amount of real-time social media data is being generated. The data is collectively known as the User-Generated Content (UGC). The UGCs reflect the pulse of society and the tone of public opinion, and help us to better understand the world around us [1]. There is thus a tremendous need to analyze these contents to offer better understanding of the state of our society and the people living within it.

Given the vast quantity of live social media streams and their impact on society, many research groups and organizations are carrying out projects to collect and analyze live UGC streams to support different applications. One key task is the generation of social media analytics with respect to a target entity, where the entity can be an organization, an event, a product, a person or a location. The reliable analysis of social media content with respect to a target entity faces a number of key challenges.

The first and most fundamental challenge is the ability to gather "representative" data about an entity from multiple sources. The gathering strategy must follow the typical usage model of users. In such model, a user may actively discuss about an entity for a while with continuous vocabulary change that often includes multimedia content. Hence a multifaceted strategy must be devised to gather data based on fixed and evolving set of keywords, representative image content, and key users. The next challenge is that the social media content is becoming increasingly multimedia often with little correlated text. This is especially true for product and brand images. Hence there is a need for deep visual analysis to infer the semantics of images towards a more complete social media analysis. The third challenge is the detection and tracking of sub-topics around the target entity, along with deep content analysis to extract named-entities appearing in the text, visual concepts present in images and sentiment analysis. We need to develop both incremental and adaptive learning techniques to identify new and track evolving sub-topics. Finally, for most entities, people would like to know what social media posts or sub-topics related to the entities are likely to become viral, and what actions they can take to address the problem. Overall, from social media analysis, we want to generate relationships a-

SAM'14, November 7, 2014, Orlando, FL, USA.
ACM 978-1-4503-3124-1/14/11.
http://dx.doi.org/10.1145/2661126.2661131.

mong/between entities/users, comparative analysis between related entities/products, and reports.

2. THE LIVE SOCIAL OBSERVATORY SYSTEM

To realize the above aim, we propose a live social observatory system named '*NExT-Live*' to mine multiple social media streams automatically [2]. The system adopts a distributed architecture. It deploys an efficient and robust set of crawlers to continually crawl online social interactions on multiple facets from various social network sites. It tracks various types of social media sites where information is of public natures, including the microblog sites such as Twitters, various blogs and forums sites, location sharing sites such as the 4Sqaure, and image/video sharing sites such as the Instagram, Flicker and YouTube, as well as the equivalent sites in China. The data crawled are stored and processed in a distributed Hadoop architecture. The system first analyzes each social media post to extract high-level attributes such as the named entities and sentiments, and then analyzes the social media streams jointly to generate high-level analytics. It also generates predictive and prescriptive analytics. This talk will describe the architecture of *NExT-Live*, and discuss the technical details and related research efforts towards the generation of various high-level analytics. It will also discuss various applications of *NExT-Live* to generate analytics for *India Election* and *Comparative Review of Products*.

3. SUMMARY

The live social observatory system will be used as the basis for education and research. It will also be used as platform to support collaborative social observatory systems, and to support deep analytics and privacy preserving research. In addition to UGC, many sensor devices are also transmitting various types of signals continually. These devices include the mobile phones and health/sports sensors that the users wear, security video sensors and various environmental sensors. They are collectively known as the Device Generated Content (DGC). The integration of UGC and DGC, as well as the structured data, will offer us a rather complete set of information of the world around us, as well as new challenges in making sense of these data sources.

4. ACKNOWLEDGMENTS

NExT-Live system is developed by NExT Search Center[1] [2]. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This work is also supported by the National Natural Science Foundation of China (No.61303075).

5. REFERENCES

[1] T.-S. Chua, H.-B. Luan, M. Sun, and S. Yang. Next: Nus-tsinghua center for extreme search of user-generated content. In *IEEE Multimedia*, 2012.

[2] H.-B. Luan, J. Li, M. Sun, and T.-S. Chua. The design of a live social observatory system. In *WWW (Companion Volume)*, 2014.

[1] http://next.comp.nus.edu.sg/