NExT-Live: A Live Observatory on Social Media

Huanbo Luan[1,2], Dejun Hou[1], and Tat-Seng Chua[1]

[1] School of Computing, National University of Singapore, Singapore
[2] Department of Computer Science and Technology, Tsinghua University, China
luanhuanbo@gmail.com, {houdj,chuats}@comp.nus.edu.sg

Abstract. This demonstration presents a live observatory system named '*NExT-Live*'. It aims to analyze live online social media data to mine social phenomena, senses, influences and geographic trends dynamically. It builds an efficient and robust set of crawlers to continually crawl online social interactions on various social networking sites, covering contents from different facets and in different medium types. It then performs analysis to fuse these social media data to generate analytics at different levels. In particular, it researches into high-level analytics to mine senses of different target entitles, including People Sense, Location Sense, Topic Sense and Organization Sense. *NExT-Live* provides a live observatory platform that enables people to know the happenings of the place in order to lead better life.

Keywords: Live, Observatory, Monitoring, Social Media, UGC, NExT.

1 Introduction

We are living in the midst of a rich social media environment. We freely and spontaneously generate contents as part of our daily activities including making comments, sharing photos, checking-in to locations, asking and answering questions. Through the wide variety of social media channels, more and more such real-time social media data, collectively known as User-Generated Content (UGC), are being generated. The contents of UGC reflect the pulse of a society and the tone of public opinion, and affect our culture and the way we communicate. Aiming to better understand and analyze live social interactions[1], social media observation and long-term digital preservation have become highly relevant and urgent. Thus there is a strong need for live data crawling, archiving, access, retrieve and analysis.

Web science research community has recently proposed the creation of a global "Web Observatory Community Group" to establish a global open data resource collaboratively by many web observatory nodes across the world [2]. On the other hand, there are some commercial social media monitoring tools and platforms that claim to be able to help track and monitor business or brand in social media channels such as Radian6, BuzzLogic, Visible Technologies, Brandwatch, Brandtology. Although some such tools show good marketing performance, they usually suffer from the problems of narrow application domain, limited data coverage and data types, in which most focus primarily on twitter data. Moreover, most such tools are not fully automated and cannot handle live data well.

S. Li et al. (Eds.): MMM 2013, Part II, LNCS 7733, pp. 514–516, 2013.

To address the above problems, we propose a livesocial observatory system named '*NExT-Live*' to mine multiple social channels automatically. It can continually crawl live and semi-structured multimedia UGC data, including text, images, videos and user-relation graphs etc. It supports real-time analysis and fusion of these data sources to generate multiple social analytics, including People Sense, Location Sense, Topic Sense and Organization Sense.

2 System Architecture and Implementation

The overall system framework of *NExT-Live* is illustrated in Fig.1. The system comprises three layers: *Live Data Crawlers, Big Data Management* and *Multi-phase Observatory*. *NExT-Live* currently runs on a cluster with 17 server nodes within the NUS campus.

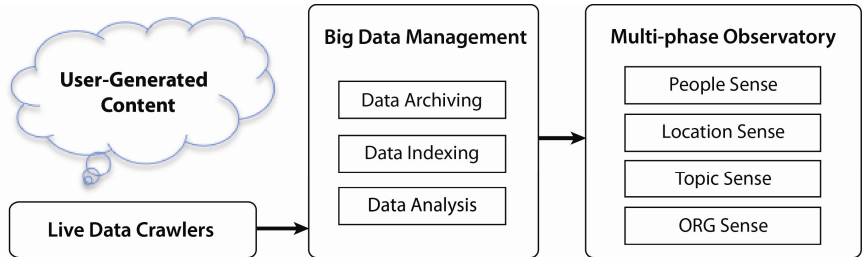

Fig. 1. The overall system architecture of *NExT-Live*

2.1 Live Data Crawler

NExT-Live tracks multiple social networking sites including *Flickr, Foursquare, Instagram, Panoramio, TecentWeibo, SinaWeibo, Twitter, Youtube, Amazon, Dianping, Fantong, as well as* some forum and blog sites. It provides the best real-time coverage of multi-modality UGC such as text posts, user comments, images, videos, user profiles and user relations. In order to ensure continual real-time crawling, we build a set of live robust crawlers that works well across different platforms, channels, and is easy to maintain and extend. The crawlers are made intelligent and robust by supporting IP proxy, heuristically crawling, noise filtering, exception handling, as well as multiple threads and distributed crawling. Table 1 presents a glimpse of the size and variety of live UGC data that we have crawled over a 4½ month period.

Table 1. The crawled User-Generated Content and their sizes (1 May 2012 - 15 Sep 2012)

Data Types	Number of Posts	Size
Micro-blog Posts	1,402,948,496	988 GB
User Comments	175,732,324	178 GB
User Profiles	132,715,138	129 GB
Images	242,913,348	21 TB
Videos	92,277	3.2 TB

2.2 Big Data Management

The live data steam is sent to *Big Data Management* module to perform:

Data Archiving: It utilizes MangoDB and NFS to store text and media data in distributed servers. MongoDB stores JSON-like documents with dynamic schemas and shows good scalability and agility in handing huge data set.

Data Indexing: It then triggers indexing function automatically to build the distributed index in real time for data access and retrieval. The text index is created with SOLR and visual index is generated with hashing and inverted files based on the extracted visual features.

Data Analysis: It carries out the analysis and fusion of multiple UGC sources to generate higher order analytics.

2.3 Multi-phase Observatory

NExT-Live offers multi-phase observatory that helps users better understand the trends and pulses of a society. It builds tools to perform content analysis, data fusion, topic mining, user community discovery, sentiment analysis, as well as the integration of multiple social signals to track and mine events and senses in society. In particular, given a target topic, it mines the evolution of relevant contents, user community and events and integrates them to infer the sense of the entity. The entity can be a person, location, topic or an organization, thus giving rise to observatory for people, location, topic and organization senses. For example, the "Organization Sense Observatory" will analyze relevant UGCs to uncover both emerging and hot events, as well as user community and key users, related to the target organization; while the "People Sense Observatory" will return and analyze what other people post and say about the target person. Collectively, it provides valuable observatories to help us better understand ourselves and the larger environment that we live in.

Acknowledgement. *NExT-Live* system is developed by NExT Search Center [3], which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

References

1. Cui, P., Wang, F., Liu, S.W., Ou, M.D., Yang, S.Q.: Who Should Share What? Item-level Social Influence Prediction for Users and Posts Ranking. In: International ACM SIGIR Conference (2011)
2. http://www.w3.org/community/webobservatory/
3. Chua, T.S., Luan, H.B., Sun, M.S., Yang, S.Q.: NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content. IEEE Multimedia 19(3), 81–87 (2012)