The Design of a Live Social Observatory System

Huanbo Luan[1,2], Juanzi Li[2], Maosong Sun[2], Tat-Seng Chua[1]

[1]School of Computing, National University of Singapore
[2]Department of Computer Science and Technology, Tsinghua University

luanhuanbo@gmail.com, chuats@comp.nus.edu.sg, {lijuanzi, sms}@tsinghua.edu.cn

ABSTRACT

With the emergence of social networks and their potential impact on society, many research groups and originations are collecting huge amount of social media data from various sites to serve different applications. These systems offer insights on different facets of society at different moments of time. Collectively they are known as social observatory systems. This paper describes the architecture and implementation of a live social observatory system named '*NExT-Live'*. It aims to analyze the live online social media data streams to mine social senses, phenomena, influences and geographical trends dynamically. It incorporates an efficient and robust set of crawlers to continually crawl online social interactions on various social network sites. The data crawled are stored and processed in a distributed Hadoop architecture. It then performs the analysis on these social media streams jointly to generate analytics at different levels. In particular, it generates high-level analytics about the sense of different target entitles, including People, Locations, Topics and Organizations. *NExT-Live* offers a live observatory platform that enables people to know the happenings of the place in order to lead better life.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing, H.4.0 [**Information Systems Applications**]: General

Keywords

Live, Observatory, Monitoring, Social Media, UGC, NExT

1. INTRODUCTION

We are living in the midst of 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, more and more such real-time social media data are being generated. These data are collectively known as the User-Generated Content (UGC). The contents of UGC reflect the pulse of society and the tone of public opinion, and help us to better understand the world around us. There is thus a tremendous need to analyze the UGC to offer better understanding of the state of our society and the people within it [3].

Given the vast quantity of live social media streams and their impact on society, many research groups and organizations are launching projects to collect and analyze live UGC streams to serve different applications. These systems offer different facets of activities of society at different moments in time. Such systems are termed "Social Observatory Systems", a term coined by a global "Web Observatory Community Group" under W3C. The group aims to establish a global open data resource collaboratively by many web observatory nodes across the world [9]. In the meantime, there are many commercial social media monitoring tools and platforms that claim to be able to help track and monitor business or brand in UGCs such as Radian6, BuzzLogic, Visible Technologies, Brandwatch and Brandtology. Although some such tools show good marketing potentials, they typically suffer from the problems of narrow application domain, limited data coverage and data types. Moreover, they tend to focus primarily on twitter data, not fully automated and cannot handle live data well.

To address the above problems, we propose a live social observatory system named '*NExT-Live'* to mine multiple social media streams automatically [3, 8]. The system adopts a distributed architecture. It deploys an efficient and robust set of crawlers to continually crawl online social interactions on various social network sites. It tracks various types of social media sites where information are of public natures, including the microblog site 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. 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. In particular, it generates high-level analytics about the sense of different target entitles, including People, Locations, Topics and Organizations. This paper describes the architecture of *NExT-Live* and discusses the technical details towards the generation of various high-level analytics.

2. SYSTEM ARCHITECTURE

The *NExT-Live* is designed to be an integrated platform that gathers, analyses and organizes live UGCs from a wide variety of social media sites. The system architecture to support this overall task is shown in Figure 1. The architecture comprises 7 layers: Live Intelligent Crawlers; Data Filtering; Data Storage; Data Analysis & Processing; Cross Indexing; Multiple Visualization;

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579214

and, Social Observatory Application. *NExT-Live* currently runs on a cluster with 17 servers (80 VM nodes) within the NUS (National University of Singapore) campus. It has been running well since May 2012 [8].

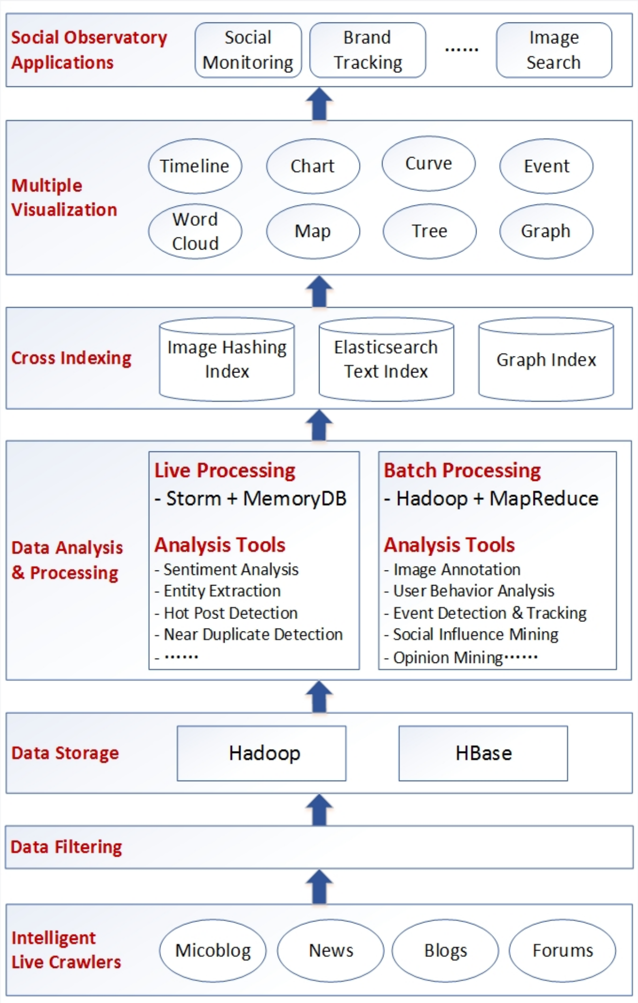
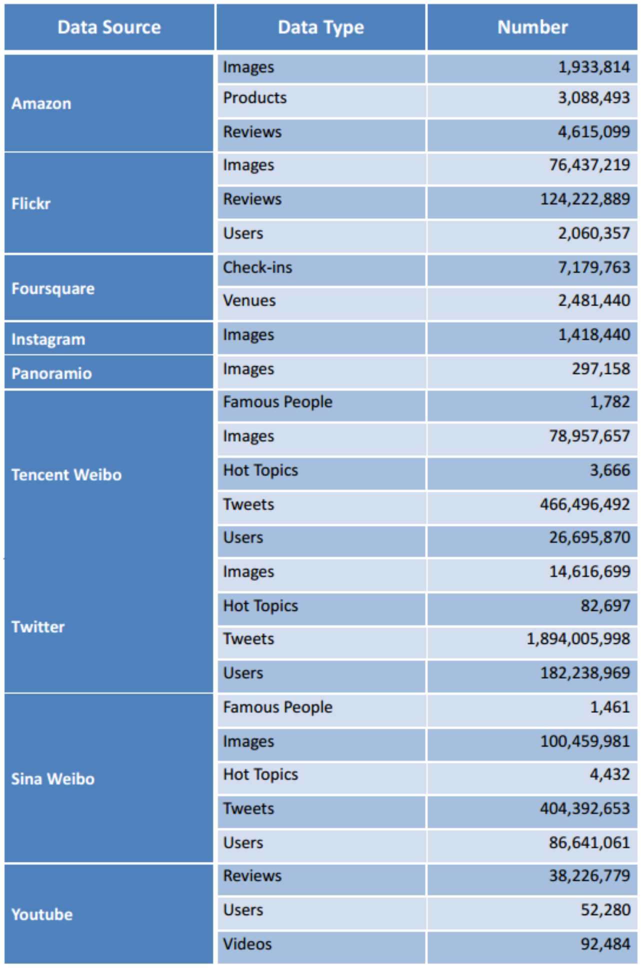


Figure 1. The overall system architecture of *NExT-Live*.

2.1 Intelligent Live Crawlers

NExT-Live tracks multiple social networking sites including *Flickr, Foursquare, Instagram, Panoramio, Tecent Weibo, Sina Weibo, Twitter, Youtube, Amazon,* as well as some forum and blog sites. It aims to offer real-time coverage of a variety 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. To ensure robustness in crawling, we support IP proxy, heuristic crawling, noise filtering, exception handling, as well as the ability to perform multiple threads and distributed crawling. Figure 2 presents a glimpse of the data sources that we have crawled and their size as of 12 Jan 2014. The total number of data records is over 2.8 billion, with close to 275 million images.



Data Source	Data Type	Number
Amazon	Images	1,933,814
	Products	3,088,493
	Reviews	4,615,099
Flickr	Images	76,437,219
	Reviews	124,222,889
	Users	2,060,357
Foursquare	Check-ins	7,179,763
	Venues	2,481,440
Instagram	Images	1,418,440
Panoramio	Images	297,158
Tencent Weibo	Famous People	1,782
	Images	78,957,657
	Hot Topics	3,666
	Tweets	466,496,492
	Users	26,695,870
Twitter	Images	14,616,699
	Hot Topics	82,697
	Tweets	1,894,005,998
	Users	182,238,969
Sina Weibo	Famous People	1,461
	Images	100,459,981
	Hot Topics	4,432
	Tweets	404,392,653
	Users	86,641,061
Youtube	Reviews	38,226,779
	Users	52,280
	Videos	92,484

Figure 2. Monitored data sources and sizes (as of 12 Jan 2014)

2.2 Data Filtering

During data collection, a lot of noisy data will be crawled at the same time. These data consumes a lot of resources in terms of storage and processing time, and will affect the accuracy of subsequent analysis step. For efficiency reason, we remove these noisy data before storage and processing. To accomplish this, we train a series of filters/classifiers to remove the obviously irrelevant data and identify the duplicate data so that only a single copy is stored.

2.3 Data Storage

The crawled live data steams are sent to big data storage module for archiving and analysis. Our system utilizes Hadoop to store the unstructured data, and HBase for structured data. In particular, HBase is deployed to store JSON-like documents with dynamic schemas and has been shown to possess good scalability and agility in handing huge data set.

2.4 Data Analysis and Processing

Given the set of social media records gathered, this layer builds a set of tools to perform a range of analysis to extract high-level attributes at the message or post level. For live data streams, it performs the analysis in real-time by using MemoryDB based on Storm. Specifically, the analysis aims to extract the sentiment, and entities present within each social media message. The entity pre-defined here includes name of location, person and organization, etc. In addition, it also determines whether an incoming message is likely to become hot (or viral) and whether it is a duplicate of an existing message. The technique for detecting viral messages is presented in Section 3.

For historical data that has already been indexed in the database, the system performs batch analysis by using MapReduce based on Hadoop. The analysis performed here is deeper and more detailed than in the live processing case. In particular, the analysis aims to uncover annotation of images, user behavior, as well as social influence and opinion of users. These tasks require extensive analysis of a large set of data over a longer period of time.

2.5 Cross Indexing

The results of data analysis are stored in a number of distributed indices to facilitate real-time data access and retrieval. The text entities are indexed using Elasticsearch, while the media entities (images/videos) are stored in our own hash-based image index. We need to build our own image index in order to index the huge amount of image/video data that we have crawled, as well as to serve as the platform to perform large-scale media content analysis such as image annotation, product/object detection etc. Our image indexing system employs the highly discriminative spatial pyramid image features [7] for images, and generates the hash-codes from these features for indexing. The combined text and image indices have been found to be effective in tacking and analyzing multimedia brand/product related events from live social media streams [5].

Finally, we set up the graph index to encode the semantic social relations.

2.6 Multiple Visualization

The indexed entities can be visualized in multiple ways, including in the form of timeline, graph or map, as word cloud, chart, tree and graph, or as events. Often, the mode of visualization depends on the data types as well as the preference of the users.

2.7 Social Observatory Application

Based on the functionalities built up, a wide range of applications can be developed at the top layer. The types of applications we have built include: live event detection with respect to an entity [2]; brand and product tracking [5]; as well as user community discovery. In the remaining of this paper, we will describe a social observatory application that we have developed – to detect events for organizations, and to detect hot tweets/events.

3. OUTBREAK DETECTION

Given the public nature of social microblogging sites such as the Twitter and Weibo in China, they have been widely used by users to propagate timely information related to events, organizations and activities. Many such microblog message quickly become viral, reaching a huge number of users within a short span of time, and have great impact to the society, organizations and their activities. In social information science, this phenomenal is known as information cascade (or herding) which occurs when people observe the actions of others and make the same choice as the others have made. It is observed that only a tiny proportion of message on such microblogging sites will break out (i.e. affecting a large population of users in the social network), while the remaining will diminish before the critical point of outbreak. Hence how to predict these rare cascading outbreaks in early stage of a microblog message is of paramount importance for social marketing and rumor prevention etc. Cascade outbreak detection has therefore attracted a lot of research efforts such as an early representative work reported in [6].

Take Twitter as an example, after a user publishes a message, some of his/her followers (or friends) will forward this message to their followers, and this message may spread out over the social network to form an information cascade, and possibly break out if a certain cascade size is reached. During the whole process, the cascading behaviors (i.e. forwarding) of the involved users cause the outbreak of this message, and clearly the importance of these users are not the same in that some user's forwarding may bring more subsequent forwarding behaviors and thus has higher correlation with outbreaks. Hence the cascading outbreak prediction problem can be simplified as the problem of measuring the user’s importance with respect to information cascade in the early stage. A naive and intuitive solution would be to select the big users (e.g. celebrities) who have many followers, as is done in other studies. However, our empirical study suggests that these topological measures are not adequate.

Our preliminary study indicates that early cascade outbreak is dependent on the involvement of some keyusers. Hence we want to develop an efficient technique to predict cascade outbreaks that involves only a minimum number of strategic users and within the shortest possible time. To this end, we propose an Orthogonal Sparse Logistic Regression (OSLOR) method, as shown in Figure 3, to select the minimum number of selected users (known as the sensors) to predict the cascade outbreaks effectively [4]. We treat the cascade outbreak prediction as a binary classification problem of predicting a cascade as outbreak or non-outbreak. The problem is thus formulated with a sparse logistic regression model, which minimizes the prediction loss with a sparse linear model, in which only a small number of least redundant users (sensors) are active.

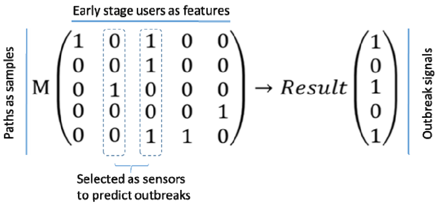


Figure 3. The proposed method for outbreak prediction. Note that the rows denote the cascades and columns indicate the user behavior. Only a limited number of users are selected as cascade sensors.

We evaluate the proposed method on a real online social network dataset collected from a Twitter style website in China. We gathered a total of 116.3 million users, 4.05 billion social relations, and 182.7 million cascades over the network during the period of 10-20 March 2011. Our experiments demonstrate that OSLOR could achieve a much higher prediction accuracy than the topological measure based methods and other feature selection based methods. Our results show that we can accurately predict over 70% of outbreaks by using only 500 sensors within 5 minutes after the occurrence of these cascades.

4. EVENT DETECTION

Besides facilitating communications among individuals, microblog services also explicitly or implicitly contain rich information about organizations, such as the banks, universities, and government organizations, etc. Many organizations are keen on continually mining and analyzing these user-generated social data to better understand the concerns of their users and to extract invaluable market insights. The primary foundation of these applications is based on topic monitoring and tracking. Specifically, organizations would like to: (1) track the evolution of any identified relevant topics about them; and (2) be informed of any new emerging topics which are fast gathering momentum in microblogs. This Section describes an application to detect event and predict hot emerging events for an organization [2]. The techniques described can be applied to other entities such as people, locations and topics, etc.

4.1 Intelligent Crawling Strategy

The first task in any event analysis is to collect a relatively complete and representative set of relevant data for the target organization. Such a task is often overwhelmed by the tremendous amount of relevant as well as irrelevant and missing data. To ensure comprehensive data collection, two interconnected observations can be made: (a) users related to organizations are more likely to post tweets related to the organization; and (b) tweets on organization often contain organization related keywords. These two observations enable us to generate descriptive keywords and cues, such as fixed keywords, dynamic keywords, known accounts, and organization keyusers. Accordingly, we design a comprehensive set of crawlers to comprehensively crawl organization relevant data from multiple aspects, as shown in Figure 4.

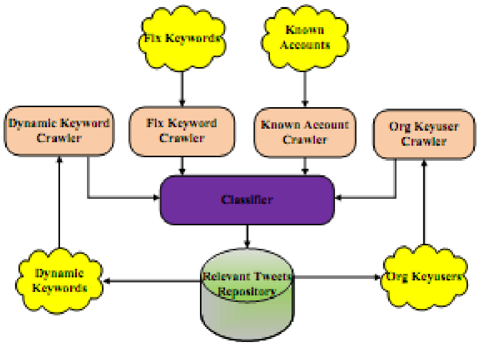


Figure 4: Crawling strategy to collect representative information about an orgaization

Given an organization, we first identify a set of organization related accounts. These are typically official accounts of the organization on the social media platform that often post relevant tweets about the target organization, such as news about the target organization. Such accounts can easily be discovered and a simple "discovery program" can be developed for this purpose. The text posted on these accounts as well as their followers are mostly relevant, and can be used as the initial list of relevant tweets and relevant users with respect to the organization. We employ the Known Account Crawler to monitor information on these accounts.

Next, we select a small set of fixed relevant keywords that can uniquely identify the organization to kick-start the crawling process. Examples of such keywords include the name of the organization, its brand names, and the name of its CEO, etc. These fixed keywords can be identified manually or automatically by mining the highly correlated and discriminative set of keywords from the initial set of relevant posts found. They are used in the streaming based Fixed Keyword Crawler.

As the microblog messages are conversational in nature and the terms used in these messages change continually over time, the use of a fixed set of keywords is inadequate. In order to elicit a live and more diverse set of relevant set of organization keywords, we extract a list of temporally relevant emerging terms about the organizations at each time point t. Emerging terms are defined as those newly introduced terms that are able to represent emerging topics about the organization. They are identified by finding those terms that are different from the existing set of terms in a previous time window. The emerging terms are used to collect more data using the Dynamic Keyword Crawlers.

The above mentioned three kinds of crawlers explicitly identify and collect data about the target organization. However, they overlook some important tweets posted by the users related to the organization. The tweets are relevant to the target organization in implicit form, i.e., they do not contain the fixed or emerging set of keywords related to the organization. To accomplish this, we also identify active users of the organization and their relationships at time t, and crawl all data sent by these users using the Org Keyuser Crawlers.

For the Dynamic and Keyuser Crawlers, we also offer theoretically proven submodularity solutions in which efficient greedy algorithms can be implemented. We conduct comprehensive experiments on a complete set of real world microblog data, and show that our combined strategy could lead to high recall and more representative crawling of live microblog data streams, as compared with existing or less comprehensive crawling strategy.

4.2 Event Detection Techniques

The set of data crawled, though comprehensive, also contains a huge amount of noise. We thus send the crawled data to a binary SVM classifier to filter out the noise. The classifier not only uses the latest set of text features, but features also leverage on the user and social relations, which are extracted from the latest organization user network based upon the existing relationships between users within the organization.

For the remaining set of relevant data, we employ the well-known incremental clustering algorithm to discover topics in real time. In particular, we employ a single-pass incremental clustering algorithm [1] with a threshold τ. The algorithm considers each tweet in turn and determines the suitable cluster assignment based on a similarity function. The algorithm tries to assign each tweet to an existing cluster C_j, and update the cluster center of this evolving event. However, if the similarity value is less than τ, this signifies the occurrence of a new event and a new emerging cluster will be generated. The details of this algorithm can be found in [2]. At the end of this process, we generate a set of evolving events and a set of emerging events.

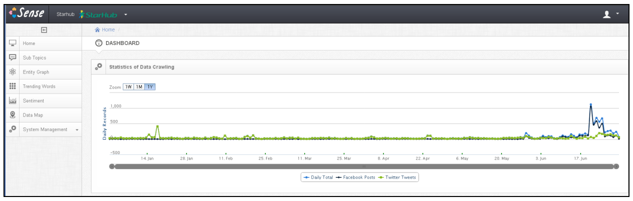
4.3 Hot Event Detection

In order to determine whether an emerging event will become viral, we analyze the features within each event. In addition to the user-based features used in hot tweet detection in Section 3, we also consider the rate-based features with respect to the event and organization. In particular, given a target organization at time t, we extract the set of keyusers and emerging keywords from the points of view of both the organization and the local emerging topics. We extract six representative features for each topic at time t to train the emerging topic learner as follows:

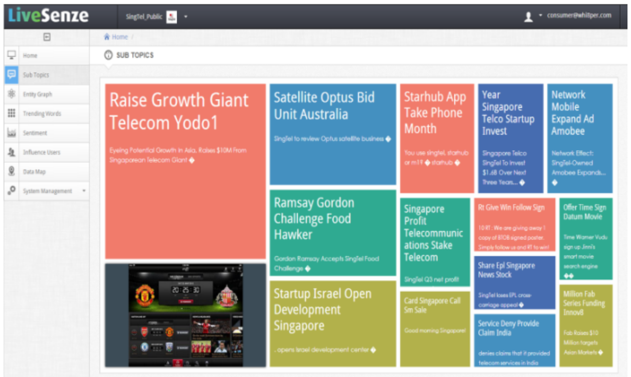
- f_1 is the rate of increase of user number;
- f_2 is the rate of increase of tweets number;
- f_3 is the rate of increase of re-tweets number;
- f_4 is the overlap between org keyusers and top N influential topic users;
- f_5 is the overlap between org keywords and top N influential topic keywords; and
- f_6 represents the rate of increase of influence of the accumulated weight of tweets.

These features take into account the number of participating users, the increasing rate of tweets, the number of retweets, and the overlap with the current org keyusers and dynamic keywords. The resulting classifier, trained based on SVM, has been found to be effective in detecting hot emerging events before they become viral with over 90% accuracy [2].

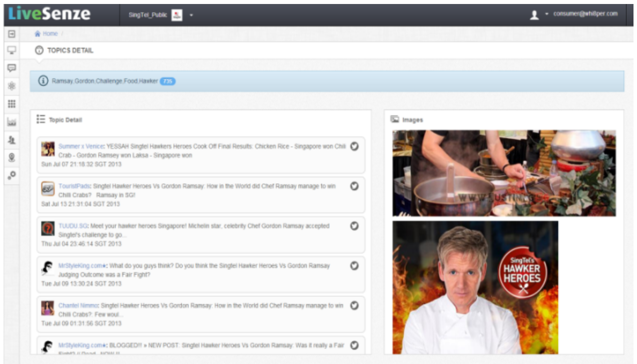
Figure 5 shows a snapshot of the tracking of an organization in Singapore and the sub-events detected by our system.



(a) A timeline view of the data stream related to the organization.


LiveSenze
SingTel_Public
consumer@whisper.com
Home
SUB TOPICS
Home
Sub Topics
Entity Graph
Trending Words
Sentiment
Influence Users
Data Map
System Management
Raise Growth Giant Telecom Yodo1
Eyeing Potential Growth In Asia, Raises $10M From Singaporean Telecom Giant
Satellite Optus Bid Unit Australia
Singtel to review Optus satellite business
Starhub App Take Phone Month
You use singtel, starhub or m1? starhub
Year Singapore Telco Startup Invest
Singapore Telco SingTel To Invest $1.6B Over Next Three Years...
Network Mobile Expand Ad Amobee
Network Effect: SingTel-Owned Amobee Expands...
Ramsay Gordon Challenge Food Hawker
Gordon Ramsay Accepts Singtel Food Challenge
Singapore Profit Telecommunications Stake Telecom
Singtel Q3 net profit
Rt Give Win Follow Sign
10 RT : We are giving away 1 copy of BTOB signed poster. Simply follow us and RT to win!
Offer Time Sign Datum Movie
Time Warner Vudu sign up Jinni's smart movie search engine
Share Epl Singapore News Stock
SingTel loses EPL cross-carriage appeal
Startup Israel Open Development Singapore
. opens Israel development center
Card Singapore Call Sm Sale
Good morning Singapore!
Service Deny Provide Claim India
denies claims that it provided telecom services in India
Million Fab Series Funding Innov8
Fab Raises $10 Million targets Asian Markets

(b) The list of sub-events detected. The size of the box indicates the importance of each sub-event based on the amount of data available.



Ramsay,Gordon,Challenge,Food,Hawker 735

Summer x Venice: YESSAH Singtel Hawkers Heroes Cook Off Final Results: Chicken Rice - Singapore won Chili Crab - Gordon Ramsey won Laksa - Singapore won
Sun Jul 07 21:18:32 SGT 2013

TouristPads: Singtel Hawker Heroes Vs Gordon Ramsay: How in the World did Chef Ramsay manage to win Chili Crabs? Ramsay in SG!
Sat Jul 13 21:31:04 SGT 2013

TUUDU.SG: Meet your hawker heroes Singapore! Michelin star, celebrity Chef Gordon Ramsay accepted Singtel's challenge to go...
Thu Jul 04 23:46:14 SGT 2013

MrStyleKing.com★: What do you guys think? Do you think the Singtel Hawker Heroes Vs Gordon Ramsay Judging Outcome was a Fair Fight?
Tue Jul 09 13:30:24 SGT 2013

Chantel Nimmo: Singtel Hawker Heroes Vs Gordon Ramsay: How in the World did Chef Ramsay manage to win Chili Crabs?: Few woul...
Tue Jul 09 01:31:56 SGT 2013

MrStyleKing.com★: BLOGGED!! » NEW POST: Singtel Hawker Heroes Vs Gordon Ramsay: Was it really a Fair

(c) The list of tweets and the accompyning images for a sub-event.

Figure 5: The event detection results for an organization in Singapore

5. CURRENT EFFORTS

Research of live social observatories is only beginning, with many centers devoting large efforts to gathering their own data and implementing their own tools and analytics to serve different applications. These observatories are developed independently, and are not connected and thus unable to share their data and analytics. Besides, there are legal and privacy issues involved that prevent the open sharing of data, especially raw data. These problems are identified by Web Science Trust (WST) [9] as key obstacles towards open social observatory framework, in which multiple observatories can share data and analytics to accomplish a larger set of common goals. A number of problems need to be tackled and framework established before true sharing, both legally and algorithmically, can take place. Here we list a number of efforts that we are working on toward open social observatory system and framework.

First, we will extend the framework to handle multimedia data, which is becoming increasingly important as more users are posting messages containing multimedia content often with no relevant text. Towards this end, we have extended the event detection framework described in Section 4 to perform crawling and analysis of images and videos. We employ our multimedia indexing and search engine developed to analyze the huge amount of multimedia data for clues of objects/events appearing in its content.

Second, we will work towards a privacy framework. One area that we will devote effort upon is on techniques to anonymize data so that the original privacy content cannot be re-constructed. This will be done at both the entity level and application level.

Third, we will work towards a common framework for sharing of analytics. Again, this is general and will need to start with a few common applications across multiple social observatory systems. Examples of such applications include environment data, data on social health and personal well-being, as well as local users' concerns and sentiments on these issues. Finally, along the third aim, we will launch a number of applications in collaboration with other social observatory systems targeting at different cities and across the world.

6. SUMMARY

In this paper, we have described the design and architecture of a large-scale live social observatory system. The system is developed under NExT, a NUS-Tsinghua Joint Center on Extreme Search [3, 8]. The Center aims at gathering, mining and analyzing social media information. The social observatory under NExT has been in operation for over a year and has collected and analyzed over 2.6 billion records, with over 250 million non-textual contents. The current system has been evolved over past year into a robust, efficient and scalable system. This paper shares our experience in realizing the system and in handling large-scale live textual and non-textual data. We believe that the framework is robust and scalable. Under the framework, a number of research efforts have been carried out, including analytics on entities, such as the organizations, people, topics and locations; differential news; as well as community discovery systems, etc.

The research is only beginning and it offers great potential for cutting edge research in social media analysis. For greater impact, however, it needs to be deployed for large-scale use, and most importantly, needs to move towards international collaboration, in which data and systems from multiple cities can be aggregated to discover social phenomena across the national boundaries.

ACKNOWLEDGMENTS

NExT-Live system is developed by NExT Search Center [3, 8]. 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).

REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, 2010.

[2] Y. Chen, H. Amiri, Z.J. Li and T. S. Chua: Emerging Topic Detection for Organizations. Proceedings of ACM SIGIR'2013, Jul 28–Aug 1 2013, Dublin, Ireland.

[3] T. S. Chua, H. B. Luan, M. S. Sun, S. Q. Yang: NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content. IEEE Multimedia 19(3): 81-87, 2012.

[4] P. Cui, S. Jin, L. Y. Yu, F. Wang, W. W. Zhu and S. Q. Yang. Cascading Outbreak Prediction in Networks: A Data-Driven Approach. KDD'13, Aug, 2013, Chicago, USA.

[5] Y. Gao, F. Wang, H.B. Luan and T.-S. Chua. Brand Data Gathering From Social Media Streams. To appear in ICMR 2014 (Glasgow, UK), 1-4 Apr 2014.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. van Briesen and N. Glance. Cost-effective outbreak detection in networks. Proceedings of KDD, 420–429, 2007.

[7] J. Yang, K. Yu, Y. Gong and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. Proceedings of CVPR, 1794–1801, 2009.

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

[9] http://www.w3.org/community/webobservatory/