uTrack: Track Yourself!
Monitoring Information on Online Social Media

Tiago Rodrigues
Universidade Federal de
Minas Gerais, UFMG
Belo Horizonte, Brazil
tiagorm@dcc.ufmg.br

Prateek Dewan
Indraprastha Institute of
Information Technology, Delhi,
IIIT-Delhi
New Delhi, India
prateekd@iiitd.ac.in

Ponnurangam
Kumaraguru
Indraprastha Institute of
Information Technology, Delhi,
IIIT-Delhi
New Delhi, India
pk@iiitd.ac.in

Raquel Melo Minardi
Universidade Federal de
Minas Gerais, UFMG
Belo Horizonte, Brazil
raquelcm@dcc.ufmg.br

Virgílio Almeida
Universidade Federal de
Minas Gerais, UFMG
Belo Horizonte, Brazil
virgilio@dcc.ufmg.br

ABSTRACT

The past one decade has witnessed an astounding outburst in the number of online social media (OSM) services, and a lot of these services have enthralled millions of users across the globe. With such tremendous number of users, the amount of content being generated and shared on OSM services is also enormous. As a result, trying to visualize all this overwhelming amount of content, and gain useful insights from it has become a challenge. In this work, we present uTrack, a personalized web service to analyze and visualize the diffusion of content shared by users across multiple OSM platforms. To the best of our knowledge, there exists no work which concentrates on monitoring information diffusion for personal accounts. Currently, uTrack monitors and supports logging in from Facebook, Twitter, and Google+. Once granted permissions by the user, uTrack monitors all URLs (like videos, photos, news articles) the user has shared in all OSM services supported, and generates useful visualizations and statistics from the collected data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms, Design, Experimentation, Measurement

Keywords

Information diffusion, Online Social Media, tracking, visualization

WWW 2013 Companion, May 13-17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

1. INTRODUCTION

"Information visualization" is the use of computer - supported, interactive, visual representations of abstract data to amplify cognition [1]. There have been multiple efforts and achievements of computer science in the developing models, algorithms and techniques to ease data analysis. This has been achievable by summarizing data and information extraction for sense making when exploring huge amounts of data. However, inspite of these achievements, traditional textual patterns are not easy to comprehend by human perceptual and cognitive system. Analyzing real world events [2] and presenting them in a succinct form to decision-makers are becoming a big need of the hour.

The success of OSM services has made them a tremendous source of information for individuals using these services. For instance, on Twitter, users share and forward various URLs with personal recommendations like "Check out this great article about the last Olympic games!" and on Facebook users share any kind of content like photos and videos from their vacation. A huge amount of content is generated and shared every day in the form of URLs. For instance, 30 billion pieces of content are shared each month on Facebook, 1 million links are shared in just 20 minutes, and 136,000 photos are uploaded every minute. The numbers are also impressive on Twitter: 1 billion tweets were sent every five days in 2011, and a significant part contained links. [1] Facebook alone has 240 billion photos in total. [2] The popularity of each of such URL is reasonably well known within a social network individually. For example, a post containing a YouTube video shared on Facebook, has a "number of shares" statistic associated with it, which is visible to a user. Similarly, a tweet comes with "number of re-tweets", which informs a user about how many users shared / posted the same content or URL. Other OSM services provide similar statistics with their content. However, what people miss

[1] http://thesocialskinny.com/100-social-media-statistics-for-2012/

[2] http://thenextweb.com/facebook/2013/01/15/facebook-our-1-billion-users-have-uploaded-240-billion-photos-made-1-trillion-connections/

is how the information they share is being diffused across multiple networks.

In this work, we present uTrack, a web service which tracks the content shared by users, across multiple OSM networks, and provides them with detailed analysis and statistics about this content. uTrack enables it's users to visualize how their content diffused across multiple services through time and space. The service currently supports logging in from Facebook, Twitter, and Google+, and provides users with statistics and visualizations about who shared their content, what is popular within their friends' network, where is their and their friends' content getting the most attention, and when is the content getting the most attention. All this information is presented to the user on a "dashboard" interface. A user can also connect multiple accounts to visualize all of her content across OSM platforms.

uTrack lets users know who is talking about their YouTube videos on Facebook; companies and marketers can gauge what consumers think about their products and marketing campaigns; journalists and bloggers will be able to analyze and understand readers' reactions on articles and blog posts. Curiosity, understanding the audience, and analysis of professional content like advertisements and marketing campaigns are some of the key benefits provided by our technology.

There are some commercial and academic tools which enable users to do social media analytics in the market. Truthy [3] is a system to analyze and visualize the diffusion of information on Twitter, developed by Indiana University Center for Complex Networks & Systems Research [5]. Marcus et al. [3] presented two systems for querying and extracting structure from Twitter-embedded data. Salesforce Marketing Cloud service [4] is a commercial platform for brands to monitor their popularity and audience on OSM networks. The service is solely for commercial purposes, and a common OSM user cannot benefit much from it. SocialMention [5], and WhosTalkin [6] are two real-time social media search and analysis services, which allow users to search for topics across multiple OSM platforms. Both these services are free, but are categorized as social media search engines, and do not track content specific to a particular user. SocialAppsHQ [7] is a social media marketing platform for Facebook. Like Radian6, SocialAppsHQ is also a paid service, and built mainly for commercial purposes. Wolfram Alpha, a "computational knowledge engine", recently came up with a new feature that allows a user to quickly get an overview of all her data on Facebook. [8] Wolfram Alpha, in general, isn't specific to social media, and aligns more with the field of intelligent computing.

In contrast to these tools, the focus of uTrack is to collect and analyze the diffusion of information in multiple OSM services. uTrack also differs from all the aforementioned services as it concentrates on user-specific content. Apart from it's ability to function as a search engine, uTrack is of special interest for common users too, since it is personalized.

2. uTrack ARCHITECTURE

uTrack has a distributed architecture designed to collect, process, and recover a huge volume of data in an efficient manner. A web interface provide users a dashboard mechanism to analyze their data. Users can log into uTrack and give permission to collect their posts through the official APIs provided by OSM services. uTrack will then look for mentions of their content (URLs) shared by other users, across several services, through their search APIs. Any and all mentions found, are then collated together, and a detailed analysis of this information and knowledge is presented to the user on her "dashboard". Figure 1 shows a high-level flow of the complete process.

To get a better understanding of how uTrack works, the complete system can be broken down into four modules: (1) the *user authentication* lets the user authenticate herself and allows uTrack to collect her posts from various OSM services; (2) the *back-end crawler* is responsible for collecting data from all OSM services through APIs; (3) the *data processing* module applies filters and stores relevant information for efficient recovery; (4) the *user interface* comprises of the "dashboard" with all its visualizations and statistics. We will now discuss each of these in more detail.

2.1 User Authentication

The URLs and user IDs fed as initial seeds to the crawler, come from uTrackâĂŹs users. In the current version, uTrack supports logging in from three OSM services, i.e., Facebook, Twitter, and Google+. Each login button redirects the user to the respective OSM services' authentication page, where the user is asked for permissions to get access to the content they post. Since all these three OSM services use the OAuth mechanism for authentication, we devised and deployed a generic authentication module, which, can be configured and used to deploy a log in / authentication mechanism for any OSM service which provides an OAuth based API. If the user grants uTrack the required permissions, uTrack creates an entry with her user ID in the database, and feeds the user ID into the crawler. She is then redirected back to the uTrack website, and lands on her dashboard page.

2.2 Backend Crawler

The crawler is the back-bone of the system, and is responsible for extracting data from various OSM platforms through their APIs. The implementation of the crawler is highly robust and distributed to take care of scalability. The crawler in itself is a complete and sophisticated data collection system capable of collecting data from multiple APIs and multiple OSM services in parallel. It follows a "*master-slave*" architecture, where different physical servers can be used as slaves to collect data, and push it to a master physical server. The crawler was conceived and prototyped as part of the master dissertation work of Tiago Rodrigues [6].

To start the crawling process, a "master" is initialized with one parameter viz. the name of the OSM service. At least one master is required to be initialized for each OSM service which needs to be crawled. Once a master is initialized, "slaves" are spawned from other servers. Each slave is configured with some initial parameters, like the master's address, and the OSM service to track. This slave then interacts with the master over the network, using TCP sockets. One master is capable of handling multiple slaves tracking the same OSM service. The master passes initial seeds along with it's

[3] http://truthy.indiana.edu/
[4] http://www.radian6.com/
[5] http://www.socialmention.com/
[6] http://www.whostalkin.com/
[7] https://www.socialappshq.com/
[8] http://www.wolframalpha.com/facebook/

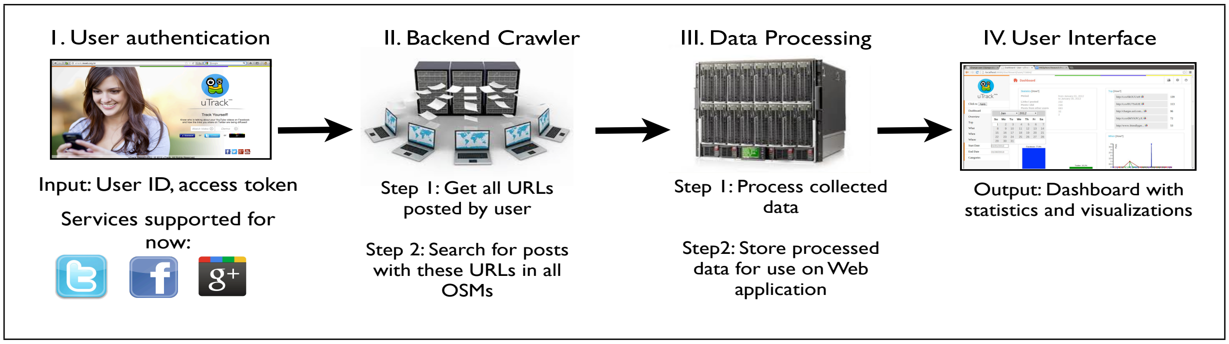

I. User authentication
uTrack
Input: User ID, access token
Services supported for now:
II. Backend Crawler
Step 1: Get all URLs posted by user
Step 2: Search for posts with these URLs in all OSMs
III. Data Processing
Step 1: Process collected data
Step2: Store processed data for use on Web application
IV. User Interface
Output: Dashboard with statistics and visualizations

Figure 1: Architecture, along with the modules involved in converting the user information to visualization of the information – User authentication, crawler, data processing, and user interface.

type, to all it's slaves. A seed can be either a user ID, or a URL. If the seed is a user ID, the crawler fetches all content / feed posted by this user, filters out all the URLs and related text, and sends this information back to the master. If the seed is a URL, the crawler queries the search API of the OSM service to look for any other users who may have posted the same URL, and sends this information back to the master. The master receives all this information and stores it into a central database. This central database is shared by all masters to syncronize data like new URLs found. Moreover, several policies were created to increase the crawler's performance (i.e., collect the highest amount of data) respecting limits imposed by the OSM services APIs.

2.3 Data Processing

This module is responsible for processing the raw data collected by the back-end crawler, and preparing the processed data for an efficient recovery when doing the analysis. Computationally intensive processes are performed in this module, like natural language processing tasks (remove stop-words, stemming, detecting languange, sentiment analysis), and detection of basic units of information (URLs, categories, users, mentions, retweets, shares, likes, location, timestamp, etc). After processing each post collected, all relevant information is saved on the database for an efficient recovery by the dashboard mechanism.

Several indexes, models, and optimizations are created to guarantee a good performance which will let our users analyze their data and gain insights on how their information is being diffused in the OSM.

2.4 User Interface

uTrack's user interface and visualizations have been given special attention to ensure that it's users are able to see exactly what they want. The user interface has been kept fairly simple to reduce the visual cognitive load on the user [4]. uTrack presents the users with an easy to navigate dashboard view, and opens with the *Overview* dashboard. This dashboard shows a summary of all the data and it's visualizations to the users. The user can navigate to the other views / dashboards through the links provided on a frame on the left side of the page. This frame also contains a date-picker which can be used to select the time period for which the users want to view their data. Users can also filter content into categories like photos, videos, links, check-ins, music, etc. and see analysis and visualizations only for selected type of content.

Apart from the overview dashboard, the main user interface has been divided into five dashboards, and each dashboard presents a different set of analysis to the user. The *Top* dashboard shows the most popular objects among the user data, like top URLs and users mentioned, for instance. The *What* dashboard presents what contents have being shared, like the last posts and their sentiments. The *When* dashboard demonstrates a comprehensive temporal analysis of the content shared by the user. Finally, the *Where* dashboard shows in which OSM services the URLs are being diffused, as well as the places from where they have been posted.

Graphs in both *When* and *Where* dashboards are divided into divided into two categories, i.e., "Posts made by the user", and "Posts made by others". This simple division enable users to contrast their actions with other users. Our goal is to provide users a rich set of statistics and visualizations, allowing them to analyze and understand several aspects about the diffusion of their information. For instance, our users are be able to know what content was shared by who, when it was shared and from where. Understanding the audience and analysis of professional content like advertisements and marketing campaigns are examples of key benefits provided by our technology.

The web application is build in Django [9], a Python web framework. We also used other technologies to build our user interface like Javascript [10], JQuery [11], and D3 [12].

3. DEMONSTRATION

In order to create a demo for uTrack, some of the most famous celebrities in the world were manually selected and inserted to be monitored. uTrack is able to get all public posts made by these users through the OSM services APIs. Table 1 shows some statistics from the pop singers Justin Bieber, Britney Spears, and Rihanna. A snapshot of the overview dashboard for Justin Bieber is shown on Figure 2.

Justin Bieber, a popular pop singer all over the world, has posted 46 posts containing 14 URLs from January 26th, 2013 to February 2nd, 2013. In total, these URLs have been posted 22,635 times by other users on Twitter (98%), Facebook (1.9%), and Google+ (0.1%). Moreover, Justin Bieber

[9] https://www.djangoproject.com/

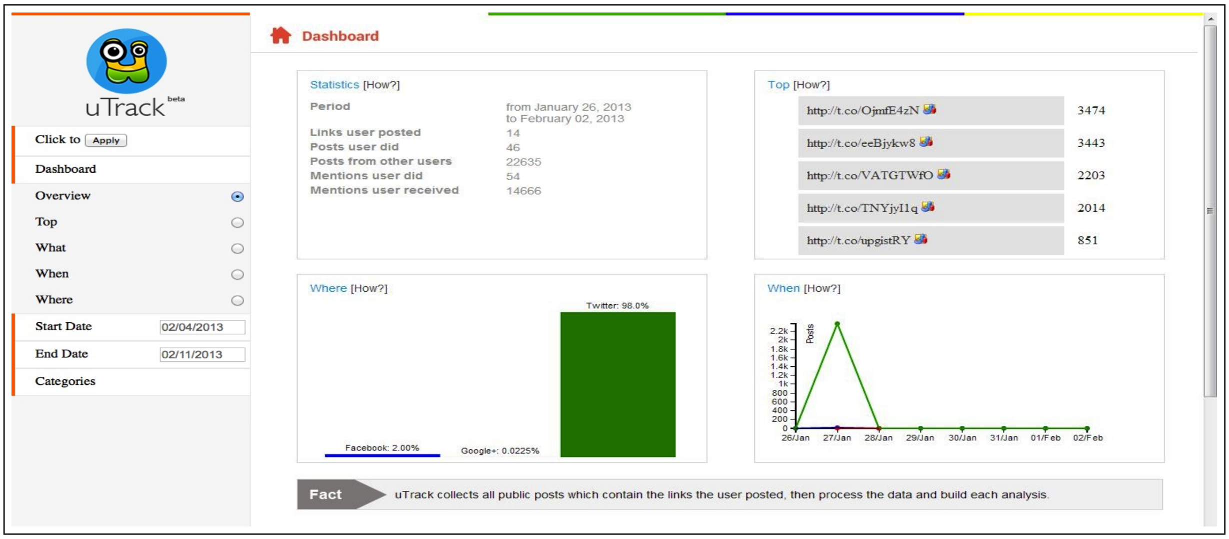
[10] http://en.wikipedia.org/wiki/JavaScript

[11] http://jquery.com/

[12] http://d3js.org/

User	#users	#posts	#posts others	#mentions	#mentioned	Twitter	Facebook	Google+
Justin Bieber	14	46	22,635	54	14,666	98.0%	1.9%	0.1%
Britney Spears	3	6	10,535	13	10,513	98.8%	1.0%	0.2%
Rihanna	7	7	15,142	1	15,025	99.5%	0.5%	-

Table 1: Statistics for a few famous uTrack users, from January 26th, 2013 to February 2nd, 2013.



Statistics [How?]

Period	from January 26, 2013 to February 02, 2013
Links user posted	14
Posts user did	46
Posts from other users	22635
Mentions user did	54
Mentions user received	14666

Top [How?]

http://t.co/OjmfE4zN	3474
http://t.co/eeBjykw8	3443
http://t.co/VATGTWfO	2203
http://t.co/TNYjyI1q	2014
http://t.co/upgistRY	851

Fact uTrack collects all public posts which contain the links the user posted, then process the data and build each analysis.

Figure 2: uTrack overview dashboard for user "Justin Bieber".

has mentioned other users 54 times in this period, and has been mentioned 14,666 times by other users. As can be seen on Figure 2, the most popular URL posted by Justin Bieber in this period is *http://t.co/OjmfE4zN*, which is a photo from one of his concerts. This URL was re-posted 3,474 times.

At present, uTrack has 56 beta testers. In total, 1,039 users are being monitored (including famous accounts manually added for testing purposes), and 23,709,183 posts have been collected, with 1,076,026 unique URLs. One of the most popular URLs (11,580 posts) is the video clip of the song "Gangnam Style" [13], by Psy. uTrack was visited by 207 unique users, in a total of 7,925 pageviews, from September, 2012, to February, 2013.

4. DISCUSSION AND FUTURE WORK

Users share a lot of contents every day in OSM services. What users miss is an easy way to monitor how information is being diffused among other users and services. We fill this gap by creating a web service which will provide users a rich set of statistics and visualizations, allowing our users to analyze and understand several aspects about the diffusion of their information.

Extracting knowledge from large datasets and dynamic networks is a challenge today. Filtering high quality content, processing the data in real time, finding influential users, and direct adverstisements according to user tastes and needs, are some examples of important tasks in which we could apply the technology we are building in this research project.

As future work, we plan to increase the number of services supported by uTrack. We are working towards adding support for other famous platforms like LinkedIn, FourSquare, and MySpace. We also plan to extend support for blogs and implement a mobile version of the user interface. As future research directions directly applicable to our platform, we plan to study recommendation of contents and automatical categorization of links and content. New features like a social search engine are also among our plans, which might also require some research. Moreover, we plan to open a start-up using our technology.

5. REFERENCES

[1] S. Card, J. Mckinlay, and B. Schneiderman. *Readings in information visualization: using vision to think.* Academic Press, 1999.

[2] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. *Workshop on Privacy and Security in Online Social Media, Co-located with WWW*, 2012.

[3] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller. Processing and visualizing the data in tweets. *ACM SIGMOD Record*, 40(4):21–27, 2012.

[4] R. Mayer and R. Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52, 2003.

[5] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[6] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011.

[13] http://www.youtube.com/watch?v=9bZkp7q19f0