Multiple Media Analysis and Visualization for Understanding Social Activities

Masashi Toyoda
Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku
Tokyo, Japan
mtoyoda@acm.org

ABSTRACT

The Web has involved diverse media services, such as blogs, photo/video/link sharing, social networks, and microblogs. These Web media react to and affect realworld events, while the mass media still has big influence on social activities.

The Web and mass media now affect each other. Our use of media has evolved dynamically in the last decade, and this affects our societal behavior. For instance, the first photo of a plane crash landing during the "Miracle on the Hudson" on January 15, 2009 appeared and spread on Twitter and was then used in TV news. During the "Chelyabinsk Meteor" incident on February 15, 2013, many people reported videos of the incident on YouTube then mass media reused them on TV programs.

Large scale collection, analysis, and visualization of those multiple media are strongly required for sociology, linguistics, risk management, and marketing researches. We are building a huge scale Japanese web archive, and various analytics engines with a large-scale display wall. Our archive consists of 30 billion web pages crawled for 14 years, 1 billion blog posts for 7 years, and 15 billion tweets for 3 years.

In this talk, I present several analysis and visualization systems based on network analysis, natural language processing, image processing, and 3 dimensional visualization. First, I introduce a visualization system for monitoring the information diffusion in Twitter. It can visualizes large scale hierarchical dynamic graphs with temporal animations. Tweets are clustered by their topics, and retweet and mention relationships are represented as graphs. We can easily grasp the appearance of topics and influential users in each topic. I will show some information diffusion patterns under the Great East Japan Earthquake.

Next, I introduce a system for analyzing temporal changes in the activities and interests of bloggers through a 3D visualization of phrase dependency structures in sentences. This system enables us to find events about a topic, and drill down details of the events. It also enables us to compare events with different timings and on multiple topics.

Finally, I demonstrate our framework for inter-media analysis through image flows extracted from blogs and TV to understand societal behaviors. Both blogs and TV generate a huge amount of image flows every day. Comparing such large image flows is one of our biggest challenges to capture activities over multiple media. This framework is based on our web archive and a broadcast news video archive that includes news videos on six TV channels over nineteen months collected by by National Institute of Informatics. To compare the occurrence frequencies of these images in both blogs and news videos, we use a scalable shot retrieval index that can search similar shots in news videos from blog images. Images extracted from blogs and TV are visualized in 3D space. For each topic and medium, images are piled up like a time series histogram and are arranged so that the user can easily compare differences in exposure and timing. We also provide a dynamic query function that helps us to explore images with interesting characteristics from visualized images in 3D space.

By extracting similar images on blogs as on TV from huge media archives, we can observe various inter-media phoenomenon. For a given topic, we can find out which medium first provided the information. If images captured from TV spread on the web, we can confirm when and how many times these images were broadcasted on TV. If images first appeared on the Web are used on TV programs, we can investigate when these images appeared and how many users shared them.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]

General Terms

Experimentation

Keywords

Web archive, Multiple media analysis, Visualization

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579040 .