Getting the Bigger Picture – Extracting Media Items Covering Events from Multiple Social Networks

Thomas Steiner
Univ. Politècnica de Catalunya
Department LSI
08034 Barcelona, Spain,
tsteiner@lsi.upc.edu

Raphaël Troncy
EURECOM
06560 Sophia Antipolis
France
rtroncy@eurecom.fr

Ruben Verborgh
Ghent University – IBBT, ELIS
Multimedia Lab
9050 Ghent, Belgium
ruben.verborgh@ughent.be

Joaquim Gabarró Vallés
Univ. Politècnica de Catalunya
Department LSI
08034 Barcelona, Spain,
gabarro@lsi.upc.edu

ABSTRACT
Core contributions: Social network and media hosting platform agnostic media item search, and search results alignment. Search results semantic enrichment by putting microposts and media items in relation. Cross-channel popularity analysis. Visual clustering, photos contained in videos.

Categories and Subject Descriptors
H.3.4 [**Information Systems**]: Information Storage and Retrieval—*World Wide Web*; H.3.5 [**Online Information Services**]: Web-based services

Keywords
{**TODO**: Keywords}

1. INTRODUCTION
1.1 Definitions
1.1.1 Media Item
1.1.2 Micropost
1.1.3 Event
2. SOCIAL NETWORKS AND MEDIA ITEMS
2.1 Social Networks vs. Hosting Platforms
The boundary between social networks and media hosting platforms is fluid. There are media hosting platforms where people can upload their content with optionally content viewers – in a relation to the original uploader or not – having the possibility to react in form of comments, or in form of likes or dislikes. An example is YouTube. There are social networks where people can update their status, post links to stories, or upload their content with necessarily viewers – standing in a relation or not – having the option to react. An example is Facebook. Finally, there are hybrid forms, where social networks – typically via third party applications – integrate with media hosting platforms. An example is the TweetDeck for Twitter application and their integration with TwitPic.

2.2 Media Item Extraction
Talk about media item recall (ratio items that match the query vs. items that match the query with media items).

2.3 API Access vs. Web Scraping
An *Application Programming Interface (API)* in the sense of Web-based API is a programmatic specification intended to be used as an interface by software components on client and server to communicate with each other.

Web scraping is the process of automatically extracting information from websites. Web scraping involves practical solutions based on existing technologies that are often entirely ad hoc. Examples of such technologies are regular expressions or DOM parsing of Web pages into a DOM tree.

The difference to the somewhat related concept of *screen scraping* is that screen scraping relies on the visual layout of a website, whereas Web scraping relies on the textual and/or hierarchical structure of websites.

Social networks today are very much perceived as "walled gardens", excellently illustrated by a cartoon by David Simonds (Figure 1). As with Orwell, where some animals are more equal than others, some social networks are more walled than others. Some social networks have full read and write access via specified APIs. An example is Twitter. Other social networks have read access via APIs. An example is Google+. Interestingly, some media hosting platforms have just write API support, however, no read support, which requires us to fall back to Web scraping the website in order to retrieve data. An example is Img.ly.

3. IMPLEMENTATION DETAILS
3.1 Data Structure
3.2 Media Item Collectors
3.3 Machine Translation
3.4 Part of Speech Tagging
3.5 Named Entity Disambiguation
3.6 Irrelevant Media Item Pruning
4. EXPERIMENTS
For our experiments, we have taken into account several events that happened on, or were still ongoing the morning of January 10, 2012, CET, and thus the subject of discussion

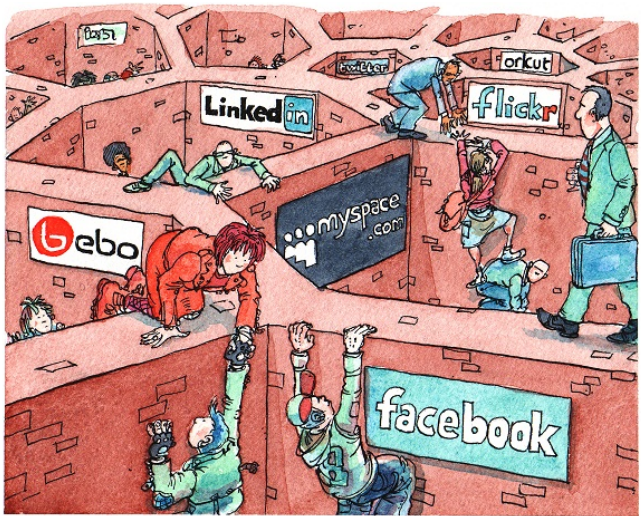

Figure 1: David Simonds illustrates social networks as walled gardens due to their (by design) lock-in effects [3].

on various social networks. We have captured event-related media items and microposts, and made it available online[1].

4.1 Considered Events

In this Subsection, we give an objective and short overview on the context of the considered events in order to give the reader the necessary background knowledge.

4.1.1 Assad Speech

On January 10, 2012, Syrian President Bashar al-Assad delivered a lengthy televised talk strongly defending his government's actions and motivations, despite world pressure on his embattled government for its 10-month crackdown on protesters[2].

4.1.2 CES Las Vegas

The International Consumer Electronics Show (CES) is a major technology-related trade show held each January in the Las Vegas Convention Center. Not open to the public, the Consumer Electronics Association-sponsored show typically hosts previews of products and new product announcements[3].

4.1.3 Christian Wulff Case

Since December 2011, German President Christian Wulff faces controversy over discrepancies in statements about a loan while governor of Lower Saxony. When the affair settled down, it was revealed that he had applied pressure on Springer Press to delay revelations on the issue until he was back from a visit abroad. When Wulff found out that a tabloid was going to break the story, he left a message on the voice mail of the editor-in-chief in which he threatened to take legal action[4].

4.1.4 Cut the Rope Launch

On January 10, 2012 during Microsoft's keynote at CES, the HTML5 version of the popular mobile game *Cut the Rope* was announced[5].

4.1.5 Dixville Notch

Dixville Notch is an unincorporated village in Dixville township of Coos County, New Hampshire, USA, best known in connection with its longstanding middle-of-the-night vote in the U.S. presidential election. In a tradition that started in the 1960 election, all the eligible voters in Dixville Notch gather at midnight in the ballroom of The Balsams. This year, on January 10, 2012, the voters cast their ballots and the polls officially closed one minute later[6].

4.1.6 Free Mobile Launch

Free Mobile is a French mobile broadband company, part of the Iliad group. On January 10, 2012, a long-awaited mobile phone package for €19.99 with calls included to 40 countries, texts, multimedia messages and Internet was announced by the Iliad group's Chief Strategy Officer, Xavier Niel[7].

4.1.7 Ubuntu TV Launch

Ubuntu TV by Canonical, based on the user interface Unity, is a variant of the Ubuntu operating system, designed to be a Linux distribution specially adapted for embedded systems in televisions. It was announced by Canonical on January 10, 2012, at CES[8].

4.2 Discussion

5. RELATED WORK

[1] Just Flickr and YouTube. Use clustering algorithm for textual content. Duplicate detection via Color and Edge Directivity Descriptor (CEDD). Restrict selection to geolocated content and time period.

[2] Objective is to aggregate heterogeneous event information sources using Linked Data principles. Just Flickr and YouTube, with additional media from event information sources. Query with "what", "where", "when", uses "title"+"time" and "geotag"+"time". Use machine-tagged media items to find visually similar media items.

[1]Event-related media items and microposts: `http://www.lsi.upc.edu/~tsteiner/experiments/icmr2012/`

[2]Assad Speech: `http://www.cnn.com/2012/01/10/world/meast/syria-unrest/`

[3]CES Las Vegas: `http://www.cesweb.org/aboutcea.asp`

[4]Christian Wulff Case: `http://www.spiegel.de/international/germany/0,1518,804631,00.html`

[5]Cut the Rope Launch: `http://ces.cnet.com/8301-33377_1-57356403/`

[6]Dixville Notch: `http://www.washingtonpost.com/2012/01/09/gIQANslKnP_story.html`

[7]Free Mobile Launch: `http://www.nytimes.com/2012/01/11/technology/iliad-takes-aim-at-top-mobile-operators-in-france.html`

[8]Ubuntu TV: `http://www.theverge.com/2012/1/9/2695387/ubuntu-tv-video-hands-on`

6. FUTURE WORK

Privacy and licensing issues. Support more social networks and improve Web scrapers.

7. CONCLUSION

8. ACKNOWLEDGMENTS

This work was partially supported by the European Commission under Grant No. 248296 FP7 I-SEARCH project.

9. REFERENCES

[1] M. del Fabro and L. Böszörményi. Summarization and Presentation of Real-Life Events Using Community-Contributed Content. In *MMM*, pages 630–632, 2012.

[2] X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 58:1–58:8, New York, NY, USA, 2011. ACM.

[3] D. Simonds. Walled Gardens. Cartoon in *The Economist*. Taken from "WWW and Hopes for the Future" by T. Berners-Lee, Feb. 2011. Available at `http://www.w3.org/2011/Talks/0222-saudi-tbl/`.