Citizen Sensor Data Mining, Social Media Analytics and Development Centric Web Applications

Meena Nagarajan
IBM Almaden Research Center
San Jose, CA, USA
http://bit.ly/m33naa
MeenaNagarajan@us.ibm.com

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, OH, USA
http://knoesis.org/amit
amit@knoesis.org

Selvam Velmurugan
eMoksha and Kiirti
Seattle, WA, USA
http://emoksha.org
selvam@emoksha.org

ABSTRACT

With the rapid rise in the popularity of social media (500M+ Facebook users, 100M+ twitter users), and near ubiquitous mobile access (4.1 billion actively-used mobile phones), the sharing of observations and opinions has become common-place (nearly 100M tweets a day, 1.8 trillion SMSs in US last year). This has given us an unprecedented access to the pulse of a populace and the ability to perform analytics on social data to support a variety of socially intelligent applications -- be it towards targeted online content delivery, crisis management, organizing revolutions or promoting social development in underdeveloped and developing countries.

This tutorial will address challenges and techniques for building applications that support a broad variety of users and types of social media. This tutorial will focus on social intelligence applications for social development, and cover the following research efforts in sufficient depth: 1) understanding and analysis of informal text, esp. microblogs (e.g., issues of cultural entity extraction and role of semantic/background knowledge enhanced techniques), and 2) building social media analytics platforms. Technical insights will be coupled with identification of computational techniques and real-world examples.

Categories and Subject Descriptors

H.3.5 On-line Information Service, H.3.1 Content Analysis and Indexing H.3.3 Information Search and Retrieval

General Terms

Algorithms, Economics, Human Factors, Languages.

Keywords

Citizen sensing, social signals, user generated content, social media analysis, semantic social mashup, semantic social web, social development application, people-content-network view of social media, mobile development application.

INTRODUCTION

This tutorial weaves together three themes and respective topics:

1. Opportunity to exploit the massive amounts of social data that have resulted from the participation of millions of users through a wide variety of on-line interaction mechanisms and types of social data. Some examples are SMS through simple mobile phones, tweets with time and location coordinates, as well as multi-sensory metadata from smartphones.
2. People-Content-Network view of analyzing social media use, including technical challenges and recent research efforts in extracting metadata from casual/informal user-generated content on social media platforms.
3. Experiences in building robust and scalable platforms and applications to serve the needs of (a) users at the top of the pyramid (1.5 billion users with smart phones and modern network access) as well as (b) users in the middle of the pyramid (3 billion users with mobile phones without access to Internet or modern networks).[1]

Three parts of the tutorial are described next.

PART I: SOCIAL MEDIA & CITIZEN SENSING

We will review patterns of participation along three facets: Technology, Diversity, and Purpose.

1. Technological: Web, Mobile including SMS and Smartphones with sensors

2.Diversity of Social Media: variety of technical capabilities and features, as well as diversity in demographics across different platforms yielding a variety of content (e.g., well-written blogs vs poorly structured comments on MySpace)

3. Purpose: personal/informational, transactional, broadcast/sharing, short-term events, long-term situations, science and development. Examples of citizen sensing: citizen science, political events, natural disasters and activism.

PART 2: SOCIAL DEVELOPMENT CENTRIC PLATFORMS AND APPLICATIONS:

We will discuss system architecture, scalability, robustness, deployment, etc. [1] along with the challenges and experiences in building systems that exploit user-generated data for diverse

[1] This use of pyramid courtesy of Ramesh Jain. http://bit.ly/i8na79

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

classes of social development centric applications. Some examples are as follows:

(a) Ushahidi (ushahi.org), eMoksha (emoksha.org), Kiirti (kiirti.org), Sahana (sahanafoundation.org), and other real-world platforms for development-centric applications and crisis management in developing countries. We will borrow from case studies at Harvard, Columbia and institutions worldwide on the use of eMoksha/Kiirti projects involving social media and technologies for development. [4]

(c) Twitris —application that supports spatio-temporal-thematic analysis and extraction of social signals from microblogs, also complemented by news, Web 2.0, and image/video sources (http://twitris.knoesis.org) [3]

(d) Twarql -annotations and management of streaming Tweets, encoding information from microblog posts as Linked Open Data for collectively analyzing microblog data for sensemaking (http://wiki.knoesis.org/index.php/Twarql) [5]

PART 3: METADATA EXTRACTION FROM USER GENERATED DATA AND ANALYSIS TECHNIQUES

(a) Different types of data and metadata:

- Structured and unstructured component of data sources Structured metadata: spatio, temporal, thematic data, profile/user demographic data, attention metadata (likes, views, listens, clicks, tags etc.)
- Unstructured component of data: user-generated textual content ranging from relatively context-sparse microblogs, forum messages, to context-rich blogs.

(b) Resources used in metadata extraction: Task independent domain knowledge, task specific domain information, linguistic and cultural features, network features, temporal and spatial properties of data.

(c) Metadata extraction from unstructured textual social data (in-depth treatment): Extraction of different types of metadata relevant to social development centric applications - entity, sentiment, intention etc.

(d) Continuous semantics and real-time data annotation/extraction of real-time data/streams (with the example of Twarql, dynamic model creation, and real-time analysis, http://bit.ly/c-sem).

A variety of techniques used to exploit metadata, support analysis and achieve insights will also be discussed with a related literature review. This will include the unique role of Semantic Web standards, technologies and techniques for building advanced capabilities. [2]

INTENDED AUDIENCE AND COVERAGE

The tutorial will interest both researchers (primarily related to tracks identified below) as well as technologists. The talk covers relatively new frontiers in research demonstrated via examples and also covers research efforts that have matured to large-scale deployments. Although the content analysis portion of the presentation will be discussed in-depth, attendees are not expected to have any specialized background and/or knowledge (use of examples will make it easier to gain technical insights). Those with a background in more traditional text analysis will be able to understand the novel technical challenges dealing with short and informal text as well as other aspects of social data..

From subject/area perspective, this tutorial will interest participants from multiple tracks:
■ Social Systems and Graph Mining — e.g., issues of information diffusion
■ Bridging Structured and Unstructured Data — e.g., extensive and synergistic use of both structured metadata, Web of data, background knowledge and unstructured data
■ Content Analysis — e.g., spatio-temporal-thematic-sentiment-intension analysis with specific challenges in social data
■ Semantic Web — e.g., how use of background knowledge improves machine learning and statistical NLP techniques resulting in improved NER on informal text
■ Web of Emerging Regions — coverage of not only twitter and Web/Mobile Web but also SMS extensively used in developing nations, including applications related to development and crisis management which is of special relevance to emerging regions

The tutorial will provide breadth and depth of research on some topics matched by real-world applications as well as a combination of perspectives from academic, industry and non-governmental organizations.

REFERENCES

[1] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, A. Sheth, Multimodal Social Intelligence in a Real-Time Dashboard System, special issue on 'Data Management and Mining for Social Networks and Social Media', the VLDB Journal, 19 (6), 2010. http://portal.acm.org/citation.cfm?id=1921807

[2] A. Sheth, Citizen Sensing, Social Signals, and Enriching Human Experience, IEEE Internet Computing, July/August 2009, pp. 80-85. http://www.computer.org/portal/web/csdl/doi/10.1109/MIC.2009.77

[3] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju and A. Jadhav, Spatio-Temporal-Thematic Analysis of Citizen-Sensor Data - Challenges and Experiences, 10th Intl Conf on Web Information Systems Engineering, Oct 5-7, 2009, pp. 539 - 553. http://www.springerlink.com/content/q682443076047208/

[4] OSI, Harvard and Columbia Case Studies covering project carried out by Selvam's NGOs: see http://emoksha.org

[5] P. Mendes, A. Passant, P. Kapanipathi, A. Sheth, "Linked Open Social Signals," WI2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI-10), Toronto, Canada, Aug. 31 to Sep. 3, 2010.