Notes From the Social Media Workshop

Building Communities for Transforming Social Media Research Through New Approaches for Collecting, Analyzing, and Exploring Social Media Data

April 10-11, 2014
DIMACS Center, Rutgers University

Organizers

Chirag Shah[1], Mor Naaman[2], Winter Mason[3]
1. Rutgers University, chirags@rutgers.edu
2. Cornell Tech, mor.naaman@cornell.edu
3. Stevens Institute, winter.mason@stevens.edu

This Workshop was presented under the auspices of the DIMACS Special Focus on Information Sharing and Dynamic Data Analysis with support from DIMACS and its partners, and from the National Science Foundation under grant #1144502.

Tools

SOCRATES

SOCIAL & CROWDSOURCED ACTIVITIES EXTRACTION SYSTEM

Friends

INFOSEEKING

PEOPLE ANALYTICS

NSF

Executive Summary

The Social Media Workshop, "*Building Communities for Transforming Social Media Research Through New Approaches for Collecting, Analyzing, and Exploring Social Media Data,*" held at the DIMACS Center at Rutgers University, New Jersey, on April 10-11, 2014, brought together scholars and practitioners across various disciplines, who are working on social media related problems. The purposes of the workshop were (i) to establish a diverse research community, (ii) to share their experience and findings from studies, and (iii) to discuss problems and challenges when studying social media, which can be applied to the design and implementation of the SOCRATES.

Organization of the Workshop
The format of the workshop was organized to enhance sharing knowledge and experience, and discussing relevant challenges. On the first day, twelve 15-min short talks were presented, which provided broad and worthy information from various disciplines. The short talks and the following discussions, which include the topics of what kind of information social media is producing, how the information on social media can be observed and visualized, and what issues the researchers should consider well, became a good opportunity for participants to share their wisdom and opinions. In-depth discussion and social networking continued through following reception and dinner.

At the beginning of the second day session, introduction and demonstration of SOCRATES was presented to the participants. The social-computational platform, which aims to provide researchers with the effective system that enable seamless social media research, was first publicly announced and received attendants' attention.

During the following sessions, three broad topics covering collection, analysis, and exploration of social media data were discussed. The diversity of the participants' backgrounds helped the discussions carry out interesting outcomes. Practical and ethical concerns while conducting social media study are addressed by participants. Also, data collecting, analyzing, and visualizing tools they are currently using were introduced; and suggestions for more effective tools and systems were made through the discussion. The lively and active discussions, with open-ended questions and flexible format, took place among the researchers and concluded with future works and suggestions to the workshop.

Acknowledgements
The organizers and the participants of the workshop are very grateful to the National Science Foundation (NSF) for the extraordinary opportunity to share and discuss the research on social media, and to DIMACS for sponsoring and hosting the workshop. The organizers would also like to thank the NSF for supporting the development of SOCRATES system through BCC-SBE award no. 1244704.

DIMACS Workshop on Building Communities for Transforming Social Media Research Through New Approaches for Collecting, Analyzing, and Exploring Social Media Data

April 10 - 11, 2014
DIMACS Center, CoRE Building, Rutgers University

Organizers:
Chirag Shah, Rutgers University, chirags@rutgers.edu
Mor Naaman, Cornell Tech, mor.naaman@cornell.edu
Winter Mason, Stevens Institute, winter.mason@stevens.edu

Presented under the auspices of the DIMACS Special Focus on Information Sharing and Dynamic Data Analysis.

http://dimacs.rutgers.edu/Workshops/SocialMedia
Social media hashtag: **#smworkshop2014**

Introduction

Social media platforms and services such as Facebook, Twitter, YouTube, FourSquare, Flickr, WordPress, and Tumblr make more data available regarding people's lives, intentions, thoughts, activities, and attitudes than ever before. As these services are shifting the social and communication infrastructure of our society, having proper tools and techniques to study these platforms becomes ever more critical for understanding social activities, public opinion, political action, and more. At the same time, while new social media have been seized upon by researchers in marketing, advertising, and information science researchers, many fields still lack the tools and expertise to use these new social media in their research. Meanwhile, transformative social and political norms are emerging in online social practices and their off-line corollaries. The pace of change and innovation is a significant challenge for researchers armed only with traditional tools and techniques (e.g., spreadsheets). These researchers currently lack the tools to collect and analyze social media at scale.

This workshop brought together researchers and practitioners working on various social media problems across multiple disciplines to better understand various issues experienced while working with social media data. Several of these researchers gave short talks about their fields of interests and findings that relate to their use of social media.

The workshop also brought up lively and active discussions on three broad topics covering collection, analysis, and exploration of social media data.

First Day: Thursday, 4/10/14

We started with a luncheon at noon, followed by gathering at the CoRE building ground floor auditorium for introduction and short talks. After a welcome message from the DIMACS director, Dr. Rebecca Wright (Rutgers), quick introductions of everyone present were made, followed by an introduction message to the workshop by Dr. Chirag Shah (Rutgers). We then proceeded with the scheduled short talk.

The presentations are available to view at:
http://dimacs.rutgers.edu/Workshops/SocialMedia/Slides/slides.html

Short talks – see presentations online for more details

Is it an Epidemic of GIGOitis?
Leonard Hirsch, Smithsonian Institute

terasaur: Gigabytes to Terabytes
Paul Jones, University of North Carolina at Chapel Hill

Wiki Surveys: Open and Quantifiable Social Data Collection
Matthew J. Salganik, Microsoft Research and Princeton University

Coding the Twitter Sphere: Humans and Machines Learning Together
Stuart Shulman, Texifter

Sensing, Understanding, and Shaping Human Behavior
Vivek Singh, Massachusetts Institute of Technology

A Wordcount Approach to Assessing the Moral Color of Old & New Media
John Voiklis, Brown University and Harmony Institute

Automated Discovery and Visualization of Communication Networks from Social Media
Anatoliy Gruzd, Dalhousie University, Canada

Collecting and Connecting On and Offline Political Network Data
Libby Hemphill, Illinois Institute of Technology

Is Content Really King? An Objective Analysis of the Public's Response to Medical Videos on YouTube
Tejas Desai, East Carolina University

Searching for Information in Online Health Communities
Yan Zhang, University of Texas at Austin
Text-mining Social Media to Study Mental and Physical Health
Lyle Ungar, University of Pennsylvania

Taming the Long Tail: Identifying Filtering in Social Media
Augustin Chaintreau, Columbia University

Second Day: Friday, 4/11/14

The second day of the conference started out in the morning with breakfast, followed by an introduction and demonstration of SOCRATES by Kevin Albertson and Ziad Matni (Rutgers).

Subsequently, we convened for breakout sessions where Mor Naaman (Cornell Tech) and Winter Mason (Stevens Institute) led group discussions on three broad topics of collection, analysis, and exploration of social media data via question and group answers.

Q1: If you could have a tool that could collect ANY data in the world, what would it do?

- Physical object -> digital format (data availability)
- Standardizing data
- Contextualizing data
- Making data useful for people (i.e., health)
- All interactions between people (but, privacy issue)
- Cross-platform
- Do
    - Find causality
    - What data I need or when to get it
- What
    - How people feel everyday
    - Read minds
    - Health data from China
    - What people are saying where and then

Q2: Name all the social media sites that you know of to obtain data, in order of importance. Among the list, what is the social media you don't use for research? And why?

NOTES:

a) ***If there's a x*n *next to a name, then* n *indicates the number of times the website was mentioned.***
b) ***Websites* highlighted *in orange are the ones that no one in the audience claimed to use.***

1. Twitter x4
2. Facebook x4
3. Instagram x4
4. Quora
5. Reddit x2
6. LinkedIn x4
7. YouTube x2
8. Pinterest x4
9. Perl Trees
10. Github
11. Adhoc crowdsourcing sites (SBTF)
12. Renren
13. Weibo x2
14. Live Journal
15. Tumblr x2
16. Flickr x2
17. Yelp x4
18. Airbnb
19. Couchsurfing
20. 4chan
21. Meetup x2
22. Free Republic
23. Tagged
24. Ask.fm
25. Spotify
26. Pandora

27. Behance
28. Scratch
29. Snapchat x2
30. WhatsApp x2
31. Google Plus x2
32. Patients Like Me x2
33. Diabetic Connect
34. 23 and Me
35. Skype
36. Orkut
37. Amazon reviews x2
38. Ancestry.com
39. Angie's List
40. WebMD comment
41. Disqus
42. Glympse
43. Grindr
44. Tinder
45. We Chat
46. Vine
47. OK Cupid
48. Craigslist
49. Jezebel comments
50. NYT comments
51. ESPN comments
52. Goodreads
53. Academia.edu
54. Mendeley
55. Figshare
56. Research Gate
57. Stumble Upon
58. eLance
59. Freelancer
60. eBay
61. Netflix
62. Ravelry
63. eHarmony.com
64. MySpace
65. Daily Candy
66. Everyblock
67. Nextdoor
68. Second Life
69. Wikipedia ?
70. Yahoo! Answers ?
71. Wikianswers ?
72. Fark.com ?
73. icanhascheeseburger ?
74. Foursquare
75. Ratemyprofessors.com

Why are some sites not used in research?

- No access
- Lack of good data on them
- Too niche
- Where the people are or aren't
- Availability of tools
- Lack of awareness because of lack of overlap with academia

Q3. What types of data might you want to get from these sites?

- Who's watching - demographics (esp. YouTube)
- Cross-platform user IDs
- Ability to go back and re-collect data
- Online experimentation / A/B testing
- Access to email & SMS data
- Access to historical data
- Content of (dead) links
- Connecting with offline data (importing data from GSS, census, etc...)
- More accurate sentiment
- Measures of trust
- Self-censored texts
- Trustworthiness/accuracy of online or self-reported data?
- Find multiple online personas?
- Time-related data:
    - How people leave & join communities & why do they leave?

    - What is new vs. “repost”
    - When does a group/collaborators become inactive
- Lurkers on the site
- What do people pay attention to?
- What signals of interaction/engagement are meaningful?
- Motivational factors
- Informal communications, esp. in health situations
- How did companies use (my) data?

Q4. What are the ethical concerns researchers should have when collecting publicly available data?

- When do you need consent to use public data?
- How does the data or the user of data have potential to cause harm?
- What are the cultural norms and differences in expectations of what is public vs. private?

- Re: responsibility of researchers to respect the privacy of the subject. What’s good for the subject? Benefits vs. risks.

- It is our ethical responsibility to:
    - understand implicit vs. explicit permissions
    - protect identifying information that “slips out”
    - reveal the context in which we collected data (disclosure)
        - filtering? when did you collect the data? etc…
    - comply with rules for deletion of data

- Re: justice in the studies. How do we ensure that we are doing a just study?
    - Equitable distribution of benefits & costs
    - Limit generalizations to actual study population
    - An unethically conducted study should not be published
    - Attempt to collect data on all populations

- Accessibility of data for replication
- Privacy vs. public need-to-know
- “Ethically sourced data”

Q5. If you could have any set of people to annotate any set of data, what set of people would you like to use?

- People who know information/prediction markets
- Expertise: how do we identify it & is it necessary?
- Does connecting disparate data sources counts as annotation?
    - Would you collect data re: online interactions - is that annotating?
- Get a diverse set of annotators to match the diversity of the problem. Examples:
    - Annotating fine art by experts, as opposed to novices
    - Using undergrads to annotate middle-aged FB users (one group -> other group for perceptions)
- Wisdom of the “wise crowd”
- Sometimes we need people with soft knowledge for annotators

- All possible research projects annotate if it's ethical
- Every Twitter user for every tweet
    - do you understand it? (i.e. generational gap)
    - have you heard about it before? is it new to you?
    - is it interesting to you?
- Coding to classify data vs. to understand data

- Use of domain experts
- Legal perspective for ethics and legality issues
- Societal perspective (sociological)
- Natural Language expert
- Self-annotation (or within group)
- Producers vs. consumers of data

Q6. What mechanisms (i.e. computational methods) have you used for SM data that worked well? That didn't work well? Why? What would you like to do computationally?

- Quantitative: word count, regression, classification, clustering, semantic analysis, exploration by SPSS, cleaning & filtering, correlation, network analysis
- Qualitative: Nvivo, survery, interviews
- Mixed: log data
- Problems in general:
    - Doing without understanding meaning behind it
    - Define v. clearly between "good" & "bad" metrics
    - Problems with evaluation

- Tools:
    - Mechanical Turk
    - Grad students!
    - LIWC - Collection of words including
    - LDA - topics
    - R for analysis
        - SNA, iGraph, ggplot
        - graphics and display
    - Machine learning tools
        - lightside
    - scikit/ : python
    - NetworkX: python lib
    - Gephi (standalone tool)
    - Node XL
    - panda
    - Visualization
        - D3: javascript library
        - Many Eyes (IBM service)
    - Amazon EC2 & RDS
    - Hive/PIG
    - Hadoop
    - DB: MongoDB, MySQL
    - Weka
    - Mallet

    - VW
    - A/B testing -> correl vs. causality
    - Plan Out

- Not v. good with:
    - Sentimental analysis
    - behavioral analysis
    - telling me “how”
    - doing mixed methods

- It would be helpful to have:
    - A mixed method approaches
    - Tools/methods to complement existing/preferences
        - i.e. finding ways to study pictures, video, other non-text data

Q7. What tools are useful? Not useful? Will never be useful?

- HARD
    - Humor
    - Sarcasm
    - Jargon, slang
    - Causality
    - Emotion (sentimental)
    - Intention
    - Expectation
    - Motivation
    - Non-textual expression from people
    - Dealing with difficult text data (handwriting)
    - Extracting contextual metadata (from other than text: such as conversation analysis)
- EASY
    - Lobbying strategy (rhetorical strategies)
    - Topic detection (if in structured environment, low impact of failure)

Q8. What is your favorite instance of data visualization? Why?

- Simple visuals that send clear message and reduce complexity
    - Venn diagrams
    - Single points – non-complex
- Live plasma: artists influence on other artists

- Examples:
    - The ones used in “An Inconvenient Truth” (Al Gore)
    - Baby names - states over time (ManyEyes)
    - Sensible city
    - Raining cabs
    - Sunlight foundation
    - AT&T network data (Global network)
    - 100’ * 50’ (proprietary tech in Basking Ridge, NJ)
    - observe/control traffic during disasters
    - http://hint.fm/wind: map of wind across US
    - NYT visualizations
    - visualcomplexity.com

- Books:
    - Design for information
    - Flowing data
    - Functional art
    - The laws of simplicity
- Visualizations - what is it that makes them useful?
    - Data over time
    - Data with geographic components
    - Interactive for big data

Final thoughts

In reviewing our discussions of the day, we reflected on what we liked or enjoyed, what bothered us in general about this research area, what we would like to see more of, and so on.

We liked the discussion regarding the collection of social media (SM) tools that the workshop participants were using and felt might be useful to them. Some of the more qualitative-oriented participants thought they should explore the quantitative side of things. Some of the more highly mentioned tools from the presentations that people mentioned were the Wikisurvey and tools related to health informatics. The ethics discussion was also mentioned as one most of us enjoyed.

However, we would still like to explore certain questions, like:

- How can you tell if a SM project will succeed?
- How can you tell if SM tool will be re-appropriated for something its not intended for by design?
- What makes the field of SM studies unique?
- Can we identify the core problems?
- What is the best way that can help us characterize the field? Winter Mason suggests that what binds this field is an object of study, as opposed to a method of study.
- How do we best detect misinformation in SM?
- How do we best study communities and detecting communities in SM?
- In educational settings, how do we know if a SM tool is effective in an educational setting?
- As SM has “changed the social contract”, what is the impact of SM on society, health, and education?

Likewise, we wanted to explore more possible tools or methodologies, for example having common datasets to compare different studies against each other. We need to build some system(s) to do SM research with: are researchers’ basic needs being met?

Nice to have for next time: collaboration between academia and industry and government agencies.

We asked if we should do this workshop or something like it in the future? Some suggestions:

- More time for Q&A
- 1.5 days was a good amount of time
- Re: breakout sessions - a little more structure might help
- Having more niche discussions that don’t involve everyone with everything

- Nice to have someone come up with a framework, like from Association of Internet Researchers
- Focus on different domains
- “Storify!”
- Have workshop events around conferences.

We are still grappling with the ethical issues. There are still some grey areas and we need to address them inevitably. This ties into the definition of privacy that seems to be in flux today.