Methods of Social Media Research:
Data Collection & Use in Social Media

Florida State University
College of Communication and Information
Sanghee Oh

shoh@cci.fsu.edu
Overview

• Introduction to myself
• Introduction to social media
• Social media and LIS research
• Social media research methods
  – Data collection from human subjects
  – Data collection from social media
• Exercise
• Data Analysis
Introduction

• Sanghee Oh, PhD
• Assistant professor at Florida State University

• Social informatics
• Health informatics
• Human Computer Interaction
• Human Information Interaction
Social Media Research Projects

- Quality of Health Answers in Social Q&A (Online Evaluation Tool, Librarians, Nurses, and Yahoo! Answers Users)
- Motivations for Information Sharing in Social Media (Surveys, Amazon MTurk Users)
- Use of Social Media in Health (Surveys, undergraduate students, nurses)
- Disease-specific Information Needs & Use (Content Analysis & Data/Text Mining)
- Health-related Web Resource Selection Behaviors of Social Media Users (Data Mining)
Introduction to Social Media

- Web 2.0 applications which support user content-creation and collaboration
- People seek and share ideas, information, experiences, expertise, opinions, and emotion with both acquaintance and strangers on the Internet, based on the effect of the Wisdom of Crowds (Surowiecki, 2004).

- Definition of Social Network Site (SNS):
  - “A web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.” (Boyd & Ellison, 2007, p. 211)
Social Media Demographics (USA, 2012) by Online MBA

- Facebook: 845 Million Act. Users
- Linkedin: 150 Million Reg. Users
- Twitter: 127 Million Act. Users
- Google+: 90 Million Uniq. Visitors
- Pinterest: 21 Million Uniq. Visitors
Social Media

- collaborative filtering
- social tagging
- file sharing
- instant messaging
- podcasting, and multi-player games
What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.
- People create, publish or distribute information in the form of text, photos, video, audio, or tweets.
- People share bookmarks, presentation slides, or other files.
- People provide feedback on or rate others’ information.
- People create social tags or folksonomies.
Why People Use Social Media (U.S. only)

- Stay in touch with friends: 67%
- Stay in touch with family: 64%
- Reconnect with old friends: 50%
- Connect with people who share similar interests: 14%
- Make new friends: 9%
- Read comments from celebrities, athletes, or politicians: 5%
- Find potential romantic partners: 3%
Why People Share Information with Others in Social Media (Oh, 2011)

- Enjoyment
- Efficacy
- Learning
- Personal Gain
- Altruism
- Empathy
- Community Interest
- Social Engagement
- Reputation
- Reciprocity

- Other studies about motivations (Nam, Ackerman, & Adamic, 2009; Raban & Harper, 2008)
Social Media in LIS Research

• Research has been actively conducted since 2006.

• About 500 articles were retrieved by Titles and Abstracts search of peer-reviewed journal articles about “social media” from two major LIS databases -- Library, Information Science, Technology & Technology (LISTA), and Library Literature & Information Science (LL).
Tag Cloud from Social Media Papers

Applications
Tools
Technology
Services
Media
Software

Open
Participation
Collaborative
Sharing
Communication
Networking

Create
Posted
Published
Review

Health
Medical
Policy
Learning
Marketing
Business
Organization
University

Blogs
Twitter
Facebook
YouTube
2011 Queensland Floods (QFs)

  - Surveyed 428 people from several Facebook groups about QFs

  - Interviewed 4 people in QUT about their experiences of using social media during the floods.

- Bruns, A., Burgess, J. E., Crawford, K., & Shaw, F. (2012) #qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods. ARC Centre of Excellence for Creative Industries and Innovation, QUT, Brisbane QLD Australia.
  - Captured 35,000 tweets using the #qldfloods hashtags posted during 10-16, Jan 2011
Social Media Research Methods

• Interviews
• Surveys
• Content analysis
• Data mining

• Data from human subjects
• Data from social media
Data from Human Subjects (Interviews & Surveys)

- Teenagers
- College students
- Nurses
- Patients
- Librarians

What are benefits and challenges of doing interviews or surveys in social media research?
# Data from Human Subjects

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Easy to recruit participants</td>
<td>• Hard to reach social media users in some cases</td>
</tr>
<tr>
<td>• Easy to obtain responses directly from personal experiences</td>
<td>• Memories from personal experiences can be manipulated.</td>
</tr>
<tr>
<td>• Good to have exploratory findings for a follow-up study</td>
<td>• Time-consuming</td>
</tr>
<tr>
<td>• Good to find out about why</td>
<td>• Expensive</td>
</tr>
</tbody>
</table>
Example Study 1: Use of Social Media

- Populations
  - Undergraduate students
  - Nurses

- Asked if they use social media for health information (or not)
- Generated associated questions to social media chosen by the participant
Select all social media you use for health information.

- Social networking sites (e.g., Facebook, Google+, MySpace)
- Social Q&A (e.g., Yahoo! Answers, AnswerBag, WikiAnswers)
- Twitter
- Blogs
- Podcasts/Videocasts (e.g., YouTube, Bing Video, iTune Podcasts, NPR Podcasts, etc)
- Mobile Apps
Why do you not use social media? Mark all that apply.

☐ Do not know how
☐ Do not trust
☐ No access to devices such as computer, tablet PC, and smartphone
☐ No access to Internet
☐ Not interested
☐ Not useful
☐ Privacy concern
☐ Unreliable resources
☐ Waste of time
☐ Other reasons.
   Please explain.
Example Study 2: Motivation Study

• What are the motivations of people for sharing information in social media?

• Social Media Users
  – Yahoo! Answers
  – Delicious
  – Twitter
  – Facebook
  – YouTube
  – Flickr

• How do you reach and contact the social media users in order to recruit them to your surveys or interviews?
Yahoo! Answers Users

C.M. C
Member since: October 09, 2008
Add Contact  Block
View your Yahoo! profile
Email C.M. C
20% Best Answer

fruitsalad
Member since: March 07, 2008
Add Contact  Block
View your Yahoo! profile
17% Best Answer

About me
39 year old Australian optometrist
Currently studying construction management for a career change
Mother of two boys
Keen netball player
<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoho</td>
<td>226 saves <a href="https://www.zoho.com/">https://www.zoho.com/</a></td>
</tr>
<tr>
<td>Apache OpenOffice</td>
<td>4 saves <a href="http://www.openoffice.org/why/">http://www.openoffice.org/why/</a></td>
</tr>
<tr>
<td>National Guideline Clearinghouse</td>
<td>Home</td>
</tr>
</tbody>
</table>
Amazon Mechanical Turk (MTurk)

- Amazon Web Services
- Workers (Survey Panel)
- Requesters (Researchers)
- HITs (Human Intelligent Tasks)
- The workers primarily locate in the U.S., but could be any country. It is known that the demographics of workers are similar to the overall Internet population in the U.S.

The Turk
An automated chess playing machine of the 18th centuries in Europe

An automated chess playing machine of the 18th centuries in Europe
MTurk Interface Example

Manage HITs

Managing Your HITs

Your HITs are available for Workers to complete until all assignments are completed, the HIT expires, or you decide to expire your HIT early.

The HIT management interface allows you to review and approve or reject completed assignments, download results, and extend your HITs by adding more assignments or extending the expiration time.

A Short Survey on Motivations for using Delicious

Requester:  slis cua
HIT Expiration Date:  Jul 13 2012, 08:45 PM PDT
Reward:  $0.25
Assignments Requested:  223

Assignments Pending Review:  0
Reviewed Assignments:  212
Remaining Assignments:  11
Remaining Time:  Expired  Add time

Download results
Other ways to contact social media users

• Snowball sampling or chain sampling
Data from Social Media
(Content Analysis & Text Mining)

• Facebook walls
• Tweets
• Delicious bookmarks
• YouTube comments
• Flickr comments, tags
• Yahoo! Answers questions, answers, comments, ratings
• Blog articles
• Wiki articles

• What are benefits of conducting content analysis or text mining in social media research?
Data from Social Media

**Benefits**

• Obtain natural/raw data on human information behaviors
• Easy to access massive amount of data to analyze
• Good to conduct a longitudinal study, if timestamp is available
• Easy to set parameters for the purposes of the study

**Challenges**

• Hard to collect data (Need to know tools and scripts)
• Need to clean data before use (Time consuming)
Example Study 3: Information Needs on STDs

• Health topics
  – Sexually Transmitted Disease(s)
  – Cancer

• How are the information needs articulated in questions of social Q&A?

• Analyzed questions (and answers) on certain topic.
• Randomly selected questions (and answers) from the primary data by limiting time (the most recent) or other conditions, such as rates, best answer status, and number of answers.
Example Study 3: Web Resource Selection

• What kinds of web resources do social media users -- Yahoo! Answers, Delicious, and Twitter -- choose, collect and share with others?

• Extracted URLs from contents in social media
Yahoo! Answers

- How do you want to collect data from Yahoo! Answers?

Yahoo! Answers API (Application Program Interface) allows to request 5,000 queries per day (basically it collects 5,000 questions and associated answers per day)

Yahoo! Answers allow you to collect, question, answers, user info, ratings, comments, time stamp, topics, sub-topics, number of answers, etc.
Yahoo! Answers API

Yahoo! Answers is a place where people ask and answer questions on any topic. The Answers API lets you tap into the collective knowledge of millions of Yahoo! users. Search for expert advice on any topic, watch for new questions in the Answers categories of your choice, and keep track of fresh content from your favorite Answers experts.

Sample data through YQL

```
select * from answers.getbycategory where category_id=2115500137 and type="resolved"
```

```
<?xml version="1.0" encoding="UTF-8"?>
<query xmlns="http://www.yahoolinks.com/v1/base.rng"
      xmlns:yahoo="http://www.yahooapis.com/v1/base.rng"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

<diagnostics>
  <proxy>DEFAULT</proxy>
  <![CDATA[http://answers.yahooapis.com/AnswersService/V1/getByCategory?category_id=2115500137&type=resolved&start=0&results=10]]>
  <url execution-start-time="26" execution-stop-time="792" execution-time="766"
       proxy="DEFAULT">
    <user-time>793</user-time>
    <service-time>766</service-time>
    <build-version>20958</build-version>
  </url>
</diagnostics>

<results>
  <Question xmlns="urn:yahoo:answers" id="20120717181149AAxzs88" type="Answered">
    <Subject>Who is willing to to help with writing a Ballad on Iliad, Homer?</Subject>
    <Answer>Can you please provide more information about the ballad you want to write? Are you looking for assistance on how to structure the story or do you have specific ideas about the plot and characters? Providing more details will help me give you better advice.
  </Question>

```
```
Data from Yahoo! Answers API

• Return all answered questions from a category (getBycategory)
  o question ID, category, question, question description, posting date, number of answers,
  o questioner ID, questioner nickname,
  o chosen answer (best answer), chosen answerer ID, chosen answerer nickname, chosen answer posting date, chosen answer rating

• Return all answers and information from a specific question
  o question ID, answer, sources, posting date,
  o answerer ID, answerer nickname

• Return all questions from a specific user
```
use socialqa;
select * from qdataH Limit 10;
```
use socialqa;
select * from qdataH Limit 10;
```sql
select * from adataH Limit 10;
```
More tools are available.

- **Twitter**
  - yourTwapperKeeper: PHP scripts to run data collection on a server space

- **Delicious**
  - Developer Page ([http://delicious.com/developers](http://delicious.com/developers))
    - APIs, or RSS & JASON Feeds

- **YouTube**
  - Developer Page ([https://developers.google.com/youtube/](https://developers.google.com/youtube/))

- **Flickr**
XML and JSON

- JSON: JavaScript Object Notation (http://json.org/)
- a lightweight data-interchange format
- easy for humans to read and write
- easy for machines to parse and generate
- a text format that is completely program language independent
XML Example

Sample data through YQL

```
select * from answers.getbycategory where category_id=2115500137 and type="resolved"
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<query xmlns:yahoo="http://www.yahooapis.com/v1/base.rng"
    <diagnostics>
        <publiclyCallable>true</publiclyCallable>
        <url execution-start-time="60" execution-stop-time="829"
            execution-time="769"
            proxy="DEFAULT"><![CDATA[http://answers.yahooapis.com/AnswersService/V1/getByCategory?category_id=2115500137&type=resolved&start=0&results=10]]></url>
        <user-time>830</user-time>
        <service-time>803</service-time>
        <build-version>28958</build-version>
    </diagnostics>
    <results>
        <Question xmlns="urn:yahoo:answers" id="20120718132150A0CPIK" type="Answered">
            <Subject>Is it possible to delete a resolved question?</Subject>
            <Content>I had a question I posted get deleted and it had been resolved for a few weeks. I have tried to delete other questions just to see if it is possible to delete resolved questions but I either don't know how or maybe they can't be deleted. It was not reported or if it was I have not received any a mails and there is no reason it would have been removed except...
            </Content>
        </Question>
    </results>
</query>
```
JSON Example

Sample data through YQL

```
select * from answers.getbycategory where category_id=2115500137 and type="resolved"
```

```
"build-version": "28958",
"results": {
"Question": [
{
"id": "20120718132150AA0CPIK",
"type": "Answered",
"Subject": "Is it possible to delete a resolved question?",
"Content": "I had a question I posted get deleted and it had been resolved for",
"Date": "2012-07-18 13:21:51",
"Timestamp": "1342642911",
"Link": "http://answers.yahoo.com/question/?qid=20120718132150AA0CPIK",
"Category": {
"id": "2115500137",
"content": "Poetry"
},
"UserId": "OP2Zc5thaa",
"UserName": null,
"UserPhotoURL": null,
"NumAnswers": "3",
"NumComments": "0",
"ChosenAnswer": "The only way I know of to have resolved questions deleted is "
}
}
```
Exercise

• Find data collection (data archiving) methods for Twitter, Delicious, YouTube, and Flickr.

• Review the assigned social media thoroughly and identify data entries that you would like to observe.

• Discuss your findings and experiences with others.
diabetes – One of the widespread disease and medical emergency

Created on Sat, 19 May 2012 14:02:46 -0400 and total number of tweets = 378013

tags:

@alexiagironp RT @soniayates: ya no tendre anemia ahora diabetes ays!
Wed, 18 Jul 2012 04:55:43 +0000 - tweet id 2.23499498024706+17 - #1

@edisontuazon Time to get fit.. Bka kse manahin ko ung diabetes ng tatay ko :]
Wed, 18 Jul 2012 04:38:39 +0000 - tweet id 2.23499482062287E+17 - #2
@lisforakyma RT @A_for_Anetta: boladona na diabetes

Sun, 15 Jul 2012 03:39:33 +0000 - tweet id 2.2435246813747741e+17 - #1

@peakwriter RT @FingersFlying: The 'Pre-Existing Condition' Hoax in #ObamaTax #ObamaCare #Obamacare #health #illness #cancer #Diabetes #insurance

Sun, 15 Jul 2012 03:59:31 +0000 - tweet id 2.2435246777579678e+17 - #2

@RetaVerca Diabetes RT @TwitPelajaran @dindaa12_ Hipoglikemi adalah obat yang digunakan untuk ... #MEDIS

Sun, 15 Jul 2012 03:59:31 +0000 - tweet id 2.24352468100388e+17 - #3

@ChristinaJBS5 @emmymee6 not enough fladidos...... #tapeworm #diabetes

Sun, 15 Jul 2012 03:59:31 +0000 - tweet id 2.24352468100388e+17 - #4

@SarahThompson42 RT @baileystanfield: This one time... at diabetes comp.. #sweeneyprobs

Sun, 15 Jul 2012 03:59:32 +0000 - tweet id 2.2435243116371e+17 - #5
Data Analysis

• Content Analysis
  – Need to understand from the contexts
  – Need to understand language that people use in social context

• Data/Text Mining
  – Need to develop a customized dictionary
  – Need to develop a customized categories for keyword searching and analysis
# Content Analysis

<table>
<thead>
<tr>
<th>Core Categories</th>
<th>Definitions &amp; Descriptions</th>
<th>Sub-Categories</th>
<th>Cancer</th>
<th>STDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Information</td>
<td>• Demographic information of patients and care givers</td>
<td>Sex, age, relationship between questions and patients or potential patients in questions</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Disease-specific Information</td>
<td>• Disease-related information in which people explained about the conditions of patients or potential patients</td>
<td>Prevention/causes, symptoms, tests, treatments, diagnoses</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Provided)</td>
<td></td>
<td>Medical records, family medical history, food/exercise, smoking/alcohol, environmental factors</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Other information</td>
<td>• Information that people share about their personal records or lifestyles for maintaining health</td>
<td>Prevention, symptoms, diagnoses, tests, treatment, prognoses</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Provided)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diseases-specific Information</td>
<td>• Information asked for by patients or potential patients about the disease that they are suspicious about having or for which they are already diagnosed</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Asked)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socio-emotional Information</td>
<td>• Information asked for by patients or potential patients about the ways of handling their situations emotionally</td>
<td>Isolation, acceptance, social supports, coping</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Asked)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Life Information</td>
<td>• Information asked for by patients or potential patients for maintaining healthy lives</td>
<td>Alcohol, smoking, environmental factors,</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Asked)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk Factors</td>
<td>• Risk factors causing STDs (sexual intercourse or other behaviors)</td>
<td>Unprotected sex, multiple partners, early age of sexual onset, alcohol use, illicit drug use, etc.</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>
Data/Text Mining

- SPSS Modeler / SAS Enterprise Miner
- Rapid-I
- AlchemyAPI
  (http://www.alchemyapi.com/api/demo.html)

- Word Frequency
  - Basic Resources
  - MeSH
  - Custom categories
- Tag Cloud
- Info Visualization
Thank you!

• Sanghee Oh, Assistant Professor at Florida State University (FSU)

• Contact Information
  o Office: 1-850-645-2493
  o Email: shoh@cci.fsu.edu
  o Personal Website: http://shoh.cci.fsu.edu
  o Research Website: http://socialqa.cci.fsu.edu