2009

User-Generated Content on Social Media
Challenges, Opportunities

Meenakshi Nagarajan
Wright State University - Main Campus

Follow this and additional works at: https://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
https://corescholar.libraries.wright.edu/knoesis/92

This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
User-Generated Content on Social Media Challenges, Opportunities

Meena Nagarajan, KNO.E.SIS, Wright State
meena@knoesis.org, http://knoesis.wright.edu/researchers/meena/
The Shift in the rules of the game

- Online Media: Packaged Goods Media to a Conversational Media
- Variety of networked interactions, many in near real-time
- Information economy: from dearth of signals to plenty much!
Social Media Investigations

• **Network**: Social structure emerges from the aggregate of relationships (ties)

• **People**: poster identities, the active effort of accomplishing interaction

• **Content**: studying the content of communication. "Who says what, to whom, why, to what extent and with what effect?" [Laswell]
Effects of Networked Publics

• Certain social phenomenon admittedly more complex
  • begs for a people-content-network confluence

• Micro-level variations of Content-people-network on macro-level features
  • “How do the topic of discussion, emotional charge of a conversation, poster characteristics & network connections affect ....?”
People-Content-Network - Possibilities

• Emerging social order in online conversations

• How are the people-content-network dynamics shaping online conversations?

• Can we understand the Influentials theory, information diffusion properties in networks (etc.) while taking people and content into account?
• The point is that we need a strong grasp on the micro-level variables of the content, people and network dimensions to begin explaining what they are doing to any social phenomenon...

• My focus is on the micro-level variables in the content dimension.
Mapping User-generated Content to Context
Dimensions Of Analysis

WHAT

- What are the Named Entities and topics that people are making references to?
- How are they interpreting any situation in local contexts and supporting them in their variable observations?

- Named Entity Identification and Disambiguation
- Cultural Named Entities
- Music artist, track named entities (IBM) [ISWC09a, VLDB09], Movie named entities (MSR) [WWW2010]
- Summaries of user perceptions behind real-time events from Twitter
Dimensions Of Analysis

• What are the Named Entities and topics that people are making references to?

• How are they interpreting any situation in local contexts and supporting them in their variable observations?
Dimensions Of Analysis

• What are the diverse intentions that produce the diverse content on social media?
• Why we share by looking at what we predominantly do with the medium. Value derived, repurposing..
• Emotion, sentiment expressions..
Dimensions Of Analysis

- Mapping User Intentions
- Information Seeking, Sharing, Transactional intents \([\text{WI09}]\)
- What is the intention landscape of social media
- where is the monetization potential
Dimensions Of Analysis

- What do word usages tell us about an active population, or about the medium?
- Dynamics of a conversation - snubs, flaming words, coordination.. or lack thereof!
- Self-presentation in Online Dating Profiles (with Prof. Marti Hearst, UC Berkeley) [ICWSM09]
Dimensions Of Analysis

• What do word usages tell us about an active population?

Self-presentation

Dynamics of a conversation - snubs, flaming words, coordination.. or lack thereof!

The Social Media Content Landscape..
Population, Medium Diversity
Some mediation
Rate of exchange (asynchronous, synchronous)
Many-to-many reach
Shared Contexts
Slangs, abbreviations, grammar, spelling, media-specific vocabulary
Interpersonal interactions
Variety & Formality

Formality Score = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100)/2

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>FORMALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nat Broadcast Reportage</td>
<td>62.2</td>
</tr>
<tr>
<td>Informational writing</td>
<td>61</td>
</tr>
<tr>
<td>Academic Social Science</td>
<td>60.6</td>
</tr>
<tr>
<td>Writing</td>
<td>58</td>
</tr>
<tr>
<td>Professional Letters</td>
<td>57.5</td>
</tr>
<tr>
<td>Non Acad Social Science</td>
<td>56.9</td>
</tr>
<tr>
<td>Broadcasts</td>
<td>55</td>
</tr>
<tr>
<td>Blog corpus</td>
<td>53.3</td>
</tr>
<tr>
<td>Scripted Speech</td>
<td>53</td>
</tr>
<tr>
<td>Email Corpus</td>
<td>50.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>FORMALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepared speeches</td>
<td>50</td>
</tr>
<tr>
<td>Personal Letters</td>
<td>49.7</td>
</tr>
<tr>
<td>Imaginative writing</td>
<td>47</td>
</tr>
<tr>
<td>Fiction Prose</td>
<td>46.3</td>
</tr>
<tr>
<td>Interviews</td>
<td>46</td>
</tr>
<tr>
<td>Unscripted Speeches</td>
<td>44.4</td>
</tr>
<tr>
<td>Spontaneous speech</td>
<td>44</td>
</tr>
<tr>
<td>Conversations</td>
<td>38</td>
</tr>
<tr>
<td>Phone Conversations</td>
<td>36</td>
</tr>
</tbody>
</table>

Weblogs, Genres and Individual Differences: How bloggers write for who they write for; Scott Nowson
Variety & Formality

Formality Score = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100)/2 *

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>FORMALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nat Broadcast Reportage</td>
<td>62.2</td>
</tr>
<tr>
<td>Informational writing</td>
<td>61</td>
</tr>
<tr>
<td>Academic Social Science Writing</td>
<td>60.6</td>
</tr>
<tr>
<td>Writing</td>
<td>58</td>
</tr>
<tr>
<td>Professional Letters</td>
<td>57.5</td>
</tr>
<tr>
<td>Non Acad Social Science Broadcasts</td>
<td>56.9</td>
</tr>
<tr>
<td>Broadcasts</td>
<td>55</td>
</tr>
<tr>
<td>Blog corpus</td>
<td>53.3</td>
</tr>
<tr>
<td>Scripted Speech</td>
<td>53</td>
</tr>
<tr>
<td>Email Corpus</td>
<td>50.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>FORMALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcasts</td>
<td>55</td>
</tr>
<tr>
<td>Blog corpus</td>
<td>53.3</td>
</tr>
<tr>
<td>Scripted Speech</td>
<td>53</td>
</tr>
<tr>
<td>Email Corpus</td>
<td>50.8</td>
</tr>
<tr>
<td>Critic Music reviews from <a href="http://www.metacritic.com/music/">www.metacritic.com/music/</a></td>
<td>50.13</td>
</tr>
<tr>
<td>Yahoo Personals AboutMe</td>
<td>50.10</td>
</tr>
<tr>
<td>MySpace About Me</td>
<td>50.07</td>
</tr>
<tr>
<td>MySpace - comments on Artist Pages</td>
<td>50.06</td>
</tr>
<tr>
<td>Prepared speeches</td>
<td>50</td>
</tr>
<tr>
<td>Personal Letters</td>
<td>49.7</td>
</tr>
<tr>
<td>Twitter</td>
<td>49.46</td>
</tr>
<tr>
<td>Facebook posts</td>
<td>48.20</td>
</tr>
<tr>
<td>Imaginative writing</td>
<td>47</td>
</tr>
<tr>
<td>Fiction Prose</td>
<td>46.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TYPE OF DATA</th>
<th>FORMALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepared speeches</td>
<td>50</td>
</tr>
<tr>
<td>Personal Letters</td>
<td>49.7</td>
</tr>
<tr>
<td>Imaginative writing</td>
<td>47</td>
</tr>
<tr>
<td>Fiction Prose</td>
<td>46.3</td>
</tr>
<tr>
<td>Interviews</td>
<td>46</td>
</tr>
<tr>
<td>Unscripted Speeches</td>
<td>44.4</td>
</tr>
<tr>
<td>Spontaneous speech</td>
<td>44</td>
</tr>
<tr>
<td>Conversations</td>
<td>38</td>
</tr>
<tr>
<td>Phone Conversations</td>
<td>36</td>
</tr>
</tbody>
</table>

Weblogs, Genres and Individual Differences: How bloggers write for who they write for; Scott Nowson
Making up for lack of context..

- Supplement what the data is showing you with what you already know..
- Statistical NLP + Contextual Knowledge
- Ontologies, Taxonomies, Dictionaries, social medium, shared spatio-temporal contexts..
Representative Efforts

WHAT

WHY

HOW
Cultural NER
It was THE HANGOVER of the year..lasted forever.. so I went to the movies..bad choice picking “GI Jane” worse now
It was THE HANGOVER of the year..lasted forever.. so I went to the movies..bad choice picking “GI Jane” worse now

LOVED UR MUSIC YESTERDAY!
I decided to check out the *Wanted* demo today even though I really did not like the movie minus Mrs Jolie a.k.a Fox of course!

*It was THE HANGOVER* of the year..lasted forever.. so I went to the movies..bad choice picking “GI Jane” worse now

LOVED UR MUSIC YESTERDAY!
It was **THE HANGOVER** of the year..lasted forever.. so I went to the movies..bad choice picking “GI Jane” worse now

I decided to check out the *Wanted* demo today even though I really did not like the movie minus Mrs Jolie a.k.a Fox of course!

**LOVED UR MUSIC YESTERDAY!**

Obama the *Dark Knight* of socialism.. the man is not as impressive as Ledger yea
Intuitions..

• Spotting and Sense Identification

• Open vs. Closed world
  • unlike person, location, named entities, contexts and senses change fairly rapidly

• We assume an open-world wrt senses
  • No comprehensive sense knowledge base
  • Reduce it to a spotting and binary sense classification problem

It was THE HANGOVER of the year..lasted forever.. so I went to the movies..bad choice picking “GI Jane” worse now
Two flavors..

- Artist and tracks spotting in MySpace music forums
  - using the MusicBrainz Taxonomy
  - with Daniel Gruhl, Jan Pieper, Christine Robson, IBM Almaden, Amit Sheth, Knoesis [ISWC09a]
  - on Thursday Oct 29, Session: Discovering Semantics

- Movie names from Weblogs
  - with Amir Padovitz, Social Streams MSR, [WWW2010]
Cultural NER in Weblogs

- Goal: Supplement classifiers with information that will help them disambiguate the reference of a term better!
- A Complexity of Extraction measure associated with an entity in target sense in a corpus
- with all cues equal, systems that are ‘complexity aware’ will treat cues differently
Measure of Extraction Complexity

- Feature extraction: Graph-based spreading activation and clustering
  - entity sense definition from Wikipedia + evidence a corpus presents for the target sense of the entity
- Ranked list speaks for itself
- More varied senses and contexts, implies higher extraction complexity
1500+ hand-labeled data points
Blue: basic features
Red: with Entropy baseline
Green: with our Complexity of Extraction feature

X axis: precision
Y axis: recall

Feature as a Prior
Decision Tree and Boosting Classifiers
As a Prior in Binary Classification

Average F-measure over 1000 decision tree, boosting models

Average Accuracy over 1000 decision tree, boosting models

1500+ hand-labeled data points
Blue: basic features
Red: with Entropy baseline
Green: with our Complexity of Extraction feature
To chew on..

• The concept of ‘Extraction Complexity’ as an additional prior is very promising

• applies to general NER
User Intention Mapping

• Unlike Web search intent, entity alone is insufficient to characterize intent here.

• Three broad intentions: information seeking, sharing, transactional, combinations thereof.

  • ‘i am thinking of getting X’ (transactional)
  • ‘i like my new X’ (information sharing)
  • ‘what do you think about X’ (information seeking)
• Resorted to ‘action patterns’ surrounding named entities
  • “where can i find a psp cam..”

• A minimally supervised bootstrapping algorithm
  • 10 seed action patterns, learn new ones from unannotated corpus, relying on a empirical and semantic similarity with seed patterns
  • semantic similarity from communicative functions of words Linguistic Inquiry Word Count (www.LIWC.net)
Information Seeking, Transactional

• Patterns learned using 8000 uncategorized posts on MySpace forums

  Sample learned patterns
  does anyone know how
  know where i can
  was wondering if someone
  Im not sure how
  someone tell me how

• Intent recognition recall using pre-classified user posts from Facebook Marketplace (to buy): 81%
Impact on Online Advertising?

• Generate ads from user profile (interests, hobbies) or from posts with monetizable intents?

Personal Information

Interests: NCAA Football, NFL, Carnatic Classical Music, Blogging, Random Computer programming stuff

Karthik Gomadam is looking for good and cheap body shops around Dayton to mount a new door
9 April 2 at 12:24am · Comment · Like
Targeted Content Delivery Platform

- Of all the ads generated using profile (hobbies, interests) information, 7% received attention.
- Ads generated using authored, monetizable posts, 59% received attention.

More at [WI09], Beyond Search and Internet Economics Workshop, MSR, Redmond, WA

Self-Presentation

• On Online-dating profiles [ICWSM09] (with Prof. Marti Hearst, UCB)

• quantifying usages of words from linguistic, personal and psychological categories in LIWC

• Exploratory Factor Analysis to identify systematic co-occurrence patterns among LIWC variables

• grouping user profiles on the basis of their shared multi-dimensional features to compare and contrast self-presentation
Imitate to Impress!? 

• More similarities than differences
• Men displaying a higher usage of tentative words (maybe, perhaps..)
  • typically attributed to feminine discourse
• Many similarities in word combinations and words used!
• Perhaps, self-expression tends towards attempting homophily in online dating..
Science, Fun and Profit
“A pioneering project to tap into the online buzz surrounding artists and songs, by leveraging several popular online sources”

De-spam, slang transliterations, entity identification, voting theory to combine multi-modal online data sources [ICSC08a, VLDB09]

Real-time user perceptions as the fulcrum for browsing the Web [ISWC09b]
Iran elections: Discussions in the US and Iran on the same day

The mystery of Soylent Green: information where you can use it
Come see & play with Twitris @ the International Semantic Web Challenge at ISWC ’09

http://twitris.knoesis.org
Thank You!

Google, Bing, Yahoo: Meena Nagarajan
meena@knoesis.org
http://knoesis.wright.edu/researchers/meena
http://knoesis.wright.edu/researchers/meena/pubs.php


[VLDB09] Daniel Gruhl, Meenakshi Nagarajan, Jan Pieper, Christine Robson, Amit Sheth, Multimodal Social Intelligence in a Realtime Dashboard System, Pending Review, VLDB Journal, Special Issue on "Data Management and Mining on Social Networks and Social Media", 2009.


