From $\text{AI}^K$ to $\text{AI}^D$: Acquiring Social Media Intelligence via `Big’ Data

Huan Liu
Thanks to Former & Current PhD Students

- Reza Zafarani, Asst Prof, Syracuse U
- Xia Hu, Asst Prof, Texas A&M U
- Magdiel Galan, Intel
- Shamanth Kumar, Castlight Health
- Pritam Gundecha, IBM Res Almaden
- Jiliang Tang, Asst Prof, MSU
- Huiji Gao, LinkedIn
- Ali Abbasi, Machine Zone
- Salem Alelyani, Asst Prof, King Khalid U
- Xufei Wang, LinkedIn
- Geoffrey Barbier, AFRL
- Lei Tang, Clari
- Zheng Zhao, Google
- Nitin Agarwal, Chair Prof, UALR
- Sai Moturu, PostDoc, MIT Media Lab
- Lei Yu, Assc Prof, Binghamton U, NY

- Robert Trevino, AFRL
- Yunzhong Liu, LeEco, US
- Somnath Shahapurkar, FICO
- Fred Morstatter, USC ISI
- Christophe Faucon
- Isaac Jones
- Suhas Ranganath
- Suhang Wang
- Tahora Nazer
- Jundong Li
- Liang Wu
- Ghazaleh Beigi
- Kai Shu
- Justin Sampson

Citations per year


Citations: 0, 1500, 3000, 4500, 6000
A Tortuous but Fortuitous Path to Social Computing

1989-1993
- Self-generating Systems
- Neural-Network Learning
- Feature-Selection
- Supervised Artificial-Intelligence
- Robotic-Grasping

1993-1999
- Self-generating Oblique-Trees
- Backpropagation Inductive-Learning
- Scalable-Feature-Selection Concept-Drifts
- Knowledge-discovery

2008-Now
- Microblogging
- Spammer-Detection
- Sentiment-Analysis
- Social-Networks
- Crowd-Detecting
- Malicious-Users

Until 1989
- Multi-dimensional task-requirements
- Dexterous Neural-Networks
- Robotic-hand-eye-coordination
- Robotic-hand-control
- Robotics Systems
- Automation grasping

1989-1993
- Neural-Network Learning
- Feature-Selection
- Supervised Artificial-Intelligence
- Robotic-Grasping

1993-1999
- Self-generating Oblique-Trees
- Backpropagation Inductive-Learning
- Scalable-Feature-Selection Concept-Drifts
- Knowledge-discovery

2008-Now
- Microblogging
- Spammer-Detection
- Sentiment-Analysis
- Social-Networks
- Crowd-Detecting
- Malicious-Users

Until 1989
- Multi-dimensional task-requirements
- Dexterous Neural-Networks
- Robotic-hand-eye-coordination
- Robotic-hand-control
- Robotics Systems
- Automation grasping

1989-1993
- Neural-Network Learning
- Feature-Selection
- Supervised Artificial-Intelligence
- Robotic-Grasping

1993-1999
- Self-generating Oblique-Trees
- Backpropagation Inductive-Learning
- Scalable-Feature-Selection Concept-Drifts
- Knowledge-discovery

2008-Now
- Microblogging
- Spammer-Detection
- Sentiment-Analysis
- Social-Networks
- Crowd-Detecting
- Malicious-Users

Until 1989
- Multi-dimensional task-requirements
- Dexterous Neural-Networks
- Robotic-hand-eye-coordination
- Robotic-hand-control
- Robotics Systems
- Automation grasping
From AI\textsuperscript{K} to AI\textsuperscript{D}

• “Knowledge is Power”: AI was then solely about \textbf{K}
  – Expert Systems or Rule-based Systems
    • “Intelligence is ten million rules.”
  – Knowledge-based Systems (Cyc)

• “Data is the New Oil”: AI is now hyped up with \textbf{D}
  – Big data is ubiquitous
  – CS, Statistics, Information Science $\rightarrow$ Data Science

• Recent surge of AI is powered by Data
  – Machine Learning (including Deep Learning)
  – For any learning algorithm to work, data is key
Big Social Media Data

- **Twitter**
  - 300 million users
  - 500 million tweets / day
  - 1% (5 million) released for research
- **Facebook**
  - 2 billion users
  - 422 million updates / day
  - 196 million photos / day
- **Instagram**
  - 700 million users
  - 80 million photos / day

![Facebook Degree Distribution](image1.png)

![Instagram Users over Time](image2.png)
Discovering Social Media Intelligence

• Graph Theories
• Network Measures and Models
• Data Mining, NLP, and Visual Analytics
• Community Detection and Analysis
• Information Diffusion
• Influence and Homophily
• Recommender Systems
• Behavior Analytics
  – Sentiment Analysis
Some Challenges in Acquiring SM Intelligence

• Social media data seems really big, but why are we often still short of data?
  – How can we make data ‘bigger’?

• Data is power, so it can produce any result
  – Can we algorithmically evaluate the results from big data?

• We don’t know what we don’t know
  – How can we know if our result of social media analysis is of any value?
Making Big Data “Bigger”

• What is big data?
  – A conventional answer is 4Vs
  – A practitioner’s answer is more nuanced

• Big data can be actually little or thin

• For machine learning or data mining to work, *the more data, the better*
  – Make little data bigger
  – Make thin data thicker
Curse of Dimensionality: Required Samples

- Sparsity becomes exponentially worse as feature dimensionality increases
  - Conventional distance metric becomes ineffective as far and near neighbors have similar distances

3 samples per unit region  1 sample per region  1/3 sample per region

http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/
Relevant, Redundant and Irrelevant Features

- Feature selection retains relevant features for learning and removes redundant or irrelevant ones.
- For a binary classification task below, $f_1$ is relevant, $f_2$ is redundant given $f_1$, and $f_3$ is irrelevant.

(a) relevant feature $f_1$  (b) redundant feature $f_2$  (c) irrelevant feature $f_3$
Feature Selection

Feature selection selects an `optimal’ subset of relevant features from the original high-dimensional data given a certain criterion.

\[ \mathbf{X} \in \mathbb{R}^{5 \times 10} \]

\[ \mathbf{X}_{\text{new}} \in \mathbb{R}^{5 \times 3} \]
Feature Selection and scikit-feature

• Feature selection can make data ‘bigger’
  – Assuming all binary attribute values in our toy example
  – Before FS, $5/2^{10} = 5/1024$, after FS, $5/2^3 = 5/8$

• Does FS always work?
  – Yes, for most high-d data

• Where can we find it?
  • scikit-feature, an open-source repository in Python
Making Thin Data Thicker

• Most people like many of us are in the long tail
  – Our data is thin or sparse
  – With little data, machine learning is powerless

• Social media data offers new opportunities
  – Multiple facets: posts, profile, linked information
  – Multiple platforms that offer different functions

• Two case studies
  – Feature selection using social network information
  – Connecting users across more than one social media site
Making Sense of Big Data

• For big social-media data, we want to automatically get a sense of what it is
  – User needs, sentiment, opinions, behavior, and trends
• A big part of big data is TEXT
• NLP and text mining can help extract *topics* from text
• If these machine-learned topics are for human consumption, are they actually comprehensible?
  – How can comprehensibility be measured?
Measuring Topic Interpretability

- How to measure interpretability of topics generated from machine learning?
- One common way is to indirectly measure predictive performance of these learned topics
  - The higher the performance (say, accuracy), the better
  - Does it really measure interpretability?
  - Human experts seem to be the best evaluator
- But involving human experts in evaluation may not be *scalable* and *reproducible*
- Hence, it is a challenging problem
Big Text Data

• Some example corpora:

<table>
<thead>
<tr>
<th>Source</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>36 million articles</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>100+ billion static web pages</td>
</tr>
<tr>
<td>Social Media</td>
<td>500 million new tweets each day</td>
</tr>
</tbody>
</table>

• Too much data to read
• How can we begin to understand all of these large bodies of text data?
Topic Models

K = 10

LDA

Document 1: board presentation massive

Document 2: american beef domestic

Document 3: basketball hall fame

<table>
<thead>
<tr>
<th>Document</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.02</td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>0.1</td>
<td>0.3</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Topic 1:
- season 2.0%
- game 1.8%
- home 1.5%
- start 1.2%
- hit 1.1%

Topic 2:
- percent 3.3%
- market 1.6%
- fell 1.3%
- shares 1.2%
- u.s. 1.2%

Topic 10:
- film 1.9%
- series 1.1%
- director 0.8%
- tv 0.8%
- movie 0.8%
Measuring Interpretability

• How do we measure the interpretability of statistical topic models

• A dilemma
  – Experts are **credible**, but **not scalable**, 
  – Crowdsourcing needs *no experts*, so **scalable**, but has *no expertise*, thus is **not credible**
A Measure of Topic Interpretability

• **Model Precision**
  • It shows a Turker 6 words in random order
    – Top 5 words from the topic
    – 1 “Intruded” word
    – Ask the Turker to identify the “Intruded” word

\[
MP_{model,topic} = \frac{\# \text{ Correct Guesses}}{\text{Total \# Guesses}}
\]

**Topic i:**

cat  dog  bird  truck  horse  snake

Observing Model Precision (MP)

What does Model Precision measure?
What doesn’t Model Precision measure?
It seems we need another measure
Measuring Coherence – Another Measure

• **Model Precision Choose Two**

• Nearly the same setup as Model Precision:
  – **Difference**: A Turker is asked to choose top two words

• Intuition: if the topic is coherent, then it would be difficult to consistently choose a second word

\[
MPCT_k^m = H\left(p_{turk}(w_{k,1}^m), \ldots, p_{turk}(w_{k,5}^m)\right)
\]
A Comparative Example

Model
Precision
Choose Two
## News Corpus for Experiments

### Yahoo! News Dataset

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>258,919</td>
</tr>
<tr>
<td>Tokens</td>
<td>6,888,693</td>
</tr>
<tr>
<td>Types</td>
<td>214,957</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Dataset</th>
<th>Strategy</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>News-010</td>
<td>News</td>
<td>LDA</td>
<td>10</td>
</tr>
<tr>
<td>News-025</td>
<td>News</td>
<td>LDA</td>
<td>25</td>
</tr>
<tr>
<td>News-050</td>
<td>News</td>
<td>LDA</td>
<td>50</td>
</tr>
<tr>
<td>News-100</td>
<td>News</td>
<td>LDA</td>
<td>100</td>
</tr>
</tbody>
</table>
Can MPCT Replace MP?

- Yahoo! News, Run with K = 10, 25, 50, 100.
- “Random” Topics
## MPCT vs. MP

### Top 5 Words

<table>
<thead>
<tr>
<th>Production, plants, provide, food, plant</th>
<th>suppressor</th>
<th>1.00</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number, system, transactions, card, money</td>
<td>flees</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Methods, data, information, analysis, large</td>
<td>diesel</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Series, fans, season, show, episode</td>
<td>leven</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Nuclear, fundamental, water, understanding, surface</td>
<td>modularity</td>
<td>0.13</td>
<td>0.92</td>
</tr>
<tr>
<td>Film, khan, ians, actor, bollywood</td>
<td>debonair</td>
<td>0.30</td>
<td>1.00</td>
</tr>
<tr>
<td>Mechanisms, pathways, involved, molecular, role</td>
<td>specialized</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Injury, left, list, return, surgery</td>
<td>tests-results</td>
<td>0.00</td>
<td>0.25</td>
</tr>
</tbody>
</table>

### MPCT Complements MP

- Both measures are needed with little extra overhead

```
0 0 | 1 0
0 1 | 1 1
```
Summary

- MPCT measures a topic’s *within*-topic distance
- MPCT complements Model Precision
- MPCT provides another dimension of topic quality
  - Low correlation with Model Precision ($\rho = 0.29$)

Addressing Don’t-Know-Don’t-Know Problems

• When collecting data, we often don’t know when we have a sufficient amount
  – We don’t know when to stop collecting, though we can’t collect forever
• A dilemma in studying migration on social media:
  – If we know its existence, no need for the study
  – If we don’t know, how can we verify the result?
Illustrative Examples of DNDN

1. When-to-Stop Dilemma: Collecting data forever vs. having credible patterns
   – How much data vs. how credible

2. Is There Migration on Social Media?
   – Users are a primary source of revenue
     • Ads, Recommendations, Brand loyalty
   – New SM sites need to attract users for expansion
   – Existing SM sites need to retain their users
   – Competiting for attention entails the discovery of migration patterns
Migration on Social Media

• Site Migration
  – Users leave a site by profile deletion or profile removal
  – Difficult to convince a user who left to return
  – Hard to study these users cross sites because we need their registration information

• Attention Migration
  – Users become inactive on a site
  – A harbinger for site migration
  – Can be detected by observing user activities across sites
  – Can take action to prevent site migration after understanding migration patterns
Patterns from Observation

(a) Delicious

(b) Digg

(e) StumbleUpon

(f) Twitter

(d) Reddit
Do We Know What We Didn’t Know?

• If a pattern is significant, it is valid
  – Significant differences observed in StumbleUpon, Twitter, and YouTube

• When to stop?
  Stop when we are certain, continue otherwise

<table>
<thead>
<tr>
<th>Site</th>
<th>Observed Coefficients</th>
<th>Shuffled Coefficients</th>
<th>p-value</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>A</td>
<td>R</td>
<td>N</td>
</tr>
<tr>
<td>Delicious</td>
<td>0.2858</td>
<td>0.4585</td>
<td>-</td>
<td>0.6029</td>
</tr>
<tr>
<td>Digg</td>
<td>0.4796</td>
<td>0.8066</td>
<td>-</td>
<td>0.52</td>
</tr>
<tr>
<td>Flickr</td>
<td>1</td>
<td>1</td>
<td>0.9797</td>
<td>0.2922</td>
</tr>
<tr>
<td>Reddit</td>
<td>0.5385</td>
<td>0.6065</td>
<td>-</td>
<td>0.4846</td>
</tr>
<tr>
<td>StumbleUpon</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0.4191</td>
</tr>
<tr>
<td>Twitter</td>
<td>0.5215</td>
<td>1</td>
<td>0.5335</td>
<td>0.2811</td>
</tr>
<tr>
<td>YouTube</td>
<td>0</td>
<td>1</td>
<td>0.1644</td>
<td>0.7219</td>
</tr>
</tbody>
</table>
1. Pop-Tarts before a hurricane (Walmart)
2. Higher crime, more Uber rides (Uber)
3. Typing with proper capitalization indicates creditworthiness (A financial services startup)
4. Users of the Chrome and Firefox browsers make better employees (An HR firm over Xerox data)
8. Female-named hurricanes are more deadly (University Researchers)

... 

Yes, they are bizarre, but are they true?
Evaluation without Ground Truth

The CACM article can be found at dl.acm.org
More Challenges Ahead

- Estimating the impact of an event
  - E.g., not all misinformation is catastrophic

- Predicting the future not the past
  - Are they two sides of the same coin?
    - Predicting general election result with Twitter data?

- Automating measures to replace crowdsourcing evaluation
  - Problems with evaluation methods involving AMT
Revisit Challenges in Acquiring SM Intelligence

• Social media data is obviously big, but why are we often still short of data?
  – How can we make data ‘bigger’?

• Data is power, so it can produce any result
  – Can we algorithmically evaluate the results from big data?

• We don’t know what we don’t know
  – How can we know if our result of social media analysis is of any value?
Repositories and Recent Books

• **scikit-feature** – an open source feature selection repository in Python
• Social Computing Repository
• Some books available for free download
Social Media Mining

An Introduction

A Textbook by Cambridge University Press

Reza Zafarani  
Mohammad Ali Abbasi  
Huan Liu

Syracuse University  
Machine Zone  
Arizona State University

Accessed 90,000+ times  
from 160+ countries and 1200+ Universities

The growth of social media over the last decade has revolutionized the way individuals interact and

http://dmml.asu.edu/smm/
Discovering Social Media Intelligence

- Graph Theories
- Network Measures and Models
- Data Mining, NLP, and Visual Analytics
- Community Detection and Analysis
- Information Diffusion
- Influence and Homophily
- Recommender Systems
- Behavior Analytics
  - Sentiment analysis

http://dmml.asu.edu/smm/
THANK YOU ALL & Conference Organizers

• for this opportunity to share our research
• Acknowledgments
  – Grants from NSF, ONR, ARO, among others
  – DMML members and project leaders
  – Collaborators: CMU (Minerva), CRA (IARPA-CAUSE)

More information by searching for “Huan Liu” or at [http://www.public.asu.edu/~huanliu](http://www.public.asu.edu/~huanliu)
Further Readings

- **Jundong Li** and Huan Liu. ``Challenges of Feature Selection for Big Data Analytics'', Special Issue on Big Data, IEEE Intelligent Systems. 32 (2), 9-15. 2017

- **Fred Morstatter** and Huan Liu. ``A Novel Measure for Coherence in Statistical Topic Models'', Association of Computational Linguistics (ACL), August 2016. Berlin, Germany


• Most people like many of us are in the long tail
  – Our data is thin or sparse
  – Without little data, machine learning is powerless
• Social media data offers new opportunities
  – Linked information
  – Multiple platforms as they offer different functions
• Two case studies
  – Feature Selection using social network information
  – Connecting users *across* more than one social media site
Use Link Information for Data Thickening

- Where can we find additional information for feature selection
- Social media data contains various types of data
  - Link information is additional
  - Other sources such as sentiment, like, etc.
- Are there theories to guide us in using link info?
  - Social influence
  - Homophily
- Extracting distinctive relations from linked data for feature selection
Representation for Social Media Data

Social Context
Relation Extraction

1. CoPost
2. CoFollowing
3. CoFollowed
4. Following
## Evaluation Results on Digg Data

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Features</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TT</td>
</tr>
<tr>
<td><strong>τ_5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>45.45</td>
<td>44.50</td>
</tr>
<tr>
<td>100</td>
<td>48.43</td>
<td>52.79</td>
</tr>
<tr>
<td>200</td>
<td>53.50</td>
<td>53.37</td>
</tr>
<tr>
<td>300</td>
<td>54.04</td>
<td>55.24</td>
</tr>
<tr>
<td><strong>τ_25</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>49.91</td>
<td>50.08</td>
</tr>
<tr>
<td>100</td>
<td>53.32</td>
<td>52.37</td>
</tr>
<tr>
<td>200</td>
<td>59.97</td>
<td>57.37</td>
</tr>
<tr>
<td>300</td>
<td>60.49</td>
<td>61.73</td>
</tr>
<tr>
<td><strong>τ_50</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50.95</td>
<td>51.06</td>
</tr>
<tr>
<td>100</td>
<td>53.60</td>
<td>53.69</td>
</tr>
<tr>
<td>200</td>
<td>59.59</td>
<td>57.78</td>
</tr>
<tr>
<td>300</td>
<td>61.47</td>
<td>62.35</td>
</tr>
<tr>
<td><strong>τ_100</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>51.74</td>
<td>56.06</td>
</tr>
<tr>
<td>100</td>
<td>55.31</td>
<td>58.69</td>
</tr>
<tr>
<td>200</td>
<td>60.49</td>
<td>62.78</td>
</tr>
<tr>
<td>300</td>
<td><strong>62.97</strong></td>
<td><strong>66.35</strong></td>
</tr>
</tbody>
</table>
Summary

• LinkedFS is evaluated under varied circumstances to understand how it works
  – Link information can help feature selection for social media data

• Unlabeled data is more often in social media, unsupervised learning is more sensible, but also more challenging

Jiliang Tang and Huan Liu. ``Unsupervised Feature Selection for Linked Social Media Data'', the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012.
Gather more Data with Little Data

• Collectively, social media data is indeed big
• For an individual, however, the data is little
  – How much activity data do we generate daily?
  – How many posts did we post this week?
  – How many friends do we have?
• When “big” social media data isn’t big,
  – Searching for more data with little data
• We use different social media services for varied purposes
  – LinkedIn, Facebook, Twitter, Instagram, YouTube, ...
An Example

- Little data about an individual
+ Many social media sites
- Partial Information
+ Complementary Information

> Better User Profiles

• Each social media site can have varied amount of user information

• Which information definitely exists for all sites?
  – **Usernames**
    – But, a user’s usernames on different sites can be different

• Our work is to connect the information of the same user provided across sites
Our Behavior Generates Information Redundancy

- Information shared across sites provides a behavioral fingerprint
  - How to capture and use differentiable attributes

MOBIUS
MOdeling Bbehavior for Identifying Users across Sites
Behavioral Modeling Approach with Learning

Generates

Captured Via

Behavior 1

Information Redundancy

Feature Set 1

Behavior 2

Information Redundancy

Feature Set 2

Behavior n

Information Redundancy

Feature Set n

Identification Function

Learning Framework

Data
Summary – Making Data Bigger

• Gathering more data is often necessary for effective data mining
• Reducing dimensionality can make data bigger
• Social media data provides unique opportunities to do so by using different sites and abundant user-generated content
• Traditionally available data can also be tapped to make thin data “thicker”