Extracting Information from Social Networks

Aggregating site information to get trends

- Not limited to social networks
- Examples
  - Google search logs: flu outbreaks
  - "We Feel Fine"
  - Bullying

Bullying

Xu, Jun, Zhu, Bellmore published 2012

- Look for Twitter posts in response to bullying
- To provide source of data for studying bullying
- Techniques used
  - natural language processing methods
  - text classifiers
  - hand labeled training data
- Data set "enriched"
  - public Twitter API
  - collect only tweets using a word-form of "bully"

Some details: 4 major tasks

1. Recognizing tweets on bullying versus other uses of word "bully"

- 1762 tweets labeled by indep. annotators
- found 684 on bullying (39%)
- tried 4 common text classifiers
- held out 262 of 1762 to test classifier
- different size training sets
- best classifier 81.3% accuracy

2. Identify roles within each bullying tweet

- labels: accuser, bully, reporter, victim, other
- label author
  - classifier 61% accurate
- label each person mentioned in tweet
  - "named entity recognition"
- annotators labeled each token in bullying tweets
  - accuser, bully, reporter, victim, other, not-person
- classify each token
- 684 bullying tweets for training and test
- best:
  87% tokens correctly labeled incl not-person
  53% tokens labeled some kind person labeled corrrectly
  42% true person tokens labeled correctly

3. sentiment analysis

- focused on detecting teasing
  "*lol stop being a cyber bully lol*" not serious bullying? coping?
- of interest to social scientists
- classifier
  - 89% accuracy for 694 test tweets but
  - accuracy of teasing tweets 53%
  - accuracy of not teasing tweets 96%

4. topic analysis

- topics of discussion in bullying tweets
- use Latent Dirichlet Allocation (LDA)
- example topics: feelings, suicide, family, school

Kamvar & Harris:“We Feel Fine”

developed 2005-06, published 2011

- extract feelings
  - not looking at statistical significance
- both art and science
- "crowdsourced qualitative research"
- graph of "frequently co-expressed emotions”
- tool "surprisingly accurate”
  - replicating results
  - suggesting hypotheses – confirmed

7

METHODS

- continuous crawl blog, micro blog, social networking sites
- 14 million expressions of emotion from 2.5 million people as of paper submission
- get info on authors from profiles
- sentence-level analysis
  - explicit use “I feel”, “I am feeling” “I felt” etc
- extract information by regular expressions
- find emotion words
  - 5000 emotion words pre-determined by hand
- index by emotions

8

Results

- associate largest image on entry with feeling
- use data:
  - feeling,
  - age,
  - gender,
  - weather,
  - location,
  - date
- produce visuals
- additional analysis thru API

9

Visuals: Art + Information

- “Madness” - swarming 1500 feelings
  - color = tone
  - click feeling: get sentence, image
- “Murmurs” - particles + scrolling list feelings
  - reverse chronological
- “Montage” – photographs
- “Mobs” displays particles organized for summary:
  - feelings- histogram
  - location – map
- “Metrics” features most differentially expressed
  - for given sub-pop against global pop.
- “Mounds” - every feeling scaled and sorted by freq.

10

[We Feel Fine: An Almanac of Human Emotion]

11

Information from social network structure

- Explore properties of graph
  - nodes
  - edges
- Interpret in context of subject of network

12

Graph measures of interest for nodes

- degree/indegree/outdegree
- pagerank
- sum of distances to all other nodes
  - Reciprocal is closeness centrality
- betweenness centrality
  - number of shortest paths in graph that go through the node
- cluster coefficient
  - fraction of pairs of neighbors of node that have edge between them

13

Uses

- Look at nodes that stand out under different measures
- Look at distribution of values of measure

14

See figure in
http://en.wikipedia.org/wiki/Centrality

15

Graph properties of interest for network

- density
  (number of edge)/(number of possible edges)
  directed vs undirected? self-edges?
- diameter
  largest shortest path
- distribution of shortest paths
  "6 degrees of separation"
- average cluster coefficient
- distribution of degrees

16

Characterizing social networks

for social network with n nodes

- average density low
- average shortest path log(n) or less
  - small world network
- form communities
- distribution of degrees follows power law
  - scale-free

17

Small world phenomena

- Travers & Milgram 1969 *Sociometry*
  - 296 letters to start; 67 reached target person
  - Mean length path followed 6.2
- Leskovec & Horvitz 2008 *WWW Conf*
  - Microsoft Instant Messenger, 240 million active users
  - Edge: two-way conversation
  - One giant component
  - Average distance 6.6
  - 90% effective diameter 7.8

18

See figure 2.11 in the textbook

Easley, David; Kleinberg, Jon. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Cambridge University Press, July 19, 2010.

19

Characterizing relationships

- Relationship: edge between two nodes
  - Consider now just undirected
  - Refer to as "neighbors"
- Would like to extract properties of the relationship from network structure.
- Measures – here are two
  - Embeddedness: number of mutual neighbors
  - Dispersion: measure of connectedness among mutual neighbors
    - Backstrom & Kleinberg, 2014

20

A network Analysis of Relationship Status on Facebook

Backstrom & Kleinberg 2014

- Observe: person's network of friends represents diverse set of relationships
- Question: Can one recognize romantic partners on Facebook from structure of friends network?
- Contributions (some)
  - Define new measure dispersion
  - Show dispersion works better that embeddedness
  - Show dispersion works pretty well
  - Show combining dispersion with many other signals via machine learning does even better

21

Dispersion Definition

- Actually define several versions
- Basic: absolute dispersion disp(u,v) for link (u,v)
  - Define G_u as the subgraph on neighbors of u
  - Define $C_{u,v}$ as the set of common neighbors of u and v
  - For s,t nodes in $C_{u,v}$, define $f_{u,v}(s,t)$ with value

    1 if s, t not neighbors and have no common neighbors in G_u other than u and v

    0 otherwise
  - $\text{disp}(u,v) = \sum_{s,t \text{ in } C_{u,v}} f_{u,v}(s,t)$

22

Experiments: Data

- Facebook users
  - At least 20 years old
  - Between 50 and 2000 friends
  - Listed spouse or relationship partner on profile
- Sample ~1.3 million of these users selected uniformly at random and their network neighborhoods (extended dataset)
  - Neighborhoods avg 291 nodes, 6652 links
  - 379 million nodes , 8.8billion links overall
- Subsample 73,000 neighborhoods (primary dataset)
  - Only neighborhoods with at most 25,000 links
  - Uniformly at random

23

Experiments: Modify definition of dispersion

- For improved results
- Normalized dispersion: disp(u,v)/emb(u,v)
  - emb(u,v) is embeddedness
- Recursive dispersion: look at neighbors of neighbors of neighbors …
  - Find best performance using 3 levels

24

See Figure 4 in the paper

Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook,

Backstrom & Kleinberg, CSCW 2014

25

Additional questions in paper

- How much better can lots of features do?
  - Combined 120 features for nodes in primary dataset
    - Combined variations of dispersion def
    - Included many other properties from user pages and behavior
  - Used machine learning classifier
    - Trained on 50% users
  - Overall precision at 1st position 0.705 (vs 0.506)

26

Additional questions in paper

- What about predicting whether in a relationship?
  - High dispersion link from u does not mean romantic relationship
    - Property is bridging groups of u's friends
      - family, close friends
  - Used machine learning yes/no classifier
    - 68.3% accuracy single vs any relationship
      - Baseline 59.8 – predict more common class
    - 79.0% accuracy single vs married
      - Baseline 56.6
  - Max over user's friends of normalized dispersion most important of network features used

27

Do all social networks, as networks, have same properties?

- Kwak, Lee, Park, Moon study Twitter (pub 2010):

  NO

28

Kwak, Lee, Park, Moon experimental set-up

- July 6-31, 2009 crawl of Twitter
  - 41.7 million user profiles,
    - compare over 500 million today
    - crawl + those refer to trending topics
  - 1.47 billion social relations,
    - started with "Paris Hilton" and crawled followers and "followings"
  - 4,262 trending topics
    - collected top ten every 5 minutes
  - 106 million tweets
    - tweets mentioning trending topics

29

Kwak, Lee, Park, Moon Findings

- # followers fits power law but
- users with > 100,000 followers have many more followers than expect
- 77.9% links one way
- shortest path between users shorter than other social networks
  - median 4.12
  - for 97.6 % pairs, path length ≤ 6

30

Kwak, Lee, Park, Moon: ranking users

- followers graph
  - number of followers
  - PageRank

  similar rankings
- retweets of user's posts
  - very different from graph measures

31

Summary: Social Networks and Obtaining Information

- Social networks provide many ways of improving our acquisition of information
- Uses still in active development

32