Extracting Information from Social Networks

1

Reminder: Social networks

- Catch-all term for
  - social networking sites
    - Facebook
  - microblogging sites
    - Twitter
  - blog sites (for some purposes)

2

Ways we can use social networks to find information

- ✓ Extract meta-information for “regular” Web search
  - site information
  - site properties
- Extract information to use directly
  - search content of social site
  - aggregate information from site content
  - information from structure of social network

3

Searching social network content

- How does searching a social network site differ from searching the Web with a SE?
- Does this affect
  - indexing?
  - query evaluation?

- social site - Facebook
- microblog site - Twitter

4

Searching Facebook

- search for objects (e.g. people) as well as information
- focused searches
  - people
  - friends
  - photos
- link structure central
  - find friends who …
- other?

5

Searching Twitter vs Web

- Study by Teevan, Ramage and Morris pub. 2010
- Experimental setup
  - data from browser logs from Bing Toolbar
  - harvest queries issued to search engines
    - “general purpose” : Bing, Google, Yahoo
    - “vertical search engines”: Twitter
    - associate with user IDs and timestamps
  - Sampled 126,316 queries to Twitter
    - subset of 33,405 users
  - 2.5 million queries by same subset users from Bing, Google, Yahoo

6

Teevan et al results

- unsurprising:
  - top 10 Web searches navigational
  - top 10 Twitter queries mixed celebrities, movies, games, memes (eg "#theresway2many"): popular items
- more surprising:
  - 23.19% Twitter queries issued only once, vs 49.73% Web
  - 55.76% Twitter queries issued more than once by same user, vs 34.71% Web

7

more results Teevan, Ramage and Morris

- temporal characteristics
  - session = series queries by user "in close succession". Use 15 min. inactive as delimiter
  - Twitter sessions shorter: 2.2 queries vs 2.88 Web
  - 9.38 sec btwn Twitter queries in session vs 13.63
- combined Twitter, Web searches
  - informational: monitor with Twitter, learn with Web
  - 61.92% of time start on Web
  - 20.56 sec. btwn queries in a session
  - 6.13 queries per session
  - 43.74% queries issued to both in one session

8

Twitter characteristics that may change search approach?

- history more important – Twitter findings
- recency more imporant – trending
- popularity more important?
- labels available – hashtags
- other?

9

Unicorn: A System for Searching the Social Graph
by many Facebook researchers (2013)

- primary backend for Facebook Graph Search
- "designed to search trillions of edges between tens of billions of users and entities and entities on thousands of commodity servers"
- thousands of edge types used
  - including obvious "friend" "like"
- graph sparse:
  - typical node < 1000 edges
  - average user has ~130 friends

10

Unicorn: graph querying

- query language on edge relationships

  "find female friends of user 6" becomes query

  `(and friend:6 gender:1)` intersection of sets
- supports queries on paths
  - rounds of basic query evaluation

    "find pages liked by friends of user 7 who like Emacs (object 42)" becomes

    `(and friend:7 likers:42)` giving {$resultID_1$, …, $resultID_k$}

    followed by

    `(or likes:resultID_1 … likes:resultID_k)`
  - does through APPLY operator

    `(apply likes: (and friend:7 likers:42) )`

11

Unicorn APPLY operator

- applies "or" to results of inner query

`(apply likes: (and friend:7 likers:42) )`
- can nest APPLY arbitrarily deep
  - friends of friends of friends of friends of user 21

`(apply friend:(apply friend:(friend 21) ) )`
- limit on number results of inner query
  - solution: drop some results
  - issue: performance
  - cut-off ~100,000 terms applied to outer query

12

Unicorn: index struture

- index represents adjacency list
- index term <edge-type>:<id>
  - `friend:5` selects list of friends of userID 5
- form of adj. list entry:
  - ((sortkey, DocID), other info)
  - nodes on adjacency list sorted first by sortkey, then by nodeID

13

Unicorn performance

query "people who like computer science"

- > 6 million results - ask for 100 returned
- run 100 times
- average performance
  - latency 11 ms
  - aggregate CPU across 37 index servers 31.22 ms

query "friends of likers of computer science"

- for APPLY with trunction limit 10^5, latency almost 2 sec.
- for APPLY with trunction limit 10^3, latency about 100ms

14

Aggregate site information to get trends

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

15

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"

16

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

17

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

18

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

19

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

20

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

21

Results

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

22

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.

23

Social network properties

- Graph measures of interest for nodes
  - pagerank
  - degree/indegree/outdegree
  - betweenness centrality
    - number of shortest paths in graph that go through the node
  - cluster coeffiient
    - fraction of pairs of neighbors of node that have edge between them
- Look at nodes that stand out under different measures

24

Social network properties

- 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

25

Characterizing social networks

for social network with n nodes

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

26

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

- Kwak, Lee, Park, Moon study Twitter (pub 2010):
  NO

27

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

28

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

29

Kwak, Lee, Park, Moon: ranking users

- followers graph
  - number of followers
  - PageRank

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

30

Summary: Social Networks and Obtaining Information

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

31