Social Data Mining in Facebook
Insights, Methods and Practice for Large-Scale Social Media Analysis

Rong Yan
Research Scientist
Facebook
Data Scale in Facebook

- 845 million monthly active users
- 100 billion friendships
- 250 million photos uploaded per day
Pervasive Influence on Daily Life

Source: stories.facebook.com
Pervasive Influence on Daily Life
Facebook is being built around people.

People make friends.

Friends communicate in the social network.

Communication determines who are being influenced.
Large-Scale Social Data Mining in Facebook

To understand users better and improve their social experience

1. Identity
2. Connection
3. Communication
4. Influence
What can we answer from the Data?

To understand users better and improve their social experience?

| Identity                  | • What kinds of ads are relevant to a user  
|                          | • Can we suggest face tags for each photo  
|                          | • How to recommend the best images / videos |
| Connection                | • Who should be suggested to you as new friends |
|                          | • How to measure and discover online communities |
| Communicate               | • What are users talking about in their news feed |
|                          | • How to analyze user rating and reputation |
|                          | • User sentiment and emotion in blogsphere |
| Influence                 | • How does friend influence populated in Facebook? |
|                          | • Will influence work for social advertising? |
IDENTITY
Facebook’s Social Identity
Social Identity consists of Many Factors
Finding Relevant Ads for Users with Interest Modeling
How big are Facebook Ads?

- **Revenue:** >4 billion USD / year
- **Coverage:** >10x billion of impressions / day
- **Reach:** >20% of 18 – 30 US male on Friday
Heterogeneous Data on Facebook

- Social graph
  - Friendship networks
  - User-ads network ...
- Text
  - News feed
  - Ads title ...
- Images
  - Uploaded photos
  - Ads figures ...
- Demographics
  - Age, occupation ...
Hierarchical Text Classification Model

Input Text

$50 off Your 1st Bonobos
cim.meebo.com


Output Labels
(Open Directory Project, ~50K labels)

Games
Shopping
Computers

Clothing
Sports
Food

Expected output:
Shopping/Clothing 0.8
Arts/Fashion/Men 0.3
Vector Space Model on Interest Topics

\[
\cos(U, A) = \frac{(\Sigma_i U_i)^T (\Sigma_j A_j)}{\|\Sigma_i U_i\| \|\Sigma_j A_j\|}
\]

- \( U_1 \) (User concept from source 1)
  - Society: 0.2
  - FIFA series: 0.2
- \( U_2 \) (User concept from source 2)
  - Society: 0.1
  - Society / Dating: 0.4
- \( A_x \) (Ad concept from source \( x \))
  - Soccer: 0.5
- \( A_{x+1} \) (Ad concept from source \( x+1 \))
  - Beer: 0.6

\[
\cos(\text{User concept}, \text{Ad concept}) = 0?
\]

- User concept
  - Society: 0.3
  - Society / Dating: 0.4
  - FIFA series: 0.2
- Ad concept
  - Soccer: 0.5
  - Beer: 0.6
Sources and Concepts are not Identical!

- Different information sources
  - They should not be associated with uniform weights
- Concept spaces are not identical
  - Users who like Skiing might like ads about Hotels, Hiking, Beer, ...

![Image of a person and a diagram]
Improved User-Ad Interest Model
[Wang, Raina, et al, SIGIR’11]

\[
\text{Score}(U, A) = \frac{(\Sigma_i p_i U_i)^T W (\Sigma_j q_j A_j)}{\|\Sigma_i p_i U_i\| \|\Sigma_j q_j A_j\|}
\]

- Learn \( p, q, W \) using generalized linear models
  - Optimized one parameter \( W = \) logistic regression
  - Optimized \( p, q \) and \( W \) with online gradient descent
  - L1 regularization to improve sparsity
# Experimental Results of Our New Methods

**Data Set:** 5M users, 80 objects per user

<table>
<thead>
<tr>
<th>Methods</th>
<th>Testset Log-loss Reduction (rel.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No learning</td>
<td>1.0</td>
</tr>
<tr>
<td>Learned $p, q$</td>
<td>3.8x</td>
</tr>
<tr>
<td>Learned $p, q, W$</td>
<td>14.3x</td>
</tr>
</tbody>
</table>
| $L_1$ Regularization | 13.7x  
(80% sparser, 53x speed up for online score computation) |
CONNECTION
Today’s Reality: Friend Connections on Facebook

Average user: 130 friends
Total: 100B social connections
Friend Suggestion in Facebook
People You May Know (PYMK) [Backstrom et al.]

- Show friend recommendations
- Let users accept/remove suggestions
Who are Candidates for Friend Suggestions?

Most friendships go to friends-of-friends

- 92% of new friendships on FB

From a practical point of view, doing beyond FoF is impossible

- Average user has over 130 friends
  - $130 \times 130 = 17K$ FoFs
  - $130 \times 130 \times 130 = 2.2M$ FoFoFs
Learning to Suggest Friends of Friends (FoF)

Challenges: # of FoF is Too large

- A typical user has >10K FoFs on average, >100K for 99th percentile!

Friends in common is a good start

- Two people are 12x more likely to become friends with 10 mutual friends than 1 mf

Other network features are helpful

- If your good friend just made a new friend, that is a good suggestion
Baseline PYMK Ranking

Examine all FoFs, compute features and rank each candidate

- Generate list of top 100 candidates
- Bagged decision trees output score for each \((u,w_i)\) pair

Poor experience: output too static

- Decision models can only be run once per 2 days

Solution: re-ranking w/ logistic regression in every impression
Putting it all together

Lars

FoF Discovery and Feature Generation

Lars, Greg:
Mutual Friends = 10,
Age(Lars) = 27, ...

Bagged Decision Trees

Score(Lars, Greg) = 0.045
Score(Lars, Shelly) = 0.021 ...

Real-Time CTR Prediction

Cache-only memcache

Impressions(Lars, Greg) = 3
Impressions(Lars, Shelly) = 2

CTR(L, G) = 0.012 ...

CTR * Value > threshold?

Outcome
Current Friending Stats for PYMK

Millions of connections made per day

Without PYMK many would not be found through other channels

- New users removed from PYMK have 27% fewer friends after 6 weeks

% Active30 vs. control (n00bs)
How many friends do you need?

Guestbook activity network in *CyWorld*, largest Korean SNS

- How does the number of friends affect sociability?
- How do activity graphs compare with friendship graphs?

[Chun et al, IMC-2008]
How many friends do you need?

Capacity cap

- Node strength (sociability) increases with the number of friends up to a limit.
- Authors argue that the limit could be connected to Dunbar’s number.
- Dunbar (1998): 150 is limit of manageable relationships.

Node strength: Sum of messages across all directed edges.
COMMUNICATION
DARPA Challenge 2009
Locating 10 Balloons using Social Network
What are People talking about in Status Update?

- Analyze the overall trends for status updates
  - Extract popular status keywords
  - Sentiment analysis
  - Visualization
  - Keyword Suggestion

- Hadoop + Hive used for text analysis
“Party” vs. “Hangover”

Search: party tonight, hangover
Suggestions: eid | american idol | vacation | batman, iron man

party tonight | hangover

New Year’s Eve vs. New Year’s Day

Sat vs. Sun

- Frequently measured using self-report
  - Diener et al. (1985): Satisfaction with Life (SWL)
  - Gallup-Healthways Well-Being Index

- Self-report is gold standard, but hard to scale
  - Gallup telephones 1,000 people daily
The Pursuit of Happiness

• **Status update** could be an measure of happiness

1. Representative (unbiased sampled population)
2. Unobtrusive (based on naturalistic behavior)
3. Computational (no human raters)
4. Correlated (co-vary with other valid measures, e.g., non-naturalistic self-reports)
5. Efficient (process millions of updates per day)
Word-Count Happiness Model

- **LIWC2007 Dictionaries**: Word categories showing psychological content (Pennebaker et al., 2007)
  - e.g. emotion words (407 positive and 506 negative)

- **Example**: “Fred hates aggressive Facebook updates, but loves irony.”
  - “hate”, “aggress” are negative emotional terms
  - “love” is a positive emotion term
  - Update is 25% negative, 12.5% positive
The Happiness Model

• Hypothesis
  • Happier people use more positive words

• Validation Method:
  • Recruited 1,341 English speakers (via a Facebook ad)
  • Standardized scores relative to their country level

\[ H_u = \frac{p_u - \mu_{pd}}{\sigma_{pd}} - \frac{n_u - \mu_{nd}}{\sigma_{nd}} \]
Validation with SWL

• Learn a hierarchical linear model between happiness score from user’s Satisfaction with Life Scale (SWL)

• Result: the happiness score is a significant predictor for SWL based on t-test, $p < 0.001$
  • SWL-happiness correlation $r = 0.17$
Facebook Gross National Happiness

- Estimate each country’s average happiness based on status updates [http://apps.facebook.com/gnh_index]

![Graph showing National Gross Happiness with holidays and weekly and long-term trends marked.](image-url)
INFLUENCE
How People Influence Each Other in FB?
[Bakshy et al., WWW’ 2012]

How information spreads in networks and how people influence one another

- Adoption of products, technologies, opinions
- Media contagion (e.g. sharing links to Web sites, videos)

To what extent, does correlated sharing behavior reflect actual influence processes?

- Who exposes you to the most relevant content?
- Who exposes you to novel information?
The Confounding Factors from Strong Ties

People who communicate frequently (strong ties) tend to be more similar (homophily)

As a result, a “strong tie” may:

- Share more relevant information to the users
- Visit the same websites as the users

This makes it difficult to measure the causal effect of social networks on information diffusion
Causal Picture of Sharing (formally)

Unknown correlation between friends' characteristics (expected to be stronger for closer friends)

Known characteristics \( X_i \) \( U_i \) \( X_j \) \( U_j \)

Unknown characteristics (e.g. Web browsing behavior)

Friend's sharing behavior \( Y_{ja}(t_0) \)

External Influence

Focal user's sharing behavior \( Y_{ia}(t_1) \)

Facebook feed \( D_{ija} \)
Strong Ties are Individually More Influential

Randomized experiments: hold out 1B news feed for display

- Users are more likely to share the same content as their strong ties - even if they didn’t see the content
- More importantly, users are more likely to share more content from strong ties
Weak Ties are Collectively More Influential

Weak ties expose friends to novel information

- Weak ties are collectively responsible for the majority of information spread

Facebook is not an echo chamber as some expects
**Summary:** Social Data Mining in Facebook

Large-scale social data analysis helps us to understand users better and improve their social experience.

1. **Identity**
2. **Connection**
3. **Communication**
4. **Influence**
Questions?

rongyan@fb.com
http://yanrong.info
Busted: PYMK unravels Bigamy

Busted: When Ellenora Fulk saw this picture on Teri Wyatt-O’Neill's Facebook page, she realised that the stranger was married to her husband, Alan O’Neill.