Community Detection on Facebook and Twitter

PROPONENTS: FERNANDEZ, RYAN AUSTIN
POBLETE, CLARISSE FELICIA M.
SAN PEDRO, MARC DOMINIC
TAN, JOHANSSON E.
ADVISER: CHARIBETH K. CHENG

Outline of the Presentation

1. Overview of Current State of Technology
2. Research Objectives and Scope and Limitations
3. Significance of the Study
4. Research Methodology

Overview of Current State of Technology

Social Network

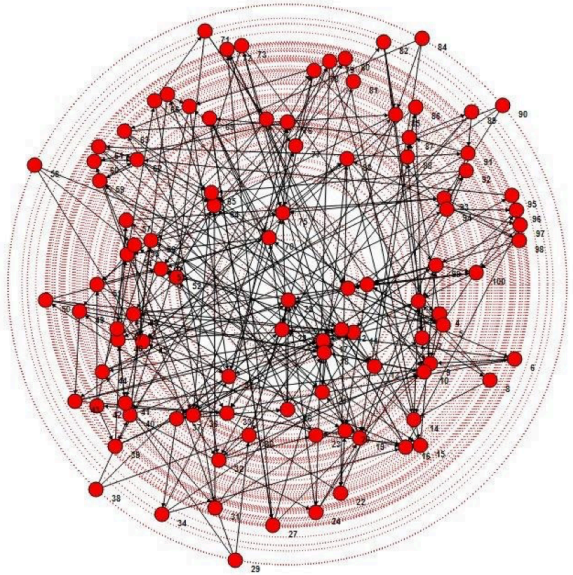
noun

a dedicated website or other application that enables users to communicate with each other by posting information, comments, messages, images, etc.

Social Network Analysis & Visualization

Gephi

SocNetV



Community Detection

- Detecting networks of users with similarity to each other
- Multiple algorithms
  - Greedy Modularity Optimization[1]
  - Clique Percolation Method[2]
  - Vertex Similarity[2]

1 Clauset, A., Newman, M. E. J., & Moore, C. (2004, Dec). Finding community structure in very large networks. Phys. Rev. E, 70 , 066111. doi: 10.1103/PhysRevE.70.066111
2 Tang, L., & Liu, H. (2010). Community detection and mining in social media. Morgan & Claypool. doi: 10.2200/S00298ED1V01Y201009DMK003

Community Detection

- Multiple algorithms (cont.)
  - Hierarchical Clustering[1]
  - Interest-based community detection[2]
  - k-means clustering[3]

1 Tang, L., & Liu, H. (2010). Community detection and mining in social media. Morgan & Claypool. doi: 10.2200/S00298ED1V01Y201009DMK003
2 Lim, K., & Datta, A. (2012). Following the follower: Detecting communities with common interests on twitter. In Proceedings of the 23rd ACM conference on hypertext and social media (ht12) (Vol. 1, pp. 317{318). Association for Computing Machinery. doi: 10.1145/2309996.2310052
3 Zhang, Y., Wu, Y., & Yang, Q. (2012). Community discovery in twitter based on user interests. Journal of Computational Information Systems, 8 (3), 991-1000.

Similarity Parameters (Sentiment Analysis)

- Formula for text similarity based on topics in the text[1]
- Similar word usage in communities vie Euclidean Distance[2]
- Naive Bayes Subjective/Objective Positive/Negative Classifier[3]
- Cosine similarity[4]

1 Zhang, Y., Wu, Y., & Yang, Q. (2012). Community discovery in twitter based on user interests. Journal of Computational Information Systems, 8 (3), 991-1000.
2 Bryden, J., Funk, S., & Jansen, V. A. (2013). Word usage mirrors community structure in the online social network twitter. EPJ Data Science, 2 (1), 1{9. doi: 10.1140/epjds15
3 Deitrick, W., & Hu, W. (2013). Mutually enhancing community detection and sentiment analysis on twitter networks.
4 Bakillah, M., Li, R.-Y., & Liang, S. H. L. (2015, February). Geo-located community detection in twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon haiyan. Int. J. Geogr. Inf. Sci., 29 (2), 258{279. doi: 10.1080/13658816.2014.964247

Similarity Parameters (Other Parameters)

- A majority of the studies dealt with Twitter
- URL Similarity[1,2]
- Hashtag Similarity[1,2]
- Following Similarity[1,2,3]
- Retweeting Similarity[1,2,3]
- Mentions[1,2]

1 Zhang, Y., Wu, Y., & Yang, Q. (2012). Community discovery in twitter based on user interests. Journal of Computational Information Systems, 8 (3), 991-1000.
2 Bakillah, M., Li, R.-Y., & Liang, S. H. L. (2015, February). Geo-located community detection in twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon haiyan. Int. J. Geogr. Inf. Sci., 29 (2), 258{279. doi: 10.1080/13658816.2014.964247
3 Darmon, D., Omodei, E., & Garland, J. (2015, 08). Followers are not enough: A multifaceted approach to community detection in online social networks. PLoS ONE, 10 (8), 1-20. doi: 10.1371/journal.pone.0134860

Evaluation Metrics

- Average number of mutual following links per user per community(FPUPC) to evaluate their communities[1]

1 Zhang, Y., Wu, Y., & Yang, Q. (2012). Community discovery in twitter based on user interests. Journal of Computational Information Systems, 8 (3), 991-1000.

Research Gap

- Most studies are about Twitter

Research Gap

- Facebook has more active users

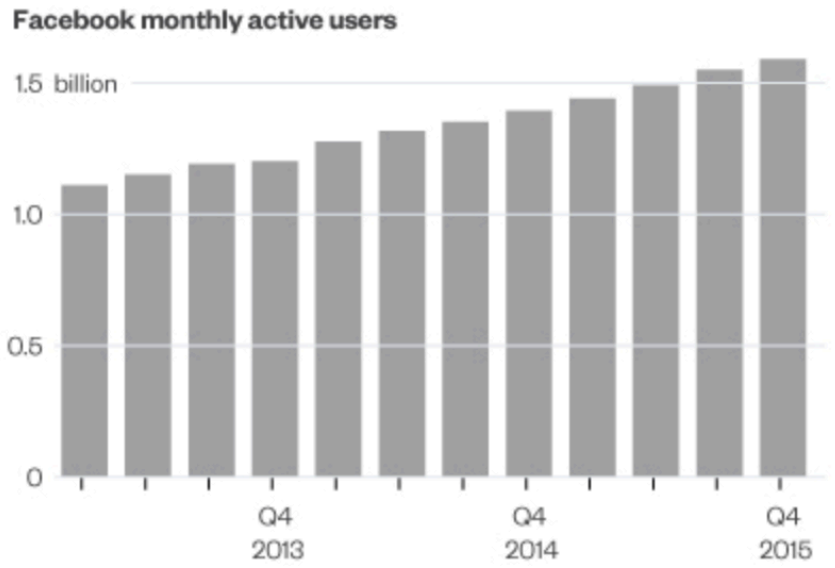
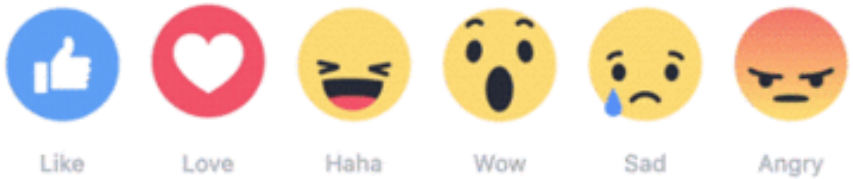


Image from http://www.bloomberg.com/gadfly/articles/2016-02-12/social-studies-comparing-twitter-with-facebook-in-charts

Research Gap

- Facebook has other features



Images from http://findicons.com, http://www.wired.com/2016/02/facebook-reactions-totally-redesigned-like-button/, http://play.google.com

Research Problem

- Determine which algorithms and features provide more accurate communities considering both Facebook and Twitter

Research Objectives, Scope, and Limitations
General Objective

- To produce a visualization of the detected communities on data found on Facebook and Twitter

Specific Objective #1

Specific Objective

- To build a corpus of social media data

Scope and Limitations

- Searching for Facebook and Twitter API's
- Collecting data from both social networks using the APIs

Specific Objective #2

Specific Objective

- To determine the various techniques and algorithms in detecting communities

Scope and Limitations

- Identify the appropriate algorithms for clustering users into communities.
- Limited to review of algorithms in RRL

Specific Objective #3

Specific Objective

- To determine the appropriate parameters to use in detecting the communities

Scope and Limitations

- Parameters that indicate user similarity
- Limited to:
  - Sentiment analysis
  - Elements which can be extracted from a user's profile/posts

Specific Objective #3

Specific Objective

- To determine the appropriate parameters to use in detecting the communities

Scope and Limitations

- Facebook specific features such as:
  - Group membership
  - Likes and reactions
  - Chat messages
  - Event participation

Specific Objective #4

Specific Objective

- To determine how to evaluate the correctness of the detected communities

Scope and Limitations

- Find appropriate metrics in determining the accuracy of detected communities

Specific Objective #5

Specific Objective

- To implement a tool for the visualization of detected communities using the gathered information

Scope and Limitations

- Visualization for Facebook and Twitter communities

Significance of the Study
Community Detection

- Determining whether Facebook data improves community detection could contribute to future research in the domain

Target Users and Domain

- This research can also be a very useful tool in the domains of:
  - Viral marketing
  - Political endorsement

Target Users and Domain

- Interested companies may use the result of this research to improve their sales and marketing.
- The government may use this to gauge
  - public opinion on certain issues
  - which geographical areas have a particular opinion.

Research Methodology

- 3 Phases
  1. Preparation
  2. Iterative Experimentation
  3. Analysis and Finalization

Preparation

- RRL
- Theoretical Framework – Implementation Details of:
  - Community Detection Algorithms
  - Similarity Parameter Computation
  - Evaluation Metric Computation
- APIs for Facebook and Twitter
- Platform Selection for Data Storage
- Programming Language Selection

Iterative Experimentation

- Different algorithm and similarity parameter per iteration
- Time: 2 – 3 weeks

Iterative Experimentation: Steps

1. Similarity Parameter Selection
   - Need not be applicable to both Facebook and Twitter
   - Reason for having multiple iterations
2. Community Detection Algorithm Selection
   - Must be compatible with the selected parameter
   - Algorithm may differ per iteration

Iterative Experimentation: Steps

3. Data Collection
   - Using API's to get data from Facebook and Twitter
   - Anonymization and data transformation
4. Model Design
   - Time: 3 – 4 days
   - First iteration includes evaluation parameters

Iterative Experimentation Steps

5. Model Implementation
   - Time: 1 – 2 weeks
   - First iteration includes evaluation metrics
6. Model Evaluation
   - Time: 2 – 3 days
   - Running the algorithms on the data
   - Using the evaluation metrics to determine accuracy

Iterative Experimentation: Steps

7. Documentation
   - Time: 1 – 2 days
   - Ensure quality and correctness of documentation of iteration
   - Retrospective

Analysis and Finalization

- Time: 2 – 3 weeks
- Revisit data collected
- Possible supplementary study
- Determine best parameter-algorithm combination
- Produce visualization

Calendar of Activities

Activities	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Data Gathering (RRL)	_●●●	●●●_											__●●	●___
Data Gathering (TF)	_●●●	●●●_	__●●											
Iterative Experimentation				●●●●	●●●●	●●__	_●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●__	
Analysis and Finalization													__●●	●___

Thank You!