CS 5604 Informational Storage & Retrieval
Spring 2015
Social Networks & Importance

04/30/2015
Bharadwaj Bulusu - bbsb08@vt.edu
Vanessa Cedeno- vcedeno@vt.edu
Islam Harb - iharb@vt.edu
Yilong Jin – jin28@vt.edu
Sai Ravi Kiran Mallampati - sairavi5@vt.edu
Social Network and Importance

• Our Role (SN Team)
  – Pick up the social non-content based features.
  – Calculate the Importance Values for Tweets and Web pages based on social network features and connectivity.
Non-Content Features

• Content-based features are out of scope.
• Features are considered with
  – Connectivity (Popularity)
  – Trustworthiness
• This metric “Trustworthiness and popularity” will be mainly reflected in the “User Importance Value (UIV)”.  

• Tweet Specific Features
  – Retweet count: The number of times the tweet has been retweeted.
  – Favorites count: The number of times the tweet has been favored.

• Account authority (User) features
  – Followers count: The number of followers a user has.
  – List Count: The number of lists a user belongs to.
UIV: User Importance Value

U_Edge Weight(1,2) = ( #Mentions of U1 made by U2) / #Mentions[U2] ....   Equation (1)
U_Edge Weight(2,1) = ( #Mentions of U2 made by U1) / #Mentions[U1] ....   Equation (2)

Where:
#Mentions[U1] : Total number of tweets posted by U1 that mention other users.
#Mentions[U2] : Total number of tweets posted by U2 that mention other users.

\[ UIV = \begin{cases} 
\frac{\sum (U_{Edge \ Weight})}{\text{number of inlink edges}} & \text{if there are no inlink edges} \\
0 & \text{otherwise}
\end{cases} \]
Tweet Importance Value

\[
TIV = \sum_{i=1}^{5} W(i) \times A(i) / \text{Number of attributes}
\]

Weights are arbitrary.
- For Simplicity, chosen to be uniform.
- E.g. \( W1 = W2 = W3 = W4 = W5 = 0.2 \).

The Five Attributes

- **A1**: Favorite Count
  \[
  \frac{\# \text{Fav}(i) - \text{Fav}(\text{min})}{\text{Fav}(\text{max}) - \text{Fav}(\text{min})}
  \]

- **A2**: Retweet Count
  \[
  \frac{\# \text{RT}(i) - \text{RT}(\text{min})}{\text{RT}(\text{max}) - \text{RT}(\text{min})}
  \]

- **A3**: List Count
  \[
  \frac{\# \text{List}(i) - \text{List}(\text{min})}{\text{List}(\text{max}) - \text{List}(\text{min})}
  \]

- **A4**: Number of Followers
  \[
  \frac{\# \text{Followers}(i) - \text{Followers}(\text{min})}{\text{Followers}(\text{max}) - \text{Followers}(\text{min})}
  \]

- **A5**: User Importance Value (UIV)
The graph

- **Users:**
  - Edge indicates a mention occurred.

- **Tweets:**
  - Dotted Edges between tweets indicates a retweet.
  - Solid Edges between Tweet and a user indicates the tweet’s ownership.

- **Web pages:**
  - Edges indicates a link/connectivity.
Webpage Importance Value

\[ WPIV_j = \frac{\sum TIV(i)}{t} \times \frac{t - \text{Minimum number of edges to a webpage} + 1}{\text{Maximum number of edges to a webpage} - \text{Minimum number of edges to a webpage} + 1} \]

where,

• “t” is the number of tweets that points to WPIV(j).

• Example:
  - For WP1, the “t” is equal to 3 (comes from T1->T3).
  - For WP2, the “t” is equal to 4 (comes from T3->T4).
Approach (Data Structure)

- JSON File
  - Avro Tweets: Comply with other teams schema
  - Avro User: Only for SN team, basically
    - Number of Followers
    - List Count.
Approach (Sequential vs. Parallel)

- Parallelizing the “Parser” in Python
  - Using Standard Python Multi-processing library.
  - Local machine (Jin’s) – 8 cores
  - ~5x improvement/speedup
GraphX

• Introduces the Resilient Distributed Property Graph: a directed multigraph with properties attached to each vertex and edge.
• RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators.
• Includes a growing collection of graph algorithms and builders to simplify graph analytics tasks
Statistics

- Very small data collection
  - 2553 tweets
  - num of user: 915
  - Reading tweets time: 0.612 second(s)
  - Reading User time: 0.066 second(s)
  - Updating Edges: 0.012 second(s)
  - Calculating UIV: 0.023 second(s) // this contains users that are not in the dataset
  - Calculating tweet score: 0.048 second(s)
Statistics

- **Small data collection**
  - 428574 tweets
  - 265410 users
  - Reading tweets time: 110.551 second(s)
  - Reading User time: 22.568 second(s)
  - Updating Edges: 3.821 second(s)
  - Calculating UIV: 0.951 second(s) // this contains users that are not in the dataset
  - Calculating tweet score: 6.383 second(s)
References


2. Tianyi Wang et al., 2011. "Understanding Graph Sampling Algorithms for Social Network Analysis", 31st International Conference on Distributed Computing Systems Workshops (ICDCSW), Pages 123-128, Minneapolis, MN, USA.


References


Appendix A

top 3 tweets

Israel will be allowed to "act defensively" on tunnels during cease-fire, U.S. says.

http://t.co/4sxohD3iXV

cnnbrk

score 0.240

UIV: None

# of followers: 17842891

list count: 155404

mentioned by users in the collection 0 times

retweet count: 0

fav_count: 0
Ebola vaccine human tests could begin as early as September, National Institutes of Health says. http://t.co/dYwYaqnsf7

cnnbrk

score 0.240

UIV: None

# of followers: 17842891

list count: 155404

mentioned by users in the collection 0 times

retweet count: 0

fav_count: 0
Appendix A

---------------------
RT @JustCallMe_Mese: Retweeting this literally takes 1.8 seconds http://t.co/7MV02iwKVj
miketanz11
score 0.200
UIV: 0
# of followers: 221
list count: 0
mentioned by users in the collection 0 times
retweet count: 55897
fav_count: 0
---------------------