Mining the Social Web

Eric Wete
ericwete@gmail.com
June 13, 2017

Outline

- **The big picture**
- **Features and methods (Political Polarization on Twitter)**
- **Summary (Political Polarization on Twitter)**
- **Features on the Demo (Political Hashtag Trends)**
- **How It Works (Political Hashtag Trends)**
- **Summary (Political Hashtag Trends)**
- **Discussions**

The big picture

- Social web : social relations linking people through World Wide Web
- Mining the social web: explore social web to get useful information
- Problem: get useful info.
- Importance : Medicine, School, Marketing, Politic,…

Source: https://makeawebsitehub.com/social-media-sites

Features and methods (Political polarization on twitter)

- Twitter Platform
- Political Content identification
- Political Communication Networks Analysis
- Cluster Analysis
- Interaction Analysis

Twitter Platform

- Interactions with (Re)tweets
- Interactions with Mentions
- Interactions with Hashtags

Political Content identification

- Political communication : any tweet with at least one politically relevant hashtag
- Relevant hashtag: sample of 2 most political hashtags #p2 (Progressives 2.0) and #tcot (Top Conservatives). For each seed we identified the set of hashtags with which it co-occurred in at least one tweet, and ranked the results using the Jaccard coefficient.

Political Content identification

- Let S be set of tweets containing seed hashtag and T set of tweets containing another hashtag. Jaccard coefficient between S and T :

If σ is big the two hashtags are deemed to be related.

$$\sigma(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

- Around 252,000 tweets

Political Communication Networks Analysis

- Tweets analysis with focus on “retweets” and “mentions”
- Retweets network:
  - an edge runs from a node A to a node B if B retweets content originally broadcast by A.
  - > 23,000 connections among tweets
- Mentions network:
  - an edge runs from node A to node B if A mentions B in a tweet
  - >10,000 (smaller than retweets network)

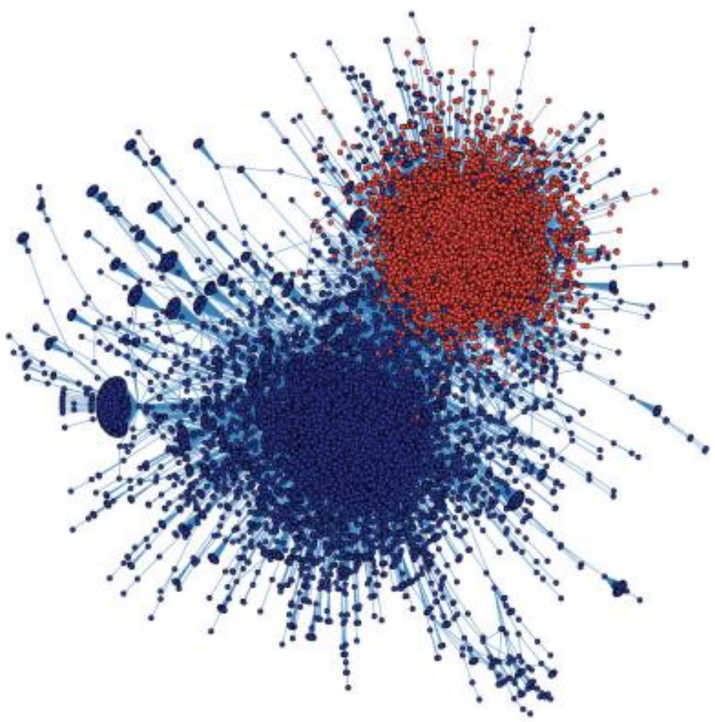
➔Users receive or spread huge amount of information

Cluster Analysis – Community structure

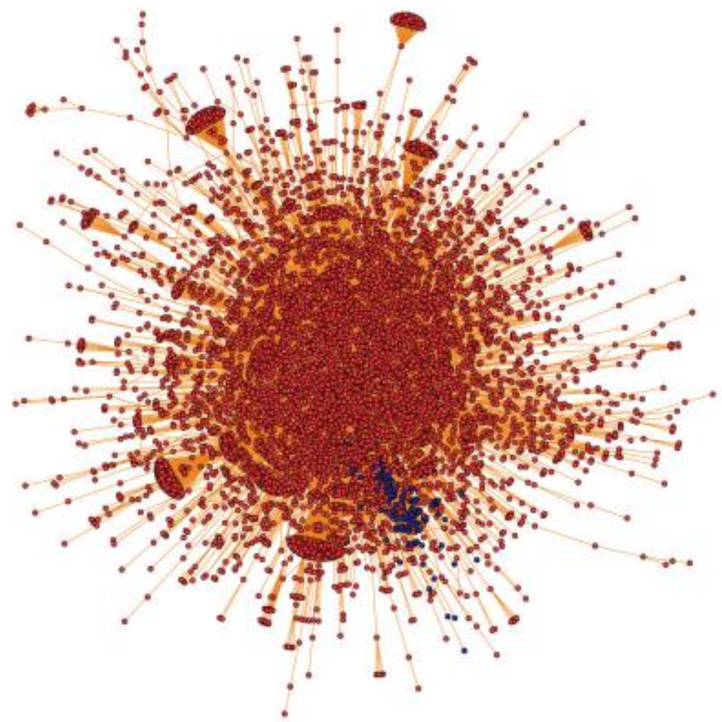
- Community detection using a **label propagation** method for two communities to establish the large-scale political structure of retweet and mention network
- **Label propagation** : assign an initial arbitrary cluster membership to each node and then iteratively updating each node's label according to the label that is shared by most of its neighbors.

Cluster Analysis – Community structure

- Retweet : Two clusters
- Mention: replies among active users



Retweet



Mention

Cluster Analysis – Content Homogeneity

Is the actual content of the discussions involved similar ?

- Association of each user with a profile vector containing all the hashtags in her tweets, weighted by their frequencies.
- Computation of cosine similarities between each pair of user profiles
- The table shows the average similarities in the retweet and mention networks for pairs of users both in cluster A, both in cluster B, and for users in different clusters.

	Retweet	Mention
A↔A	0.31	0.31
B↔B	0.20	0.22
A↔B	0.13	0.26

Cluster Analysis – Political Polarization

Correspond the cluster in retweet network to groups of users of similar political alignment ?

- Qualitative content analysis: techniques from the social sciences
- one author first annotated 1,000 random users who appeared in both the retweet and mention networks.
- 200 random users from the set of 1,000 to establish the reproducibility of this annotation scheme.

Cluster Analysis – Political Polarization

- coders' annotations agreement with an objective judge is Cohen's Kappa, defined as $$\kappa = \frac{P(\alpha) - P(\epsilon)}{1 - P(\epsilon)}$$

where P(α) is observed rate of agreement between annotators and P(ε) is expected rate of random agreement given relative

Network	Clust.	Left	Right	Undec.	Nodes
Retweet	A	1.19%	93.4%	5.36%	7,115
	B	80.1%	8.71%	11.1%	11,355
Mention	A	39.5%	52.2%	8.18%	7,021
	B	9.52%	85.7%	4.76%	154

Interaction Analysis – Cross-Ideological Interactions

- Compare the observed number of links between manually-annotated users with the value we would expect in a graph where users connect to one another without any knowledge of political alignment.

	Mention		Retweet	
	→ Left	→ Right	→ Left	→ Right
Left	1.23	0.68	1.70	0.05
Right	0.77	1.31	0.03	2.32

- For both means of communication, users are more likely to engage people with whom they agree. But it is far less pronounced in the mention network, where we observe significant amounts of cross-ideological interaction.

Interaction Analysis – Content Injection

- Any Twitter user can select arbitrary hashtags to annotate his or her tweets. One explanation might be that he seeks to expose those users to information that reinforces his political views.
- We propose that when a user is exposed to ideologically opposed content in this way, he will be unlikely to rebroadcast it, but may choose to respond directly to the originator in the form of a mention.
- Consequently, the network of retweets would exhibit ideologically segregated community structure, while the network of mentions would not.

Summary (Political polarization on twitter)

- the two major mechanisms for public political interaction on Twitter — mentions and retweets — induce distinct network topologies. The retweet network is highly polarized, while the mention network is not.
- Explanations are most based on the role of hashtags and users applying them to get communities involved.
- we know for certain that ideologically-opposed users interact with one another, either through mentions or content injection, they very rarely share information from across the divide with other members of their community.
- Dataset : http://cnets.indiana.edu/groups/nan/truthy/

Features of the Demo (Political Hashtag Trends)

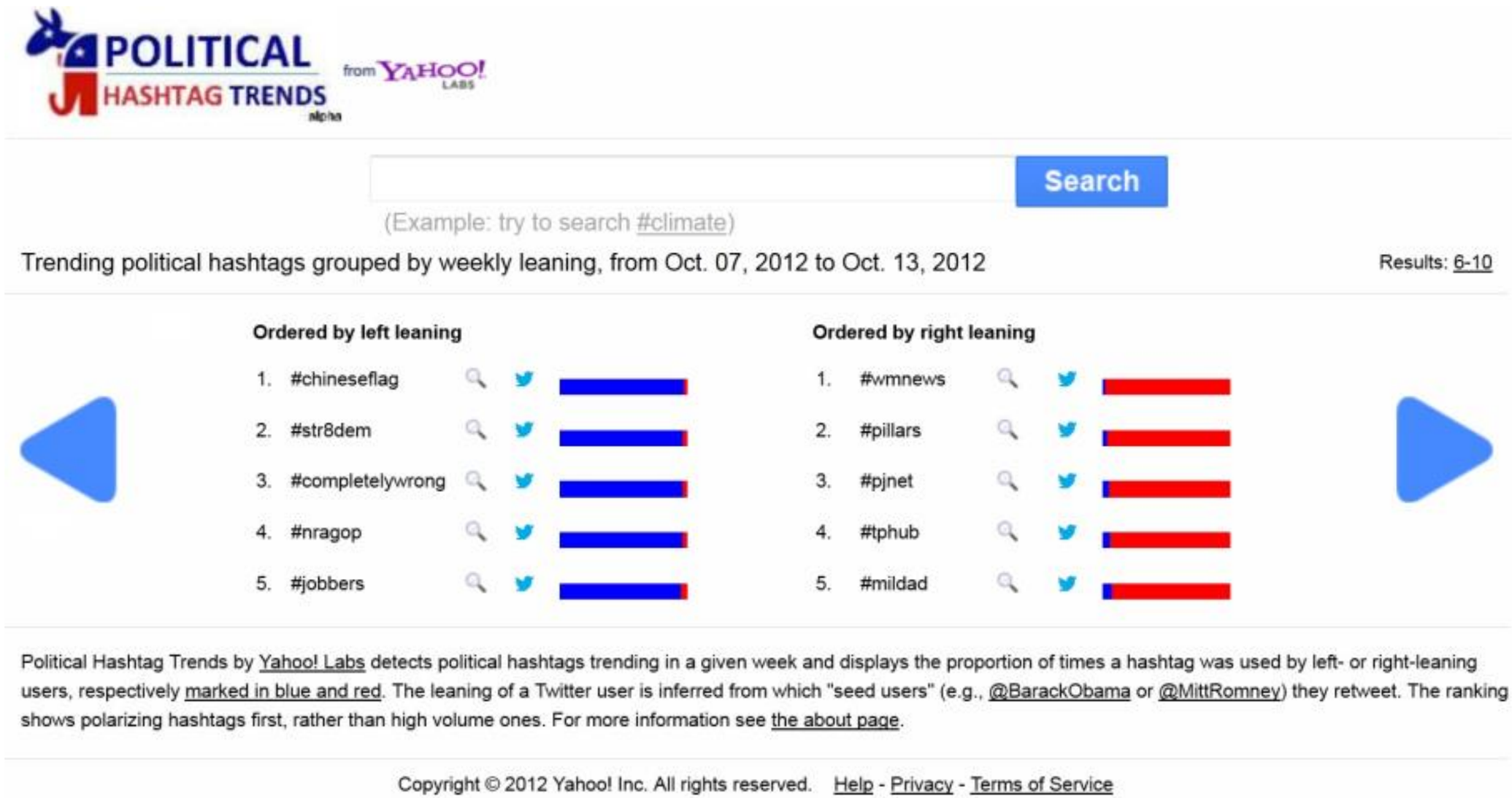
- Lean and Trend Information
- Putting content into a context

Lean and Trend Information

- Political Hashtag Trends (PHT) is an analysis tool for political leftvs.-right polarization of Twitter hashtags. PHT computes a leaning for trending, political hashtags in a given week, giving insights into the polarizing U.S. American issues on Twitter.
- Core functionality
  - Identify trending, political hashtags in a given week
  - Assign a leaning to them

Lean and Trend Information

The starting page of PHT



Putting content into Context

- Twitter current search: show recent tweets for a selected hashtag
- Twitter archive search: show tweets for the week of interest for the given hashtag
- New York Times archive search: Each hashtag is linked to a date-specific search on the New York Times archive

Features and methods

- Seed Users
- Identifying Politicized Users
- Detecting Political Hashtags
- Assigning a Leaning to Hashtags
- Assigning Trending Score to Hashtags

Seed Users

- Set of seed users with known political orientation (eg. @Barackobama, @MittRommey)
- Expansion using retweet behavior and later cleaned by limiting the geographic scope

Seed users selection (twitter account) :

- Had to belong to either a political leader in office or it hat to be an official party account
- For a person, it had to be the “personal” account rather than an office-related account
- It had to be verified twitter account

- Dataset retrieved with :Twitter REST + API Apigee
- Got: 14 accounts for the left and 19 for the right

Identifying Politicized Users

- For each seed users retrieve their publicly available tweets
- For each tweet identity up to 100 retweeters, filter on retweeters' location (here U.S.) => +110,000 users
- For each week, assign these users a fractional leaning corresponding to the ratio of their retweets for the given week
- For retweeters we obtain their public tweets for the given week

Note: This method allows a change in leaning of retweeters

Detecting Political Hashtags

- Look at co-occurrence with a set of hashtags which are deemed to be political
- This seed set include hashtags referring to main political parties and events (#p2, #tcot, #teapary, #gop…) and hashtag containing “obama”, “rommey”,…
- Compute within-week user volumes for each hashtag
- For each week and each leaning, keep 5% top hashtag in term of user volume

User’s volume can contribute fractionally to both leanings

- For each hashtag h, count the number of user who use h at least once in combination with a political seed hashtag
- Keep 25% in term of political-to-all user fractions, again for each leaning
- Merge the two lists and the resulting (h,w) pairs are used in the analysis

Assigning a Leaning to Hashtags

- Use a voting approach to compute the leaning of hashtags
- vL (vR) = aggregated user volume of h in w for the left (right) leaning
- VL (VR) = total left (right) user volume of all hashtags in w

User can contribute fractionally based on his fractional leaning in w

The leaning of h in w is defined as

$$\mathrm{Lean}(h,w) = \frac{\frac{v_L}{V_L} + \frac{2}{V_L+V_R}}{\frac{v_L}{V_L} + \frac{v_R}{V_R} + \frac{4}{V_L+V_R}}$$

Where leaning of 1.0 is fully left and 0.0 is fully right

Assigning Trending Score to Hashtags

- A trending score t(h,w) to h in w is based on the past frequencies of h and the overall frequencies in the given week.
- Sort hashtags by t(h,w) and from the top assign them to either left (Lean(h,w)>=0.5) or right (Lean(h,w)<0.5) leaning.
- For each leaning, keep top 20 in term of t(h,w) and reranked according to Lean(h,w) for the left or –Lean(h,w) for the right
- t(h,w) defined as:

$$t(h,w) = \frac{f(h,w) \Big/ \sum_{h' \in H} f(h',w)}{\sum_{u \leq w} f(h,u) \Big/ \sum_{h' \in H} \sum_{u \leq w} f(h',u)}$$

f(h,w) : user volume for h in w.

Summary (Political Hashtags Trends)

- PHT computes a leaning for trending, political hashtags in a given week, giving insights into the polarizing U.S. American issues on Twitter.
- The leaning of a hashtag is derived in two steps.
  - First, users retweeting a set of "seed users" with a known political leaning, such as Barack Obama or Mitt Romney, are identified and the corresponding leaning is assigned to retweeters.
  - Second, a hashtag is assigned a fractional leaning corresponding to which retweeting users used it. Non-political hashtags are removed by requiring certain hashtag co-occurrence patterns. PHT also offers functionality to put the results into context.
- After the “seed users” step, PHT identifies the Politicized Users, detects Political Hashtags, assign them a Lean and a Trending Score which are use for their ranking.

Discussion

- Political Polarization on Twitter : Contend based methodology (retweets & mentions) : less information here than PHT
- Political Hashtag Trends : Contend (hashtags) and time (1 week) based methodology (public available tweets), user volume and frequency based : here more available information (tweets' history and context) than the above paper

References

- Conover, Michael, et al. *Political Polarization on Twitter*. ICWSM '11.
- Weber, Ingmar, Venkata Rama Kiran Garimella, and Asmelash Teka. *Political hashtag trends*. ECIR '13.