Beyond Keywords: Tracking the development of conversations on social media through linked, nested hashtag clusters

NO AUTHOR GIVEN

In seeking to know how newsworthy events influence public and political conversations, researchers benefit from data available through Social Media. Messages on platforms such as Twitter and Facebook represent a high volume sample of the national conversation in near real time, and with the right handling can give insight into events such as elections, as demonstrated by Huberty (2013) and Tumasjan, Sprenger, Sandner, and Welpe (2010); or processes such as social movements, as demonstrated by Agarwal and Bennett (2014) and DiGrazia (2015). Standard methods of social media analysis use keyword tracking and sentiment analysis to attempt to understand the range of perceptions surrounding these events. While helpful for basic situational awareness, these methods do not help us understand how a set of narratives compete to interpret and frame an issue for action.

For example, if we are interested in understanding the national conversation in reaction to the shooting at Emanuel AME church in Charleston, South Carolina, we could plot a timeseries of the volume of tweets containing the hashtag `#charleston`, as seen in Figure 1.

Alternate methods include assessing the relative 'positive' or 'negative' sentiment present in these tweets, or using a supervised learning categorizer to group messages according to preconceived ideas about their contents, as demonstrated by Becker, Naaman, and Gravano (2011), Ritter, Cherry, and Dolan (2010), and Zubiaga, Spina, Fresno, and Martínez (2011). The addition of demographic information can help researchers understand the context of messages (McCormick et al. 2015)

These techniques, however, are unable to infer from the data coherent patterns of thought that signify interpretation of the events? deeper meanings. Interpretation depends upon making connections between the events as they happen and other concepts in the public discourse. One way to measure these connections is to look as a network of co-citations surrounding our topic of interest, as has been demonstrated by Cogan and Andrews (2012) and Smith and Rainie (2014). Representing hashtags as nodes on a network, each shared message contributes to the weight of an edge between these nodes. This type of analysis is helpful in that it helps us begin to understand the structure of the discourse. If we perform k-clique clustering on the network, we can see which sets of connections form coherent and conversations, as seen in Figure 2.

In this example there are two distinct conversations happening with regards to the event: the first a description of the shooting itself and the human elements of the tragedy. A second conversation focuses on the larger national-scale political conflicts to which the event points. While each of

the conversations is motivated by the same event, they are distinct from one another in the language they use and in the connections they make.

0.1   The contributions of this paper

We may hypothesize that how these conversations develop over time will influence the social and political response to the event. In this paper we will explore methods of identifying and tracking the development of these conversations over time.

Due to the computation-intensive nature of this analysis, we chose to implement the data manipulation algorithms in both python and Unicage shell scripts, [1] for prototyping and speed of execution, respectively. Descriptions of these scripts can be found in the appendices, along with performance comparisons between the two languages.

1.   IDENTIFYING CONVERSATION CLUSTERS

The image in Figure 2 is based upon network closeness between hashtags. Each of the hashtags present in the dataset forms a node in this network, and the relative strength of edges depends upon the number of times the pair occur together in a tweet ? their ?co-occurrence?, using the method of Marres and Gerlitz (2014).

The clusters themselves are then defined by k-clique community detection algorithms implemented in the COS Parallel library, developed by Gregori,

---

[1]For a description of Unicage development tools, see Tounaka (2013)

Lenzini, and Mainardi (2013) and their use is demonstrated in the appendices.

Every node in a cluster must be able to participate in at least one fully connected subgroup of size k with the other members of the group. Thus, the metric k determines how strict we are about closeness between keywords when defining the boundaries of a particular cluster. For example, high values of k would impose strict requirements for interconnectedness between elements of an identified conversation, leading to a smaller, more coherent identified conversation, as seen in Figure 3.

On the other hand, smaller values of k are less stringent about the requirements of connectivity they put on the elements in the cluster, leading to a larger, more loosely coupled group, as seen in Figure 4.

2. REPRESENTING CONVERSATIONAL CLUSTERS AS NESTED SETS

Tight conversational clusters (high k) must necessarily be contained within larger clusters with less stringent connection requirements (low k). Performing clustering along a range of k values allows us to place a specific conversation in context of the larger discourse. It becomes helpful to represent these clusters as nested sets, as seen in Figure 5, ignoring the node and edge construction of the network diagram in favor of something which allows us to observe the nested relationships the conversations have with one another.

In this representation, we are able to observe the tightly clustered 5-clique conversation in context of the 4-clique conversation it inhabits, and the neighboring 4-clique conversations that together inhabit the larger discourse.

3. TRACKING CONVERSATIONS CHRONOLOGICALLY

In order to track how elements of conversation weave into and out of the general discourse, we need to be able to interpret how conversational clusters identified at one point in time relate to those in subsequent intervals. We can do this in one of two ways.

The first method is to track the volume of co-citations identified in the various conversational clusters identified on the first day of the analysis, as it changes over subsequent days. This indicates how well the connections made on the first day maintain their relevance in the larger conversation. Figure 6 shows how the connections made in the conversational clusters shown above in Figure 5 fall in volume over the 10 days subsequent to the initial event, paralleling the decay in pure keyword volume seen in Figure 1.

The second method for tracking conversation volume over time takes into account the changes that happen within the conversation itself. The fundamental assumption in this analysis is that while the words and connections present in a conversation change, they do so incrementally in such a way as to allow for matching conversations during one time period with those in the following time period.

Palla, Barabási, and Vicsek (2007) discuss how communities of individuals develop over time and change. We can use the same techniques to track continuity of conversational clusters. The most basic way to do this is to count the fraction of elements of a conversational cluster at time 0 that are present in each conversational cluster at time 1, and use this fraction as the likelihood that each cluster at time 1 is an extension or contraction of the time 0 cluster in question. From this we can construct a transition matrix relating conversational clusters at time 0 with clusters at time 1, as seen in Table 1.

To improve our estimates, we can take advantage of the fact that clusters that correspond from time 0 to time 1 will participate in a larger cluster that emerges if we perform our clustering algorithm on the union of all edges from the networks at time 0 and time 1. This reduces the number of possible pairings between days, yielding more specificity in our intra-day transition matricies.

These transition matricies can be used to infer how clusters present in the first day?s analysis correspond to clusters in the second day, and so forth. These are visualized as a set of nested cluster diagrams, with traces linking likely clusters together, as seen in Figure 7. Heavier traces between clusters imply more confidence in the transition between the two sets.

The volume of messages forming the edges of each cluster are shown plotted day by day in Figure 8. As the linkage between subsequent clusters is probabilistic, in this plot so to are the links connecting volume measures,

6

with the weight of each link proportional to its likelihood.

4. CONCLUSION

The utility of social media analysis for sociological research can be extended well beyond the practice of tracking keyword volume, net post sentiment, or supervised classification. In particular, tracking conversational clusters within a network of hashtag co-citations can give both structural understanding of a conversation, and insight into how it develops over time.

5. AUTHOR'S NOTE

The full set of scripts used to generate the analysis in this paper can be found at `www.github.com\removed to anonymize`.

REFERENCES

Agarwal, S. and Bennett, W. (2014), "A model of crowd enabled organization: Theory and methods for understanding the role of twitter in the occupy protests," ... *Journal of Communication.*

Becker, H., Naaman, M., and Gravano, L. (2011), "Beyond Trending Topics: Real-World Event Identification on Twitter.," *ICWSM.*

Cogan, P. and Andrews, M. (2012), "Reconstruction and analysis of twit-

ter conversation graphs," *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research.*

DiGrazia, J. (2015), "Using Internet Search Data to Produce State-level Measures: The Case of Tea Party Mobilization," *Sociological Methods & Research*, 0049124115610348–.

Gregori, E., Lenzini, L., and Mainardi, S. (2013), "Parallel k-clique community detection on large-scale networks," *IEEE Transactions on Parallel and Distributed Systems*, 24(8).

Huberty, M. (2013), "Multi-cycle forecasting of congressional elections with social media," ... *of the 2nd workshop on Politics, Elections and Data.*

Marres, N. and Gerlitz, C. (2014), "Interface Methods: Renegotiating relations between digital research, STS and Sociology," unpublished manuscript.

Palla, G., Barabási, A., and Vicsek, T. (2007), "Quantifying social group evolution," *Nature.*

Ritter, A., Cherry, C., and Dolan, B. (2010), "Unsupervised modeling of twitter conversations," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 172–180.

Smith, M. and Rainie, L. (2014), "Mapping twitter topic networks: From polarized crowds to community clusters," unpublished manuscript.

Tounaka, N. (2013), "How to Analyze 50 Billion Records in Less than a Second without Hadoop or Big Iron,".

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010), "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.," *ICWSM*.

Zubiaga, A., Spina, D., Fresno, V., and Martínez, R. (2011), "Classifying trending topics: a typology of conversation triggers on twitter," *Proceedings of the 20th ACM ....*