COACTION PUBLISHING

Who should we be listening to? Applying models of user authority to detecting emerging topics on the EIN

Jason Fries*, Donald Curtis, Alberto Segre and Philip Polgreen

Computer Science, The University of Iowa, Iowa City, IA, USA

Objective

To explore how different models of user influence or authority perform when detecting emerging events within a small-scale community of infectious disease experts.

Introduction

Emerging event detection is the process of automatically identifying novel and emerging ideas from text with minimal human intervention. With the rise of social networks like Twitter, topic detection has begun leveraging measures of user influence to identify emerging events. Twitter's highly skewed follower/followee structure lends itself to an intuitive model of influence; yet, in a context like the Emerging Infections Network (EIN), a sentinel surveillance listserv of over 1400 infectious disease experts, developing a useful model of authority becomes less clear. Who should we listen to on the EIN? To explore this, we annotated a body of important EIN discussions and tested how well 3 models of user authority performed in identifying those discussions.

In previous work, we proposed a process by which only posts that are based on specific 'important' topics are read, thus drastically reducing the amount of posts that need to be read. The process works by finding a set of 'bellwether' users that act as indicators for 'important' topics and only posts relating to these topics are then read. This approach does not consider the text of messages, only the patterns of user participation.

Our text analysis approach follows that of Cataldi et al. (1), using the idea of semantic 'energy' to identify emerging topics within Twitter posts. Authority is calculated via PageRank and used to weight each author's contribution to the semantic energy of all terms occurring in within some interval ti. A decay parameter d defines the impact of prior time steps on the current interval.

Methods

We considered 3 models of authority: (1) (emerging topic detection) ETD uniform (i.e., equal weight); (2) ETD PageRank; and (3) bellwether. For 1 and 2, we built a directed graph of user activity, using thread coparticipation to define edges. Each EIN post was text tokenized, POS-tagged, and had stop words removed. We use a time window $t = 5$ days to aggregate messages terms and a decay parameter $d = 60$ days. We identified emerging terms using an energy threshold approach, where emerging is defined as any term where energy $> k*\mu$ energy over the interval t, where k is some constant; we used $k = 1.3$. Any term identified as emerging that occurs in the postsubject line is flagged as important. For the bellwether model, we algorithmically selected 80 and 90 authors from the year prior to the one under analysis and flagged threads as important when one of those bellwethers participated in it. We also conducted 1000 random trials of selecting subsets of 80 users to follow.

For evaluation, we used an annotated set of EIN threads identified as clinically important. We measured the total number of threads each authority model flagged for reading versus the number of actual important messages.

Table 1. Authority model performance evaulations

Model	*Threads read*	*Threads*	*Read important*	*Important*	*Read unimportant*	*Total*
ETD-Equal	877	1907	30	63	31	62
ETD-PageRank	937	1907	29	63	34	62
Bellwether-90	668	1907	30	63	18	62
Bellwether-80	190	1907	15	63	5	62
Random-80	$\mu = 750$ 49 (SD = 22.94)	1907	$\mu = 34.68$ (SD = 3.26)	63	$\mu = 25.29$ (SD = 3.36)	62

Results

The performance of each authority method, measured as threads recommended by each model and evaluated over all messages from January 2003 to March 2009 (Table 1).

Conclusions

The bellwether model performs best overall, requiring the least messages read while detecting more important threads at less cost of reading unimportant threads. The differences between 80 and 90 bellwethers reflect how parameters influence the trade-off between these 3 measures. There was no significant benefit gained from viewing the EIN network in terms of a PageRank over equal authority, which aligns with previous work identifying PageRank's limitations in identifying experts (2). Moreover, compared to randomly selecting 80 authors to follow, the performance is worse with our chosen parameters. Future work will examine incorporating bellwether authority into a textual analysis framework.

Keywords

Topic detection; text analysis; social networks; social influence analysis

References

1. Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on Twitter based on temporal and social terms evaluation, Proceedings of the Tenth International Workshop on Multimedia Data Mining. 2010:1–10.
2. Zhang J, Ackerman M, AdamicL. Expertise networks in online communities: structure and algorithms, Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada; 2007:221–30.

***Jason Fries**
E-mail: jason-fries@uiowa.edu

 Citation: Emerging Health Threats Journal 2011, **4**: 11127 - DOI: 10.3402/ehtj.v4i0.11127