People of Opposing Views can Share Common Interests

Eduardo Graells-Garrido
Web Research Group
Universitat Pompeu Fabra
Barcelona, Spain
eduard.graells@upf.edu

Mounia Lalmas
Yahoo Labs
London, UK
mounia@acm.org

Daniele Quercia
Yahoo Labs
Barcelona, Spain
dquercia@yahoo-inc.com

ABSTRACT

In online social networks, people tend to connect with like-minded people and read agreeable information. Direct recommendation of challenging content has not worked well because users do not value diversity and avoid challenging content. In this poster, we investigate the possibility of an indirect approach by introducing *intermediary topics*, which are topics that are common to people having opposing views on sensitive issues, i.e., those issues that tend to divide people. Through a case study about a sensitive issue discussed in Twitter, we show that such intermediary topics exist, opening a path for future work in recommendation promoting diversity of content to be shared.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

Social Networks; Topic Modeling; Homophily; Intermediary Topics.

1. INTRODUCTION

Social sciences research has shown that, in social networks, people tend to connect with people with similar beliefs, a phenomena known as *homophily*, *and* prefer to read only agreeable information, a phenomena known as *selective exposure*. Both phenomena have been studied in online settings [4, 3, 2]. For instance, on the abortion issue, `#prolife` users hardly interact with `#prochoice` users in a debate context, although those users could engage in conversation about other interests, such as `#musicmonday`.

Motivated by this scenario, we set out to suggest new connections among users with challenging points of view on sensitive issues, i.e., those issues which tend to divide people. We introduce the concept of *intermediary topics*, which are non-sensitive topics that might help to connect people who have opposing views on sensitives issues but have similar views on intermediary topics. Previous work has focused on a direct approach to recommendation by displaying diverse and challenging information to users, but users do not value diversity [3] and still prefer agreeable and like-minded information [2]. Hence, an indirect approach through intermediary topics could be more feasible.

In this paper, we hypothesize that these intermediary topics exist and are measurable in the microblogging platform Twitter. We demonstrate that using topic modeling on user generated content and measures of centrality and diversity, it is possible to find and quantify these intermediary topics.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577368.

2. METHODOLOGY

A *stance* is defined as a position adopted with respect to something. An *issue* is said to be *sensitive* when stances about the issue tend to divide people. For instance, abortion is a sensitive issue in many countries, whereas musical taste is usually not. For a given sensitive issue, we collect relevant tweets based on keywords and *hashtags* that are associated with it. For instance, `#prolife` and `#prochoice` represent two abortion stances.

The collected tweets are used to construct several *issue stance documents* (i.e., concatenation of tweets from an issue stance) and *user documents* (i.e., concatenation of tweets authored by an user), which we represent as vectors. For each stance, we define a *stance vector* $\vec{s}$ in which each element refers to the importance of a given word w. For each user, we define a *user vector* $\vec{u}$, where each element refers to the importance of a given word w. In both definitions, word importance is weighted using TF-IDF with respect to the corpus of *issue stances*, as stances in different sensitive issues might be related because of ideology. We define the *user stance* $\vec{v_i}$ as a vector where each element v_j corresponds to the cosine similarity between the *user vector* $\vec{u_i}$ and the *stance vector* $\vec{s_j}$.

Topic Graph. We explore the topical diversity of *user documents* by performing *Latent Dirichlet Allocation* [1] to its corpus. LDA is a generative model that, given a number of topics k and a corpus, estimates which words contribute to each topic and which topics contribute to each document. We build an undirected *topic graph* where each LDA topic is a node, two nodes are connected if the two corresponding topics contribute to the same document, and edges are weighted based on the fraction of documents that contributed to it. We filter the edges based on their weight, leaving only those in the upper 10%, and compute the *betweenness centrality* of nodes in the resulting graph. We define *topic diversity with respect to a sensitive issue* as the *Shannon entropy* $H = -\sum_{i=1}^{N} p_i \log p_i$, where p_i is the fraction of users in stance i for a given issue, and N is the number of issue stances for that particular sensitive issue. We define *intermediary topics*

as topics whose betweenness centrality and topic diversity are higher than the median of both measures in the entire topic graph.

3. CASE STUDY

Our case study is focused on intermediary topics in the Chilean population on Twitter. In Chile, abortion is a sensitive issue, as the country has one of the strictest and severe abortion laws in the world. In the context of on-going campaigns for presidential elections, we crawled tweets from July 2013 to August 2013 using the *Twitter Streaming API*.[1] Initially, we used *query keywords* about known issues and hashtags: *abortion* (issue), *education* (issue), *gay marriage* (issue), *Michelle Bachelet* (candidate), *Evelyn Matthei* (candidate), *Santiago* (location), among others. We also added emergent hashtags related to news events that happened during the crawling period. For instance, `#yoabortoel25` is about a protest held on July 25th.

In total, we crawled $4,611,998$ tweets from $768,641$ users. Of those tweets, $75,432$ from $40,201$ users were related to abortion. Of those users, only $7,518$ reported a Chilean location in their profiles. Those users authored $1,962,941$ tweets about sensitive issues. We work with this subset of the crawled dataset.

Issues and Abortion Stances. We manually selected the top 200 keywords related to sensitive issues in Chile to define a corpus of documents built with the tweets containing their corresponding hashtags and keywords. Following our methodology, we built two *stance vectors* using relevant hashtags related to `#prochoice` and `#prolife` stances. Those vectors were weighted using TF-IDF according to the corpus. We estimated the *user vectors* for these users, by weighting the words used in their tweets according to the corpus. We estimated the *user stances* on abortion by computing the cosine similarity between *user vectors* and the *stance vectors*, displayed in aggregated form in Figure 1 left: `#prochoice` similarity in x and `#prolife` similarity in y. We observe two kinds of users: polarized (those who employ vocabulary from one stance only, or few vocabulary from both) and non-polarized (those who employ vocabulary from both stances and are situated on the upper-right part of the diagonal in Figure 1 left). It is surprising that many users lie on the diagonal of the plot, as showcased by the color intensity. Those findings are discussed in Section 4.

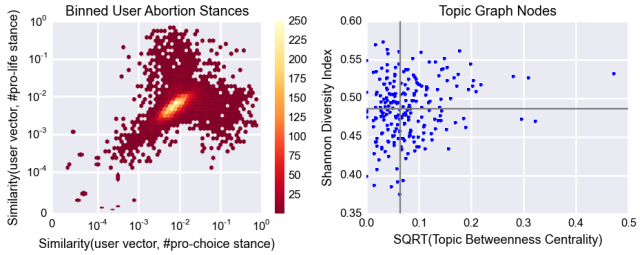

Figure 1: Left: Binned user stances: similarities between *user vectors* and *stance vectors* of `#prochoice` in x and `#prolife` in y. Right: Centrality and Shannon entropy from the *topic graph* nodes. Grey lines depict the medians of each variable.

[1] `https://dev.twitter.com/docs/streaming-api`. Data was crawled by the first author.

Intermediary Topics. In addition to our initial dataset, we crawled $2,521,113$ tweets authored or retweeted by our user pool from December 6th, 2013 until January 3th, 2014. Then, we created *user documents* by concatenating those tweets and the tweets they have published before about sensitive issues. We ran LDA with $k = 300$ and built the *topic graph*, with 234 nodes and $4,716$ edges remaining after filtering. Betweenness centrality has a mean value of 0.0098, 0.0199 std. dev and 0.0041 median, and diversity has a mean value of 0.4855 with 0.0387 std. dev., and a median of 0.4861. We estimated diversity considering a binary classification of user stances, based on the stance with the highest similarity for each user. Topics with centrality and diversity above the corresponding medians are what we define as *intermediary topics*. They are displayed in the upper-right quadrant of Figure 1 right. A manual inspection of words contributing to these topics reveals that most of them are about trending topics and non-sensitive conversations (music, places, etc.).

4. DISCUSSION AND CONCLUSIONS

Although Chile is a highly polarized country on a number of sensitive issues, in particular abortion, we found that user polarity is lower than expected, and that many users employ vocabulary related to both abortion stances. A possible explanation is that our TF-IDF weighting used in the *stance vectors* is softening the stronger stances on the issue. Nonetheless, *intermediary topics* do exist and are measurable. The existence of these topics is important as we can make use of them to enhance current social networks with mechanisms to connect users based on partial homophily while still providing a degree of diversity.

Future Work. The main question to target is for whom intermediary topics should be used. The distribution of user stances hints that intermediary topics, as defined here, can be useful to recommend content to polarized users. To confirm this, we will revisit the computation of the *stance vectors* to minimize artifacts and softening of stances. Next, we will define how to evaluate qualitatively the diversity of found intermediary topics, as popular terms might not be really diverse. Finally, evaluations of recommendations using intermediary topics will be challenging as user reception can be different than in a relevance-only context.

Acknowledgments. This work was partially funded by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain.

5. REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] Q Vera Liao and Wai-Tat Fu. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the ACM CHI*, pages 2359–2368, 2013.

[3] Sean A Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the ACM CHI*, pages 1457–1466, 2010.

[4] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM WSDM*, pages 261–270, 2010.