Ideology Detection for Twitter Users via Link Analysis

Yupeng Gu[1](✉), Ting Chen[1], Yizhou Sun[1] and Bingyu Wang[2]

[1] University of California Los Angeles, Los Angeles, USA
{ypgu,tingchen,yzsun}@cs.ucla.edu
[2] Northeastern University, Boston, USA
rainicy@ccs.neu.edu

Abstract. The problem of ideology detection is to study the latent (political) placement for people, which is traditionally studied on politicians according to their voting behaviors. Recently, more and more studies begin to address the ideology detection problem for ordinary users based on their online behaviors that can be captured by social media, e.g., Twitter. As far as we are concerned, the vast majority of the existing methods on ideology detection on social media have oversimplified the problem as a binary classification problem (i.e., liberal vs. conservative). Moreover, though social links can play a critical role in deciding one's ideology, most of the existing work ignores the heterogeneous types of links in social media. In this paper we propose to detect *numerical* ideology positions for Twitter users, according to their *follow*, *mention*, and *retweet* links to a selected set of politicians. A unified probabilistic model is proposed that can (1) integrate heterogeneous types of links together in determining people's ideology, and (2) automatically learn the quality of each type of links in deciding one's ideology. Experiments have demonstrated the advantages of our model in terms of both ranking and political leaning classification accuracy.

1 Introduction

Ideology detection, i.e., ideal point estimation, dates back to early 1980s, where political scientists first studied politicians' political affiliation using their roll call voting data [14]. Recently, more and more studies pay attention to ideology detection for users on social media, which captures rich information for ordinary citizens in addition to political figures. However, there are two major limitations of the existing literature. First, most of these approaches oversimplify the ideology detection problem as a binary classification problem (liberal/conservative), while ignoring the fact that people's ideology lies in a very broad spectrum. Second, despite the successful utilization of link information in determining one's ideology, most of the works ignore the heterogeneous link types in social media, which leads to significant information loss.

In this paper, we propose a unified probabilistic model to detect *numerical* ideology positions for Twitter users, according to their *follow*, *mention*, and *retweet* links. Although defined on Twitter network, our approach is very general to other social networks. Our approach is able to combine multiple types of links in determining people's ideology with different weights for each link type. In addition, the strength of each link type can be automatically learned according to the network. Experiments shows that (1) using multiple types of links is better than using any single type of links alone to determine one's ideology, and (2) the detected ideology for Twitter users aligns with our intuition quite well.

2 Approach

In this section, we introduce our solution to the proposed problem. We start from ideology model under a single link type, then introduce how to extend the model when multiple types of links exist, and finally introduce the learning algorithm.

2.1 Ideology Estimation Model via Single Link Type

As in traditional ideal point models, each user has an intrinsic position in a K-dimensional space $\boldsymbol{p}_i \in \mathbb{R}^K$, which represents his/her ideology. For a politics-related network, ideology can help explain the reason for link generation, which is a reflection of people's online behaviors. Take *follow* link as an example: the proximity of two users' positions in the latent ideology space indicates a high probability that they have many politician friends (followees) in common, and vice versa. Inspired by [2], we analogize the action of following others to one's voting behavior, and define the probability that user u_i follows user v_j as $p(i \to j) = \sigma(\boldsymbol{p}_i \cdot \boldsymbol{q}_j + b_j) := \sigma_{ij}$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. $\boldsymbol{q}_j$ can be interpreted as the image or impression vector of user v_j when viewed by others, and b_j can be regarded as a bias term for v_j which denotes her popularity. The dot product between two feature vectors can be regarded as a similarity measure in the vector space. Treating each link as generated by a Bernoulli distribution with parameter σ_{ij}, we are able to write down the log-likelihood of observing a network G as

$$l(G) = \log\left(\prod_{(i,j)} \sigma_{ij}^{I_{[i\to j]}}(1-\sigma_{ij})^{1-I_{[i\to j]}}\right) = \sum_{(i,j):i\to j} \log\sigma_{ij} + \sum_{(i,j):i\nrightarrow j} \log(1-\sigma_{ij}) \qquad (1)$$

where $I_{[\cdot]}$ is the indicator function. We denote the set of existing links as S_+ and sampled set of non-existing links as S_- in the remaining of this paper.

2.2 Ideology Estimation Model via Multiple Link Types

We now address the challenge of utilizing multiple types of links for ideology detection. In a heterogeneous network, nodes can be connected via different types of relations. On Twitter, people can *follow*, *mention* or *retweet* others. It naturally forms three different types of links, and different link types certainly have different interpretations. According to our previous assumption, the intrinsic ideology $\boldsymbol{p}_i$ will be consistent across all link types. However we posit that the images of users will change when observed by different types of behaviors. For example, u_i can easily decide to *follow* v_j but hesitates to *retweet* from v_j.

Therefore, $\boldsymbol{q}_j$ and b_j will be changed to relation-specific parameters $\boldsymbol{q}_j^{(r)}$ and $b_j^{(r)}$. In consideration of the heterogeneity in different types of links, we also add a relation weight w_r which represents the relative importance of the links in the corresponding relation type r. Besides, we will use the average log likelihood of each link in each type of relation in order to balance the scale of different link types. An l_2 regularization term is added on parameters to avoid overfitting.

Denoting $\boldsymbol{P} = \{\boldsymbol{p}_i\}, \boldsymbol{Q} = \{\boldsymbol{q}_j^{(r)}\}$ and $\boldsymbol{B} = \{b_j^{(r)}\}$, we define our objective as

$$\begin{aligned} l(\boldsymbol{G}|\boldsymbol{P},\boldsymbol{Q},\boldsymbol{B}) = &\sum_{r=1}^{R} w_r \cdot \frac{1}{N_r}\Big(\sum_{(i,j)\in S_{+}^{(r)}} e_{r,ij}^{+} \log \sigma_{r,ij} + \sum_{(i,j)\in S_{-}^{(r)}} e_{r,ij}^{-} \log\left(1-\sigma_{r,ij}\right)\Big) \\ &-\frac{\mu}{2}\left(||\boldsymbol{P}||_F^2 + \sum_{r=1}^{R} ||\boldsymbol{Q}^{(r)}||_F^2 + \sum_{r=1}^{R} ||\boldsymbol{b}^{(r)}||_2^2\right) \end{aligned} \quad (2)$$

where $\sigma_{r,ij} = \sigma(\boldsymbol{p}_i \cdot \boldsymbol{q}_j^{(r)} + b_j^{(r)})$ for short, N_r is the total number of links in relation r, and $\mu > 0$ is a parameter that controls the effect of regularization terms. The constraint we put on w_r is $w_r > 0,\ r = 1, \cdots, R$ and $\prod_{r=1}^{R} w_r = 1$.

3 Experiments

3.1 Data Preparation

We first collect the list of all the members of the 113^{th} U.S. congress (2013-2015). Then we use Twitter's API to collect their followees and followers. We collect at most 5,000 followers and followees for every congressman. On one hand, in order to select politics-related users, we set a threshold where we keep users who follow at least t congressmen or are followed by at least t congressmen. We choose $t = 20$ in consideration of efficiency. On the other hand, we also include around 10,000 random users who follow 3∼5 politicians as more peripheral (less politics-related) Twitter users. Our approach will be evaluated on this Twitter subnetwork with these users as vertices. Finally we collect their most recent tweets[3] up to Jan. 2016. Social networks for different relations (*follow*, *mention* and *retweet*) are built from the friend lists and extracted from one's tweets. In total, 46,477 users are involved in the dataset and the number of edges for *follow*, *mention* and *retweet* networks is 1.8M, 2.4M and 718K, respectively.

3.2 Performance Evaluation

Baseline Methods We compare our Multiple Link Types Ideal Point Estimation Model (*ML-IPM*) with the following baseline methods:

- *AVER*: the simplest baseline where the ideology of a user is the average score of her outgoing neighbors. Each Republican is assigned an ideology score of 1, and each Democrat is assigned a score of -1.
- *B-IPM* (Bayesian Ideal Point Estimation Model) [1]. Although the author does not mention their generalization to relations other than *follow*, we adopt the model for other types of links for comparison.

[3] Due to API limits, only the most recent 3,200 tweets for each user are available.

- *SL-IPM*: our Single Link Type Ideal Point Estimation Model where only one type of link is present in the social network, as introduced previously.
- *ML-IPM-fixed*: a special case of our model ML-IPM where the weights for different types of links are fixed. In this case the weights for different link types are uniformly distributed, namely $w_1 = w_2 = \cdots = w_R = 1$.

Evaluation Measures In our experiments, we will evaluate the ranking and classification accuracy to demonstrate the effectiveness of our model.

Ranking. In order to evaluate the effect of continuous ideology, we design the ranking evaluation based on 100 manually labeled users, with integer labels from 1 (most liberal/left) to 5 (most conservative/right). The manual labels are obtained by reading their profile information and tweet content, which is never used in the training stage. Here we evaluate the pairwise accuracy between Twitter users, where a pair of users is considered correct if the order of their 1-dimensional ideologies aligns with the order of manual labels. The accuracy is defined as the fraction of correct pairwise arrangements between these users. We use five different sets of random initialization for the model parameters, and report the mean and standard deviation on a total of 3,857 pairs of users in Table 1, where the relation in the bracket represents the type of link used in the corresponding method.

Classification. In the classification task, we classify users as liberal or conservative based on the ideology we have inferred from the dataset. To obtain the ground truth of some users in our dataset, we collect congress people's party affiliation as well as the political leaning for 100 popular newspaper accounts[4], and we also take advantage of the labeled users in our previous task. These multi-dimensional ideal points are used to train a logistic regression classifier. The classification performance is measured by the Area Under ROC Curve (AUC), and is averaged over 10 different runs by different samples of training data. The mean AUC and standard deviation are reported in Table 1. We select the ideology dimension to be $K = 5$ in our method.

3.3 Case Studies

We visualize the latent ideology position of Twitter users. We collect all users in our dataset who claim themselves to live in one of the 50 states in the U.S. (or Washington, D.C.), and calculate the average ideology for each area. Then we are able to map the average score to a color between red and blue. As a result, 9,362 users are identified and 29 states are labeled as red (conservative), as shown in Fig. 1. We can see that the colors of most areas agree with recent election results: states along the west coast and new England area are mostly liberal; while most conservative states lie in the midwest and south region.

4 Related Work

4.1 Ideology Detection in Roll Call Voting Data

Ideal point models attempt to estimate the position of each lawmaker in the latent political space. Legislative voting is one of the sources for quantitative

[4] Source: `http://www.mondotimes.com/newspapers/usa/usatop100.html`

Table 1. Experimental Results (test data)

Method	Ranking Accuracy (%)	Classification AUC (%)
AVER (*follow*)	42.7	52.3
AVER (*mention*)	44.6	55.8
AVER (*retweet*)	47.4	58.7
B-IPM (*follow*)	44.3 ± 10.2	86.8 ± 2.1
B-IPM (*mention*)	43.3 ± 18.3	55.8 ± 6.4
B-IPM (*retweet*)	50.1 ± 12.7	56.1 ± 6.6
SL-IPM (*follow*)	62.6 ± 1.1	95.3 ± 1.5
SL-IPM (*mention*)	62.3 ± 2.7	95.1 ± 1.8
SL-IPM (*retweet*)	63.7 ± 0.5	95.8 ± 0.5
ML-IPM-fixed	65.5 ± 0.8	93.0 ± 3.5
ML-IPM	$\mathbf{66.3 \pm 0.7}$	$\mathbf{98.6 \pm 1.3}$

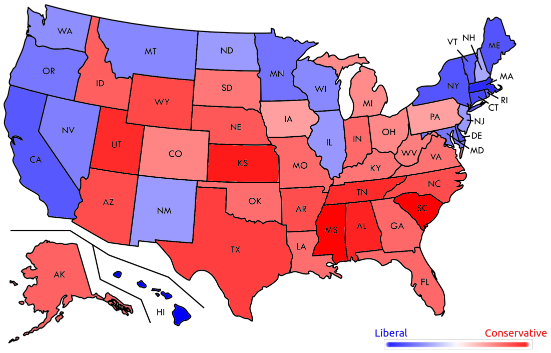

Fig. 1. Average ideology for Twitter users in each state. Darker red means more conservative, while darker blue means more liberal.

estimation of lawmakers' ideal points. Poole and Rosenthal [14] were among the first few researchers in political science domain to provide a thorough and rigorous approach for ideology estimation, which has been generalized by numerous other political science scholars [15, 8, 16, 11, 10, 2]. Researchers study the public voting record of lawmakers and model the probability of each vote, which is usually described as the interaction of the lawmaker's ideal point and the position of the bill. Along this line of research, computer science researchers extend the ideal point model to a variety of aspects, including applying natural language processing and topic modeling techniques on bills [3, 4, 7, 9, 13].

4.2 Ideology Detection in Social Networks

Apart from voting records, recently many approaches have been using information from social networks to analyze user's political leaning. Typically, inference of a user's ideal point is made by exploring her neighbors and her relationship with labeled users (e.g. politicians). Therefore, a simple yet intuitive approach would be calculating the ratio of Democrats and Republicans that a user be-friends with [5]. Wong et al. [17, 18] assume liberal people tend to tweet more about liberal events and the same for conservative users. Barberá [1] proposes a probabilistic model to describe the likelihood of the social network, where the probability of a link is defined as a function of ideal points of both users.

5 Conclusion

In this paper we present a novel approach for ideology detection on Twitter using heterogeneous types of links. Instead of predicting binary party affiliations of users, we focus on a more comprehensive task of detecting continuous ideal points for Twitter users. In addition, we improve over traditional ideology estimation models by integrating information from heterogeneous link types in social networks. Specifically, our model is able to automatically update the importance scores of various relations on Twitter. The experimental results on a subnetwork of Twitter show our advantage over the baseline methods.

References

1. P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.
2. J. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370, 2004.
3. S. Gerrish and D. M. Blei. Predicting legislative roll calls from text. In *Proc. of the 28th Int. Conf. on Machine Learning (ICML'11)*, pages 489–496, 2011.
4. S. Gerrish and D. M. Blei. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems (NIPS'12)*, pages 2762–2770, 2012.
5. J. Golbeck and D. Hansen. Computing political preference among twitter followers. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 1105–1108, 2011.
6. Groseclose, Tim and Milyo, Jeffrey A measure of media bias In *The Quarterly Journal of Economics*, pages 1191–1237, 2005.
7. Y. Gu, Y. Sun, N. Jiang, B. Wang, and T. Chen. Topic-factorized ideal point estimation model for legislative voting network. In *Proc. of the 20th Int. Conf. on Knowledge discovery and data mining (KDD'14)*, pages 183–192, 2014.
8. J. J. Heckman and J. M. S. Jr. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics*, 28:pp. S142–S189, 1997.
9. M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik. Political ideology detection using recursive neural networks. *Association for Computational Linguistics*, 2014.
10. S. Jackman. Multidimensional analysis of roll call data via bayesian simulation: identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241, 2001.
11. J. Londregan. Estimating legislators' preferred points. *Political Analysis*, 8(1):35–56, 1999.
12. J. Lott and K. Hassett Is newspaper coverage of economic events politically biased? *Public Choice*, 160 (1-2):65–108, 2014.
13. V.-A. Nguyen, J. Boyd-Graber, P. Resnik, and K. Miler. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proc. of ACL*, 2015.
14. K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384, 1985.
15. K. T. Poole and H. Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, 35(1):pp. 228–278, 1991.
16. K. T. Poole and H. Rosenthal. *Congress: A Political-Economic History of Roll Call Voting.* Oxford University Press, 1997.
17. F. Wong, C. Tan, S. Sen and M. Chiang. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*, 2013
18. F. Wong, C. Tan, S. Sen and M. Chiang. Media, pundits and the us presidential election: Quantifying political leanings from tweets. In *Proc. of the Int. Conf. on Weblogs and Social Media*, 2013
19. D, Zhou, P, Resnick and Q, Mei. Classifying the Political Leaning of News Articles and Users from User Votes. In *ICWSM*, 2011