An Integrated Model For User Attribute Discovery: A Case Study on Political Affiliation Identification

Swapna Gottipati[†], Minghui Qiu[†], Liu Yang[†,‡], Feida Zhu[†], and Jing Jiang[†]

[†] School of Information Systems, Singapore Management University
[‡] School of Software and Microelectronics, Peking University
{swapnag.2010,minghui.qiu.2010,liuyang,fdzhu,jingjiang}@smu.edu.sg

Abstract. Discovering user demographic attributes from social media is a problem of considerable interest. The problem setting can be generalized to include three components — users, topics and behaviors. In recent studies on this problem, however, the behavior between users and topics are not effectively incorporated. In our work, we proposed an integrated unsupervised model which takes into consideration all the three components integral to the task. Furthermore, our model incorporates collaborative filtering with probabilistic matrix factorization to solve the data sparsity problem, a computational challenge common to all such tasks. We evaluated our method on a case study of user political affiliation identification, and compared against state-of-the-art baselines. Our model achieved an accuracy of 70.1% for user party detection task.

Keywords: Unsupervised Integrated Model, Social/feedback networks, Probabilistic Matrix Factorization, Collaborative filtering

1 Introduction

User demographic attributes such as gender, age, financial status, region are critically important for many business intelligence applications such as targeted marketing [1] as well as social science research [2]. Unfortunately, for reasons including privacy concerns, these pieces of user information are not always available from online social media platforms. Automatic discovery of such attributes from other observable user behavior online has therefore become an important research topic, which we call the user attribute discovery problem for short.

Existing work on detecting the user demographics on datasets such as blogs, micro-blogs and web documents [3,4,5,6,7,8] have mainly adopted the supervised approach and relied on either the connections among users such as user social network, or the language aspects in the data, or both. However, in many cases the interaction between the users and the topics is not effectively incorporated.

The first contribution of our work is that we proposed an integrated unsupervised model which takes into consideration all the three components integral to the user attribute discovery problem, namely the users, the topics and the feedback behavior between the user and the topics. In particular, besides social links between users, we exploit users' feedback on topics, which gives great insight into user affiliation that cannot be modeled by current approaches. Although illustrated with the case study on political affiliation, our model by design can be generalized for most other user attributes that are associated with users' behavior including but not limited to religion affiliation, technology affiliation, political affiliation etc., We present elaborated motivation for our model in Section 2.

The second contribution of our work is that we proposed a solution to a computational challenge common to the user attribute discovery problem: data sparsity — users might not participate in all the huge number of topics and the user social network could be sparse as sites such as forums and debates are not meant for maintaining social relations but to voice out public opinions. As such standard clustering [9] and community detection algorithms [10,11] would not give satisfactory results. We adopt collaborative filtering with probabilistic matrix factorization (PMF) [12], a technique that has been successfully applied for collaborative filtering-based recommendation tasks such as social recommendation [13]. In general, the intuition behind PMF is the assumption that, if two users have same rating/stance/opinion on item/topic/user, they tend to behave the same on other items/topics/users. PMF automatically discovers a low-rank representation for both users and items based on observed rating data. We then apply clustering algorithms on users to detect the communities of users

Lastly, we evaluated our method on data set collected from the CreateDebate site[1], and compared against state-of-the-art baselines as well as degenerative versions of our model. Among the various demographic attributes, the rapidly growing attention for user political affiliation is probably the most noticeable [3,14,15]. Therefore in this paper we study our model's performance on party affiliation detection problem. Our model improves the accuracy and gives promising results for political affiliation detection task.

2 Problem Setting

To motivate our integrated model, we present an analysis of the typical problem setting of user attribute discovery from social media data. We first present an overview of the data components, followed by a motivating example to demonstrate the importance of each element as well as the insufficiency of each if used alone. We finally point out the computational challenge of data sparsity which is common for a large class of user attributes including political affiliation as well as challenge of integrating model components in a principled approach.

2.1 Data Components

We define a typical problem setting of discovering user attributes from social media data to include three components — (I)users, (II)topics and (III)behaviors — which is illustrated in Figure 1. A topic here is defined as any item like a movie or abortion that users can feedback/rate upon. Two kinds of *behaviors*

[1] `www.createdebate.com`

are usually available from the social media data, (1) social behavior between users and (2) feedback behavior of users on topics, which we detail as follows.

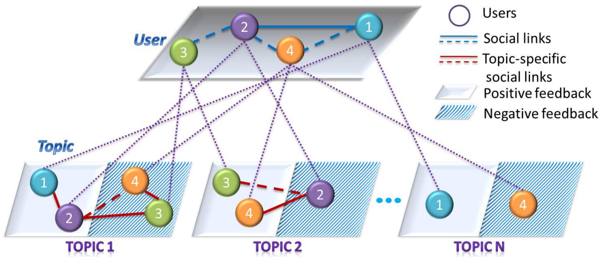


Fig. 1: An illustration of users, topics and behaviors. Social links are in user layer and topic-specific social links are in topic layers, where solid and dotted links represent positive and negative links respectively. Feedback behaviors are exhibited through user feedbacks on topics.

(1) Social behavior between users.
The social behavior can be further categorized into two types. One type is the *topic-independent* one which is usually more long-term and stable, e.g., friendship/enmity, which we represent by *User Social Matrix*. Topic-independent social behavior is an important component used in several studies for prediction, recommendation and community detection tasks. The social friendship/enmity networks can be built from the friendship information or friendship/enmity information or sender/receiver or follower/followee information depending on the type of the network structure [14,16]. In Figure 1, under "User" layer, the social links represent the social matrix.

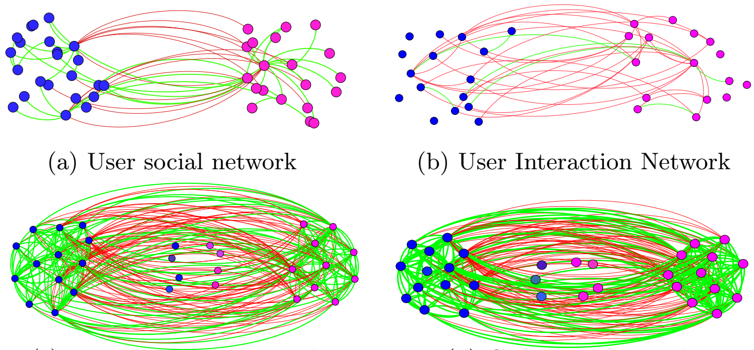
The other type is the *topic-specific* one reflecting users relationship on a particular topic, e.g., agreement/disagreement or thumbsup/thumbsdown on other user's feedback for a specific topic, which we represent by *User Interaction Matrix*. An important observation is that in forums or debate sites, users tend to dispute or agree with others on the debate issues by replying directly to the commenter. [17] observed that users not only interact with others who share same views, but also actively engage with whom they disagree. In Figure 1, under "Topic" layer, the topic-specific social links represent the interaction matrix. A pair of users exhibit different interactions across topics.

(2) Feedback behavior of users on topics.
We focus on explicit user feedbacks such as ratings or stances on topics that can be observed as user opinions towards different opinion targets, represented by a *User Feedback Matrix*. In Figure 1, feedback behaviors are exhibited in user feedbacks on topics. This model, with slightly different variations, has been adopted by many previous work in social network and media analysis [13] [18]. The difference is that, while their problem is usually social recommendation, our task here is to discover users' implicit attributes.

2.2 Correlation Analysis

In Figure 2, we show the networks from our debate dataset using Gephi[2]. We use this data as an example to illustrate two observations.

[2] `http://wiki.gephi.org`



(a) User social network (b) User Interaction Network

(c) User Feedback Network (d) Combined Network

Fig. 2: Visualization of networks from a subset of our corpus. We used ground truth to represent the nodes and edges. Blue nodes are republicans, purple nodes are democrats, green edges represent positive(friendship/agreement) links and red edges represent negative(enmity/disagreement) links. For feedback and combined networks, a link between users in the network indicates whether these two users hold the same feedback most of the time(green edge means yes, red edge means no). Modular communities and nodes in the middle represents small individual communities that are misaligned.

(I) Each type of behavior provides an important insight into users' political affiliation; For example, Figure 2a), a social network shows more friendships (green edges) within parties than inter-parties. Figure 2b), an interaction network shows large disagreements (red edges) among users but very few agreements among them. In Figure 2c), a feedback network, we observe that within parties, the support on topics (green edges) is high compared to inter-parties. These observations indicate that both behaviors provide unique and important insights to users' attribute affiliation.
(II) Each type of behavior alone is not sufficient to accurately identify users' political affiliation. For example, Figures 2a) and 2b) consists of singletons which make the task of identifying the political affiliation harder. For combined network, Figure 2d), we have fewer misaligned users compared to feedback network indicating the benefits of combining social and feedback behavior.

Combining (I) and (II), we drive home the importance of an integrated model to consider all the networks.

2.3 Computational Challenge

A common challenge in discovering user attributes from social media is data sparsity, which is actually shared by all data settings where the number of topics is huge and user participation is sparse. The debate data is a good case in point from the above analysis. Figure 2d) indicates the benefits of leveraging the feedback behavior together with social behavior. However, the current approaches do not cater such integration [19,20,10]. Hence, the second challenge that arises is integration of the model components.

To overcome the first challenge, our solution is motivated by collaborative filtering technique. The second challenge motivates the proposal of solution model that captures all these components in a principled approach. We propose a technique based on probabilistic matrix factorization (PMF), a collaborative filtering approach which could be easily extended to incorporate multiple matrices.

3 Solution

We first provide some preliminaries of our work. The corpus consists of user data (profile and network) and debate data (stances and interactions).
User Social Matrix: We use $\mathcal{S}$ to denote the user social network where each entry $s_{i,j}$ indicates the relationship between i-th user and j-th user (0 means enmity and 1 means friendship).
User Interaction Matrix: Each reply argument in our data set has an interaction information - *Disputed*, *Support* or *Clarified* to its recipient argument. We use $\mathcal{O}$ to represent agreement (*Support* and *Clarified*) and disagreement (*Disputed*) positions between users[3]. An entry $o_{i,k}$ in $\mathcal{O}$ equals to 1 means user i and user k mostly agree with each other and 0 otherwise. In case of such information is not available, we can use methods such as [21] to derive interaction network.
User Feedback Matrix: Feedback refers to user stances captured from *Side* status of a user. $\mathcal{R}$ represents the stances held by different users on the various opinion targets[4], where an entry $r_{i,m}$ is a stance score (0 for "Oppose" and 1 for "Support") indicating the i-th user's stance towards the m-th opinion target.

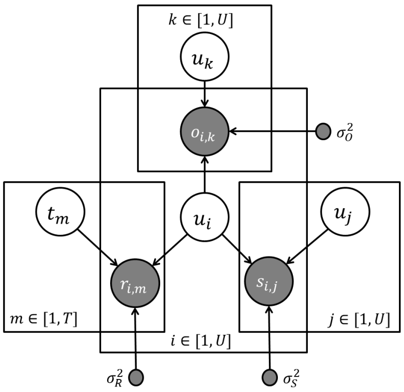
Given the matrices $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{O}$, we perform probabilistic matrix factorization to derive a low-rank vector representation and based on which we apply clustering algorithms to detect political communities.

3.1 Probabilistic Matrix Factorization

The original matrix factorization model is proposed in [12]. The model is extended in [13] to incorporate user social network for recommendation. Our model is a direct extension on the model from [13], where we add one more user interaction matrix into the model. Another difference is that the user social network and interaction work are symmetric in our model.

Figure 3 shows the plate notation for the generative model. We assume that both users and opinion targets are profiled by K latent factors. Let $u_i \in \mathbb{R}^K$ denote the vector in the latent factor space for the i-th user, and $t_m \in \mathbb{R}^K$ denote the vector for the m-th opinion target. We assume u_i and t_m are generated by Gaussian distributions as in Eqn. 1. We assume the extracted matrices $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{O}$ are generated by taking the product of the related entities. Specifically, the generation processes of the stance scores $r_{i,m}$ between i-th user and m-th opinion

[3] Users with the same stance tend to dispute with each other indicating that stance matrix and interaction matrix do not overlap.
[4] An opinion target is defined as the topic that users can express their opinions on. It can be either a controversial topic like "Abortion" or "Gun Control" with "support" or "oppose" stances, or "Does God Exist?" with "Yes" or "No" votes.



$$
\begin{aligned}
p(u_i|\sigma_U^2) &= \mathcal{N}(u_i|\mathbf{0}, \sigma_U^2\mathbf{I}), && (1)\\
p(t_m|\sigma_T^2) &= \mathcal{N}(t_m|\mathbf{0}, \sigma_T^2\mathbf{I}).\\
p(r_{i,m}|u_i, t_m, \sigma_R^2) &= \mathcal{N}(r_{i,m}|g(u_i^T t_m), \sigma_R^2), && (2)\\
p(s_{i,j}|u_i, u_j, \sigma_S^2) &= \mathcal{N}(s_{i,j}|g(u_i^T u_j), \sigma_S^2),\\
p(o_{i,k}|u_i, u_k, \sigma_O^2) &= \mathcal{N}(r_{i,k}|g(u_i^T u_k), \sigma_O^2).\\
\mathcal{N}(\cdot|\mu, \sigma^2) &: \text{normal distribution}\\
\sigma_{(\cdot)}^2 &: \text{variance parameters}\\
\mathbf{I} &: \text{identify matrix}\\
g(\cdot) &: \text{logistic function}
\end{aligned}
$$

Fig. 3: Our probabilistic matrix factorization model on user stance and social behaviors. Priors over users and opinion targets are omitted for clarity.

target, the polarity scores $s_{i,j}$ between i-th and j-th user and $o_{i,k}$ between i-th user and k-th user in the matrices $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{O}$ are in Eqn. 2.

With this generative assumption, if two users are similar in terms of their dot product in the latent factor space, then they are more likely to have positive interactions or relations. Similarly, if two users share the same stance on an opinion target, then they are similar in the latent space. The latent factors can therefore encode user preferences and similarity between two users in the latent factor space reflects whether they share similar viewpoints.

Let $\mathcal{U}(K \times U)$ and $\mathcal{T}(K \times T)$ be user and opinion target matrices. To learn $\mathcal{U}$ and $\mathcal{T}$, we need to maximize the posterior of generating all the opinion matrices $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{O}$ which is equivalent to minimize the following objective function:

$$
\begin{aligned}
&\mathcal{L}(\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{S}, \mathcal{O})\\
= &\ \frac{1}{2}\sum_{i=1}^{U}\sum_{m=1}^{T}\mathbb{I}(r_{i,m})(r_{i,m} - g(u_i^T t_m))^2 + \frac{\lambda_1}{2}\sum_{i=1}^{U}\sum_{j=1}^{U}\mathbb{I}(s_{i,j})(s_{i,j} - g(u_i^T u_j))^2\\
&+\frac{\lambda_2}{2}\sum_{i=1}^{U}\sum_{k=1}^{U}\mathbb{I}(o_{i,k})(o_{i,k} - g(u_i^T u_k))^2 + \frac{\lambda_U}{2}||\mathcal{U}||_F^2 + \frac{\lambda_T}{2}||\mathcal{T}||_F^2,
\end{aligned}
$$

where $\lambda_1 = \frac{\sigma_R^2}{\sigma_S^2}$, $\lambda_2 = \frac{\sigma_R^2}{\sigma_O^2}$, $\lambda_U = \frac{\sigma_R^2}{\sigma_U^2}$, and $\lambda_T = \frac{\sigma_R^2}{\sigma_T^2}$, $\mathbb{I}(s)$ is an indicator function which equals 1 when s is not empty and otherwise 0.

To optimize the objective function above, we can perform gradient descent on $\mathcal{U}$ and $\mathcal{T}$ to find a local optimum point. The derivation is similar to [13]. After we learn $\mathcal{U}$, we apply clustering algorithms to detect political affiliations.

Degenerative Models: To examine the effectiveness of the three extracted matrices studied in our model, we compare our model with a set of its degenerative models. We construct degenerative models by considering each matrix separately: **PMF-UT** used in [15,12], **PMF-UU** used in [12] and **PMF-AD**.

3.2 Model Generalization

The attributes supported by our model are associated with user's behavior, where the users exhibit a debatable (support/oppose) behavior on topics. For example, attributes such as religion orientation, political leaning, technology affiliation etc, can be discovered through the users behavior in several debatable topics specific to the attribute. For model generality, the feedback behavior should be domain specific and aligned to the user attribute. Similarly, the corresponding social behavior should be captured in the same settings. For example, for discovering users' technology affiliation e.g., *Apple* vs *Microsoft*, the feedback behavior should be captured for the topics related to technology, e.g., operating system, software, usability, etc., but not on lifestyle or politics. In our experiments under Section , we show our motivation for defining feedback topics for political affiliation task. To support multi-valued attributes such as multiple parties, the users can be grouped into the multiple clusters.

4 Experiments

4.1 Dataset

We collected user profile information for registered users and the corresponding sociopolitical debates' information in which these users participated in Createdebate.com. In our experiments, we focus on political affiliation discovery task.

Users	1773 (145 for evaluation - 63 Democrats and 82 Republicans))
Controversial debates	88
Controversial arguments	Abortion(1,423), gun control(1,148), same-sex marriage(1,000), death penalty(588), universal health care(438), higher taxes(418), total (5,012)
Social network links	1540 (68% friendship and 32% enmity links)
Interaction links	2830 (31% agreements and 69% disagreements)

Table 1: Some statistics of the data. Interaction links are based on controversial debates.

Testbed. The statistics of the data are also shown in the Table 1. Recall that the feedback behavior should be domain specific. For our study, we use only the two-sided debates on 6 controversial issues which are specific to political domain and motivated by party's beliefs[5] listed in Table 1. Since, the sides of debates associated with the same topic can be flipped, they should be aligned manually. We engaged two judges to manually label the sides for 88 debates as suport/oppose to the topic and both of them had perfect agreement (Cohen's kappa coefficient is 1).

Matrix generation. Recall that our solution model compromises of three matrices, **UU**, **AD** and **UT**. *User social matrix* **UU** (represents user-user matrix $\mathcal{S}$ in the solution) is generated from user friendship/enmity links. *User interaction matrix* **AD** (represents agreement/disagreement matrix $\mathcal{O}$ in the solution)

[5] `http://www.diffen.com/difference/Democrat_vs_Republican`

is generated from agreement/disagreement (interaction) links among users. *User feedback matrix* **UT** (represents user-opinion target matrix $\mathcal{R}$ in the solution) is from the user stances ("Support/Oppose") on the debates (topics).

4.2 Political Affiliation Discovery Experiments

The main goal of our study is to discover the political affiliation of the user. Through this experiment, we would like to study not only the model performance but also the performance of feedback and social behaviors independently.

Experimental settings. The ground truth on the users' political leaning is available from the users' profiles. We use all three matrices, **UU**, **AD** and **UT** described in Section 4.1. Apart from 3 degenerative baseline models described in Section 3, we consider 3 additional baselines, described as below:
Discussant Attribute Profile (DAP): A recent work [9] proposes to profile discussants by their attribute towards other targets and use standard clustering (K-Means) to cluster discussants, and achieves promising results on a similar task - subgroup detection. We thus incorporate the method on our task by profiling each user by his/her opinions towards other opinion targets and users.
Correlation Clustering (CC): Correlation clustering [20], aids to partition the corresponding signed network such that positive intra-group links and negative inter-group links are dense. It is also used in subgroup detection [19]. We use large scale correlation clustering tool [22] for our baseline.
Louvain Method (LM): Louvain method [23] is an efficient algorithm to find high modularity partitions of large networks, and is widely used in community detection. When the method produces more than two communities, we align small communities to one of the two largest communities by maximizing intra-community positive links and inter-community negative links. Since the method usually discovers too many communities when applied on sparse disconnected network (on UU and AD, more than 60 communities detected), we only apply it on UT and combined(UU+UT+AD) matrices.

In all our experiments, for our model, we set the number of latent factors to 10 as we do not observe big difference when vary the latent factor size from 10 to 50. For the other parameters in probabilistic matrix factorization methods, we select the optimal setting for each method based on the average of 10 runs. λ_1, λ_2, λ_U and λ_T are chosen from $\{0.1, 0.01\}$. We use *Purity*, *Entropy* [24], *Accuracy* and *F1 score* to evaluate the models.

Results. We first present an overall performance of all the methods on combined (UU, AD, UT) matrix and show F1 measure on the party affiliation detection task in Figure 5. We observe that our model outperforms all the baselines on the task. It achieves F1 of 74.6% and 64.5% for republicans and democrats respectively. The best baseline, correlation clustering, achieves F1 of 66.2% and 60.2% for republicans and democrats respectively. In comparison, our model has 8.4% higher performance for republicans and 4.3% higher for democrats.

We present the detailed clustering results in Table 2. We observe that the combined matrix, (UU+UT+AD) has the highest performance for most of the

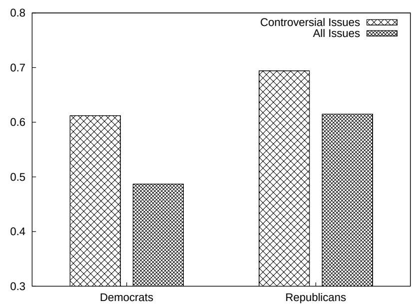


Fig. 4: F1 measure for controversial issues Vs all issues on PMF-UT.

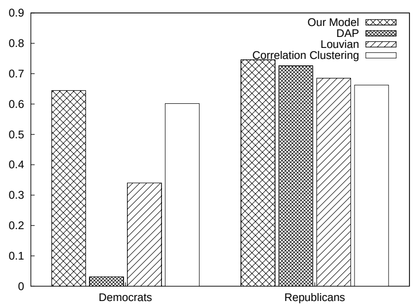


Fig. 5: F1 measure on party detection evaluation on all models using combined matrix (UU+UT+AD).

baselines and our model. Our model outperforms all the baseline models with balanced clusters. From these results, it is evident that combining all the matrices is important for the political affiliation detection. Also, it is evident that the feedback behavior plays an important role in this task. In particular, from these results we observe that for data sets such as debates where social relations are sparse, the feedback behavior of participants aids to bridge the gaps and performs efficiently in political affiliation discovery.

Method	UU			UT			AD			Combined		
	P	E	A	P	E	A	P	E	A	P	E	A
CC	0.57	0.99	0.56	0.57	0.98	0.52	**0.57**	0.99	0.55	0.64	0.91	0.64
DAP	0.57	0.99	0.56	0.57	0.98	0.51	**0.57**	**0.98**	**0.57**	0.57	0.98	0.57
LM	N/A	N/A	N/A	0.61	0.95	0.59	N/A	N/A	N/A	0.64	0.93	0.63
PMF	**0.58**	**0.96**	**0.58**	**0.65**	**0.93**	**0.65**	0.56	0.98	0.54	**0.70**	**0.88**	**0.70**

Table 2: Clustering results for political affiliation detection. Combined represents UU+UT+AD. P, E and A refer to Purity, Entropy and Accuracy, respectively.

Summary. Our model performs with promising results compared to baselines and original PMF model with an accuracy of 70.1%. We further experimented on three more degenerated baseline versions of our model, PMF-UUUT, PMF-ADUT and PMF-UUAD. In each of the baseline version, we remove one matrix from the original model to learn the latent factor representation. For PMF-UUUT, we choose UU and UT to learn the latent factor representation. This model is similar to the one used in [13]. For PMF-ADUT, matrices AD and UT are used and for PMF-UUAD we use UU and AD matrices. Our model on combined matrix still outperforms all these baseline degenerated models. Due to space constraints, we skip the details from Table 2.

4.3 Threats to Validity

Similar to other empirical studies, there are several threats to validity in interpreting the results. One such threat corresponds to the ability to link the stance(feedback) behavior to political affiliation. Our experimental results supports that leveraging stance behavior aids in political affiliation discovery and we used standard metrics for evaluation.

Another threat corresponds to the topics chosen for the feedback. "Do all issues that users participate can aid in detecting affiliation?" We crawled 1,332 debates and corresponding arguments, 10,833 on all issues. We study the performance of our model using users' leaning on controversial issues versus all issues and Figure 4 shows F1 measure on the affiliation detection. We observe that **UT** (controversial topics) outperforms $\mathbf{UT}_{all}$ (all topics). For republicans, F1 is 68% which is 6.9% higher than $\mathbf{UT}_{all}$ and for democrats, F1 is 60% which is 11.5% higher than $\mathbf{UT}_{all}$. The results indicate that a user's stances towards controversial topics have strong correlations with their political affiliation.

Another threat corresponds to the ability to generalize our results. For data generality, we evaluated with 145 users on political debates which is no way near the number of users available in all online debates. In future, we plan to extend our case study to include more users, debates and demographics. For attributes such as age, gender etc., the controversial topics should be carefully engineered and the current model cannot be applied directly. For model generality, an interesting future work is to study the cases where users can fall into more than one group and multi-party situations.

5 Related Work

User profiling studies examine users' interests, gender, sex, age, geo-localization, and other characteristics of the user profile. [25] aggregated social activity data from multiple social networks to study the users' online behavior. For user profiling, many studies took a supervised approach on various datasets; gender classification on blog data [6], age prediction on social networks [7] and location of origin prediction in twitter [8].

Similar research to user profiling studies is community or subgroup detection. [9] proposed a system that uses linguistic analysis to generate attitude vectors on ideological datasets. [23] used Louvain method which is based on modularity optimization to find high modularity partitions of large networks. Subgroup detection is studied in [19] using clustering based techniques. [26] studied both textual content and social interactions to find opposing network from online forums. In our work, besides user-user social links, we use feedback behavior which cannot be modeled by current community detection approaches.

Our proposed technique which is based on probabilistic matrix factorization (PMF) [12], a collaborative filtering based method originally used for recommendation tasks. The PMF method has been applied on social recommendation [13], news article [18] recommendation, relation prediction [27] [28] and modeling friendship-interest propagations [16]. In particular, [12] proposed a PMF model that combines social network information with rating data to perform social recommendation and [28] extended PMF for relation prediction task. Our model is a direct extension on [13] where we model three components: social, interaction and feedback. Besides this, our model assumes symmetric user social behavior.

Similar to our political affiliation task, a line of research was devoted to discover the political affiliations of informal web-based contents like news articles [29], weblogs [4], political speeches [30] and web documents [3]. Political datasets such as debates and tweets are explored for classifying user stances[31]. These applications are similar to our task as they are focussed on political content and relies on left-right beliefs.

For users' political affiliation identification on Twitter, using supervised approaches, [32,14,5] achieved high accuracy and [33] using semi-supervised label propagation method, achieved high accuracy. These studies report high performance just based on textual content or hashtags with strong nuances on political affiliations which are unique Twitter properties. Where as, we proposed an unsupervised approach and studied on data without special text characteristics. In our previous work [15], we exploited feedback behavior for the same task. However, the model performance degrades with high sparsity rate. In this work, we proposed a principled way to integrate social links and user's feedback.

6 Conclusion

In this paper, we proposed an unsupervised integrated approach based on probabilistic matrix factorization that combines social and feedback behavior features in a principled way to cater two major challenges - combining integral data components and data sparsity. Interesting future work is to study multiple-party cases and user demographics discovery such as technology or religion.
Acknowledgements. This research/project is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

References

1. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction **12**(4) (2002) 331–370
2. Behrman, J.R., Behrman, J., Perez, N.M.: Out of Sync? Demographic and other social science research on health conditions in developing countries. Demographic Research **24**(2) (2011) 45–78
3. Efron, M.: Using cocitation information to estimate political orientation in web documents. Knowl. Inf. Syst. **9**(4) (2006)
4. Durant, K.T., Smith, M.D.: Mining sentiment classification from political web logs. (WebKDD 2006)
5. Pennacchiotti, M., Popescu, A.M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: KDD '11. (2011) 430–438
6. Yan, X., Yan, L.: Gender classification of weblog authors. (AAAI 2006) 228–230
7. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: SMUC. (2011) 37–44
8. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. (SMUC '2010) 37–44
9. Abu-Jbara, A., Diab, M., Dasigi, P., Radev, D.: Subgroup detection in ideological discussions. (ACL 2012) 399–409
10. Blondel, V.D., loup Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Statistical Mechanics (2008)

11. Traag, V., Bruggeman, J.: Community detection in networks with positive and negative links. Physical Review E **80**(3) (sep 2009) 036115
12. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems (NIPS). Volume 20. (2008)
13. Ma, H., Yang, H., Lyu, M.R., King, I.: Sorec: Social recommendation using probabilistic matrix factorization. In: Proc. of CIKM. (2008)
14. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. In: ICWSM. (2011)
15. Gottipati, S., Qiu, M., Yang, L., Zhu, F., Jiang, J.: Predicting user's political party using ideological stances. In: SocInfo. (2013) 177–191
16. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. (WWW'11)
17. Yardi, S., Boyd, D.: Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. Bulletin of Science, Technology & Society **30**(5) (2010) 316–327
18. Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: In ICDM 2008. (2008)
19. Abu-Jbara, A., Radev, D.: Subgroup detector: a system for detecting subgroups in online discussions. (ACL 2012) 133–138
20. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. In: MACHINE LEARNING. (2002) 238–247
21. Galley, M., McKeown, K., Hirschberg, J., Shriberg, E.: Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. (ACL 2004)
22. Bagon, S., Galun, M.: Large scale correlation clustering optimization. CoRR (2011)
23. Traag, V., Bruggeman, J.: Community detection in networks with positive and negative links. Physical Review E **80**(3) (2009) 036115
24. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
25. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. (ACM SIGCOMM 2009) 49–62
26. Lu, Y., Wang, H., Zhai, C., Roth, D.: Unsupervised discovery of opposing opinion networks from forum discussions. (CIKM 2012) 1642–1646
27. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. (KDD 2008) 650–658
28. Qiu, M., Yang, L., Jiang, J.: Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. In: NAACL. (2013)
29. Zhou, D.X., Resnick, P., Mei, Q.: Classifying the political leaning of news articles and users from user votes. In: ICWSM. (2011)
30. Dahllf, M.: Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches - a comparative study of classifiability. LLC **27**(2) (2012) 139–153
31. Somasundaran, S., Wiebe, J.: Recognizing stances in ideological on-line debates. (NAACL HLT 2010) 116–124
32. Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. (SocialCom 2011)
33. Boutet, A., Kim, H.: What's in Twitter? I Know What Parties are Popular and Who You are Supporting Now! Number 2 (ASONAM 2012)