Predicting User's Political Party using Ideological Stances

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. Predicting users political party in social media has important impacts on many real world applications such as targeted advertising, recommendation and personalization. Several political research studies on it indicate that political parties' ideological beliefs on sociopolitical issues may influence the users political leaning. In our work, we exploit users' ideological stances on controversial issues to predict political party of online users. We propose a collaborative filtering approach to solve the data sparsity problem of users stances on ideological topics and apply clustering method to group the users with the same party. We evaluated several state-of-the-art methods for party prediction task on debate.org dataset. The experiments show that using ideological stances with Probabilistic Matrix Factorization (PMF) technique achieves a high accuracy of 88.9% at 22.9% data sparsity rate and 80.5% at 70% data sparsity rate on users' party prediction task.

Keywords: Collaborative Filtering, Ideological Stances, Memory-based CF, Model-based CF, Probabilistic Matrix Factorization

1 Introduction

Social media provides ample opportunities for citizens to participate in the political campaigns through forums, facebook, twitter and debates. These sites provide a testbed for analyzing voters behavior. Among such tasks is political affiliation detection which has been gaining much attention in recent years [1] [2] [3]. This has important consequences in targeted advertising, recommendation and personalization [4].

In our research for party prediction task, we first study American politics and draw inspiration from various political science studies on the behavior of Democrats and Republicans on social and political aspects [5] [6] [7]. These studies demonstrate the fact that the political parties take positions towards critical policies and sociopolitical issues which can ultimately lead to great differences in philosophies and ideal. Subsequently, a citizen leans towards the party that is very close to his ideological beliefs [8]. Table 1 shows the ideological beliefs of the two major parties on major social and political issues[1]. Henceforth, a users'

[1] http://www.diffen.com/difference/Democrat_vs_Republican

political affiliation is majorally dependent on his stances on the major social and political issues. We refer to such stances on controversial sociopolitical issues as *Ideological Stances*. For example, a user who supports *abortion* and is against *gun rights* is more likely a *Democrat*. His/her other stances on issues like gay marriage, health care, flat tax, death penalty, etc. can aid in detecting his party affiliation with high accuracy. In this paper, we focus on the problem of how users' ideological stances on the sociopolitical issues impacts their political affiliation and aid in detecting his party. There are other attributes such as gender, income, education, etc., that may be correlated with the party, but preliminary analysis shows the ideological beliefs is more correlated to user politics. Hence the focus of our paper is on investigating the relationship between ideological beliefs and politics.

	Death Penalty	Gay Marriage	National Healthcare	Flat Taxes	Gun Rights	Abortion
Republicans	Support (some disagree)	Oppose (some disagree)	Oppose	Support	Support	Oppose (some disagree)
Democrats	Oppose (substantial disagree)	Support (some disagree)	Support	Oppose	Oppose (some disagree)	Support (some disagree)

Table 1: Ideological stances of Republicans and Democrats in US politics.

However, the approach of using ideological stances for party prediction poses two main challenges, i.e. *data collection* and *data sparsity*.
Data collection: Gathering user's ideological stances can be on one hand a trivial problem, where a survey methods can be used or on the other hand very challenging problem, where stances are hidden in user generated content in the form of debates. Ideological belief of a user is exhibited during his/her participation in the forums or debates related to sociopolitical issues [9]. Studies such as [9] [10] can aid in generating the users' ideological stances on the controversial issues. In our work, therefore, we focus on the data sparsity problem.
Data Sparsity: The main challenge we face with the real world data is the sparsity of users' ideological stances on the controversial issues. Not all users may provide their stances on all the issues. In our corpus, the data *sparsity rate*[2] is 22.95% for six controversial issues shown in Table 1. With sparse data, standard clustering techniques would not give satisfactory results. To tackle this problem, we propose collaborative filtering based approach which has been applied successfully for recommendation tasks [11] [12].

In this paper, we present a collaborative filtering approach for predicting users party. We assume that the parties' ideological stances are known, as shown in Table 1. We also assume that users provide stances on some of these controversial issues (incomplete user-ideological stance matrix). For predicting the remaining stances, we use collaborative filtering techniques. Collaborative filtering methods have been used successfully to estimate the user-items rating for the missing ratings in the user-item matrix [13] [11]. Clustering the users based on the ideological beliefs aids in detecting users within the same party. Finally, to

[2] The percentage of missing values in a matrix.

label the clusters, the distance between the party ideology and average cluster ideology can be estimated with the standard similarity techniques.

Our Contribution: First, to the best of our knowledge, we propose the first study on the impact of ideological stances on party affiliation of online users. Traditional studies rely on social network structure or text to predict the party affiliation [4] [14]. Whereas, we claim that exploiting the ideological beliefs of users suffice the party prediction task and propose a collaborative filtering method to handle the data sparsity challenges. Second, we design our experiments to evaluate intermediate results on the stance prediction and the party prediction results. Our evaluation results show that PMF achieves a high accuracy of 88.9% over state-of-the-art methods.

The rest of our paper is organized as follows. Section 2 presents related works. Section 3 presents our problem setting and followed by our solution in Section 4. We describe experiments in Section 5. Section 6 concludes the paper.

2 Related Work

User Profiling. Party prediction is among many user profiling studies which examine users' interests, gender, age, geo-localization, and other characteristics of the user profile. [15] [2] proposed supervised approach for gender prediction. Other similar approaches are taken for age prediction on social networks [3] and location of origin prediction in twitter [14]. Aggregating social activity data from multiple social networks to study the users' online behavior [1] shows promising results. In this paper, we study the problem of party prediction. In our approach, we also exploits users online behavior but with the focus on the ideology belief correlated with party leaning.

Political Affiliation Prediction. Some studies focussed on discovering political affiliations of informal web-based contents like news articles [16], political speeches [17] and web documents [18][19][20]. Political datasets such as debates and tweets are explored for classifying user stances [10][9] and also for predicting election results [21]. Closer to these studies is subgroup detection [22][23][24][25]. These works exploit content and other corpus specific properties such as hashtags, social networks etc., for predicting tasks. Some studies are motivated with the fact that the users are influenced by the community in the social network [1][2][3]. Such peer influence may impact a user on his/her political leaning. In real situations, a user social network can be sparse and politically opposing users can be friends. This can limit the performance of the existing methods. In our approach, we use ideological beliefs for party prediction and study if it has high impact on prediction task. Other factors such as hashtags, social networks can be a complimentary to our method.

Memory-based and Model-based Collaborative Filtering. Memory-based techniques have been proposed for recommendation tasks [26]. However, due to their limitation, model-based techniques have been more popular recently. PMF has been applied on social recommendation [11][27], news article [12] recommendation, relation prediction [13][28] and modeling friendship-interest prop-

agations [29].Inspired by these works, we propose an approach based on PMF for users' party prediction task.

Our work was also inspired by some observations from [14] [4], who exploited corpus specific text properties such as hashtags on twitter data for party prediction. Some hashtags such as; *#gay*, *#dadt*, #912 etc., represent the controversial issues. However, in many case the stances of users cannot be captured by hashtags alone and a need of other methods arises to detect the stances. Similar to them, in our work, we exploited the controversial issues, but we used the stances of the users to predict the party affiliation.

3 Problem Setting

To formally define our problem, we first introduce a few basic concepts.

1. **Issue**: We refer issue to a controversial sociopolitical topics such as like "Abortion" or "Gun Control" or "Gay Marriage" etc., The controversial issues are those that segregates the political parties with great differences in ideals. In our work, we use major issues[3] studied by [5] [6] [9] as shown in Table 1.

2. **Stance**: Users express their positions as "Support" or "Oppose" to the issues related to sociopolitical context. Such positions are referred to as stances. Stances can be of pro/con/neural. In general, we observe a binary pattern for the issues shown in Table 1.

3. **Ideological Stance**: Debates on the controversial issues are referred to as ideological debates [9] and a user's stance on such issues is referred to as ideological stance.

The problem space can be formulated as a set of 2 matrices:

4. **Party Ideology Matrix**: Matrix of political party versus sociopolitical issues, denoted by $\mathcal{P}$, with each cell representing the stance of the political party on a specific issue. This matrix can be generated from Table 1.

5. **User Ideology Matrix**: Matrix of users versus sociopolitical issues, denoted by $\mathcal{R}$, with each cell representing a user's ideological stance on a specific issue. Table 2 shows a simplified example of a user-stance matrix where users take pro/con positions towards controversial sociopolitical issues.

	Death Penalty	Gay Marriage	National Healthcare	Flat Taxes	Gun Rights	Abortion
User1	Pro	Con	Con	Pro	?	?
User2	Pro	Con	Pro	?	?	Con
User3	?	Pro	?	Pro	Pro	Con
User4	Con	?	Pro	Pro	?	Pro
User5	Con	Con	Con	Pro	Con	?

Table 2: This is an example of a user ideology matrix where each filled cell represents a user's ideological stance for an issue. The collaborative filtering technique attempts to provide a prediction for missing stances.

User party prediction: The main task now is to predict the political party of the users who belong to the matrix $\mathcal{R}$. Clustering methods can be applied for grouping users and labeling them using $\mathcal{P}$. However, in real world, this matrix

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

is generally very sparse, since each user will only have positioned themselves for a small percentage of the total number of issues. The challenge of sparse data can degrade the performance of the clustering algorithms. Hence, we propose collaborative filtering based approach to predict the missing user stances and then apply clustering techniques for party prediction. In the next section, we explain the details of collaborative filtering based approach for prediction task.

4 Solution

Our approach takes two step processing for the party prediction. In the first step, we predict missing users' ideological stances in $\mathcal{R}$ using collaborative filtering method. In the second step, we use the predicted ideological matrix to group users using clustering technique and label the groups in a principle manner. We explain the details below.

4.1 Ideological Stance Prediction

Under this formulation, the problem is to predict the values for specific empty cells of $\mathcal{R}$ (i.e. predict a user's stance for an issue). In a typical CF setting, we have a list of n users, $\mathbf{U}$ and a list of m issues, $\mathbf{I}$. Each user, u takes position as pro/con or 1/0 for each issue, i. The task of CF algorithm aims at predicting the missing value, $\hat{r}_{u,i}$ which indicates user u's position likeliness for an issue i. $\overline{r}_u$ denotes mean rating value for user u and $\overline{r}_i$ denotes mean rating value for issue i. CF algorithms can be divided into to main categories: *memory-based* and *model-based* algorithms[30]. We explain the most popular CF algorithms in this section.

Memory-based CF Algorithms: Memory-based methods utilize the entire user ideology data to calculate the similarity or weight between users or issues and make predictions according to those calculated similarity values. We explore three popular memory-based models: user-based, item-based and slope-one method.

a. User-based: User-based method predicts missing ratings by firstly finding similar users [26]. The similarity between two users, u and v using Pearson correlation is given by Equation 1. Predicted values are computed using Equation 2.

$$sim(u,v) = \frac{\sum_{i\in\mathcal{I}}(r_{u,i}-\overline{r}_u)(r_{v,i}-\overline{r}_v)}{\sqrt{\sum_{i\in\mathcal{I}}(r_{u,i}-\overline{r}_u)^2}\sqrt{\sum_{i\in\mathcal{I}}(r_{v,i}-\overline{r}_v)^2}}, \tag{1}$$

where $i \in \mathcal{I}$ summations over the issues that both the users u and v have rated.

$$\hat{r}_{u,i} = \overline{r}_u + \frac{\sum_{v\in\mathcal{U}}(r_{v,i}-\overline{r}_v)sim(u,v)}{\sum_{v\in\mathcal{U}}|sim(u,v)|} \tag{2}$$

b. Item-based: Item-based method predicts missing ratings by first finding similar items [31]. The similarity between two items, i and j using Pearson correlation is given by Equation 3 and the predicted values are given by Equation 4.

$$sim(i,j) = \frac{\sum_{u \in \mathcal{U}}(r_{u,i} - \overline{r}_i)(r_{u,j} - \overline{r}_j)}{\sqrt{\sum_{u \in \mathcal{U}}(r_{u,i} - \overline{r}_i)^2}\sqrt{\sum_{u \in \mathcal{U}}(r_{u,j} - \overline{r}_j)^2}}, \tag{3}$$

where $u \in U$ denote users who have rated both the items, i and j.

$$\hat{r}_{u,i} = \overline{r}_i + \frac{\sum_{j \in \mathcal{I}}(r_{u,j} - \overline{r}_j)sim(i,j)}{\sum_{j \in \mathcal{I}}|sim(i,j)|} \tag{4}$$

c. Slope-One: The main idea of the Slope One algorithms is to use the difference between User A's ratings of two items X and Y and User B's rating of item X in common to predict User B's unknown rating of item Y [32]. To predict missing values we first compute the average deviation,$dev_{i,j}$ of each pair of items given by Equation 5. The predicted values are given by Equation 6.

$$dev_{i,j} = \frac{\sum_{v \in \mathcal{U}}(r_{v,i} - r_{v,j})}{|\mathcal{U}|} \tag{5}$$

where $u \in \mathcal{U}$ denote users who have rated both the items, i and j.

$$\hat{r}_{u,i} = \frac{\sum_{j \in \mathcal{I}}(dev_{i,j} + r_{u,j})}{|\mathcal{I}|} \tag{6}$$

where $j \in \mathcal{I}$ denote issues for which $dev_{i,j} \neq 0$.

Model-based CF Algorithms: Different from memory-based algorithms, model based CF techniques utilize user item matrix (the pure stance data in our case) to estimate or learn a model offline to make predictions. Among the model based CF techniques, matrix factorization models have been mostly studied and successfully applied in real recommender systems [33]. The matrix factorization models try to explain the ratings by characterizing both items and users on a latent factor space, such that user-item ratings are based on the inner products of them in the factor space. We describe two popular model based models: Singular Value Decomposition (SVD) and Probabilistic Matrix Factorization (PMF).

a. Singular value decomposition: SVD is a well-established technique for finding latent factors in information retrieval. Conventional SVD is undefined when there are missing entries in the matrix. It is a matrix factorization technique commonly used for producing low-rank approximations [33] [34]. Singular value decomposition for user ideology matrix $\mathcal{R}$, $SVD(\mathcal{R})$ is defined as

$$SVD(\mathcal{R}) = ASV^T \tag{7}$$

where A, S and V are of dimensions $n \times d$, $d \times d$ and $d \times m$. The S diagonal matrix contains singular values, which are positive and always in decreasing order (singular matrix). For low-dimensional representation we retain only $k \ll d$ entries for all matrices. $\hat{\mathcal{R}}_k = A_k.S_k.{V_k}^T$ is the rank-k matrix that is closest approximation of $\mathcal{R}$. SVD produces a set of uncorrelated eigenvectors used in collaborative filtering process. Each user and issue is represented by its corresponding eigenvector. The process of dimensionality reduction may help user who rated similar issues (but not exactly the same issues) to be mapped into the

space spanned by the same eigenvectors. Once the matrix is decomposed, the prediction can be generated by computing the cosine similarities (dot products) between n pseudo-users and m pseudo-issues. The predicted values are given by,

$$\hat{r}_{u,i} = \overline{r}_u + A_k.\sqrt{S_k}^T(u).\sqrt{S_k}.{V_k}^T(i) \tag{8}$$

SVD finds A_k, S_k and V_k to obtain a approximation of $\mathcal{R}$, which requires $\mathcal{R}$ matrix to be complete. In real cases, the rating matrix $\mathcal{R}$ is sparse which make the conventional SVD not suitable. A solution approach is to minimize the squared error with the target $\mathcal{R}$ only for the observed entries of the target matrix R. This will result in a difficult non-convex optimization problem as discussed in [35]. Moreover, the SVD method does not scale well with the number of observations and is highly prone to overfitting on sparse data.

b. PMF: Different from SVD, PMF [27] is a probabilistic algorithms that scale linearly with the number of observations and perform well on sparse data. In PMF, we assume that there are K latent factors with which both users and items can be represented. The generative process of the user u and the item i are as follows.

$$p(u|\sigma_U^2) = \mathcal{N}(u|\mathbf{0}, \sigma_U^2\mathbf{I}), \tag{9}$$

$$p(i|\sigma_I^2) = \mathcal{N}(i|\mathbf{0}, \sigma_I^2\mathbf{I}), \tag{10}$$

where σ_U^2 and σ_I^2 are two variance parameters for users and items, respectively, $\mathbf{I}$ is the identify matrix, and $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

Furthermore, the rating score $r_{u,i}$ between user u and item i is generated as:

$$p(r_{u,i}|u, i, \sigma_1^2) = \mathcal{N}(r_{u,i}|g(u^T i), \sigma_1^2), \tag{11}$$

where σ_1^2 is variance parameter and $g(\cdot)$ the logistic function.

In PMF, we seek to minimize the regularized estimation error as follows:

$$\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}} \mathbb{I}(r_{u,i})(r_{u,i} - g(u^T i))^2 + \lambda_U||\mathcal{U}||_F^2 + \lambda_I||\mathcal{I}||_F^2, \tag{12}$$

where $\mathbb{I}(s)$ is an indicator function which equals 1 when s is not empty and otherwise 0. $g(\cdot)$ is the logistic function.

To optimize the objective function above, we can perform gradient descent on $\mathcal{U}$ and $\mathcal{I}$ to find a local optimum point. This method is efficient even on large data sets as it does not need to infer the full posterior distribution over the model parameters and observations. Nevertheless, a fully Bayesian treatment of the PMF model can further boost the model performance [36].

After profiling users and items in the latent factor space, to predict a user u's rating $\hat{r}_{u,i}$ on an given item i, we simply take the dot product of the user and item vector: $\hat{r}_{u,i} = g(u^T i)$.

4.2 Party Prediction

Clustering algorithm on the predicted user stance matrix generates groups of users. We propose simple K-means for clustering the users. To label the clusters

to the respective parties, we compute the similarity distance between the cluster's average ideology and the party ideology using the similarity distance techniques like Hamming distance or Euclidean.

5 Experiments

In this section we want to answer the following research questions:

RQ1: How accurate is the ideological stance prediction and by which model?
RQ2: Which model performs better for stance prediction on sparse data?
RQ3: Is our proposed approach effective in users' party prediction?
RQ4: Does ideological belief aids in users' party prediction?

We answer the first two research questions through ideology stance prediction experiments and the last two questions through the party prediction experiments. We first describe the data set we used to evaluate our approach and our evaluation criteria. We next describe our ideological stance prediction experiments and results of collaborative filtering methods. Finally, we describe our party prediction experimental study that shows the effectiveness of our approach in predicting users political party.

5.1 Dataset

Our dataset is constructed by crawling the data from debate.org. We collected 1000 user profiles who provided the party affiliation information. The data consists of user personal details such as gender, age, political party, income, occupation, religion and other demographics. Apart from the demographics, users also provide their stances on several social, political and economical issues. After the clean up, we collected the stances for the issues that match with those in Table 1. With some preliminary experiments, we found that these issues are useful

Users	1000
Democrats	519
Republicans	481
Issues	death penalty, gay marriage, national health care, flat taxes, gun rights and abortion
User-ideology matrix sparsity rate	22.95% (percentage of missing values)

Table 3: Statistics of our dataset.

in party prediction. We first generated User Ideology Matrix, $\mathcal{R}$ described in Section 3 from the stance data. We use $\mathcal{R}$ for our stance prediction experiments. We use users' political party information as gold truth for our party prediction experiments.

5.2 Evaluation Criteria

We use standard metrics from information retrieval for evaluating the performance of the models. We compute *Accuracy* (the higher the better) for all the

models at all sparsity rates to evaluate the stance prediction performance on collaborating models. We also use *Precision* , *recall* and *F1 score* to evaluate the performance of PMF model on stance prediction. We use *Purity* (the higher the better) and *Entropy* (the lower the better) and *Rand Index* (the higher the better) to evaluate the performance of political affiliation prediction [37]. We also use *Accuracy* by computing the percentage of users that are "classified" correctly after labeling the clusters using external information, $\mathcal{P}$. We also compute *F1 score* for each political party to evaluate the models.

5.3 Ideological Stance Prediction Experiments

Recall that our first step in our solution model is to predict the missing ideological stances of users. We use $\mathcal{R}$ for these experiments. To generate the sparsity on the data, we hide the users stances randomly and predict the hidden data using CF models. We use the same hidden matrix across all models for unbiased evaluation. Through these experiments, we will answer RQ1 and RQ2.

Experimental Settings: We compare User based, Item based, Slope one, SVD and PMF models for evaluation. For matrix sparsity, we hide stances in $\mathcal{R}$ to obtain sparsity rates(percentage of missing stances) of 30%, 40%, 50%, 60% and 70%. We took an average of three random matrices for each sparse matrix rate. We use Accuracy, Precision, Recall and F1 score for comparison.

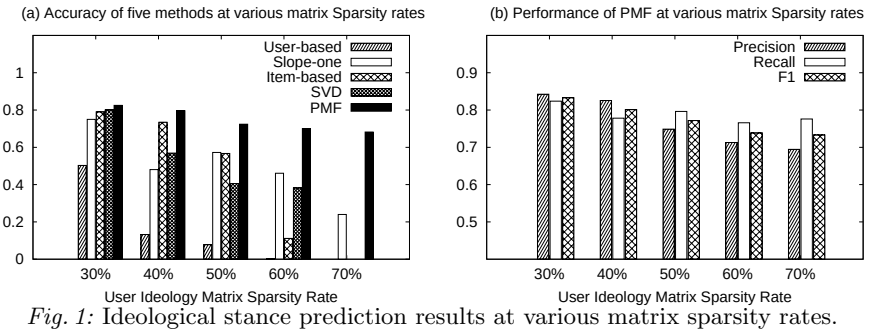

Fig. 1: Ideological stance prediction results at various matrix sparsity rates.

Results: The stance prediction accuracy results for all the collaborative model is shown in Figure 1(a). We observe that PMF model outperforms all the other collaborative techniques. At 30% sparsity rate PMF model has an accuracy of 82.5% which is 2.5% higher than second best mode, SVD which has an accuracy of 80%. In case of sparse data, user-based, item-based and SVD completely fail on the model. PMF model still performs the best with an accuracy of 68.2%. Slope-one has an accuracy of only 24% at a sparsity rate of 70%.

We further show precision, recall and F1 scores of PMF model breakdown at various sparsity rates in Figure 1(b). We observe that PMF has a F1 score of 83.3% at 30% sparsity rate. At 70% sparsity rate, the model still performs well with an accuracy of 73.3%. With these observations, we choose PMF for the party prediction experiments which we explain in the next subsection.

Summary: As an answer to RQ1, we show that PMF model outperforms other collaborative models in stance prediction with an accuracy of 82.5%. As an answer to RQ2, our experiments show that PMF has better performance than others on sparse data with an accuracy of 73.3% at a sparsity rate of 70%.

5.4 Party Prediction Experiments

The main goal of our study is to discover the political party of the user. Through this experiment, we would like to study not only the model performance but also the importance of ideological belief in party detection. We will answer RQ3 and RQ4.

Experiment settings: We conduct experiments on various sparsity rates similar to previous experiments. We use the same hidden matrix across all models for unbiased evaluation. The ground truth on the users' political leaning is available from the users' profiles. For baseline model, we use a recent work [38].

Baseline: A direct approach for party prediction can be achieved by measuring the similarity between the user vector and party vector using Hamming distance and assign the user to the party with low Hamming distance.

Discussant Attribute Profile (DAP): [38] 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 ideologies towards issues stated in Table 1.

Probabilistic Matrix Factorization (PMF): We apply PMF on $\mathcal{R}$ and then cluster users into two clusters. 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, we select the optimal setting based on the average of 10 runs. λ is chosen from $\{0.1, 0.01\}$.

As discussed in Section 3, the resulting clusters are labeled by using party ideology matrix $\mathcal{P}$. We first calculate the average ideology of each cluster and measure the distance of cluster ideology to party ideology using Hamming distance. The closer the distance to the party, all the users in that cluster are labeled with the corresponding party. We use metrics Purity, Entropy, Accuracy, RandIndex and F1 score for evaluation.

Results: We first present the detailed clustering results in the Table 4. At all sparsity rates PMF outperforms DAP and Baseline on all metrics. We also observed that for Baseline, for some users the Hamming distance to both the parties is equal(when the matrix is sparse) and hence are assigned randomly to one of the major parties. Another observation is that, higher the sparsity rates, larger the number of unassigned users. At higher sparsity rates Baseline performs better than DAP as DAP tends to generate unbalanced clusters. On original data all the models have reasonably high accuracies: Baseline has an accuracy of 83.3%, DAP has 85.6% accuracy and PMF has an accuracy of 88.9%. At 30% DAP has an accuracy of 85.8% and PMF has an accuracy of 88.9% which is 4.1% higher. This shows the importance of ideological belief in party prediction task. We observe that the accuracy drops drastically for DAP with the sparsity

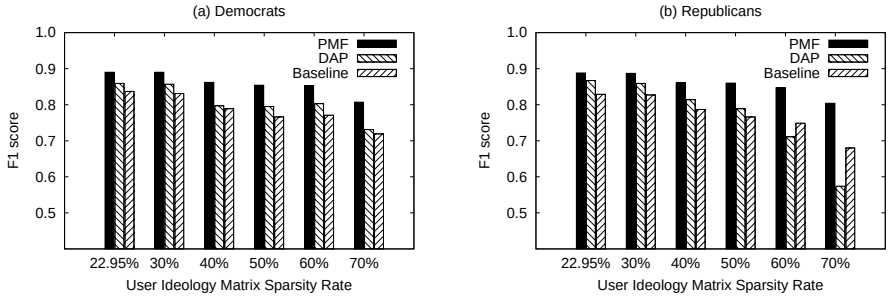

Fig. 2: Party prediction F1 results break down by parties at various matrix sparsity rates. 22.95% is the original sparsity rate of the matrix $\mathcal{R}$.

of the data yielding 67% at 70% sparsity rate, whereas for PMF the accuracy is 80.5% which is 13.5% higher. Even though Baseline also degrades at higher sparsity rates, it perform better than DAP but 10.5% lower than PMF. From our previous experiments, we observe that PMF aids in better prediction of missing ideology stances when compared to other collaborative methods. Such behavior aided in high accuracy by PMF when compared to Baseline and DAP.

Method	Metric	22.95%	30%	40%	50%	60%	70%
Baseline	P	0.833	0.829	0.788	0.766	0.761	0.701
	E	0.649	0.657	0.740	0.782	0.793	0.880
	A	0.833	0.829	0.788	0.766	0.761	0.701
	R	0.722	0.716	0.666	0.641	0.636	0.580
DAP	P	0.856	0.858	0.806	0.792	0.766	0.670
	E	0.523	0.581	0.694	0.735	0.759	0.904
	A	0.856	0.858	0.806	0.792	0.766	0.670
	R	0.753	0.756	0.687	0.670	0.641	0.557
PMF	P	**0.889**	**0.889**	**0.861**	**0.857**	**0.850**	**0.805**
	E	**0.498**	**0.499**	**0.573**	**0.578**	**0.608**	**0.707**
	A	**0.889**	**0.889**	**0.861**	**0.857**	**0.850**	**0.805**
	R	**0.809**	**0.802**	**0.761**	**0.755**	**0.745**	**0.686**

Table 4: Clustering results for political affiliation detection on all models. 22.95% is the original sparsity rate of the matrix $\mathcal{R}$.

We further studied the performance of all the models at the party breakdown. We show F1 scores break down by parties at various matrix sparsity rates for both the models in Fig. 2. We observe that PMF and Baseline has balanced F1 scores for both clusters (Democrats and Republicans) at all sparsity rates, whereas DAP shows high F1 for Democrats than Republicans at 70% sparsity rate. At 30% sparsity rate, Baseline has 83.7% and 82.9%, DAP has 85.7% and 85.9% and PMF has 89% and 88.7% accuracy for Democrats and Republicans respectively. At 70% sparsity rate, Baseline has 71.9% and 68%, DAP has 73.1% and 57.4% and PMF has 80.7% and 80.4% accuracy for Democrats and Republicans respectively. For sparse data, DAP tends to cluster users to the larger cluster, in this case, it is Democrats.

Summary: To answer RQ3, our approach of using collaborative filtering method for prediction task outperforms standard clustering techniques with and accuracy of 88.9%. As an answer to RQ4, users' ideological stances plays vital role in users' party prediction and even standard clustering techniques achieves a high accuracy of 85.5%.

5.5 Discussion

Our experiments demonstrates the benefit of ideological stances of users in predicting their party leaning. While the results are very promising, it is interesting to study the reliability of the issues that we chose for determining the ideological belief of users. For our studies, we used only 6 out of 46 issues[4] for which users provide stances. We collected stances for 46 issues for all users in our corpus and we observed that the sparsity rate of the issue matrix is 52.94%. We then tested the model with all 46 issues and achieved an accuracy of 88.1% which is 0.8% lower than our previous results. Our results with 6 issues are very close to results with 46 issues, which shows that the 6 issues in Figure 1 should suffice for determining the ideology of a user. Furthermore, with high dimensionality where the sparsity rate is high, the model may fail to predict the missing stances [11]. It might be interesting to study which combination of issues (features) have greater impact on the party and we leave it for future studies.

Further, we studied the impact of religion dimension on party prediction task. Some studies observed a correlation between religion and party affiliation such as republicans exhibit greater religiosity compared to democrats [39]. We conducted some experiments where we assign the Christians to Republicans and others to Democrats. We achieved an accuracy of 68.2% for party prediction task, which shows that religion dimension is insufficient for party prediction as religiosity cannot be captured from the religion demographic. It is interesting to study the correlation of other demographics with the party and we leave it for future work. Also another extension can be study of combining demographics with ideological beliefs, exploiting ideological stances with social networks or corpus content such as text, hashtags etc.,

6 Conclusion

In this paper, we studied the problem of predicting users political party in social media. In our approach, we exploited users' ideological stances to predict their political party and we proposed a collaborative filtering approach to solve the data sparsity problem of users stances on controversial sociopolitical issues and apply clustering method to group the users with the same party. Our experiment results proves that user's ideological belief is highly correlated with the party affiliation. Evaluation results show that using ideological stances with PMF achieves a high accuracy of 88.9% at 22.9% data sparsity rate and 80.5% at 70% data sparsity rate.

[4] `http://www.debate.org/big-issues/`

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. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. (2009) 49–62
2. Yan, X., Yan, L.: Gender classification of weblog authors. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. (2006) 228–230
3. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: SMUC. (2011) 37–44
4. Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Proceedings of 3rd IEEE Conference on Social Computing (SocialCom). (2011)
5. Speel, R.W.: The evolution of republican and democratic ideologies. Journal of Policy History **12** (7 2000) 413–416
6. Saunders, K., Abramowitz, A.: Ideological realignment and active partisans in the american electorate. In: American Politics Research. (2004) vol. 32 no. 3 285–309
7. Fiorina, M.P., Abrams, S.J.: Political polarization in the american public. Annual Review of Political Science **11**(1) (2008) 563–588
8. Killian, M., Wilcox, C.: Do abortion attitudes lead to party switching? In: Political Research Quarterly Vol. 61, No. 4. (2008) 561–573
9. Somasundaran, S., Wiebe, J.: Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. (2010) 116–124
10. Walker, M.A., Anand, P., Abbott, R., Tree, J.E.F., Martell, C., King, J.: That is your evidence?: Classifying stance in online political debate. Decis. Support Syst. **53**(4) (2012) 719–729
11. Ma, H., Yang, H., Lyu, M.R., King, I.: Sorec: Social recommendation using probabilistic matrix factorization. In: Proc. of CIKM. (2008)
12. Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: In ICDM 2008. (2008)
13. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: Proceedings of the 14th KDD. (2008) 650–658
14. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. (2010) 37–44
15. Mukherjee, A., 0001, B.L.: Improving gender classification of blog authors. In: EMNLP. (2010) 207–217
16. Zhou, D.X., Resnick, P., Mei, Q.: Classifying the political leaning of news articles and users from user votes. In: ICWSM. (2011)
17. 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
18. Durant, K.T., Smith, M.D.: Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In: WebKDD'06. (2006) 187–206

19. Durant, K.T., Smith, M.D.: Mining sentiment classification from political web logs. In: In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12 th SIGKDD (WebKDD-2006). (2006)
20. Efron, M.: Using cocitation information to estimate political orientation in web documents. Knowl. Inf. Syst. **9**(4) (2006)
21. Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls : Linking text sentiment to public opinion time series. In: Proceedings of ICWSM. (2010)
22. Abu-Jbara, A., Radev, D.: Subgroup detector: a system for detecting subgroups in online discussions. In: Proc. of the ACL'12 Demo. (2012) 133–138
23. Hassan, A., Abu-Jbara, A., Radev, D.: Detecting subgroups in online discussions by modeling positive and negative relations among participants. In: Proceedings of the 2012 Joint Conference on EMNLP and CoNLL. (2012)
24. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **Vol. 2008, No. 10** (2008)
25. Traag, V., Bruggeman, J.: Community detection in networks with positive and negative links. Physical Review E **80**(3) (2009) 036115
26. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: CSCW. (1994) 175–186
27. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems (NIPS). Volume 20. (2008)
28. Qiu, M., Yang, L., Jiang, J.: Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. In: NAACL, Association for Computational Linguistics (2013) 401–410
29. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: WWW. (2011)
30. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. in Artif. Intell. **2009** (January 2009) 4:2–4:2
31. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Itembased collaborative filtering recommendation algorithms. In: WWW. (2001) 285–295
32. Lemire, D., Maclachlan, A.: Slope one predictors for online rating-based collaborative filtering. In: Proceedings of SIAM Data Mining (SDM'05). (2005)
33. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8) (August 2009) 30–37
34. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Incremental singular value decomposition algorithms for highly scalable recommender systems. In: Fifth International Conference on Computer and Information Science. (2002) 27–28
35. Srebro, N., Jaakkola, T.: Weighted low rank approximation. In: ICML. (2003)
36. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using markov chain monte carlo. In: ICML. (2008)
37. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
38. Abu-Jbara, A., Diab, M., Dasigi, P., Radev, D.: Subgroup detection in ideological discussions. In: Proceedings of the 50th ACL. (2012) 399–409
39. Glaeser, E.L., Ponzetto, G.A.M., Shapiro, J.M.: Strategic extremism: Why republicans and democrats divide on religious values. The Quarterly Journal of Economics **120**(4) (2005) 1283–1330