4-2015

Multi-roles affiliation model for general user profiling

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Heyan HUANG
Beijing Institute of Technology

Yashen WANG
Beijing Institute of Technology

Published in Database Systems for Advanced Applications: DASFAA 2015 International Workshops, SeCoP, BDMS, Hanoi, Vietnam, April 20-23, revised selected papers. pp. 227-233. https://doi.org/10.1007/978-3-319-22324-7_20

Multi-roles Affiliation Model for General User Profiling

Lizi Liao[1,2(✉)], Heyan Huang[1], and Yashen Wang[1]

[1] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Institute of Technology, Beijing, China
liaolizi.llz@gmail.com
[2] Living Analytics Research Center, Singapore Management University, Singapore City, Singapore
{hhy63,yswang}@bit.edu.cn

Abstract. Online social networks release user attributes, which is important for many applications. Due to the sparsity of such user attributes online, many works focus on profiling user attributes automatically. However, in order to profile a specific user attribute, an unique model is built and such model usually does not fit other profiling tasks. In our work, we design a novel, flexible general user profiling model which naturally models users' friendships with user attributes. Experiments show that our method simultaneously profile multiple attributes with better performance.

Keywords: General user profiling · Multi-roles affiliation model · Social networks

1 Introduction

The rapid growth of social network websites such as Facebook, LinkedIn and Twitter attracts a large number of Internet users. However, only a small proportion of these users intentionally or unintentionally disclose their attributes like occupation, education and interests, which are important to many online applications, such as recommendation, personalized search, and targeted advertisement. Research on user profiling has focused on various kinds of user attributes, ranging from demographic information like gender [2,8], age [5,10] and location [1–3], to user preference information like political orientation or interests.

In most of these works, in order to profile a specific user attribute, a unique model is built and such model usually does not fit other profiling tasks. In our paper, we propose a general user profiling model to profile multiple user attributes simultaneously. Since social network users connect to other users regularly, many works [1,4,11] leverage the principle of homophily [7] to profile attributes via social connections. The basic assumption is that users are more likely to connect with those sharing same attribute values. Based on the observed

features from the connected friends, user's attributes can be obtained by directly applying a majority vote or its variations [6].

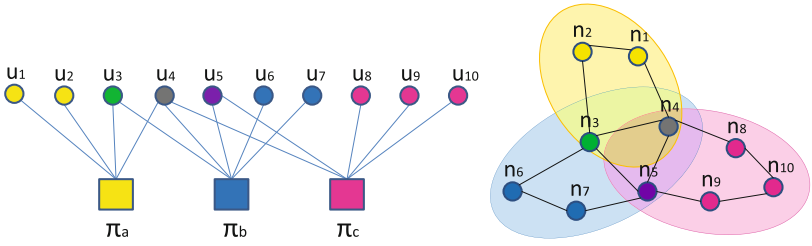
However, this assumption oversimplifies the complexity of online social networks. In real life scenarios, users become friends only because of certain attributes and those attributes make different degrees of contribution. Thus, as our second contribution, we quantify the different linking factors for each attribute entry. For instance, compared to both being *Democrats*, both working at *Google* might be more likely to link two users together. That is to say, the linking factor of attribute entry *Google* is larger than that of attribute entry *Democrat*. Note that which attributes are more likely to link users are automatically inferred from data rather than pre-defined.

2 Model

In this section, we proceed to introduce our multi-roles affiliation model(MRA). Figure 1 illustrates the essence of our model. We start with a bipartite graph $\mathcal{A}$ where the nodes in the top represent users, the nodes in the bottom represent attribute values(or roles), and the edges indicate attribute affiliations. We also observe the social network $\mathcal{G}$ of those users. In our model, there are two important intuitions.

First, each attribute entry of the users corresponds to a specific role. MRA models users' preference towards each role with a bipartite attributes affiliation network as Fig. 1(a). Formally, we assume that there is a set of N users. Each user $u = 1, 2, \cdots, N$ has a latent group membership indicator $z_{uk} \in \{0, 1\}$ for each attribute entry (or role) $k = 1, 2, \cdots, K$. In Fig. 1(a), roles are indicated as a, b, c. Note that each indicator z_{uk} of user u is independent. Each user can belong to multiple roles simultaneously.

Second, we use a set of link factors $\boldsymbol{\pi}$ to capture the probability that users sharing a certain attribute value are linked together. For example, π_k is the probability of users taking the same role k to be linked together. Note that for different roles $k1$ and $k2$, their contribution to the link formation is different,



(a) Attributes affiliation graph $\mathcal{A}$ (b) Social network $\mathcal{G}$

Fig. 1. (a) Bipartite attributes affiliation graph. Squares: attribute values, Circles: users. (b) Social network of users.

which is quantitatively measured by π_{k1} and π_{k2}. As in [14], we define the probability of creating an edge (u, v) between a pair of users u, v as:

$$\delta_{uv} = 1 - \prod_{k \in \{C_{uv}\}} (1 - \pi_k) \tag{1}$$

where C_{uv} is the set of attribute entries u and v share (or roles they both take). We can see that the equation above already ensures that pairs of users that share more attributes are more likely to link together. To allow for edges between users who do not share any attribute, we also introduce an additional role, called the ϵ-role, which connects any pair of users with a very small probability ϵ. We simply set it to be the random link probability.

3 Inference

Given partially observed binary user attribute entries $F = \{f_{uk} : u \in \{1, \cdots, N\}; k \in \{1, \cdots, K\}\}$ and the user social network $\mathcal{G}$, we aim to find the full attributes affiliation graph $\mathcal{A}$ and link factors $\boldsymbol{\pi} = \boldsymbol{\pi}_k : k = 1, \cdots, K\}$. We apply the maximum likelihood estimation, which finds the optimal values of $\boldsymbol{\pi}$ and graph $\mathcal{A}$ so that they maximize the likelihood $L(\mathcal{A}, \boldsymbol{\pi}) = P(\mathcal{G}, F|\mathcal{A}, \boldsymbol{\pi})$:

$$\arg\max_{\mathcal{A}, \boldsymbol{\pi}} L(\mathcal{A}, \boldsymbol{\pi}) = \prod_{(u,v) \in E} p(u, v) \prod_{(u,v) \notin E} (1 - p(u, v)) \tag{2}$$

We employ the coordinate ascent algorithm to solve the above optimization problem. The algorithm iterate the following two steps. First, we update $\boldsymbol{\pi}$ by keeping $\mathcal{A}$ fixed. Then we update $\mathcal{A}$ while keeping π fixed. To start the process, we need to initialize $\mathcal{A}$. Note that $\mathcal{A}$ is indeed a set of latent group membership indicator $z_{uk} \in \{0, 1\}$. For those partially observed binary user attribute entries F, we keep those z_{uk} to be the same as f_{uk}. For others, we randomly generate z_{uk} by using the ratio calculated from F and $\mathcal{G}$.

3.1 Update of Link Factors $\boldsymbol{\pi}$

By keeping the attributes affiliation graph $\mathcal{A}$ fixed, we aim to find $\boldsymbol{\pi}$ by solving the following optimization problem:

$$\arg\max_{\boldsymbol{\pi}} \prod_{(u,v) \in E} (1 - \prod_{k \in \{C_{uv}\}} (1 - \pi_k)) \prod_{(u,v) \notin E} (\prod_{k \in \{C_{uv}\}} (1 - \pi_k)) \tag{3}$$

where the constraints are $0 \leq \pi_k \leq 1$. We transform this non-convex problem into a convex optimization problem. We maximiza the logarithm of the likelihood and change the variables $e^{-x_k} = 1 - \pi_k$:

$$\arg\max_{\boldsymbol{x}} \sum_{(u,v) \in E} log(1 - e^{-\sum_{k \in C_{uv}} x_k}) - \sum_{(u,v) \notin E} \sum_{k \in C_{uv}} x_k \tag{4}$$

where the constraints $0 \leq \pi_k \leq 1$ become $x_k \geq 0$. This problem is a convex optimization of $\boldsymbol{x}$. We can solve it by gradient descent.

3.2 Update of Attributes Affiliation Graph $\mathcal{A}$

Given the link factors $\boldsymbol{\pi}$, we aim to find appropriate attributes affiliation graph $\mathcal{A}$ for all the users. We use the Metropolis-Hastings algorithm [9] where we stochastically update $\mathcal{A}$ using a set of 'transitions'. Given the current attributes affiliation graph $\mathcal{A}$, we consider two kinds of transitions to generate a new attributes affiliation graph $\mathcal{A}'$. One is that a latent group membership indicator z_{uk} change from 1 to 0. The other is that a latent group membership indicator z_{uk} change from 0 to 1. Note that we fix z_{uk} for those already observed binary user attribute entries F. Once we have generated new attributes affiliation graph $\mathcal{A}'$, we accept $\mathcal{A}'$ with probability :

$$Min(1, L(\mathcal{A}', \boldsymbol{\pi})/L(\mathcal{A}, \boldsymbol{\pi})). \tag{5}$$

In other words, we initialize $\mathcal{A}_1 = \mathcal{A}$. We start the process with some $\mathcal{A}$ and then perform a large number of steps, where each step i we take $\mathcal{A}_i$ and apply a random 'transition' generating a new attributes affiliation graph $\mathcal{A}'_i$. At each step, we accept the transition probabilistically based on the ratio of log-likelihoods. If the transition is not accepted, we do not update $\mathcal{A}_i$.

4 Experiments

4.1 Experimental Setup

We use the Facebook networks of 4 colleges and universities: Georgetown, Oklahoma and Princeton and UNC Chapel Hill from a date in Sept. 2005 [12]. The edges here are only intra-school links between users. In these Facebook datasets, there are 8 user attributes which are ID, student/faculty status, gender, major, dorm or house and high school. Since ID information is very user specific as well as high school, we ignore these two attributes. We run experiments on these data sets.

Next, we introduce natural baseline as well as the state-of-the-art method. Taking the homophily phenomenon into consideration, we use random guess within direct neighbors (RA) as our natural baselines. Thus, we predict the k-th missing attribute entry value by randomly selecting a value of the k-th attribute entry from the users neighbors. Another baseline is CESNA [13]. This method detects overlapping communities in networks with node attributes. It statistically models the interaction between the network structure and the node attributes , which makes it capable of determining community membership as well as recover missing attributes. Using the community membership result and the association weights between attributes and each community, we obtain the final probability of each missing attribute for each user. Since attributes assignment probability obtained from CESNA are continuous values varying from 0 to 1. We choose the threshold as the one which gives us the largest F1 value. Then we treat this F1 value as the CESNA results.

4.2 Results and Discussion

We first focus on comparing our model MRA with natural baseline RA here in Table 1. In order to gain a comprehensive view of the performance, we randomly hide 10 % of user attributes to 50 % of user attributes in a gradual way. Then we take an average result for comparison. Note that most of the user attribute entries are 0, which is determined by the flattening procedure. Also, people care more about the entries with value 1 in real-world. Thus we provide the precision, recall and F1 value for value 1 respectively. From the Table 1, we observe that MRA can achieve higher performance.

Table 1. Results comparing with natural baseline

	Georgetown			Oklahoma			Princeton			UNC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RA	0.34	0.42	0.38	0.25	0.41	0.31	0.30	0.41	0.35	0.30	0.41	0.35
MRA	0.52	0.36	**0.43**	0.36	0.37	**0.36**	0.43	0.36	**0.39**	0.47	0.37	**0.41**

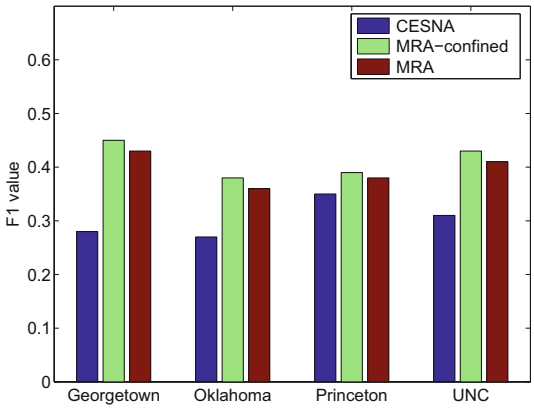


Fig. 2. Experiment results on 4 universities datasets.

Figure 2 shows the comparison between MRA and CESNA in terms of the F1 value. The x-axis refers to different datasets as mentioned above. Note that CESNA ignore those users who have no link with others. In the prediction results, the performance of CESNA is only based on those users who eventually get community assignment, which is a subset of the whole input users. To compare performance with CESNA on the same set of users who obtain community assignment successfully, the result of our model is named as MRA-confined. Since our model is able to handle both loosely connected users as well as unconnected users, we also give the performance detail named as MRA for the whole set of input users. We can see that the performance starts to deteriorate in some amount due to those uninformative users, which conforms the reality. However, those results are still better than the baseline results.

5 Conclusion

In this paper, we developed a general user profiling framework to simultaneously profile multiple attributes. Our multi-roles affiliation model (MRA) naturally captures the relationship between users friendship links and user attributes. It effectively profiles missing attributes for social network users. Moreover, the way we treat each attribute entry enables our model easily adapting to various kinds of attributes profiling.

Acknowledgments. This research is supported in part by Chinese National Program on Key Basic Research Project (Grant No. 2013CB329605). This research is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

1. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, pp. 61–70. ACM (2010)
2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309. Association for Computational Linguistics (2011)
3. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1277–1287. Association for Computational Linguistics (2010)
4. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1023–1031. ACM (2012)
5. Liao, L., Jiang, J., Ding, Y., Huang, H., Lim, E.P.: Lifetime lexical variation in social media. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
6. Liao, L., Jiang, J., Lim, E.P., Huang, H.: A study of age gaps between online friends. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 98–106. ACM (2014)
7. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**, 415–444 (2001)
8. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 207–217. Association for Computational Linguistics (2010)
9. Newman, M.E., Barkema, G.T., Newman, M.: Monte Carlo Methods in Statistical Physics, vol. 13. Clarendon Press, Oxford, Wotton-under-Edge (1999)
10. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think i am?" a study of language and age in twitter. In: ICWSM (2013)
11. Pennacchiotti, M., Popescu, A.M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 430–438. ACM (2011)

12. Traud, A.L., Kelsic, E.D., Mucha, P.J., Porter, M.A.: Comparing community structure to characteristics in online collegiate social networks. SIAM Rev. **53**(3), 526–543 (2011)
13. Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In: 2012 IEEE 12th International Conference on Data Mining (ICDM), pp. 1170–1175. IEEE (2012)
14. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 1151–1156. IEEE (2013)