Identifying Similar People in Professional Social Networks with Discriminative Probabilistic Models [*]

Suleyman Cetintas[†]
Dept. of Computer Sciences
Purdue University
scetinta@cs.purdue.edu

Monica Rogati
LinkedIn Corp.
Mountain View, CA
monica@rogati.com

Luo Si, Yi Fang
Dept. of Computer Sciences
Purdue University
{lsi,fangy}@cs.purdue.edu

ABSTRACT

Identifying similar professionals is an important task for many core services in professional social networks. Information about users can be obtained from heterogeneous information sources, and different sources provide different insights on user similarity.

This paper proposes a discriminative probabilistic model that identifies latent content and graph classes for people with similar profile content and social graph similarity patterns, and learns a specialized similarity model for each latent class. To the best of our knowledge, this is the first work on identifying similar professionals in professional social networks, and the first work that identifies latent classes to learn a separate similarity model for each latent class. Experiments on a real-world dataset demonstrate the effectiveness of the proposed discriminative learning model.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]
General Terms: Experimentation, Performance, Theory
Keywords: Social Networks, Similar People, Discriminative Learning

1. INTRODUCTION

Professional social networks (PSNs) are business oriented social networks where many core services such as recruiting, job seeking, expert/profile search, item recommendation, ad-targeting, etc. rely on successful identification of similar people. Due to the variety of available activities in PSNs, information about their users can be obtained from many heterogeneous sources such as profile content (e.g., users' job titles, their industries, specialties/skills, etc.), social graph (e.g., connections, group memberships, etc.) and user activities on the website (e.g., search and profile viewing, news sharing, etc.). Different sources provide different insights on user similarity, therefore information coming from these sources needs to be weighted and combined carefully. For instance; for a pair who do not have similar profile content while having very similar connections (due to common friends, etc.), information from their profile content can be used to decrease the importance of the information from their connectivity for estimating the overall similarity.

To the best of our knowledge, there is no prior work on identifying similar people in professional social networks. Most existing research on identifying similar users or matching people (e.g., online dating systems [3]) either used information from a single/homogeneous source (e.g., user profile content [3]), or used information from heterogeneous sources, but did not differentiate the information from these sources in a principled way (e.g., [4]).

In this paper, we propose a discriminative probabilistic model to identify similar professionals in professional social networks. Unlike the prior work, the proposed model identifies latent content and graph classes for people with similar profile content and social graph similarity patterns, and learns a specialized similarity model for each latent class. Experiments on real-world proprietary data from LinkedIn show the effectiveness of the proposed discriminative model.

2. METHOD

In a PSN, similarity features indicating whether two people are similar or not can mainly be grouped into 3 heterogenous categories: i) profile content features c; ii) social graph features g; iii) PSN usage features u. In particular, 5 content similarity features (comparing users' titles, industries, skills, specialties, associations), 3 social graph features (utilizing users' common connections, common groups, and whether their profiles are co-viewed or not), and 5 website usage features (utilizing the similarity in profile, search, inbox, news/sharing, accounts/settings page usage patterns) are used in this work. Given the set of similarity features f^v to denote all the similarity features of a people pair v, i.e., $f^v = \{c^v, g^v, u^v\}$, the proposed discriminative probabilistic model can be constructed as follows:

$$P(s^v = 1|f^v) = \sum_{t=1}^{N_t} \sum_{z=1}^{N_z} P(t|g^v) P(z|c^v) P(s^v|z, t, f^v) \quad (1)$$

where $s^v \in \{1, -1\}$ indicates whether the v^{th} pair is similar or not, $P(z|c^v)$ and $P(t|g^v)$ denote the mixing coefficients which are the probabilities of choosing latent content and graph classes z and t given content and graph features c^v and g^v. N_z and N_t are the number of latent content and graph classes and are chosen to be 3 and 2 respectively by AIC [1]. $P(s^v|z, t, f^v)$ can be modeled by logistic function as $P(s^v|z, t, f^v) = \sigma(\sum_i \lambda_i f^v_{zti})$ where $\sigma(x) = 1/(1 + \exp(-x))$ is the standard logistic function and λ_{zti}

[*] This research was partially supported by the NSF research grants IIS-0746830, CNS-1012208, and IIS-1017837.

[†] Part of this work was done when the first author was on a summer internship with LinkedIn.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

is the weight for the i^{th} feature f_i^v under the latent classes z and t. The mixing proportions $P(z|c^v;\alpha)$ can be modeled by a soft-max function $\frac{1}{Z_{c^v}}\exp(\sum_j \alpha_{zj}c_j^v)$ where Z_{c^v} is the normalization factor that scales the exponential function to be a proper probability distribution (i.e., $Z_{c^v} = \sum_z \exp(\sum_j \alpha_{zj}c_j^v)$). $P(t|g^v;\beta)$ can also be modeled by a soft-max function $\frac{1}{Z_{g^v}}\exp(\sum_q \beta_{tq}g_q^v)$.

The parameters of the model in Eqn.(1) (λ, α, β) can be estimated by the EM algorithm [2]. The E-step can be derived by computing the posterior probability of z and t, i.e. $P(z,t|f^v)$. By optimizing the auxiliary Q-function, we can derive the following M-step update rules:

$$\lambda^*_{zt\cdot} = \arg\max_{\lambda_{zt\cdot}} \sum_v P(z,t|f^v)\log\Big(\sigma\big(\sum_i s^v\lambda_{zti}f_i^v\big)\Big) \quad (2)$$

$$\alpha^*_{z\cdot} = \arg\max_{\alpha_{z\cdot}} \sum_v \big(\sum_t P(z,t|f^v)\big)\log\Big(\frac{1}{Z_{c^v}}\exp\big(\sum_j \alpha_{zj}c_j^v\big)\Big) \quad (3)$$

The update rule for (β) can be achieved similarly with Eqn.(3). Eqns.(2,3) can be optimized by any gradient descent method. In particular, we use the Quasi-Newton method. The proposed discriminative probabilistic model that has latent content and social graph classes is referred as *Latent_CG_Mod*.

For better comparison, similar models that only model latent content or social graph classes (namely *Latent_C_Mod* and *Latent_G_Mod* respectively) are constructed (each with 3 classes chosen by AIC) as well. All of those discriminative models are compared to a baseline method without any latent class. The baseline method is modeled with logistic regression and is referred as *LogReg_Mod*.

3. EXPERIMENTS

We evaluate the effectiveness of the proposed models on proprietary data from LinkedIn. As the first step of creating a dataset of similar people, a set of 2200 key profiles (target people for whom similar people will be searched) are selected by intersecting two sets: a) profiles popular among recruiters, and b) profiles popular within general users (sets identified by mining the large-scale recruiter/user activity logs for a time period of 6 months). Next; using Lucene[1], public profiles in the PSN are indexed, and each key profile is used as a structured query to retrieve top 100 candidate profiles for the key profile. From these 100 profiles, 10 candidate profiles are selected under three strategies: i) top 10, ii) bottom 10; and iii) sampled 10 (i.e., between ranks 1-10, 91-100, and randomly). A total of 22000 profile pairs are identified for annotation. Finally, the CrowdFlower[2] crowdsourcing service is employed to annotate each people pair with a similarity rating from 1 (being the least similar) to 4 by three annotators. The final rating for each pair is calculated by the average rating weighted by annotator trust.

Pairs with a similarity rating > 2.5 are regarded as similar people (with a total of 4633 similar pairs). Due to the candidate profile selection strategy, all people pairs in the annotated dataset are above some similarity level, which introduces bias in the set of negative (dissimilar) data instances. Therefore, a set of 2419 pairs randomly selected from the set of annotated pairs with similarity rating ≤ 2, is combined with another set of 2373 (less similar) pairs randomly selected from the set of public member profiles. A total of 4792 pairs of profiles are collected as the set of dissimilar people, resulting in a dataset of 9425 profile pairs balanced between positive and negative examples. Two thirds of the data (6283 pairs) are used for training the models, and one third of the data (3142 pairs) is used for testing.

Table 1: Experiment results of the proposed discriminative model. The † symbol indicates statistical significance with p-value < 0.1 (paired t-tests).

	F_1
LogReg_Mod	0.7720
Latent_G_Mod	0.7831
Latent_C_Mod	0.7859
Latent_CG_Mod	0.7921†

Experiments are conducted on the constructed dataset to evaluate the proposed model. The models are evaluated by the common F_1 measure as precision and recall are both important. It can be seen in Table 1 that both *Latent_C_Mod* & *Latent_G_Mod* achieve improvements over *LogReg_Mod*; and the improvement of *Latent_C_Mod* is comparable with *Latent_G_Mod* (while slightly better). This set of experiments shows that by having higher flexibility via introducing a latent class and allowing the combination weights vary accordingly, improvements can be achieved. Table 1 also shows that the proposed probabilistic discriminative model *Latent_CG_Mod* achieves the best performance by modeling the latent content and graph classes that provide more flexibility than *Latent_C_Mod* and *Latent_G_Mod* models, and much more flexibility than the *LogReg_Mod* model, leading to its superior performance. This explicitly shows that differentiating the pairs with different profile content and social graph similarity patterns, and specializing the similarity model for different pairs of people that share similar similarity patterns is important for achieving higher similarity accuracy.

4. CONCLUSIONS

Identifying similar people is an important task for professional social networks. Different people pairs have different profile content and social graph similarity patterns, and it is important to learn specialized similarity models for people with different similarity patterns. This paper proposes a discriminative probabilistic model that identifies latent content and graph classes for people with similar profile content and social graph similarity patterns, and learns a specialized similarity model for each latent class. Experiments on real-world data show the effectiveness of the proposed discriminative probabilistic model.

5. REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Auto. Control*, 1974.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*, 1977.

[3] F. Diaz, D. Metzler, and S. Amer-Yahia. Relevance and ranking in online dating systems. In *ACM SIGIR'10.*

[4] I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel. Same places, same things, same people?: mining user similarity on social media. In *ACM CSCW'10.*

[1] lucene.apache.org

[2] www.crowdflower.com