User Profile Completion with Online Social Circles

Hailong Qin¹, Jing Liu², Chin-Yew Lin², Ting Liu¹

¹(Harbin Institute of Technology, China)
²(Microsoft Research Asia, China)

Abstract: User profile is a kind of most effective feature and cue of user behavior analysis. However, most of users' profiles are incomplete and untruthful, and users' tag and profile data are rather sparse on social media. It makes profile completion become a popular research topic. Users have many strong relationships online, and these relationships form many social circles, such as classmate, family and colleague. A user's different online social circles can represent this user's various social attributes. We focus the task of user profile completion in this paper. We propose an algorithm of user profile completion via users' social circles by non-negative matrix factorization. The method can detect a user's multi-dimension social characteristics from various social circles. We also transfer the concept of social circle to academic networks. A paper may cite other papers with several intentions. Some of the references may be relevant according to the research problems and other references may be relevant according to the methodologies. The references of different topics form different "reference circles". By an academic paper's different "reference circles", our model can detect the paper's keywords in different views. The experiment on Facebook LinkedIn and Microsoft AcademicSearch show our method has improved performance than state-of-art methods.

Keywords - User Profile Completion, Social Circle, Non-negative Matrix Factorization

I. Introduction

Following the development of Web 2.0, more and more people communicate with their real friends on social network service (SNS), such as Facebook LinkedIn and Twitter. These users publish and share many feeds on social network. Both these content and users' profiles on SNS can be useful features of user behavior analysis. However, a main characteristic of on-line social media is that it is easy to provide false profiles such as gender, age, and hometown and so on. So it is difficult to detect these users' true identities and their profiles. For solving this problem, user profile completion becomes a popular research topic in the area of social network analysis.

Figure 1. A user's profiles on Facebook.

A user's profiles usually have many items, such as gender, age and school name (Fig 1). For example, the user's major field can show users' professions and the middle school field can indicate a user's hometown and school name generally. Most of existed works about user profile completion focus on predicting user's age...
or gender via text feature. In general, these simple profiles can be inferred from users' language styles. A lot of methods classify different age groups and genders with their posts \cite{1, 2, 3, 4}. However, it is difficult to detect users' other kinds of profiles from text feature. Few users explicitly refer their major or hometown in their posts frequently. On research of user behavior analysis, these items are usually more informative features. And it is also difficult to indicate exact classification for these items by texts.

![Figure 2](image)

**Figure 2.** A people's social circles

A user has many categories of strong relationships on social media. These strong relationships are constituted of users' online social circles. A user has several social circles on social media, such as classmates in a school, and colleagues in a company and so on (Figure 2). Social circle reflects an individual's social environment that can often be leveraged to infer important information about that individuals' attitudes, behaviors, and decisions \cite{5, 6, 7, 8, 9, 10, 11}. However, this type of social circle is different from traditional community. In the view of graph, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups. This feature of a real networks is called community structure \cite{12}. A node of a community may refer to people, and it also could be a computer in Internet or a gene in a gene network. Firstly, a traditional community is larger than a general social circle which may just has three members (such as three family members). On the other hand, a residential area can be a community in a mobile communication network. And the area may have several thousand people, but it is not anyone's social circle. Secondly, although there are some common tags or profiles of each member in same community, these members may don't know each other. In contrast, there are strong relationships between each member in the same social circle. Namely, a community of mobile network depends on the location, not explicit social relationships. As suggested above, community detection is focused on finding arbitrary highly interconnected subgraphs within larger networks, and social circle detection will instead discover several groups of strong social relationships including one or more specific individuals \cite{13}.

Users often communicates with their different social circles (such as families and classmates) on SNS and members of social circles are usually users' strong relationships. These social circles can represent a user's multi-dimension social relationships and social attributes. For example, the common characteristics of family circle are relevant to a user's hometown. And more information about a user's company can be detected from colleague circles. So members in a same social circle can supplement profiles each other. Obviously, a user usually has some common profile items with her/his every friend. But Li \cite{14} proposes that profile completion should only depends on users' close relationships. That means profiles between friends are more likely similar and valuable. However, Li \cite{14} completes profiles by users' labeled friends, and they are not ground-truths of users' social circles. Members in a same social circle have most possible to have same profile items, and the different kinds of profile items can represent a user's different characteristics. So it need an algorithm to complete user profiles with real social circles. For this purpose, we propose an algorithm by non-negative matrix factorization. As social circle also has similar situations and features on academic networks. Our algorithm can also be applied to academic networks for keywords detection of papers.

We hypothesis a user's all profile items are unknown and this user's several social circles are known. Our task is completing the user's profiles by her/his social circles on social network. For collecting more user profiles in the view of users' real social circles, this paper proposes a user profile completion model with members of users' social circles via non-negative matrix factorization. With features of users' circles and all closed friends' profiles, the algorithm can complete the center user's (The user who own these social circles.)
profiles. Our algorithm conducts on Facebook LinkedIn and Microsoft Academic Research. Results of experiment prove that our method are more effective than baselines.

The rest of this paper is organized as follows: in the next section, we describe related work. Then we introduce our methodology of user profile completion via non-negative matrix factorization in section 3. We describe datasets and experiment in section 4. Finally, we conclude our work and point out avenues for future research.

II. Related Work

2.1 Profile Completion

User profile completion is a popular area of user analysis. [1,2,3,4,15] predict users' gender and age according their tweets. [16]infers users' political orientation or ethnicity from various user information. [17]classifies users' four attributes in Twitter. [18]adds the features of authors' name to profiles completion. All their ideas are based user classification according user profiles. [19]extracts users' jobs educations and spouses in Google Plus and Freebase. According to the homophily between users, [20]completes users' profile via their neighbors' information. [21]infers users' locations based on social relationships. Bergsma [22]proposes different types of users have different attributes. Liu [23] completes user profiles via their labeled friends. However, all of these work do not refer to real social circles.

2.2 Social Circle

Social circle is closely relevant to the term of community which has two interpretations: one is the geographical notion of community and another is relational. The second one is concerned with the quality of the character of human relationships, without reference to location [24]. The online social circles carry the second meaning and do not concern physical locations. The detection of social circles is a new research area which emerging with the popularity of social networks. It is a clustering problem within ego network, members within a same social circle do not only have dense relationships, but also have some same profiles. So a social circle is the group with a specified social meaning.

A growing number of scholars study their subjects in view of social circles. The recommendation algorithm based on users' circles performs no worse than those based on the full network [25]. Question recommendation, question popularity analysis and prediction based on social circles get better performance [26,23]. It is also a main feature of linking users across online social networks [27].

Bernado proposes that it is necessary to mine users' real friends [28]. The evolution of online social groups is analyzed and predicted by Kairam [29]. Qu and Liu propose a semi-supervised method to detect social circles in twitter [30]. But some members in user groups are not users' real friends in Twitter. A lot of groups just classify different types of followees, many of these followees are not users' bilateral friends. Since 2012, some specific social circle detection algorithms have been proposed by Leskovec and Qin [31, 32]. Burton [13] introduces a detection algorithm of local social circle when whole network is unknown. Liu [33] develops an algorithm for mining tags of social circles. However, all works are about social circles detection and mining, there is few work of user analysis based on online social circles.

2.3 Keywords Detection


III. User Profile Completion (11 Bold)

3.1 Principle

Every user has several items in her/his profiles and different items can represent different social attributes. On the other hand, a user has many social circles on SNS. Most members among a social circle are strong relationships. And their common profiles can likely represent attributes of their social circle. In the point
of social circle, attributes of a user’s different social circles likely constitute the user’s multi-dimension profiles. So users’ profiles can be supplemented with the help of different social circles.

In this work, we select non-negative matrix factorization since matrices can easily embed relationships of user-circle and user-profile. Our algorithm completes users’ profiles by main attributes of users’ different social circles via NMF. This model gets great performance and also has strong interpretability.

Circle-Profile matrix can be factorized into the product of circle-user matrix and user-profile matrix. As observed matrix, circle-profile matrix represents weights of all members’ profile items for every circle. A center user may have several social circles and members of these social circles are different. So contributions of social circles for user profile completion are also different. The user-circle matrix can represent different levels of contribution. And user-profile matrix represents weights of all profile items for every center user.

3.2 Algorithm

Definition I: Matrix $CP$ is a circle-profile matrix. $CP_{ij}$ is weights of profile $j$ for circle $i$. All circles and their members’ profile of every dataset are represented in a same matrix. (The matrix do not include the center user who own this social circle.)

Definition II: Matrix $CU$ is a circle-user matrix. Every row represents a circle in dataset, and every column represents a center user who owns several social circles. If the center user owns this circle, the value is the weight of the center user in the circle. Otherwise, the value is 0 in matrix. All center users and their circles of every dataset are represented in a same matrix. The matrix is initialized by random values.

Definition III: Matrix $UP$ is a user-profile matrix. The value represents weights of profile items to center users. All center users and their profiles of every dataset are represented in a same matrix. The matrix is initialized by random values.

We will complete users’ profiles according to weights of profiles in matrix $UP$.

$$CP \approx CU \times UP$$  

According to the principle of matrix factorization, $CP$ can be decomposed into the product of $CU$ and $UP$. We can get users’ social circles and their friends’ profiles, and construct matrix $CP$. This kind of factorization accord with the hypothesis of a people’s different social circles can represent the people’s different aspects of characteristics.

The NMF is carried out by minimizing the difference between $CP$ and $CU \times UP$. We take Euclidean Distance (Equation 2) as our loss function to measure the difference between $CP$ and $CU \times UP$.

$$J(CU, CP) = \frac{1}{2} \sum_{ij} (CP_{ij} - CU_{ij} \times UP_{ij})$$

As $CP$ are fixed, we find non-negative $CU$ and $UP$ by gradient descent (Equation 3 and 4).

$$CU_{ik} = CU_{ik} - \alpha_1 \times [(CP \times UP^T)_{ik} - (CU \times UP \times UP^T)_{ik}]$$

$$UP_{kj} = UP_{kj} - \alpha_2 \times [(CU^T \times CP)_{kj} - (CU^T \times CU \times UP)_{kj}]$$

Then we set $\alpha_1$ and $\alpha_2$ as Equation 5 and Equation 6.

$$\alpha_1 = \frac{CU_{ik}}{(CU^T \times CU \times UP)_{ik}}$$

$$\alpha_2 = \frac{UP_{kj}}{(CU \times UP \times UP^T)_{kj}}$$

So Equation 7 and Equation 8 are final iteration rules.

$$CU_{ik} = CU_{ik} \frac{(CP \times UP^T)_{ik}}{(CU \times UP \times UP^T)_{ik}}$$

$$UP_{kj} = UP_{kj} \frac{(CU^T \times CP)_{kj}}{(CU^T \times CU \times UP)_{kj}}$$

Matrix $UP$ represents weights of profile items for every user. The high weight items have more possibility to be center users’ profiles.
IV. Experiment

4.1 Dataset

4.1.1 Facebook

The dataset of Facebook is released in Kaggle\(^1\). It includes 60 users and their social circles. There are 17,115 friends in all social circles, while every user has 19.73 social circle and every social circle has 28.91 friends averagely. Every center users averagely has 16.97 profile items. The task is completing center users’ profiles in this dataset.

4.1.2 LinkedIn

The LinkedIn Dataset is released by University of Illinois at Urbana-Champaign\(^2\). It includes 187 users and their labeled friends. Different from Facebook dataset, labeled friends are not ground-truths of social circle. However, there are common characteristics among friends having same labels. So the algorithm completes a user’s profiles by her/his different labeled friends.

4.1.3 Microsoft Academic Search

Academic network is paper-based network. Most academic papers of computer science propose an effective methodology to solve a specified problems. Every paper will cite many relevant papers as its references. The network is consist of relationships of citation. The references of one paper will consist of different reference circles (Fig 4). Some paper circles may represent paper groups of similar research problems, and some circles may represent previous relevant methodology. According to citation relationships and academic meanings of a paper's reference circles, we can infer keywords of the paper collaboratively.

The dataset of academic paper is extracted from Microsoft Academic Search\(^3\). There are 25 papers in dataset. The part of Related Work is divided into many sections in these papers. All references of every section are relevant to a similar research problem or a similar research methodology. So these references can be regarded as ground-truth of reference-circles. The task is detecting center papers' keywords in academic dataset. We annotate technology terminologies in titles and abstracts of these papers, these terminologies are candidate keywords of center papers.

![A generative blog post retrieval model that uses query expansion based on external collections](image)

**Figure 4.** A paper's references are divided several parts, such as Query Modeling and External Expansion.

4.2 Baseline

4.2.1 TF-IDF

We regard every social circle as a document, and all members’ profiles are word in this document. And compute every profile item \(p\)'s TF value in \(Circle_i\) (Equation 9) and \(p\)'s IDF value (Equation 10) in the circle, then get \(TF \times IDF\) values. Finally, select top profile items as profiles of the social circle.

\[
TF_{pi} = \frac{\text{Count } (p \text{ in Circle}_i)}{\sum_{m \in \text{Circle}_i} \text{ProfileCount } (m)} \quad (9)
\]
User Profile Completion with Online Social Circles

\[
\text{IDF}(P) = \log \frac{|\text{Circle}|}{\sum_{\text{circle having } P}}
\]

4.2.2 Co-Profile

Different from age and gender prediction, detecting multi-dimension profiles is not a classification problem. We extract users' profiles from friends of social circles. [14] proves that only close friends are effective for profile completion. So we select co-profile method in [14] as our strong baseline. By the same motivation, co-profile also completes users' profile by close friends. However, this method detects users' social circles at the same time of profile completion. So their social circles in the LinkedIn dataset are not ground-truths. Our model is more persuasive for effect of real social circles.

As we have ground-truths of social circles in our dataset, we do not learn circles in co-profile method. Equation 11 is the cost function of co-profile. \( v \) is a center user's friend and \( C \) is the circle. \( f \) represents a friend's profile vector, \( f_0 \) represents a center user's profile vector, The profile vectors update by Equation 12 and 13.

\[
\lambda_1 \sum_{t=1}^{K} \left( \left( \sum_{v_i, v_j \in C_t} \omega_{t} \times (f_i - f_j) \right)^2 + \left( \sum_{v_i \in C_t} \omega_{t} \times (f_0 - f_i) \right)^2 \right) + \lambda_2 \times t=1KW\times(\omega\times f^{t-1})
\]

\[
f_{i,y} = \frac{f_{0,y} \sum_{v_j \in C_t} f_{j,y}}{1+ \sum_{j \in C_t} f_{j,y}}, \omega_{t,y} = 1 \quad (12)
\]

\[
f_{0,y} = \frac{\sum_{t=1,y=1}^{\omega_{t,y}=1} \sum_{v_j \in C_t} f_{j,y}}{\sum_{t=1,y=1}^{\omega_{t,y}=1} \sum_{v_j \in C_t} 1}, \omega_{t,y} = 1, \forall t = 1 ... K \quad (13)
\]

4.3 Result Analysis

Table 1. Overview of Dataset

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>LinkedIn</th>
<th>Microsoft Academic Search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Center Users</strong></td>
<td>60</td>
<td>187</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total Friends</strong></td>
<td>17,115</td>
<td>5455</td>
<td>625</td>
</tr>
<tr>
<td><strong>Avg Circles Per Center User</strong></td>
<td>19.73</td>
<td>3.24</td>
<td>2.56</td>
</tr>
<tr>
<td><strong>AvgProﬁles Per User</strong></td>
<td>16.97</td>
<td>3.57</td>
<td>7.31</td>
</tr>
</tbody>
</table>

Table 2. Precision of User Proﬁles Completion

<table>
<thead>
<tr>
<th>Profile Type</th>
<th>TF-IDF</th>
<th>Co-Profile</th>
<th>Social Circle-Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook(P@N)</td>
<td>N/A</td>
<td>21.66%</td>
<td>19.98%</td>
</tr>
<tr>
<td>LinkedIn(P@10)</td>
<td>Employer</td>
<td>27.05%</td>
<td>60.00%</td>
</tr>
<tr>
<td></td>
<td>College</td>
<td>60.13%</td>
<td>61.00%</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>46.70%</td>
<td>68.00%</td>
</tr>
</tbody>
</table>

TF-IDF method is an effective and simple method of extracting main profiles and keywords. But different social circles have different importance for a person. TF-IDF method cannot combine with these relationships. So TF-IDF gets low performance when completing users' profiles.

In datasets of Facebook and academic data, our method is improved than TF-IDF and co-profile obviously. Because our method can extract key profiles and keywords of different social circles and different parts of academic related work. That proves that performances of our work are good in both problems of user profile completion and academic keywords detection.

In the LinkedIn dataset, only the performance of college profile is better than co-profile. After data analysis, our algorithm strongly depend on user friends' profile. The lack of friends' profiles will lead bad performance. The profile items about college in a kind of labeled friends are more identical in the dataset. As we don't know ground-truths of LinkedIn social circles, we use every user's labeled friends as a kind of social circle. Different from our model, co-profile includes the process of social circle detection based on labeled friends, so our friends sets for profile completion are different from baseline. On the other hand, both our experiment and the baseline do not use ground-truths of social circles in LinkedIn. However, other two datasets prove our model has better performance for ground-truth of user social circles.

V. Conclusion

The paper proposes an algorithm of user profile completion with the feature of social circle. In the
aspect of social relationships, we combine social memberships and user friends’ profiles in a model of non-negative matrix factorization. The model can supplement users’ multi-dimension profiles by social circles and has strong interpretability.

That is a new viewpoint of user profile completion with user strong relationships. At the same time, we also apply the concept of social circle to reference category of academic paper. Excepting completing user profile, the algorithm is also an effective keywords detection method on academic network. The experiment shows our method has improved performance than state-of-art methods. In future, we will detect multi-dimension user attributes from users’ posts, and try to complete user profile more precisely by combining features of social circles and user texts. We will also try to explore author circles on academic networks.

Acknowledgements

This work was supported by the National Basic Research Program (973 Program) of China (No. 2014CB340503), National Natural Science Foundation of China (No. 61133012 and No. 61472107).

References


