What Are You Known For? Learning User Topical Profiles with Implicit and Explicit Footprints

Cheng Cao*, Hancheng Ge*, Haokai Lu, Xia Hu, and James Caverlee
Department of Computer Science and Engineering
Texas A&M University
chengcao@tamu.edu,{hge,hlu,hu,caverlee}@cse.tamu.edu

ABSTRACT

User interests and expertise are valuable but often hidden resources on social media. For example, Twitter Lists and LinkedIn's Skill Tags provide a partial perspective on what users are known for (by aggregating crowd tagging knowledge), but the vast majority of users are untagged; their interests and expertise are essentially hidden from important applications such as personalized recommendation, community detection, and expert mining. A natural approach to overcome these limitations is to intelligently learn user topical profiles by exploiting information from multiple, heterogeneous footprints: for instance, Twitter users who post similar hashtags may have similar interests, and YouTube users who upvote the same videos may have similar preferences. And yet identifying "similar" users by exploiting similarity in such a footprint space often provides conflicting evidence, leading to poor-quality user profiles. In this paper, we propose a unified model for learning user topical profiles that simultaneously considers multiple footprints. We show how these footprints can be embedded in a generalized optimization framework that takes into account pairwise relations among all footprints for robustly learning user profiles. Through extensive experiments, we find the proposed model is capable of learning high-quality user topical profiles, and leads to a 10-15% improvement in precision and mean average error versus a cross-triadic factorization state-of-the-art baseline.

CCS CONCEPTS

•**Information systems** → *Social tagging systems; Collaborative filtering; Users and interactive retrieval;*

KEYWORDS

Social Media; User Profile; User Behavior

ACM Reference format:
Cheng Cao[1], Hancheng Ge[1], Haokai Lu, Xia Hu, and James Caverlee. 2017. What Are You Known For? Learning User Topical Profiles with Implicit and Explicit Footprints. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan, ,* 10 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080820

*Equal contribution

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3077136.3080820

1 INTRODUCTION

In social media systems, *demographic profiles* — often including name, age, gender, and location — provide an important first step toward creating rich user models for information personalization. For example, a user's location can be a signal to surface local content in the Facebook newsfeed. These demographic profiles typically reveal very little about a user's topical interests (what she likes) or expertise (what she is known for). Hence, there is great effort toward building high-quality *user topical profiles*, toward improving user experience and powering important applications like personalized web search [42], recommendation system [13, 33], expert mining [11], and community detection [53].

Indeed, there are two major approaches to build the topical profiles for social media users. One thread of methods seeks to uncover latent factors that may be descriptive of a user. For example, running Latent Dirichlet Allocation (LDA) over a user's posts in social media can reveal the topics of interest of the user [28, 34, 48]; similarly, matrix factorization approaches have proven popular at capturing user factors, often for personalization purposes [14, 15, 21, 33, 42, 51, 56]. Aside from such recommendation applications, latent factor models have also been used to find influential users, mine communities, and predict review quality [31, 48, 53]. Another thread of methods seeks to encourage social media users to directly assess each other's interests and expertise, providing a partial perspective on user topical profiles. For example, LinkedIn users can choose *skill tags* for their own profiles and can endorse these tags on the profiles of others. *Twitter Lists* allow users to organize others according to user-selected keywords, e.g., placing a group of popular chefs on the list "Top Chefs". In this way, some list names can be viewed as a topical tag for list members. In the aggregate, this crowd-contributed tagging knowledge can be viewed as explicit evidence for capturing user interests and expertise [4, 11, 39].

Both approaches, however, face great challenges. Approaches that identify latent topics (often, as a distribution over features in some lower dimensional space) are typically trained only over content (ignoring other important footprints) and are difficult to directly interpret. Methods that only use crowdsourced tags typically suffer from limited coverage; that is, while the hand-curated tags may be of high-quality, very few users actually have descriptive topical tags associated with them. For example, in a random sample of 3.5 million Twitter users, we find that only 2% have been labeled with a topical tag (more details in Section 5). Moreover, to better understand user topical interests and expertise, a more comprehensive profiling framework is necessary. For instance, it is unclear what kind of evidence is useful for user topical profiling. And how can such potentially heterogeneous evidence be modeled for user topical profiling?

Hence, in this paper, we propose to exploit heterogeneous footprints (e.g., tags, friends, interests, behavior) for intelligently learning user topical profiles. Based on a small set of explicit user tags, our goal is to extend this known set to the wider space of users who have no explicit tags. The key intuition is to identify "similar" users in terms of their topical profiles by exploiting their similarity in a *footprint space*. For instance, Twitter users who post similar hashtags may have similar interests, and YouTube users who upvote the same videos may have similar preferences. Such evidence of *homophily* has been widely studied in the sociological literature [35] and repeatedly observed in online social media, e.g., [5, 7, 26, 46, 49]. But what footprint spaces are appropriate for finding this homophily? What impact do they have on the discovery of user topical profiles? And which footprints are more effective at uncovering topical profiles?

Toward answering these questions, the rest of this paper makes the following main contributions:

- First, we formulate the problem of learning user topical profiles in social media, with a focus on leveraging heterogeneous footprints.
- Second, we demonstrate how to model different footprints (e.g., like interests, social, and behavioral footprints) under this framework, and we present a unified 2-D factorization model in which we simultaneously consider all of these footprints (called *UTop*).
- Third, we then extend this initial approach through a generalized model that integrates the pairwise relations across all potential footprints via a tensor-based model (called *UTop+*), which provides a more robust framework for user profile learning.
- Finally, through extensive experiments, we find the proposed UTop+ model is capable of learning high-quality user topical profiles, and leads to a 10-15% improvement in precision and mean average error versus a state-of-the-art baseline. We find that behavioral footprints are the single strongest factor, but that intelligent integration of multiple footprints leads to the best overall performance.

2 RELATED WORK

Finding User Interests and Expertise. Finding user interests and expertise has numerous applications, and one of the most popular tasks is personalized search and recommendation. Considerable research [13, 14, 21, 33, 34, 42, 54, 56] has been dedicated to uncover users' latent interests or expertise as their personal preferences for building recommender systems in different domains, such as web search [34, 42], web content [56], rating systems [20, 33], and social media [14, 15, 21, 54].

For social media research, the latent factor model is a state-of-the-art method for user recommendation. Interpreting the latent factors as topics, approaches based on such a model usually avoid explicitly identifying user interests but instead integrate the factors into a recommendation task. For example, Hong et al. applied matrix factorization on both users and tweets and focused on recommending user's retweeting behavior [15]. Similarly, Jiang et al. presented a probabilistic matrix factorization method to recommend whether a user adopts an item on a social network [21]. Zhong et al. collected user's webpage views to build a matrix factorization profile for web content recommendation [56].

Leveraging Footprints. A sequence of research has focused on using various footprints to learn user interests. One of the most traditional approaches is to model text-based footprints to obtain users' latent topical preferences, as in the case of PLSA and LDA [37, 38, 48, 53]. Another popular footprint is social (often via friendships) [21, 32, 36], with the natural homophily assumption that friends tend to have similar profiles. In addition, behavioral footprints have become a newer factor; for example, Guy et al. used a user's tagging behavior as evidence for content recommendation [13]. Lappas et al. considered *user endorsement* as a behavioral signal [27]. In [54], Zhao et al. focused on the behaviors of commenting, "+1", and "like" on Google+. Some of the other footprints that have been explored in previous works include user's emotions and sentiment, geo-location, temporal context and linguistic activity. For example, Hu et al. [17, 18] proposed an unsupervised factorization approach for user sentiment analysis through emotional signals. Lu et al. [30] considered user's geographical footpints to discover what people are known-for. Yin et al. [51] proposed a probabilistic graphical method to model user's temporal interest for item recommendation. Hu et al [19] applied a factorization method to infer linguistic properties of user's documents. However, typically, these different footprints have been treated separately.

Factorization Models. Technically, it is challenging to embed users' heterogeneous footprints into a factorization model. A handful of studies have adopted a regularization model [29, 31, 33] for personalized recommendation, though typically focusing on only one footprint. In [20], latent spaces are learned separately for each footprint through probablisitc matrix factorization assuming they are not independent. Tensor-based factorization methods [2, 8] have been used in many applications such as behavior modeling, healthcare, and urban planning [22, 47, 55]. A more comprehensive survey of tensor factorization and its applications can be found in [24]. In contrast, we first propose a factorization model in which we simultaneously consider multiple contexts via linearly weighted regularization. We then extend the model with a generalized tensor-based factorization so that not only different types of footprints can be considered together but their multi-linear interactions with each other can be exploited.

Several studies have focused on heterogeneous domains or entities, instead of contexts. Yu et al. put mutiple types of entities into a heterogeneous network and used a Bayesian ranking process to estimate user preferences [52]. Similarly, Hu et al. looked into a traditional user-item recommendation problem, presenting a factorization model across heterogeneous items. However, the network will quickly grow when users and items increase. Singh and Gordon proposed a framework to learn different types of relations, where they iteratively do matrix factorization between all pairs of domains [43]. Hu et al. [16] adopted the existing PARAFAC2 factorization algorithm on a tensor model, which is obtained by combining user ratings of different merchandises like book, music, and movie. Zhong et al. [56] directly applies a matrix factorization model on Web users and their clicked content items. However, in this work, we focus on learning user topical profiles rather than recommending item ratings for users.

Personalized Tag Recommendation. Another related research line focuses on personalized tag recommendation for users in social tagging systems [9, 25, 40, 41, 45, 50]. For example, Rendle et al. [40, 41] proposed tensor factorization to suggest tags to users for annotation on different items. Feng et al. [9] modeled social tagging as a multi-type graph and proposed random walk with restart for tag recommendation. Konstas et al. [25] also proposed a modified random walk with restart by exploiting social relationships and tagging for item recommendation. Our work is different from personalized tag recommendation in two aspects. The first is that we use crowdsourced tags to represent user's interests and expertise instead of annotating items in social systems. The second is that our problem is to infer users' topical profiles through tags for unknown users based on their different footprints rather than recommend tags based on partial knowledge of a user's profile.

3 PRELIMINARIES

Explicit Footprints. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be a set of users where N is the number of users, and $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ is a set of M tags each of which is associated with a particular topic. Suppose we have a subset of users $\mathcal{S} \subset \mathcal{U}$ where each user $u_i \in \mathcal{S}$ has been labeled with a subset of $\mathcal{T}$, typically based on the collective efforts of the crowd. In this paper, we refer to such labels as *explicit footprints*. Practical examples of explicit footprints include LinkedIn Skill Tags and Twitter Lists, wherein users can provide a crowdsourced summary of a user's interests and expertise [4, 11, 39]. We denote the explicit footprints as the user-tag matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}|\times M}$ in which element $\boldsymbol{P}(i, j)$ represents the number of times u_i is labeled by t_j.

Learning User Topical Profiles. Given a set of users $\mathcal{U}$, a set of tags $\mathcal{T}$, and a subset of users $\mathcal{S} \subset \mathcal{U}$ for whom we know their user topical profiles $\boldsymbol{P}$, the problem of *Learning User Topical Profiles* is the task of inferring the unknown tags from $\mathcal{T}$ for users in $\mathcal{U} - \mathcal{S}$.

An Initial Attempt with Explicit Footprints Only. A natural choice for attacking the challenge of learning user topical profiles is the matrix completion approach, which has been adopted in many related works [15, 43, 52, 56]. Under a matrix completion approach, we can extend $\boldsymbol{P}$ to a larger matrix $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ by including all users of $\mathcal{U}$. Then, we can formulate the learning user topical profiles problem as a matrix completion problem:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V}} \quad & \frac{1}{2}\|\Omega \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2, \\ \text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \end{aligned} \tag{1}$$

where $\boldsymbol{X}$ is a user-tag matrix, and $\boldsymbol{U} \in \mathbb{R}^{N\times K}$ and $\boldsymbol{V} \in \mathbb{R}^{M\times K}$ are latent representations of users and tags, respectively. $K \ll min(N, M)$ is the number of latent dimensions. Since the given $\boldsymbol{X}$ is naturally non-negative, we add the same constraints for $\boldsymbol{U}$ and $\boldsymbol{V}$ so that we can better interpret the values in them. Ω is a non-negative matrix with the same size of $\boldsymbol{X}$:

$$\Omega(i, j) = \begin{cases} 1 & \text{if } \boldsymbol{X}(i, j) \text{ is observed,} \\ 0 & \text{if } \boldsymbol{X}(i, j) \text{ is unobserved.} \end{cases}$$

The basic matrix completion model above learns an optimal set of $\{\boldsymbol{U}, \boldsymbol{V}\}$ to approximate the original matrix $\boldsymbol{X}$, estimating for unobserved users through observed user-tag pairs. However, as in many linear-inverse problems, there may not be sufficient information to estimate the original matrix $\boldsymbol{X}$ based only on the partially observed data. The problem of learning user topical profiles is one such case, since most of our target users do not have any partially explicit footprint.

Implicit Footprints. With the scarcity of explicit footprints in mind, we are interested to explore the potential of *implicit footprints* for learning unknown user topical profiles. Implicit footprints may indirectly reflect user interests or expertise. Typical implicit footprints, for example, could include user behaviors, the social circle of a user, sentiment-based features of a user's posts, the geo-location of a user, emotional cues, and temporal dynamics, among many others [13, 17–19, 21, 27, 30, 32, 36, 51, 54]. The key intuition is to identify "similar" users in terms of their topical profiles by exploiting their similarity via these implicit footprints. Since evidence from these heterogeneous implicit footprints may provide conflicting evidence, potentially leading to lower quality user profiles than considering footprints in isolation, we propose a generalized optimization framework that takes into account pairwise relations among all possible implicit footprints for learning user profiles. In this way, the benefits of each footprint may be intelligently combined to find the best evidence across multiple implicit footprints for learning high-quality user profiles.

4 LEARNING USER TOPICAL PROFILES

We turn in this section to propose a generalized model for learning user topical profiles. We first identify multiple implicit footprints and demonstrate how to model them. We then introduce a matrix factorization based approach — called UTop, before extending this version to a more general tensor-based approach — called UTop+.

4.1 Modeling Implicit Footprints

We aim to integrate many different kinds of implicit footprints into the framework for learning user topical profiles. For the concreteness in our discussion, we focus in this section on three specific types of implicit footprints that capture three different perspectives on user topical profiles. The three footprints are: *social*, based on the friends (via the social graph) around the user; *interest*, based on the text posts made by the user; and *behavioral*, based on the URL sharing activities of the user. The intuition is that these varied implicit footprints can connect related users, such that user topical profiles can be propagated from user to user. But how should we model these kinds of implicit footprints? And how can we integrate them into a matrix completion model? Note that the proposed model can be easily extended to incorporate additional footprints.

Social Footprints. Social footprints — directly suggested by homophily — naturally indicate that connected users may share common interests, and hence can be used for inferring user topical profiles [13, 21, 32, 36]. For example, if Carol and David are following each other on Twitter, the social footprint suggests that it is more likely for them to share common interests.

These social network connections between users can be naturally modeled as a matrix. We denote the matrix as $\boldsymbol{E} \in \mathbb{R}^{N\times N}$ in which the binary element $E(i, j)$ represents if user u_i and user u_j have a connection on a social network. We can model this social footprint

(a) Text-based Interest Footprint

(b) Behavioral Footprint

Figure 1: Examples of Different Implicit Footprints on Learning User Topical Profiles

as a regularization term:

$$\mathcal{L}_1 = \frac{1}{2}\|E - UU^T\|_F^2.$$

Our goal is to optimize the user latent matrix U in order to minimize $\mathcal{L}_1$, with the intuition that friends are likely to have similar profiles. Of course, users may form relationships in social media for many diverse reasons, and so these relationships may not be appropriate for inferring similar topical profiles. As one example, family members may be "friends" in a social network but can have distinct topical profiles (e.g., sister vs brother, grandson vs grandfather). Hence, we next consider additional implicit footprints that may serve to mitigate these challenges.

Interest Footprints. The second footprint we consider is based on user interests. Texts posted by users can semantically reflect related subjects associated with their interests or expertise. Thus, many studies have directly applied LDA on posted texts, assuming the (latent) topics in user's posts are their topical profiles [28, 34, 48]. In Figure 1a, Alice is a basketball fan and she has posted many tweets talking about the upcoming NBA all-star game. We find Bob's tweets share many of the same words as Alice's. Hence, their posted texts demonstrate their shared interests in basketball, suggesting that Alice's user topical profile may be similar to Bob's.

We can model this text-based interest footprint like so: let $\boldsymbol{w} = \{w_1, w_2, \ldots, w_L\}$ be the set of words, where L denotes the number of words. $\boldsymbol{A} \in \mathbb{R}^{N\times L}$ is a user-word matrix in which $\boldsymbol{A}(i,j)$ is the frequency of word w_j appearing in user u_i's posts. Similarly, $\boldsymbol{B} \in \mathbb{R}^{M\times L}$ is a tag-word matrix where $\boldsymbol{B}(i,j)$ represents the frequency of word w_j posted by all users who have tag t_i. We propose to leverage a user's interest footprint as the following loss function:

$$\mathcal{L}_2 = \frac{1}{2}\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2 + \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T\|_F^2,$$

where $\boldsymbol{W} \in \mathbb{R}^{L\times K}$ represents word's latent topics. Our goal is to minimize $\mathcal{L}_2$ so that two users who are "nearby" in the interest footprint space tend to have similar topical profiles. However, a user's posts are often short (like on Twitter) and may contain many nonsense or off-topic texts, which can interfere with clearly revealing user topical profiles. Hence, we next turn to a third footprint for overcoming these issues.

Behavioral Footprints. Finally, we propose to augment the social and textual footprints with behavioral footprints [13, 27, 54]. According to the homophily evidence in the behavior dimension [35], for instance, two YouTube users may have close tastes if they usually "like" or "dislike" the same videos. A retweet on Twitter is a strong indication of the retweeter's personal endorsement, so two users can have similar preferences if they often retweet the same tweets. Hence, these behavioral footprints may provide strong evidence beyond who users are connected to (social) and what they post (interests).

In this paper, we adopt *URL sharing* as a public, observable behavior that may serve as a first step toward improving the learning of user topical profiles. Other behavioral footprints are possible, and we anticipate revisiting these in our future work. URL sharing behavior for topical profiles has received some attention in social media research. Previous work looked into why and what content people share via URLs in social media [3, 44]. Some other work has mentioned the role of URL sharing in social spamming [10]. Through URL sharing, users can concisely express their viewpoints, interests, and professional expertise. For instance, a person who works in the IT industry may usually post URLs linking to engadget.com. A user who likes sports may often share URLs of espn.com. In Figure 1b, Carol is a political journalist so she regularly posts some URLs linking to huffingtonpost.com, and we see David also usually shares the same URLs. In this case we may infer politics-relevant tags for David.

Concretely, let $\boldsymbol{Z} = \{z_1, z_2, \ldots, z_P\}$ be the set of URLs posted by users. Similar to the interest footprint, we define $\boldsymbol{C} \in \mathbb{R}^{N\times P}$ as a user-URL matrix where $\boldsymbol{C}(i,j)$ is the frequency of URL z_j posted by user u_i. Also, $\boldsymbol{D} \in \mathbb{R}^{M\times P}$ is a tag-URL matrix with $\boldsymbol{D}(i,j)$ as the frequency of URL z_j appearing in all posts from users having tag t_i. As a result, we leverage URL sharing via the following loss function:

$$\mathcal{L}_3 = \frac{1}{2}\|\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T\|_F^2 + \frac{1}{2}\|\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T\|_F^2,$$

where $G \in \mathbb{R}^{P \times K}$ represents URL's latent topical spaces. Our goal is to minimize $\mathcal{L}_3$, with the idea that users may have similar topical profiles if they behave similarly when posting URLs.

4.2 Learning User Topical Profiles: A 2-D Model

Since evidence from multiple implicit footprints may provide conflicting evidence, potentially leading to lower quality user profiles than considering footprints in isolation, we turn in this section to developing a unified model that can integrate all possible heterogeneous footprints together into a matrix (2-D) completion model. Since all implicit footprints are modeled as regularization terms in Section 4.1, intuitively we can linearly incorporate them into the proposed *UTop* model. Again, recall that we focus our presentation here on those three specific footprints (social, interest, behavioral), but the model is designed to generalize to more alternative footprints as well.

Figure 2 gives an overview of UTop. In general, we factorize each of the social, interest, and behavioral footprint matrices, and assume that the objective user-tag matrix shares the same latent user dimensions with them. This is the fundamental assumption in most factorization-based methods for solving matrix completion problems. We also consider explicit footprints. Similarly, we collect each tag's latent representation, and multiply them with each user's latent factor for estimating the objective matrix.

Concretely, we formulate the following optimization problem as following:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{G}} \quad & \mathcal{F} = \frac{1}{2}\|\Omega \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2 \\ & + \frac{\lambda}{2}(\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2 + \|\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T\|_F^2) \\ & + \frac{\gamma}{2}(\|\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T\|_F^2 + \|\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T\|_F^2) \\ & + \frac{\delta}{2}\|\boldsymbol{E} - \boldsymbol{U}\boldsymbol{U}^T\|_F^2 \\ & + \frac{\alpha}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 + \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{G}\|_F^2) \\ \text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \boldsymbol{W} \geq 0, \boldsymbol{G} \geq 0, \end{aligned} \tag{2}$$

where λ, γ, δ and α are positive regularization parameters controlling the contributions of different implicit footprints. $\|\boldsymbol{U}\|_F^2$, $\|\boldsymbol{V}\|_F^2$, $\|\boldsymbol{W}\|_F^2$ and $\|\boldsymbol{G}\|_F^2$ are deployed to avoid overfitting. Similar to Equation 1, we insert the non-negative constraints for $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$, and $\boldsymbol{G}$.

The derivation of the objective function in Eq.(2) regarding four variables $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$ and $\boldsymbol{G}$ are demonstrated as:

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}} &= -\Omega \odot \Omega \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} - \lambda(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T) \\ &\quad - \gamma(\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T) - 2\delta(\boldsymbol{E} - \boldsymbol{U}\boldsymbol{U}^T) + \alpha\boldsymbol{U}, \\ \frac{\partial \mathcal{F}}{\partial \boldsymbol{V}} &= -\Omega^T \odot \Omega^T \odot (\boldsymbol{X}^T - \boldsymbol{V}\boldsymbol{U}^T)\boldsymbol{U} - \lambda(\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T) \\ &\quad - \gamma(\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T) + \alpha\boldsymbol{V}, \\ \frac{\partial \mathcal{F}}{\partial \boldsymbol{W}} &= -\lambda(\boldsymbol{A}^T - \boldsymbol{W}\boldsymbol{U}^T)\boldsymbol{U} - \lambda(\boldsymbol{B}^T - \boldsymbol{W}\boldsymbol{V}^T)\boldsymbol{V} + \alpha\boldsymbol{W}, \\ \frac{\partial \mathcal{F}}{\partial \boldsymbol{G}} &= -\gamma(\boldsymbol{C}^T - \boldsymbol{G}\boldsymbol{U}^T)\boldsymbol{U} - \gamma(\boldsymbol{D}^T - \boldsymbol{G}\boldsymbol{V}^T)\boldsymbol{V} + \alpha\boldsymbol{G}. \end{aligned} \tag{3}$$

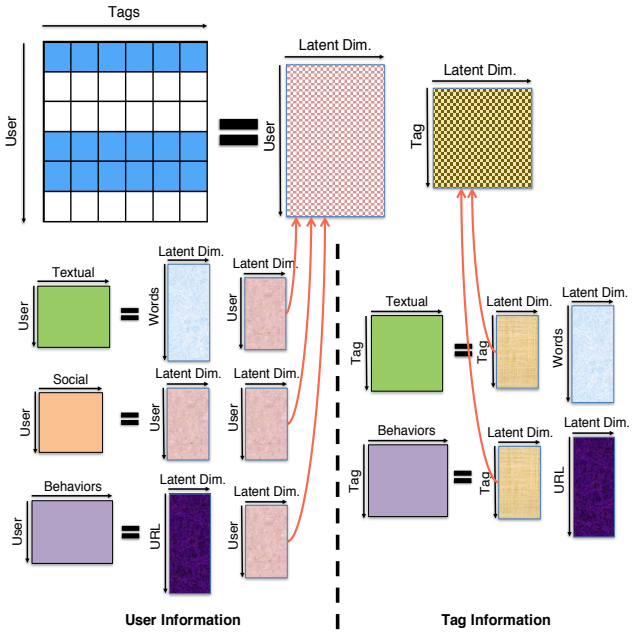


Figure 2: An Overview of the 2-D Model (UTop)

Based upon these derivations, we then apply stochastic gradient descent to iteratively update each variable by taking a step η along its gradient ascending. The algorithm details are presented in **Algorithm 1** in which learning steps η_u, η_v, η_w and η_g are chosen based upon the Goldstein Conditions [12]. We implement the non-negative constraints on $\boldsymbol{U}$ and $\boldsymbol{V}$ through forcing their negative values to 0 in each iteration. As shown, this algorithm considers all three footprints together to estimate the topical profiles for each user.

Algorithm 1: UTop Solver

Input: user-tag matrix $\boldsymbol{X}$, user-word matrix $\boldsymbol{A}$, tag-word matrix $\boldsymbol{B}$, user-url matrix $\boldsymbol{C}$, tag-url matrix $\boldsymbol{D}$, user friendship matrix $\boldsymbol{E}$, observation indication matrix Ω and parameters $\{\lambda, \gamma, \delta, \rho, \eta\}$

Output: $\boldsymbol{U}, \boldsymbol{V}$

Initialize $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$ and $\boldsymbol{G}$ randomly, $t = 0$

while Not Converged **do**

  Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, $\frac{\partial \mathcal{F}}{\partial \boldsymbol{V}}$, $\frac{\partial \mathcal{F}}{\partial \boldsymbol{W}}$ and $\frac{\partial \mathcal{F}}{\partial \boldsymbol{G}}$ in Eq.(3)

  Update $\boldsymbol{U}_{t+1} \leftarrow \max(\boldsymbol{U}_t - \eta_u \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  Update $\boldsymbol{V}_{t+1} \leftarrow \max(\boldsymbol{V}_t - \eta_v \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  Update $\boldsymbol{W}_{t+1} \leftarrow \max(\boldsymbol{W}_t - \eta_w \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  Update $\boldsymbol{G}_{t+1} \leftarrow \max(\boldsymbol{G}_t - \eta_g \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  $t = t + 1$

return $\boldsymbol{U}$ and $\boldsymbol{V}$

Though unifying all three heterogeneous implicit footprints, this initial UTop approach has two main drawbacks. First, it will become complex if we introduce additional footprints, as we bring in more controlling parameters of new footprints to be tuned. In addition,

UTop does not take into account the relations between those heterogeneous footprints which could be jointly explored in the latent space. Given these concerns, can we find a generalized model that can jointly leverage all potential heterogeneous footprints? We turn in the following section to answering this question.

4.3 Learning User Topical Profiles: A Generalized Model

In this section, we augment UTop with a generalized approach toward jointly exploring the relationships across footprints for more robust user topical profile learning. First, to relieve the dramatic increase of parameters when introducing more regularization terms, we need to replace the linear combination model in UTop by a more compact factorization model without manually tuning tradeoff parameters from different new footprints. Second, such a compact factorization model should consider all possible pairwise interactions between footprints to exploit their multi-linear relationships. Therefore, we adopt a *tensor factorization* model which explicitly takes into account the multi-way structure of data. Moreover, the factorization will only happen once even if we introduce additional heterogeneous footprints.

Figure 3 shows an overview of UTop+. In general, we model all implicit footprints in one tensor via calculating the user similarity in each footprint space. There can be many options for measuring the user similarity in every footprint space. We test many of them and report the one providing the best performance in Section 5. Then, we factorize the tensor and obtain a matrix of latent representations for all users, upon which we extract a user similarity matrix to estimate the original user-tag matrix.

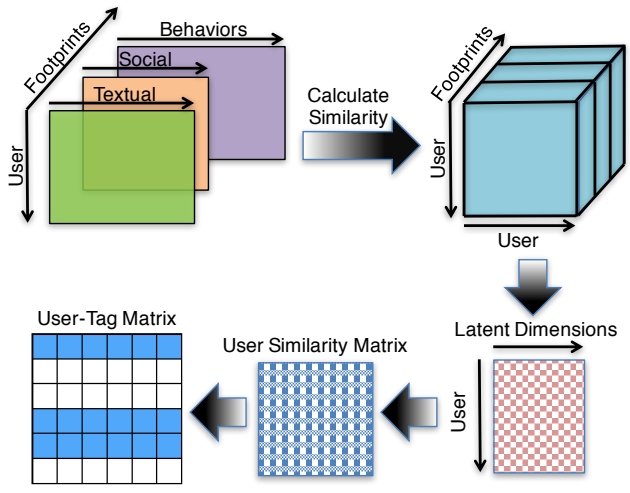


Figure 3: An Overview of the Generalized Model (UTop+)

Concretely, we denote the tensor as $\mathcal{C} \in \mathbb{R}^{N\times N\times R}$ which is a multidimensional array where R is the number of implicit footprints and N is the size of the user set. We can factorize the tensor $\mathcal{C}$ to one latent user matrices $\mathcal{Q} \in \mathbb{R}^{N\times K}$ and one latent context matrix $\mathcal{Y} \in \mathbb{R}^{R\times K}$, where K is the number of latent dimensions. The tensor factorization is to solve the optimization problem defined below:

$$\min_{\mathcal{Q},\mathcal{Y}} \quad \frac{1}{2}\|\mathcal{C} - [\![\mathcal{Q},\mathcal{Q},\mathcal{Y}]\!]\|_F^2 + \frac{\alpha}{2}(\|\mathcal{Q}\|_F^2 + \|\mathcal{Y}\|_F^2), \tag{4}$$

where $[\![\mathcal{Q},\mathcal{Q},\mathcal{Y}]\!] \in \mathbb{R}^{N\times N\times R}$ is given by

$$[\![\mathcal{Q},\mathcal{Q},\mathcal{Y}]\!] = \sum_{k=1}^{K} \boldsymbol{q}_k \circ \boldsymbol{q}_k \circ \boldsymbol{y}_k.$$

Here $\boldsymbol{q}_k$ and $\boldsymbol{y}_k$ are the k^{th} column vectors of $\mathcal{Q}$ and $\mathcal{Y}$, respectively. To solve Eq.(4), we adopt the existing CPOPT method [1] — a fitting approach for the CP (Canonical-decomposition / Parallel-factor-analysis (PARAFAC)) model. The latent footprint matrix $\mathcal{Y}$ represents the contribution of each type of footprint to latent dimensions.

The next natural question is how to leverage the new latent space $\mathcal{Q}$ of all users. The basic idea is that two users tend to have similar topical profiles if they have similar latent representations derived by jointly considering all their implicit footprints. Thus, we first calculate the user similarity matrix denoted as $\boldsymbol{\Psi}$ computed from latent features of users $\mathcal{Q}$ by the cosine similarity. We can see $\mathcal{Q}$ as a "new footprint" and formulate it as the new loss function:

$$\begin{aligned}
\boldsymbol{\Theta} &= \frac{1}{2}\sum_{i,j}\Psi(i,j)\|\boldsymbol{U}_i - \boldsymbol{U}_j\|^2 \\
&= \sum_{i,j}\boldsymbol{U}_i\Psi(i,j)\boldsymbol{U}_i{}^T - \sum_{i,j}\boldsymbol{U}_i\Psi(i,j)\boldsymbol{U}_j{}^T \\
&= \sum_{i}\boldsymbol{U}_i\boldsymbol{D}(i,i)\boldsymbol{U}_i{}^T - \sum_{i,j}\boldsymbol{U}_i\Psi(i,j)\boldsymbol{U}_j{}^T \\
&= \mathrm{tr}(\boldsymbol{U}^T(\boldsymbol{D}-\boldsymbol{\Psi})\boldsymbol{U}) \\
&= \mathrm{tr}(\boldsymbol{U}^T\boldsymbol{\mathcal{L}}\boldsymbol{U}),
\end{aligned} \tag{5}$$

where $\boldsymbol{U}_i$ is the ith row of $\boldsymbol{U}$, $tr(\cdot)$ denotes the matrix trace, and $\boldsymbol{D}$ is a diagonal matrix in which $\boldsymbol{D}(i,i) = \sum_j \boldsymbol{\Psi}(i,j)$, and $\boldsymbol{\mathcal{L}} = \boldsymbol{D} - \boldsymbol{\Psi}$ is the graph Laplacian of the user similarity matrix $\boldsymbol{\Psi}$.

How can we utilize the new implicit footprint $\boldsymbol{\Theta}$ to learn user topical profiles? Similarly, we are able to use $\boldsymbol{\Theta}$ to regulate latent representations of two similar users to make them as close as possible. Hence, we can build the generalized UTop+ by solving the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{U},\boldsymbol{V}} \quad & \frac{1}{2}\|\Omega \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2 + \frac{\beta}{2}\mathrm{tr}(\boldsymbol{U}^T\boldsymbol{\mathcal{L}}\boldsymbol{U}) \\
& + \frac{\alpha}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2), \\
\text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0,
\end{aligned} \tag{6}$$

where β is the controlling parameter. This optimization problem can be solved similarly as introduced in Section 4.2. The detailed solver is presented in **Algorithm 2**.

In summary, we first present a 2-D model for learning user topical profiles (called UTop) in which each of three heterogeneous implicit footprints is modeled as regularization terms. We provide Algorithm 1 to solve the optimization problem in Equation 2. Then we extend UTop to a compact generalized model (called UTop+). Based on a tensor decomposition method, UTop+ can jointly handle relationships across multiple footprints without introducing new

Algorithm 2: UTop+ Solver

Input: user-tag matrix $\boldsymbol{X}$, user-word matrix $\boldsymbol{A}$, user-url matrix $\boldsymbol{C}$, user friendship matrix $\boldsymbol{E}$, observation indication matrix Ω and parameters $\{\alpha, \beta, \eta_u, \eta_v\}$

Output: $\boldsymbol{U}, \boldsymbol{V}$

Calculate the tensor $\boldsymbol{\mathcal{C}}$ from $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{E}$

Calculate $[\boldsymbol{Q}, \boldsymbol{Y}] \leftarrow \text{CPOPT}(\boldsymbol{\mathcal{C}})$

Calculate the user similarity matrix $\boldsymbol{\Psi}$ based on $\boldsymbol{Q}$

Construct the graph Laplacian matrix $\boldsymbol{\mathcal{L}}$ for $\boldsymbol{\Psi}$

Initialize $\boldsymbol{U}$ and $\boldsymbol{V}$, randomly, $t = 0$

while Not Converged **do**

  Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{U}} = -(\Omega \odot \Omega)(\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T))\boldsymbol{V} + \beta \boldsymbol{\mathcal{L}}\boldsymbol{U}$

  Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{V}} = -(\Omega^T \odot \Omega^T)(\boldsymbol{X}^T - \boldsymbol{V}\boldsymbol{U}^T))\boldsymbol{U}$

  Update $\boldsymbol{U}_{t+1} \leftarrow \max(\boldsymbol{U}_t - \eta_u \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  Update $\boldsymbol{V}_{t+1} \leftarrow \max(\boldsymbol{V}_t - \eta_v \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}, 0)$

  $t = t + 1$

return $\boldsymbol{U}$ and $\boldsymbol{V}$

parameters. The complete overview of UTop+ is shown in Figure 3, and we propose Algorithm 2 to solve Equation 6.

5 EXPERIMENTS

In this section, we conduct a series of experiments to answer the following questions: (i) How well do the proposed UTop and UTop+ models work? (ii) Which implicit footprints are most effective? (iii) How does UTop+ compare with other alternatives? Does it really improve upon the simpler UTop approach? (iv) How do the proposed approaches compare to other variants?; and (v) What impact do the model parameters have on the ultimate performance? We begin by introducing the experimental setup including dataset collection and evaluation method.

5.1 Experiment Setup

In this section, we start with describing the data we collect. Next, we introduce the metrics we use for evaluation. Finally, we provide the details of three baselines, and show the parameter settings we choose in our proposed models.

Twitter Lists. We adopt Twitter Lists, a large publicly-accessible collection of crowd-contributed tagging knowledge for social media users. Recall that these lists allow one user to annotate another with a list name (or tag), e.g., politics, music, art. Via the public Twitter API, we randomly sample a set of 3.468 million Twitter users, and crawl the list membership information for each of them. We identify 977,000 users who have ever been included in some list, but we find a huge amount of noise. For instance, nonsense tags (like numbers, unicode characters, single letters) take up a major proportion. Many tags (e.g., "friend", "love", and "amigo") are not reflective of topical profiles. Also, there exist many near-synonyms and variants such as "writer-author" and "news-noticia". To obtain high-quality tags for our problem, we rank all tags by the number of labeled users, and manually curate the top-500 tags through merging variants and filtering noise.

Implicit Footprints. For interest footprints, we aggregate all terms each user has posted and adopt the standard LDA topic model after filtering stopwords and stemming. We further measure user similarity by calculating the pairwise Jensen-Shannon divergence. For social footprints, we crawl the friendship connection information for each user. Following a user can be quite casual on Twitter, so we focus on mutual followings as the basis of user similarity in the social footprint space. For behavioral footprints, we aggregate all URLs a user has posted in her tweets and obtain the posting counts. We resolve all crawled URLs (most are shortened) to take care of URL variants, and focus on the URL domain name which conceptually represents a website. For quantifying similar URL sharing patterns, we test a set of measurements (e.g., intersection, cosine, jaccard) and find the one in [6] works best.

Users. We collect a set of 72,096 users who have all those three types of implicit footprints and have been labeled by at least one of the candidate tags. Since many of them have sparse tagging information, we rank all users by the number of tags they have. We look into the top 50,000 users, and randomly select 10,000 users for training and evaluation.

In our proposed models, we end up with scores of all candidate tags for each user. Since we should take those most associated tags as user topical profiles, we rank them in descending order and focus on the top-k ranked tags. Our evaluation is based on ten-fold cross validation.

Metrics. We pick several metrics which can cover different evaluation aspects. On the one side, we would like to see the ratio of correct inferences for learning user topical profiles. And on the other side, we want to measure the prediction error. Thus, we adopt *precision@k* which measures the percentage of correctly estimated top-k tags, and *Mean Absolute Error (MAE)* which quantifies the prediction quality in terms of errors. Note that a lower MAE means a better performance.

Furthermore, besides the absolute measurement in accuracy, the relative ranking order is another important perspective, especially in some recommendation scenarios. The rank correlation coefficients of both *Kendall's* τ and *Spearman's* ρ are two prevalent metrics for measuring rank-based agreement across two lists. We use them both to measure the number of pairs of tags that are correctly ordered from our results. Their values both range from -1 to 1, with the higher the more relevant.

Baselines. We select three baselines as alternatives to the proposed UTop+ approach. To be fair, we incorporate all three proposed footprints and maintain the same experimental setup for all the following approaches:

- **Nearest Neighborhood (NN).** An intuitive solution is based on the traditional nearest neighborhood model. A user is modeled by a vector extracting from the corresponding row in the context matrix, i.e., $\boldsymbol{A}$, $\boldsymbol{C}$, or $\boldsymbol{E}$. Then, for each target user, we separately find a set of closest seed users in each context, and pick the intersected neighbors from whom we propagate their tags and scores and take the average for each tag.
- **Cross-domain Triadic Factorization (CTF)** [16]. This state-of-the-art method directly combines user ratings of different merchandise (e.g., book, music, movie) into one tensor model,

Table 1: The Impact of Different Implicit Footprints for Learning User Topical Profiles

Method	Precision			MAE			Kendall's τ			Spearman's ρ		
	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15	Top 5	Top 10	Top 15
NN (T)	0.2113	0.2356	0.2673	0.2914	0.2692	0.2432	0.2460	0.1687	0.1531	0.3054	0.2262	0.1784
NN (S)	0.1920	0.2153	0.2330	0.3048	0.2791	0.2642	0.2110	0.1420	0.1289	0.2670	0.1852	0.1682
NN (B)	0.2423	0.2629	0.3155	0.2650	0.2342	0.2110	0.2826	0.2044	0.1834	0.3314	0.2429	0.2106
UTop (T)	0.3438	0.3791	0.4668	0.2264	0.2069	0.1897	0.3221	0.2464	0.2031	0.4163	0.2987	0.2409
UTop (S)	0.3390	0.3837	0.4561	0.2298	0.2093	0.1887	0.3172	0.2421	0.2003	0.4135	0.2916	0.2341
UTop (B)	0.3556	0.3980	0.4733	0.2275	0.1982	0.1699	0.3286	0.2557	0.2067	0.4302	0.3015	0.2426
UTop (T+S)	0.3494	0.3847	0.4657	0.2300	0.2107	0.1872	0.3205	0.2516	0.2085	0.4189	0.2970	0.2378
UTop (T+B)	0.3587	0.4132	0.4758	0.2193	0.1894	0.1909	0.3329	0.2606	0.2197	0.4348	0.3071	0.2535
UTop (S+B)	0.3544	0.4069	0.4729	0.2238	0.1930	0.1852	0.3272	0.2588	0.2185	0.4322	0.3054	0.2561
UTop (T+S+B)	**0.3616**	**0.4189**	**0.4931**	**0.2137**	**0.1861**	**0.1772**	**0.3403**	**0.2746**	**0.2267**	**0.4414**	**0.3104**	**0.2682**

in which all the values are user ratings. Then, it extends the existing PARAFAC2 model [23] that transforms heterogeneous user-rating matrices of different lengths into one cubical tensor and factorizes it. Here in our problem setting, this approach can also be applied on those heterogeneous user-footprint matrices; the subsequent steps follow Equation 5 in order to solve Equation 6.

- **UTop.** Introduced in Section 4.2, this model is a basic version that considers each footprint as a regularization term and linearly adds them together.

Parameter Settings. To determine the number of latent dimensions in both UTop and UTop+, we experiment with a sequence of settings $\{5, 10, 20, 30, 40, 50, 100\}$ and empirically select 20 for both UTop and UTop+, as a trade-off between accuracy and efficiency. In Algorithm 1, there are five parameters λ, γ, δ, α, and η. The first four parameters are used to control the contributions of various footprints. The last one is a step along its gradient ascending. As is commonly done, we iteratively employ cross-validation to tune these parameters. Specifically, we empirically set $\lambda = 0.02$, $\gamma = 0.7$, $\beta = 0.1$, $\alpha = 0.4$ and $\eta = 0.05$ for general experiments, respectively. In UTop+, we choose 10 for the number of latent dimension in tensor factorization. The step size η is set to 0.05. In addition, two positive parameters α and β in Eq. (6) are involved in the experiments. Concretely, we empirically set $\alpha = 0.3$ and $\beta = 0.02$ via cross-validation.

5.2 The Impact of Different Footprints

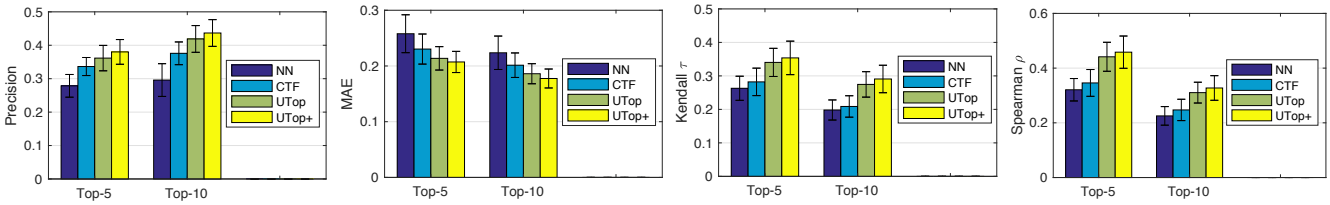
In general, interest, social, and behavioral footprints have different emphases on user topical profiles. Hence, which footprints work better (or best) is one of the most compelling questions to answer. Hence, we compare different combinations of all footprints in both NN and UTop. The reason we do not test them in UTop+ is that the multi-way manner of UTop+ may not clearly tell which footprint contributes more. We show the results in Table 1 in which T is for text-based interest, S is for social, and B is for behavioral.

When individually using each implicit footprint, we find the behavioral footprint (URL sharing) always performs the best in any setting. Moreover, combining it with other footprints always bring the biggest improvement in these experiments. For instance, within the NN method, the behavioral footprint has up to 24% larger Spearman correlation than the social footprint. In UTop, the MAE@10 decreases by 8% when the behavioral footprint is added with the interest footprint. These results indicate the importance of capturing actual user behaviors as a critical step for identifying user topical profiles (in contrast, to relying purely on social connections or on the content of what users post). These results support the intuition that social footprints may capture spurious user similarities (e.g., linking two very different users) and that text-based interest footprints may insert noise into learning user topical profiles. In contrast, behavioral cues provide a clearer perspective on user's interests and expertise.

What if behavioral data is scarce? URL sharing is one of the few publicly-available sources of behavioral information, but sometimes it can still be a scarce resource because not all users will share many URLs on social media. In contrast, social and interest-based footprints are typically more universally available. We see in Table 1 that interest and social footprints can still work well even without access to behavioral footprints. For example, in UTop, the interest footprint is only 5% behind behavioral in precision@10, and the social footprint has just 1% larger MAE@5 than behavioral. These observations show that our model can still achieve a good performance even when we have scarce behavioral evidence. But that together, the three different footprints can complement each other, leading to even better user topical profiles.

5.3 Evaluating UTop and UTop+

Given the evidence of the importance of different footprints, we now turn to evaluating the two proposed models — UTop and UTop+ — versus alternatives. As we can see in Figure 4, both UTop and UTop+ perform better than the Nearest Neighbor (NN) and the Cross-domain Triadic Factorization (CTF) across all four evaluation metrics. For precision@5, UTop+ is 36% and 13% better than NN and CTF with p-values of 0.001 and 0.003 under McNemar's test, respectively. For MAE@10, UTop+ outperforms NN by 20% with the p-value of 0.002 and CTF by 11.8% with the p-value of 0.001. The gaps become even larger for the two ranking correlation coefficients, as we can see in Figure 4c and 4d. These results suggest that the proposed learning models can better leverage all footprints together than either the neighborhood-based propagation or the immediate tensor decomposition. Note that the CTF method is fundamentally different from our problem setting where we cannot simply put together all heterogeneous footprints. In contrast, we exploit latent factors to build a user similar matrix and find its graph Laplacian as



(a) Precision **(b) MAE** **(c) Kendall's τ** **(d) Spearman's ρ**

Figure 4: Comparisons Between Proposed Models and Alternative Baselines

a new regularization term. We show the effectiveness of this step in Section 5.4.

Recall that we introduced UTop+ as an extension to UTop to provide a more compact factorization and to jointly handle relationships across multiple implicit footprints. In Figure 4 we find UTop+ surpasses UTop in all settings. UTop+ has an improvement of 4.2% in precision@10, 3% in MAE@5, 5.9% in Kendall correlation@10, and 3.8% in Spearman correlation@5. These findings indicate that the proposed UTop+ can better exploit the joint correlations between all heterogeneous footprints for improved learning of user topical profiles. All these finding are conducted under McNemar's test with p-values less than 0.01.

5.4 Considering Other Variants

Why We Need Regularization? A natural question is why we need a regularization model. Why not just put all footprints into one large matrix and directly apply state-of-the-art matrix factorization methods? To investigate this question, we put them into one matrix upon which we adopt the standard factorization technique, where we denote such a method MF. We do normalization for the data of each footprint since their values can have distinct scales. We follow the same evaluation methodology and show the comparisons in Figure 5. All results are measured at the top-10. We clearly see the proposed UTop results in better performances than MF in every metric. These results suggest that heterogeneous footprints require careful integration, and that the proposed UTop approach is a good solution in comparison.

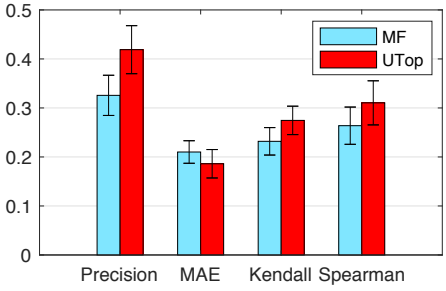


Figure 5: Comparisons Between UTop and Standard MF

Why We Do Regularization After Tensor Factorization? In UTop+, after having the latent factors of users from tensor factorization, we build a user similar matrix and find its graph Laplacian as the new regularization term. Why not just directly replace the user's latent matrix $\boldsymbol{U}$ in $\boldsymbol{X}$ after factorizing the tensor? We call such a scheme Tensor Factorization-based Matrix Factorization (TFMF), and we show the comparison results in Figure 6 for all

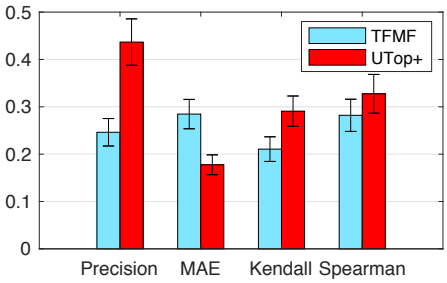
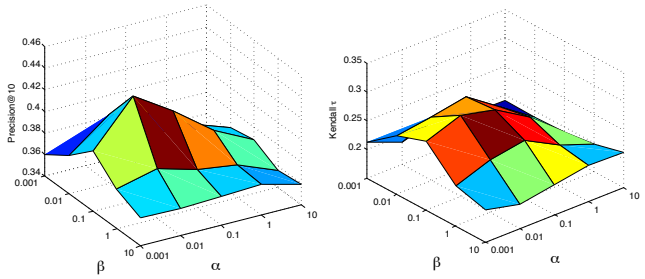


Figure 6: Comparisons between UTop+ and TFMF

metrics at top 10. Our UTop+ outperforms TFMF in all settings (e.g., 68% precison, 38% MAE, 45% Kendall correlation). These outcomes show that regularization after tensor factorization can significantly improve the performance.



(a) Precision@10 **(b) Kendall's τ**

Figure 7: Impact of α and β on UTop+

Impact of Parameters. Finally, two critical parameters in UTop+ are α and β. Recall that α is used to avoid overfitting; β is to control the contribution of the user similarity derived from three types of footprints. In order to better understand the impacts of these two parameters, we evaluate the performance of UTop+ across various parameter settings. We vary values of these parameters in [0.001 0.01, 0.1, 1, 10] and present the results of precision and Kendall's τ in Figure 7 for learning the top-10 tags. As we can see, UTop+ achieves relatively consistent performance across a wide range. Particularly, we find the setting $\alpha = 0.1$ and $\beta = 0.01$ gives the best performance. These results indicate the stability of UTop+ to these parameters.

6 CONCLUSION

Mining user's topical profiles (e.g., user interests and expertise) has important applications in diverse domains such as personalized search and recommendation, as well as expert detection. In this

paper, we tackled the problem of learning user topical profiles. In particular, we investigated how to leverage user-generated information in heterogeneous and diverse footprints. Concretely, we proposed UTop+ — a generalized model that integrates multiple implicit footprints with explicit footprints for learning high-quality user topical profiles. By taking into account pairwise relations among multiple footprints, the proposed UTop+ intelligently combines the potential benefits of each footprint to find the best evidence across footprints for learning high-quality user profiles. And indeed, extensive experiments demonstrate the effectiveness of UTop+. For instance, it surpasses other alternatives up to 36% in precision@5 and 20% in MAE@10. URL sharing, as one type of publicly-accessible user behavior, brings better results than other implicit footprints in every evaluation setting. Moreover, compared with other variants in terms of modeling, our model also has the best performances, e.g., up to 68% for precision@10 and Kendall correlation@10.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants IIS-1149383 and IIS-1657196. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

REFERENCES

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, February 2011.
[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *JMLR*, 2014.
[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.
[4] P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly, and K. P. Gummadi. Deep twitter diving: Exploring topical groups in microblogs at scale. In *CSCW*, 2014.
[5] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial life*, 2011.
[6] C. Cao, J. Caverlee, K. Lee, H. Ge, and J. Chung. Organic or organized?: Exploring url sharing behavior. In *CIKM*, 2015.
[7] D. Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 2011.
[8] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIMAX*, 2000.
[9] W. Feng and J. Wang. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *SIGKDD*, 2012.
[10] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *SIGCOMM*, 2010.
[11] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, 2012.
[12] A. A. Goldstein. *Constructive real analysis*. Courier Corporation, 2013.
[13] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel. Social media recommendation based on people and tags. In *SIGIR*, 2010.
[14] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys*, 2010.
[15] L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *WSDM*, 2013.
[16] L. Hu, J. Cao, G. Xu, L. Cao, Z. Gu, and C. Zhu. Personalized recommendation via cross-domain triadic factorization. In *WWW*, 2013.
[17] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *WWW*, 2013.
[18] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, 2013.
[19] Y. Hu, K. Talamadupula, S. Kambhampati, et al. Dude, srsly?: The surprisingly formal nature of twitter's language. In *ICWSM*, 2013.
[20] M. Jamali and L. Lakshmanan. Heteromf: recommendation in heterogeneous information networks using context dependent factor models. In *WWW*, 2013.
[21] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang. Social contextual recommendation. In *CIKM*, 2012.
[22] M. Jiang, P. Cui, F. Wang, X. Xu, W. Zhu, and S. Yang. Fema: Flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery. In *SIGKDD*, 2014.
[23] H. A. Kiers, J. M. Ten Berge, and R. Bro. Parafac2-part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13(3-4):275–294, 1999.
[24] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 2009.
[25] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR*, 2009.
[26] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.
[27] T. Lappas, K. Punera, and T. Sarlos. Mining tags using social endorsement networks. In *SIGIR*, 2011.
[28] Q. Liu, E. Chen, H. Xiong, C. H. Ding, and J. Chen. Enhancing collaborative filtering by user interest expansion via personalized ranking. *IEEE SMC*, 2012.
[29] X. Liu and K. Aberer. Soco: a social network aided context-aware recommender system. In *WWW*, 2013.
[30] H. Lu, J. Caverlee, and W. Niu. Discovering what you're known for: A contextual poisson factorization approach. In *RecSys*, 2016.
[31] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW*, 2010.
[32] H. Ma. On measuring social friend interest similarities in recommender systems. In *SIGIR*, 2014.
[33] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM*, 2011.
[34] A. Majumder and N. Shrivastava. Know your personalization: learning topic level personalization in online services. In *WWW*, 2013.
[35] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001.
[36] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, 2010.
[37] R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets: Investigating how users behave across image- and text-based social networks. *ICWSM*, 2014.
[38] M. Qiu, F. Zhu, and J. Jiang. It is not just what we say, but how we say them: Lda-based behavior-topic model. In *SDM*, 2013.
[39] V. Rakesh, D. Singh, B. Vinzamuri, and C. K. Reddy. Personalized recommendation of twitter lists using content and network information. In *AAAI*, 2014.
[40] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *SIGKDD*, 2009.
[41] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, 2010.
[42] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *CIKM*, 2007.
[43] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *SIGKDD*, 2008.
[44] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*, 2010.
[45] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Tag recommendations based on tensor dimensionality reduction. In *RecSys*, 2008.
[46] J. Tang, Y. Chang, and H. Liu. Mining social media with social theories: a survey. *SIGKDD Explorations Newsletter*, 2014.
[47] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *SIGKDD*, 2015.
[48] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.
[49] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, 2010.
[50] D. Yin, Z. Xue, L. Hong, and B. D. Davison. A probabilistic model for personalized tag prediction. In *SIGKDD*, 2010.
[51] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang. A temporal context-aware model for user behavior modeling in social media systems. In *SIGMOD*, 2014.
[52] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han. Personalized entity recommendation: A heterogeneous information network approach. In *WSDM*, 2014.
[53] X. Zhang, J. Cheng, T. Yuan, B. Niu, and H. Lu. Toprec: domain-specific recommendation through community topic mining in social network. In *WWW*, 2013.
[54] Z. Zhao, Z. Cheng, L. Hong, and E. H. Chi. Improving user topic interest profiles by behavior factorization. In *WWW*, 2015.
[55] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *TIST*, 2014.
[56] E. Zhong, N. Liu, Y. Shi, and S. Rajan. Building discriminative user profiles for large-scale content recommendation. In *SIGKDD*, 2015.