# Recommending Tumblr Blogs to Follow with Inductive Matrix Completion

Donghyuk Shin*
Dept. of Computer Science
University of Texas at Austin
Austin, TX, 78721
dshin@cs.utexas.edu

Suleyman Cetintas
Yahoo Labs
Sunnyvale, CA, 94089
cetintas@yahoo-inc.com

Kuang-Chih Lee
Yahoo Labs
Sunnyvale, CA, 94089
kclee@yahoo-inc.com

## ABSTRACT

In microblogging sites, recommending blogs (users) to follow is one of the core tasks for enhancing user experience. In this paper, we propose a novel inductive matrix completion based blog recommendation method to effectively utilize multiple rich sources of evidence such as the social network and the content as well as the activity data from users and blogs. Experiments on a large-scale real-world dataset from Tumblr show the effectiveness of the proposed blog recommendation method.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Search and Retrieval**]: Information Filtering

## Keywords

Blog Recommendation, Inductive Matrix Completion, SVD

## 1. INTRODUCTION

Tumblr[1] is one of the most popular microblogging services where users can create and share posts with the followers of their blogs. Similar to Twitter[2] and different from Facebook[3], connections in Tumblr are unidirectional. Unlike Twitter, users can create longer, richer and higher quality content in the form of several post types such as text, photo, quote, link, chat, audio, and video [4]. Tumblr also supports liking and reblogging a post as well as attaching tags to it.

One of the core problems in microblogging sites is predicting whether a user will follow a blog or not. In addition to the user-item interactions (user-item matrix), vast majority of the existing work either used the social network information or the textual content/profiles of users and items, and did not effectively utilize both types of information. A recent comprehensive survey of the state-of-the-art methods can be found in [10]. In the case of Tumblr, in addition to the social network information, rich user and blog features can be extracted from high quality posts as well as like and reblog graphs from user activities.

In this paper, we propose a novel inductive matrix completion (IMC) based blog recommendation system that effectively utilizes the social network as well as the rich content and activity (network) data from users and blogs in a unified framework. Although, IMC has been shown to be highly effective in other domains such as bioinformatics (e.g., the gene-disease association problem), it has not been utilized for recommendation tasks [9]. However, IMC has several advantages over the standard low rank matrix completion approaches. Specifically, it overcomes extreme sparsity in the data by incorporating diverse features of users and blogs obtained through various sources. Furthermore, it is capable of making new recommendations even for users or blogs with very few or unknown following information. Experiments on real-world proprietary data from Tumblr show that IMC significantly outperforms standard methods for the blog recommendation task.

## 2. METHOD

Formally, let $R \in \mathbb{R}^{m \times n}$ be the user-blog follower matrix, where each row corresponds to a user and each column corresponds to a blog, such that $R_{ij} = 1$, if user $i$ is following blog $j$ and 0 otherwise. The low rank matrix completion approach is one of the most popular and successful collaborative filtering methods for recommender systems [7]. The goal is to recover the underlying low rank matrix by using the observed entries of $R$, which is typically formulated as follows:

$$\min_{W,H} \sum_{(i,j)\in\Omega} (R_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2),$$

where $W \in \mathbb{R}^{m \times r}$ and $H \in \mathbb{R}^{n \times r}$ with $r$ being the dimension of the latent feature space; $\Omega$ is the set of observed entries; $\lambda$ is a regularization parameter.

However, the standard formulation is restricted to the transductive setting, i.e., predictions can only be made to existing users and items, and suffers performance with extreme sparsity in the data. Recently, a novel inductive matrix completion (IMC) approach was proposed by [6] that incorporates side information of users and items given in the form of feature vectors alleviating data sparsity issues and enabling predictions for new users and items. Mathe-

*Work was done when the first author was on a summer internship with Yahoo Labs.
[1]www.tumblr.com
[2]www.twitter.com
[3]www.facebook.com


*RecSys 2014* Poster Proceedings, October 6-10, 2014, Foster City, Silicon Valley, USA.

matically, IMC is formulated as follows:

$$\min_{W,H} \sum_{(i,j)\in\Omega} (R_{ij} - \mathbf{x}_i^T W H^T \mathbf{y}_j)^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2),$$

where $\mathbf{x}_i \in \mathbb{R}^{f_u}$ and $\mathbf{y}_j \in \mathbb{R}^{f_b}$ are feature vectors for user $i$ and item $j$, respectively (i.e., $W \in \mathbb{R}^{f_u \times r}$ and $H \in \mathbb{R}^{f_b \times r}$). Given a new blog $\bar{j}$, the predictions $R_{i\bar{j}}$ for each user $i$ can be calculated with the feature vector $\mathbf{y}_{\bar{j}}$ available. Note that the number of parameters to learn is $(f_u + f_b) \times r$, which depends only on the number of user and item features, whereas there is $(m + n) \times r$ parameters in the standard matrix completion.

# 3. EXPERIMENTS

We evaluated IMC for blog recommendation, where additional side information of both users and blogs are available. We compare IMC against the standard matrix completion formulation (MC) as well as the Singular Value Decomposition (SVD), which has been shown to perform well for top-$N$ recommendation tasks [5]. As a baseline, we also report results of using a simple global popularity ranking (Global) for recommendation, where blogs are ranked by the number of followers. We randomly sampled 1 million users and 100 thousand blogs from the Tumblr follower graph resulting in 12.5 million follows, i.e., nonzero elements in $R$. Note that we retain users and blogs with at least 5 followees and followers, respectively, and use 10-fold cross-validation for evaluation.

For additional features, we use like, reblog, and tags information collected over a period of 1 month from Tumblr. Both like and reblog activities can be represented as a graph similar to the follower graph $R$. One way to obtain useful and robust features is to consider the principal components of the adjacency matrix corresponding to the like and reblog graphs. That is, we compute $p$ principal components and use them as latent user and blog features for IMC. For the tags used in the posts of each blog, we first compute vector representations of each tag using the continuous skip-gram model [8], which in turn are used to cluster the tags into $c$ clusters by the $k$-means algorithm. Using the cluster information, we create a histogram of the tag's cluster for each blog as a compact representation of tags used in that blog. Thus, we have $f_u = p$ features for users and $f_b = p + c$ features for blogs, where we use $p = 500$ and $c = 1000$ in the experiments. We use rank $r = 100$ for SVD and MC, rank $r = 10$ for IMC, and set $\lambda = 0.1$, which are determined using cross-validation.

We measure the recommendation performance using the $F_1$ score for the top-20 recommendations generated by each method, which is the region of partical interest for recommender systems. We also report the AUC (area under the curve) measured from the precision-vs-recall plot. Results of the proposed IMC method are shown in comparison to the baselines, Global, SVD, and MC for both metrics in Table 1. Note that we present normalized (relative) results in both metrics using the MC method as the baseline.

It is very interesting to see in Table 1 that the simple Global method ourperforms both SVD and MC baselines. This can be explained by the facts that most users follow highly popular blogs such as institutions or celebrities [4] and that both SVD and MC suffer significantly from data sparsity. Table 1 also shows that IMC is the best performing method out of all methods. This set of results explicitly shows that IMC succesfully handles data sparsity by incorporating the rich user and blog data, and significantly improves over MC as well as other methods.

**Table 1: (Normalized) Results of the proposed Inductive Matrix Completion method (i.e., IMC) in comparison to several baselines.**

| Method | $F_1$@20 | AUC |
|---|---|---|
| Global | 1.3825 | 1.2313 |
| SVD | 1.0955 | 1.0612 |
| MC | 1.0000 | 1.0000 |
| IMC | 1.4443 | 1.2653 |

# 4. CONCLUSIONS

Recommending blogs (users) to follow is one of the core tasks for enhancing user experience in online microblogging sites such as Tumblr. In addition to the social network information, it is very important to effectively utilize the rich user and blog content (e.g., tags) as well as users' activities such as like and reblog. This paper proposes a novel inductive matrix completion based blog recommendation method that effectively utilizes the social network as well as rich content and activity data from users and blogs. Experiments on large-scale real-world data from Tumblr show the effectiveness of the proposed blog recommendation method.

Future work will mainly be conducted in i) utilizing additional information from users and blogs such as the rich visual features from posts in Tumblr as well as the (sparser) textual features, and ii) using probabilistic latent-class or mixed-membership approaches as shown in [2, 1, 3].

# 5. REFERENCES


[1] S. Cetintas, D. Chen, and L. Si. Forecasting user visits in online display advertising. *Journal of Information Retrieval*, 16:369–390, 2013.

[2] S. Cetintas, M. Rogati, L. Si, and Y. Fang. Identifying similar people in professional social networks with discriminative probabilistic models. In *SIGIR*, pages 1209–1210, 2011.

[3] S. Cetintas, L. Si, Y. P. Xin, and R. Tzur. Probabilistic latent class models for predicting student performance. In *CIKM*, pages 1513–1516, 2013.

[4] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu. What is tumblr: A statistical overview and comparison. *CoRR*, abs/1403.5206, 2014.

[5] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-N recommendation tasks. In *RecSys*, pages 39–46, 2010.

[6] P. Jain and I. S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

[7] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[9] N. Natarajan and I. S. Dhillon. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

[10] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, 2014.