# A New Hybrid Popular Model for Personalized Tag Recommendation

Jing Wang*, Nianlong Luo

Information Technology Center, Tsinghua University, Beijing, 100084, China.

* Corresponding author. Tel.: +86 152-1056-0386; email: wj152444521@163.com

Manuscript submitted March 14, 2015; accepted July 3, 2015.
doi: 10.17706/jcp.11.2.116-123

**Abstract:** Tagging systems have been playing an important role in many websites during the web2.0 age. Users use a social tagging system to effectively mark web resources with personalized tags, which can help them organize and share their items easily. In this work, we put forward a new model IBHP to recommend personalized tags for users. We evaluate our model on a real-world dataset collected on Delicious[1]. Data tests show that our model can get better performance than currently widely used popularity-based methods, which also use the same available information: <user, item, tag> ternary relations.

**Key words:** Hybrid popular, recommender, tagging system, collaborative filtering.

## 1. Introduction

Nowadays, there are a great amount of resources on the Internet. Users are easily lost in their finding what they really want. From the web 2.0, there are more and more social websites where users can share and manage resources online, users are simultaneously consumers as well as producers of resources. It's difficult for users to seek the resources effectively. Therefore, recommender systems are applied in social websites [1]. In order to manage and share the resources conveniently, tags are introduced to annotate resources. Tag is a kind of label which is able to present the content or related information of resources. Social tagging systems took a great step in the web2.0 that transferred taxonomic hierarchies of internet resources from specialists to common users. It allowed users to freely provide metadata to describe the content of the resource on the internet. In this way, users can easily organize the resources and share with their friends on social media sites. For example, users can annotate and share their online bookmarks on Delicious, images on Flickr[2], clips on Youtube[3], publications on CiteULike[4].

Personalized tag recommendation is the main part in a tagging system. Because users add tags totally freely, maybe many bad tags are used. For example, users may use misspelled words as tags, or the tags are not able to describe the content of the resource. Bischoff [2] found the fact that more than 60% tags are used by a single user. These tags make noises and will reduce the efficiency of the system. As a result, tag recommenders[3] are used in systems to alleviate the problem. The tag recommender predicts and provides users a set of tags that they are most likely to use. Every tagging behavior generates a ternary relation <Users (*U*), Items (*I*), Tags (*T*)> [4]. When Tag *t* is recommended for User *u* to annotate Item *i*, *t*

[1] www.delicious.com
[2] www.flickr.com
[3] www.youtube.com
[4] www.citeulike.org

serves two purposes for $u$ and $i$ [5]. 1. $t$ has highly relevant relation with $i$, which means $t$ is able to represent main content of $i$. 2. $t$ is highly related to $u$, because different people have different ways of tagging. Recommended tags need to meet the requirements of presenting item content and satisfying user preference. Tag recommendation is a process of tag prediction. As shown in Fig. 1, what we should do is to predict which tags user1 will use by using the information of tagging history.

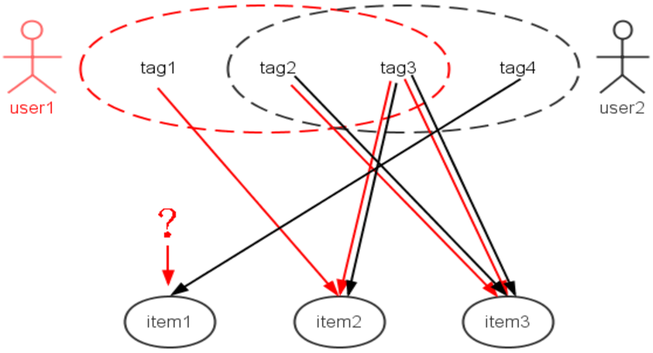


Fig. 1. Tag prediction.

In the conventional recommendation systems, items are recommended for users. Many methods being successfully applied in the conventional recommender are introduced into tag recommendation, like collaborative filtering ([6]-[8]), association-rules ([9], [10]), latent semantic analysis, topic models [11], [12]). So, we can take the ideas in these methods into tag recommendation models for improvement.

In this paper, our contributions are:

- We present a new model for personalized tag recommendation.
- We conduct experiments on real-world dataset and demonstrate that our model outperforms other popularity-based models.

The remainder of this paper is structured as follows: In Section 2, we review and discuss the related work, followed by defining the problem in Section 3. We completely describe and formulate our IBHP model in Section 4. Section 5 reports result and analysis of data tests. Section 6 comes to conclusion and outlines future work.

## 2. Related Works

The personalized tag recommendation system is a hotspot in recommender system study over the years. Huberman *et al*. [13] and Marlow *et al.* [14] gave details of different types of systems.

From the information that tag recommenders use, tag recommendation methods usually fall into two categories: content-based approaches and graph-based approaches.

Content-based methods, which usually extract information from context of items (e.g., web pages, anchor text, academic papers or other textual items) to build a user model or an item model, can predict tags even for cold starting. Dawei Yin [15] proposed a probabilistic model that represents an item by a bag of words on the word-independence assumption. In this model, tag distribution is dependent on the words distribution. The work by Yuta Liu [16] proposed a content-based approach motivated by the observation that similar items tend to get same tags. Xiance Si [17] described a fast tag suggestion method named Feature-Driven Tagging which indexes tags by features. The feature can be a word, an id or other context information. Although content-based approaches outperform non-content-based approaches in most cases because of the formers' utilization of more information, they can't be applied to situations that items are nonstructural resources like movies, songs, etc.

On the other hand, graph-based methods, which are concentrated on the relations between users, items and tags, mostly have lower computation complexity without parsing item content. Adriana [18] identified a core principle underlying tag propagation and presented a new model to generate tags along the edges of graph which relates similar documents. This model that takes items as nodes, and item relations as edges is available for directed graph (edge for webpage link, academic reference, etc.). FolkRank, an adaption of PageRank which is effective and has been applied in search engines successfully, was introduced by Hotho [19]. FolkRank is shown to generate better recommendation than the model bases on collaborative filtering model. Wei Feng [20] proposed an optimization framework called OptRank. OptRank incorporates heterogeneous information to represent edges and nodes with features. OptRank can get better results than methods in which only <user, tag, item> relation is available, but the training is very time-consuming for overmany parameters. Krestel [21] introduced an approach based on Latent Dirichlet Allocation on the hypothesis that item content consists of certain topics. Each topic fits certain tag distribution, tags can be predicted according to latent topics based on posterior probability. Method Higher Order Singular Value Decomposition (HOSVD) is based on the Tucker Decomposition (TD) in [22]. Steffen [23] presented a special TD model called pairwise interaction tensor factorization (PITF) with linear runtime for learning and prediction. On real world dataset, their model outperforms TD largely in runtime and gets better prediction as well. Since methods based on Factorization or Randwalk generally have higher complexity, they can hardly apply to real systems.

In the fields of item recommendation like e-commerce or movie, music or photo websites, graph methods based on collaborative filtering have been widely used successfully. [22] showed the item based collaborative filtering can quickly produce high quality recommendations even for large scale problems.

There are some methods considering problems from other perspectives. Jin Yanan [24] proposed an orientation of motivation discrimination model (OMDM), they used five indices to measure the motivation of users and items. [25] takes location information into consideration to recommend tags for photos. Temporal information is also introduced to build some models ([26], [27]). Song Yang [28] proposed a multi-class sparse Gaussian process classification framework (SGPS) which is able to classify data with very few training instances.

## 3. Definition

When users are to post tags for items, the recommendation system will recommend some tags users want to use. The system uses users' behavior history and item context for the task. Here, we only use relations between users, items, tags. Methods using only the relations can be applicable no matter what kind of items is.

To formalize tag recommendation, we use the following notations. Let $U$ be the set of all users, $I$ be the set of items, and $T$ be the set of all tags. We annotate all history tagging behaviors with $S \subseteq U \times I \times T$. So, an element $(u, i, t) \in S$ means user $u$ annotated item $i$ with tag $t$. For tag recommender, the task is to give a specific pair $(u, i)$ a list of tags which the user is mostly like to use to annotate the item. Here, we define distinct user-item pairs $Ps$:

$$P_s = \{(u, i)\exists t \in T, (u, i, t) \in s\} \tag{1}$$

All methods presented in this paper, give a scoring function: $Y: U \times I \times T \mapsto \mathbb{R}$ to derive tag ranking list. The Top-$N$ tags list is:

$$\mathrm{Top}(u, i, N) = \underset{t \in \mathrm{T}}{\overset{N}{\mathrm{argmax}}} \mathrm{y}_{u,i,t} \tag{2}$$

Here, $N$ is the number of tags in the ranking list.

## 4. Tag Recommendation

Many methods based on popularity are widely used in real systems for lower computation complexity than Factorization models [23] or RandWalk models [20]. In the following, we will describe three common popularity models [29] and our IBHP model.

### 4.1. Item Popular (IP) model

For a pair (*u*, *i*), the item popular method is to give the top frequency tags of *i* to *u*. Here, tags of *i* come from all other users who have tagged *i*. To score *t* for the pair (*u*, *i*), we compute the scores as:

$$y_{u,i,t} = \sum_{(u',i,t)\in Ps} 1 \tag{3}$$

$$IT_{u,i} = (y_{u,i,t1}, y_{u,i,t2}, ...) \tag{4}$$

The IP method is easy and can show the item content and social effect directly. However, it gives the same tag list to everyone who is annotating the item. Like other non-personalized tag recommenders (e.g. [24], [25]), user preference is missing in the model.

### 4.2. User Popular (UP) Model

For a pair (*u*, *i*), user popular method is to give the top frequency tags of u to *u*. Here, tags of *u* come from all tags that have been used by *u*. To score t for the pair (*u*, *i*), we compute the scores as follows:

$$y_{u,i,t} = \sum_{(u,i',t)\in Ps} 1 \tag{5}$$

$$UT_{u,i} = (y_{u,i,t1}, y_{u,i,t2}, ...) \tag{6}$$

UP method is also easy, but it can neither give tags for new users nor give users new tags.

### 4.3. Hybrid Popular (HP) Model

HP model, the combination of UP model and IP model, takes both item content and user preference into consideration. HP model is formalized as:

$$HT_{u,i} = \alpha \times I\hat{T}_{u,i} + (1-\alpha) \times U\hat{T}_{u,i} \tag{7}$$

$$I\hat{T}_{u,i} = \frac{IT_{u,i}}{\max_{y_{u,i,t}\in IT_{u,i}} (y_{u,i,t})} \tag{8}$$

$$U\hat{T}_{u,i} = \frac{UT_{u,i}}{\max_{y_{u,i,t}\in UT_{u,i}} (y_{u,i,t})} \tag{9}$$

Here, $\widehat{IT}_{u,i}$ and $\widehat{UT}_{u,i}$ are normalized according to the corresponding maximum value so that we can control the influence of the two factors properly. Neither the influence of user preference will decrease for items being too hot, nor the influence of item content will decrease for user being too active.

### 4.4. Item Based Hybrid Popular (IBHP) Model

As we mentioned above, tags need to be related to item content and user preference. HP meets the tagging requirements by combining UP and IP directly. However, HP model has the same influence of UP (that means tag weights of user preference) as the user is tagging different items, but this is not proper. For example, after user u has annotated many books about computer science and few books about literature, his normalized user preference may be $\widehat{UT}_{u,i}$<DM (1), ML (0.8), network (0.4), poem (0.2), Shakespeare (0.1)>. Empirically, if then u is tagging a literary book, u will tend to use those (like poem, Shakespeare ) about literature instead of those (like DM, ML, network) about computer science even though tags about computer science are of greater weight. So, through analyzing the user preference part of HP model, we think if the user preference $\widehat{UT}_{u,i}$ enhances the weight of the used tags annotating similar items and decreases the weight of the used tags annotating dissimilar items, the user preference will be more reasonable.

In our IBHP model, our user preference varies according to the target item. Our method is described as follows:

1) Build normalized tag weights vector for each item.

$$T_{\text{item}_i} = (t_{i1}, t_{i2}, t_{i3}, \ldots) \Big/ \sum_j t_{ij} \tag{10}$$

Here, $t_{ij}$ is the frequency that tag *j* is used to annotate item *i* by all users.

2) Calculate similarity between two items. $T_{item_i}$ is tag probability distribution of i. Here, we take the ratio of the overlapping area to the total area of two items' tag probability distribution graph in probability histogram.

$$\text{sim}(\text{item}_i, \text{item}_j) = (\sum_k \min(t_{ik}, t_{jk})) \Big/ \sum_k \max(t_{ik}, t_{jk}) \tag{11}$$

3) Our user preference based on items for this model is:

$$UT_{u,i} = \sum_{j \in I(u)} T(u, j) \quad \text{sim}(\text{item}_i, \text{item}_j) \tag{12}$$

Here, $I(u)$ is item set annotated by $u$, $T(u, j)$ is the tag distribution which was used to annotate *j* by *u*.

4) Finally, we derive our IBHP model:

$$IBHT_{u,i} = \alpha \times \hat{IT}_{u,i} + (1 - \alpha) \times \hat{UT}_{u,i} \tag{13}$$

Here $\widehat{UT}_{u,i}$ the normalize user preference, is calculated according to (12) and (9). $\widehat{IT}_{u,i}$ is calculated by (8).

## 5. Experiment

In our experiment, we test our model on publicly available bookmark datasets on Delicious. In the raw data, there are 437590 annotating records of 1867 users, 69223 items and 40897 tags. The raw data is too sparse so that we select part of them as our dataset. First, we select top 400 tags in terms of annotating frequency as our tag set, and then, select top 500 items in terms of annotating frequency as our item set. We select 10960 annotating records each of whose tag is in the tag set and item is in item set at the same time. Finally, there are 242 distinct users in the record set.

To use 5-fold cross validation, we randomly split the record set into meaningful training and test sets ($S_{\text{train}}$ 90%, $S_{\text{test}}$ 10%) five times. We use common precision and recall to measure the performance.

$\alpha$ in equation (7) and (13) controls the ratio of item content to user preference. We evaluate how precision changes when alpha is assigned from 0.2 to 0.9 in model HP and IBHP. From Fig. 2, when alpha is between 0.75 and 0.85, the precision seems the highest. Finally, we set α to be 0.8.

We compare results of IBHP with other three models. In Fig. 3, results of IP, UP, HP and IBHP on our experimental data are shown. Generally, IBHP results in the best performance.

First, the precision and recall of UP is much lower than the other three, for a user can be of wide interests, but some tags he often used may be totally irrelevant to the target item. Next, HP and IBHP outperform IP apparently. So, considering user preference into the model can indeed improve result.

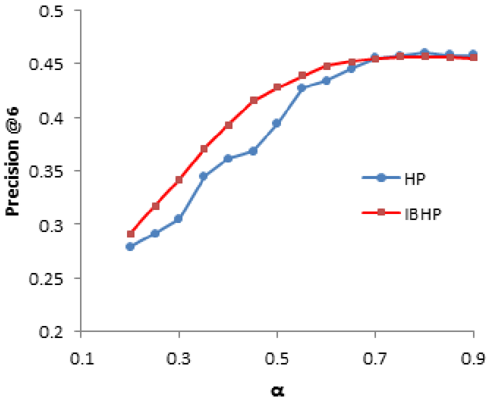


Fig. 2. $\alpha$ test.

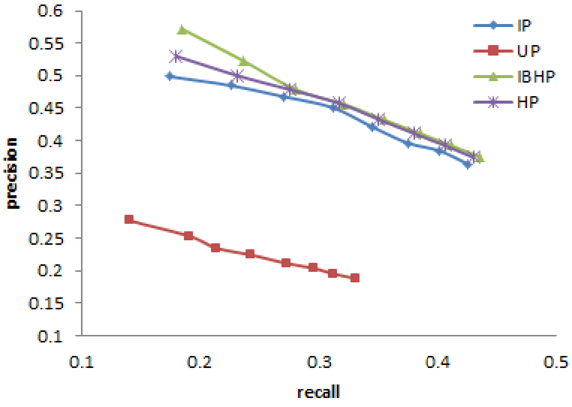


Fig. 3. Performance of 4 models.

Finally, in Fig. 2 and Table 1, comparing the performances of HP and IBHP, we can see that when *N* is small (like 3 or 4), IBHP shows much better precision and recall. As *N* increases, the advantage of IBHP is getting weaker.

Table 1. Precision @ *N*

| N | UP | IP | HP | IBHP |
|---|---|---|---|---|
| 3 | 0.276752636 | 0.498579993 | 0.530143 | 0.571795 |
| 4 | 0.253163937 | 0.484700065 | 0.500376 | 0.523231 |
| 5 | 0.234807031 | 0.467626901 | 0.478647 | 0.479162 |
| 6 | 0.224363539 | 0.450874433 | 0.457791 | 0.457607 |
| 7 | 0.211177477 | 0.421564581 | 0.433334 | 0.433561 |
| 8 | 0.203449766 | 0.396343831 | 0.411263 | 0.412420 |

## 6. Conclusion and Future Work

In this paper, we proposed a new model IBHP for personalized tag recommending, that combines item-based collaborative filtering and hybrid popular method. User preference in our model can vary properly with different target item. Finally, we empirically showed that our model outperforms some other popularity-based models.

In future work, one can analyze the effects of neighboring users. We believe a user tend to use tags which his similar users have used. Adding this factor to the current IBHP model may make further improvements.

## References

[1] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109-132.

[2] Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 193-202).

[3] Belém, F. M., Martins, E. F., Almeida, J. M., & Gonçalves, M. A. (2014). Personalized and object-centered tag recommendation methods for Web 2.0 applications. *Information Processing & Management*, *50(4)*, 524-553.

[4] Heymann, P., Ramage, D., & Garcia-Molina, H. (2008). Social tag prediction. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 531-538).

[5] Qiao, L., Y., *et al*. Review of tag recommendation method based on folksonomy in China. *Journal of Information Resources Management*, *4*, 41-46.

[6] Lu, C., Hu, X., Park, J. R., & Huang, J. (2011). Post-based collaborative filtering for personalized tag recommendation. *Proceedings of the 2011 iConference* (pp. 561-568).

[7] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web* (pp. 285-295).

[8] Ungar, L. H., & Foster, D. P. (1998). Clustering methods for collaborative filtering. *Proceedings of AAAI Workshop on Recommendation Systems*: *Vol. 1* (pp. 114-129).

[9] Wang, J., Hong, L., & Davison, B. D. (2009). Rsdc'09: Tag recommendation using keywords and association rules. *ECML PKDD Discovery Challenge,* 261-274.

[10] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *Proceedings of the 2nd ACM Conference On Electronic Commerce* (pp. 158-167).

[11] Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, *22(1)*, 89-115.

[12] Tuarob, S., Pouchard, L. C., & Giles, C. L. (2013). Automatic tag recommendation for metadata annotation using probabilistic topic modeling. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 239-248).

[13] Huberman, S. G. B. A. (2005). *The Structure of Collaborative Tagging Systems*.

[14] Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 31-40).

[15] Yin, D., Xue, Z., Hong, L., & Davison, B. D. (2010). A probabilistic model for personalized tag prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 959-968).

[16] Lu, Y. T., Yu, S. I., Chang, T. C., & Hsu, J. Y. J. (2009). A content-based method to enhance tag recommendation. *IJCAI*, 9, 2064-2069.

[17] Si, X., Liu, Z., Li, P., Jiang, Q., & Sun, M. (2009). Content-based and graph-based tag suggestion. *Proceedings of the ECML PKDD Discovery Challenge* (pp. 243-260).

[18] Budura, A., Michel, S., Cudré-Mauroux, P., & Aberer, K. (2009). Neighborhood-based tag prediction. *The Semantic Web: Research and Applications*, 608-622.

[19] Hotho, A., *et al*. (2006). *Information Retrieval in Folksonomies: Search and Ranking*. Springer Berlin Heidelberg.

[20] Feng, W., & Wang, J. (2012). Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1276-1284).

[21] Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems* (pp. 61-68).

[22] Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2008). Tag recommendations based on tensor dimensionality reduction. *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 43-50).

[23] Rendle, S., & Schmidt-Thieme, L. (2010). Pairwise interaction tensor factorization for personalized tag recommendation. *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 81-90).

[24] Van-An, J. (2013). Approach for tag recommendation based on orientation of motivation. *Application Research of Computers*, *30(1)*, 72-77.

[25] Liu, J., Li, Z., Tang, J., Jiang, Y., & Lu, H. (2014). Personalized geo-specific tag recommendation for photos

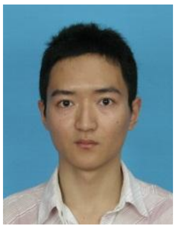
on social websites. *IEEE Transactions on Multimedia, 16(3)*, 588-600.

[26] Yin, D., *et al*. (2011). *Temporal Dynamics of User Interests in Tagging Systems*.

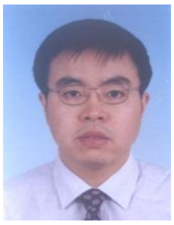
[27] Luo, J., Pan, X., & Zhu, X. (2012). Time-aware user profiling for tag recommendation. *Advances in Information Sciences & Service Sciences*, *4(19)*.

[28] Song, Y., Zhang, L., & Giles, C. L. (2008). A sparse Gaussian processes classification framework for fast tag suggestions. *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 93-102).

[29] *The Practice of Recommender System*. (2012). Beijing: Post & Telecom Press.



**Jing Wang** received his B.S. degree in the Department of Electronic Engineering from Tsinghua University, Beijing, China. Currently, he is a postgraduate student in the Department of Computer Science and Technology, Tsinghua University, China. His research interests include data-mining and machine learning.



**Nianlong Luo** was born in 1971. He received the Ph.D. degree in the School of Economics and Management at Tsinghua University, Beijing, China. Currently, he is an associate researcher working in the Information Technology Center at Tsinghua University, his research interests include information management, educational informationalization and data-mining.