Personalized Search on Flickr based on Searcher's Preference Prediction

Dongyuan Lu, Qiudan Li
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{dongyuan.lu, qiudan.li}@ia.ac.cn

ABSTRACT
In this paper, we propose a personalized search model to assist users in obtaining interested photos on Flickr, which exploits the favorite marks of the searcher's friends to predict the searcher's preference on the returned photos. The proposed model utilizes a co-clustering method to extract latent interest dimensions from users' implicit interests, and employs a discriminative learning method to predict searcher's preference on the returned photos. Preliminary experiments demonstrate the improvement of the proposed model compared to existing one-fit-all methods and a user-based collaborative filtering method.

Categories and Subject Descriptors: H.3.3 [**Information storage and retrieval**]: Information search and retrieval

General Terms: Algorithms, Design, Experimentation.

Keywords: Personalized search, Preference prediction, Latent interest dimension, Discriminative learning.

1. INTRODUCTION
Flickr.com, as one of the most popular photo-sharing and social-networking websites, has been hosting over 5 billion photos since September 2010 [1]. When searching photos by submitting a query, a user may receive hundreds or thousands of returned results, e.g., 118,147 photos are returned by searching with "*Great Wall*". Obviously, users need a tool to assist them in getting access to interested photos more easily. Personalized search serves as such a tool which rearranges the returned results based on the preference of the searcher. Flickr encourages users to perform various activities such as sharing photos with tags, joining in interested groups, contacting other users with similar interest as friends, as well as expressing their preference on photos by adding favorite marks. These social activities offer valuable information for solving personalized search problem. Existing studies have focused on building associations between the searcher and photos by performing tag analysis [2, 3]. These approaches are beneficial when sufficient tags are available. However, there are still plenty of photos with deficient tags. This paper thus offers a new perspective, which explores the favorite marks of the searcher's friends to predict searcher's preference on returned photos for personalized search.

Typically, users are interested in more than one field, and the searcher may share different interests with different friends. The variety of users' implicit interests can be mined and encoded into the latent interest dimensions. Friends may contribute differently to searcher's preference prediction according to the submitted query and the interest distribution. For example, a friend distributed consistently with the searcher on the latent dimensions related to *Travel* and *Landscape* will contribute much to a query like '*Great Wall*'. Therefore, determining the relevant dimensions for a specific query is essential to accurately predict the searcher's preference on returned photos. The proposed model utilizes a co-clustering method to extract latent interest dimensions from users' implicit interests, and employs a discriminative learning method to predict the searcher's preference on photos, which can automatically select query-dependent interest dimensions for prediction. In the following of this paper, the details of this model will be illustrated, along with the experimental results.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03..

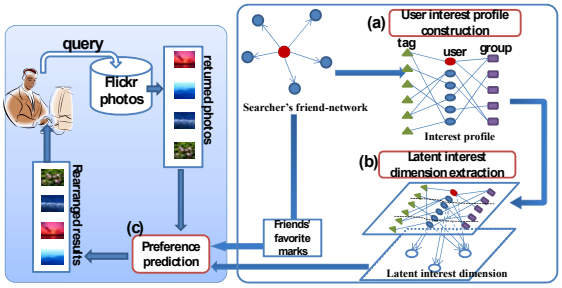

Fig. 1. The framework of the personalized search model

2. METHOD
The framework of the proposed model is shown in Fig. 1. The personalized search model firstly constructs users' interest profiles to represent their implicit interests. And then the latent interest dimensions are extracted to obtain users' interest distributions. Finally, it rearranges the returned photos for the searcher by predicting his preference on these photos.

User Interest Profile Construction Uploading photos with attached tags and joining in interested groups are two main manners for users to share personal photos, which reflect their implicit interests. Both tag and group are thus used to construct a user's interest profile. We refer to $U=\{u_1, u_2,\ldots, u_M\}$ as the set of users including the searcher and his friends, and T as the bag-of-tags utilized by the users, and G as the bag-of-groups the users join in. As illustrated in Fig. 1 (a), the relations of user-tag and user-group construct two correlated bipartite graphs, denoted as $\mathbb{G}$ (U, T, G, W_{UT}, W_{UG}), where W_{UT} denotes the edge weights between U and T, and W_{UG} denotes the edge weights between U and G. The simple yet widely used *tf-idf* is employed to set the edge weights. Therefore, a user's interest profile is simultaneously represented by a weighted tag-vector and a weighted group-vector.

Latent Interest Dimension Extraction With users' interest profiles represented as two correlated bipartite graphs, a soft clustering method is performed to extract the latent interest dimensions. We utilize the clustering method proposed in [4], which aims to co-

cluster high-order correlated bipartite graphs based on spectral graph partitioning (as illustrated in Fig. 1 (b)). The basic idea of spectral graph partitioning focuses on minimizing a cost function. In this study, we adopt *Ncut* which simultaneously minimizes the between-cluster similarities and maximizes the within-cluster similarities. By introducing a parameter λ to balance the costs on both graphs, the co-clustering task can be turned into the optimization problem of minimizing the following generalized cost function:

$$Cost = \lambda \sum_{i=1}^{h} Ncut\left(S_i^U, S_i^T\right) + (1-\lambda) \sum_{i=1}^{h} Ncut\left(S_i^U, S_i^G\right) \qquad (1)$$

where S_i^U, S_i^T and S_i^G denote the *i*-th partition in *U*, *T* and *G*, respectively. And *h* is the desired cluster number, i.e., the size of latent interest dimension. The optimal cluster partitions can be achieved by solving a constrained quadratic programming problem [4]. Consequently, each user is represented in terms of a *h*-dimension latent interest feature vector.

Preference Prediction Suppose there are *R* returned photos after the searcher submitting a query. Given users' latent interest features and their favorite marks on these photos, let $\mathbb{N}=(U, F, Y)$ denote the prediction problem. *U* is the user set and $F=\{f_1, f_2, \ldots, f_M\}$ is the corresponding user interest feature set. $Y=\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ denotes the favorite mark set, where $\boldsymbol{y}_i \in \{0, 1\}^R$ is a *R*-dimension indication vector representing user *i*'s preference status on the *R* returned photos, and $y_i^j=1$ represents user *i* favoring photo *j*, $y_i^j=0$ otherwise. Given known values of $\boldsymbol{y}_i$ for the searcher's friends, we formulate the preference prediction task as an ordinal regression problem. Instead of pursuing absolute preference value for each photo, what we care is the predicted relative preference relationship. From this perspective, we adopt Rank-SVM [5] to rearrange the returned photos for the searcher based on the estimated score of each returned photo.

3. EXPERIMENTS

3.1 Dataset and Parameter Settings

To evaluate the performance of the proposed model, a dataset was collected using Flickr API. We manually selected three popular subjects [6] that concern *Nature*, *Travel* and *Landscape* and chose 18 popular tags as queries. Through finding the overlapping users in groups concerning these subjects, 402 users were collected as searchers. For each searcher, the contacted users were crawled as his friends. The assigned tags, the interested groups users join in and the favorite marks of all the users were crawled. The dataset finally consists of 81,288 users with 0.9 million unique tags, 230,900 distinct groups and 112.5 million favorite marks. The parameter λ in interest dimension extraction was empirically set to be 0.5 to evenly weight the tag and group components. We also tuned the parameter *h* that denotes the interest dimension size from 5 to 20 with a step of 5, and found the best performance at $h=10$. Therefore, the parameter *h* was empirically set to be 10 in our experiments.

3.2 Results and Discussions

The information retrieval metric NDCG@*k* (Normalized Discount Cumulative Gain of top *k* results) was utilized to evaluate the performance. Following the evaluation framework in [7], any photo historically favored by the searcher was considered relevant for personalization. For each query, we rearranged the top 50

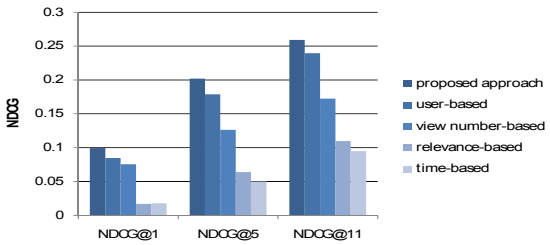

Fig. 2. Evaluation results of comparing five search methods

non-personalized Flickr results sorted by the option *Interesting* for each searcher, and averaged the evaluated scores over queries and searchers. To demonstrate the validity of the proposed model, four additional search methods were conducted as baselines. Three of them are one-fit-all retrieval methods depending on relevance, view number and timeliness, respectively. The other one is the user-based collaborative filtering [8]. We utilized Euclidean distance to measure the similarities between the searcher and his friends, and used the similarities as weights to sum the friends' favorite marks as searcher's preferences.

Fig. 2 demonstrates the evaluation results, from which we can see the two personalization methods outperform the three one-fit-all retrieval methods. It reveals the validity of inferring the searcher's preference from his friends' favorite marks by mining users' latent interest dimensions. Moreover, our approach achieves better personalization performance than the user-based collaborative filtering. It may be due to the leverage of discriminative learning that automatically selects the relevant interest dimensions for inference.

4. CONCLUSION

In this paper, we propose a personalized search model to assist users in getting access to their interested photos by predicting the searcher's preference on returned photos. Preliminary experiments have proved the validity of the proposed model. In the future, we would like to further improve the efficiency of the proposed mechanism.

5. ACKNOWLEDGMENTS

This research is supported by the projects 863 (No. 2006AA010106), 973 (No. 2007CB311007), NSFC (No. 60703085).

6. REFERENCES

[1] http://blog.flickr.net/en/2010/09/19/5000000000/.
[2] Lerman, K., Plangprasopchok, A. and Wong, C., 2008. Personalizing image search results on Flickr. In AAAI workshop on Information Integration'08.
[3] Bender, M., Crecelius, T., Kacimi, M. and Michel, S. 2008. Exploiting social relations for query expansion and result ranking. In ICDE workshop'08, pp 501-506.
[4] Zhou, D., Councill, Isaac., Zha, H. and Lee Giles, C. 2007. Discovering temporal communities from social network documents. In ICDM'07, pp745-750.
[5] Elisseeff, A., Weston, J. 2002. A kernel method for multi-labelled classification. In Advances in Neural Information Processing Systems 14, Cambridge, MA: MIT Press, pp 681-687.
[6] http://www.flickr.com/photos/tags/.
[7] Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. 2008. Exploring folksonomy for personalized search. In SIGIR'08, pp 155-162.
[8] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. 2000. Analysis of recommendation algorithms for e-commerce. In EC'00, pp285-295.