PROMPT: Personalized User Tag Recommendation for Social Media Photos Leveraging Personal and Social Contexts

Rajiv Ratn Shah
National University of Singapore
Singapore
rajiv@comp.nus.edu.sg

Anupam Samanta
Indian Institute of Technology Dhanbad
India
samanta.anupam.19@gmail.com

Deepak Gupta
Indian Institute of Technology Dhanbad
India
deepakismcse@gmail.com

Yi Yu
National Institute of Informatics
Japan
yiyu@nii.ac.jp

Suhua Tang
The University of Electro-Communications
Japan
shtang@uec.ac.jp

Roger Zimmermann
National University of Singapore
Singapore
rogerz@comp.nus.edu.sg

***Abstract*—Social media platforms such as Flickr allow users to annotate photos with descriptive keywords, called, tags with the goal of making multimedia content easily understandable, searchable, and discoverable. However, manual annotation is very time-consuming and cumbersome for most users, which makes it difficult to search relevant photos. Moreover, predicted tags for a photo are not necessarily relevant to users' interests. Thus, it necessitates for an automatic tag prediction system that considers users' interests and describes objective aspects of the photo such as visual content and activities. To this end, this paper presents a tag recommendation system, called, PROMPT, that recommends personalized tags for a given photo leveraging multimodal information. Specifically, first, we determine a group of users who have similar tagging behavior as the user of the photo, which is very useful in recommending personalized tags. Next, we find candidate tags from visual content, textual metadata, and tags of neighboring photos, and recommends five most suitable tags. We initialize scores of the candidate tags using asymmetric tag co-occurrence probabilities and normalized scores of tags after neighbor voting, and later perform random walk to promote the tags that have many close neighbors and weaken isolated tags. Finally, we recommend top five user tags to the given photo. Experimental results on a Flickr dataset (46,700 photos in the test set and 28 million photos in the train set) with 1,540 unique user tags confirm that the proposed algorithm outperforms state-of-the-arts.**

***Keywords*-Tag recommendation; personalized user tags; social media; multimodal information; multimodal analysis**

I. Introduction

PROMPT stands for a personalized user tag recommendation for social media photos leveraging multimodal information. It leverages knowledge structures from multiple modalities such as the visual content, textual metadata, user details, and semantically similar neighbors of a given photo to predict personalized user tags. Since the number of photos on social media platforms has increased rapidly (*e.g.*, Flickr has over ten billion photos) due to advancements in smartphone and digital camera technologies, it requires an automatic tag recommendation system for an efficient multimedia search and retrieval. User tags are very helpful in providing several significant multimedia-related applications such as a landmark recognition [7] and a tag-based photo search and group recommendation [9]. A few automatic photo annotation systems based on visual concept recognition algorithms are proposed [2], [8]. However, they have limited performance because classes (tags) used in the training of deep neural networks to predict tags for a photo are restricted and defined by a few researchers and not by actual users. Thus, it necessitates a tag recommendation system that exploits tagging behaviors of other similar users.

The PROMPT system enables people to automatically generate user tags for a given photo leveraging knowledge structures from visual content, textual metadata, spatial information, and semantically similar neighboring photos. Moreover, it exploits information from past photos annotated by a user to understand the tagging behavior of the user, which is useful in recommending personalized user tags. In this study, we consider the 1,540 most frequent user tags from the YFCC100M dataset, a collection of 100 million media records from Flickr (see Section IV for details), for the tag prediction task. We construct a 1,540-dimensional feature vector, called, the UTB vector, to represent a user's tagging behavior using the bag-of-words model (see Section III for details). We cluster users and their photos in the train set with 28 million photos into several groups based on cosine similarities among UTB vectors during pre-processing. Moreover, we construct a 1,540-dimensional feature vector for a given photo, called, the photo description (PD) vector, using the bag-of-the-words model, to compute the photo's N nearest semantically similar neighbors. UTB and PD vectors help PROMPT to find an appropriate set of candidate photos and tags for the given photo. Since PROMPT focuses on candidate photos instead of all photos in the train set for tag prediction, it is relatively fast. We adopt the following approaches for tag recommendation.

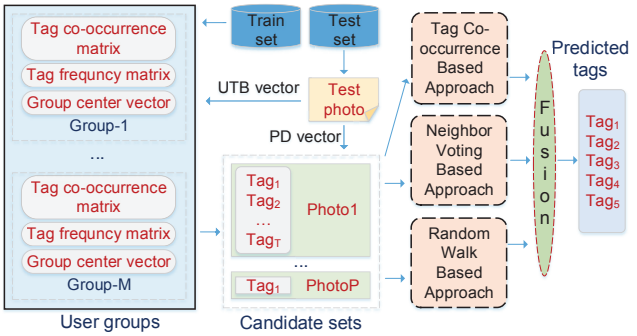

Figure 1. System overview of the PROMPT system.

- Often a photo consists of several objects and it is described by several semantically related concurrent tags (*e.g.*, beach and sea) [24]. Thus, our first approach is inspired by employing asymmetric tag co-occurrences in learning tag relevance for a given photo.
- Many times users describe similar objects in their photos using the same descriptive keywords (tags) [9]. Hence, our second approach for tag recommendation is inspired by employing neighbor voting schemes.
- Random walk is frequently performed to promote tags that have many close neighbors and weaken isolated tags [10]. Therefore, our third approach is based on performing random walk on candidate tags.
- Finally, we fuse knowledge structures derived from different approaches to recommend the top five personalized user tags for the given photo.

In the first approach, the PROMPT system first determines seed tags from visual tags (see Section IV for more details on visual tags) and textual metadata (excluding user tags) such as the title and description of a given photo. Next, we compute top five semantically related tags with the highest asymmetric co-occurrence scores for all seed tags, and add them to the candidate set of the given photo. Next, we combine all seed tags and their co-occurred tags in the candidate set using a sum method (*i.e.*, if some tags appear more than once then their relevance scores are accumulated). Finally, the top five tags with the highest relevance scores are predicted for the given photo. In the second approach, our tag recommendation system first determines the closest user group for the user of the given photo based on the user's past annotated photos. Next, it computes the N semantically similar nearest neighbors for the given photo based on the PD vector constructed from textual metadata (excluding user tags) and visual tags. Finally, we accumulate tags from all such neighbors and compute their relevance scores based on their vote counts and prior frequency in the train set, and recommend the top five tags to the given photo.

In our third approach, we perform a random walk on candidate tags derived from visual tags and textual metadata. The random walk helps in updating scores of candidate tags iteratively leveraging exemplar and concurrent similarities. Next, we recommend the top five user tags when the random walk converge. Finally, we investigate the effect of fusion by combining candidate tags derived from different approaches and next perform a random walk to recommend the top five user tags to the given photo. Experimental results on a test set of 46,700 Flickr photos (see Section IV for details) confirm that our proposed approaches outperform state-of-the-arts in terms of precision, recall, and accuracy scores. Steps of recommending user tags by PROMPT for a social media photo is summarized as follows (see Figure 1).

- It determines a group of users from the train set with 259,149 unique users, having similar tagging behavior as the user of the photo, based on cosine similarity.
- The candidate sets of photos and tags for the photo are computed from the selected user group.
- Relevance scores are computed for candidate tags using our proposed approaches. Finally, the top five user tags are recommended to the photo.

The paper is organized as follows. In Section II, we review related work and Section III describes the PROMPT system. The evaluation results are presented in Section IV. Finally, we conclude the paper with a summary in Section V.

II. Related Work

Our purpose is to automatically predict personalized user tags for a given photo. The steps of such a process is described as follows: (i) learn personal interests of a user from photos uploaded in past by the user, (ii) determine a group of users who have similar tagging behavior as the user, and (iii) recommend tags for the photo leveraging knowledge structures derived from its personal and social contexts. In this section, we briefly provide some recent progress on computing tag relevance for social media photos.

Learning tag relevance for a photo based on accumulating votes from visually similar neighboring photos is a popular approach [9], [18]. Research works in multimedia analytics suggest that multimodal information is very useful in several significant social media related applications and services [15], [16]. For instance, multimodal information is very useful in semantics and sentics understanding of user-generated content [3], [12], [20], [25]. Shah *et al.* [17], [19] leveraged visual tags and textual metadata (*e.g.*, title, description, and user tags) to build semantics and sentics engines. Sigurbjörnsson and Van Zwol [21] presented a tag recommendation system to predict tags based on tag co-occurrence for each user input tag and merge them into a single candidate list using their proposed aggregate (Vote or Sum) and promote (descriptive, stability and rank promotion) methods. Anderson *et al.* [1] presented a tag prediction system for Flickr photos which combines both linguistic and visual features of a photo. Further, Rae *et al.* [13] proposed an extendable framework that can recommend additional tags to partially annotated photos using a combination of different personalised and collective contexts such as (i) all

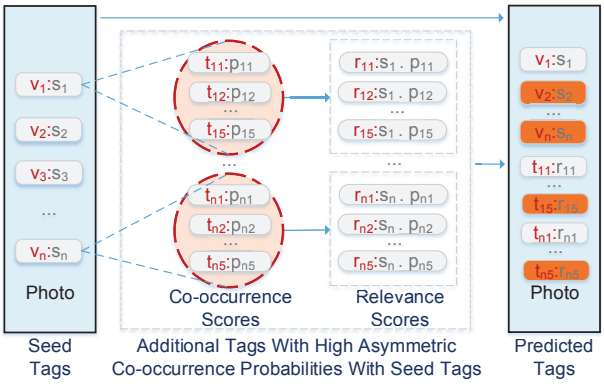

Figure 2. System framework of the tag prediction system based on asymmetric co-occurrence scores.

photos in the system, (ii) a user's own photos, (iii) photos of the user's social contacts, and (iv) photos posted in groups of which the user is a member. Garg and Weber [4] proposed a system that suggests related tags to user, based on the tags that she or other people have used in the past along with (some of) the tags already entered. The suggested tags are dynamically updated with every additional tag entered/selected. However, these approaches are not fully automatic since they expect a user to input (annotate) a few initial tags.

Wu *et al.* [24] formulated tag recommendation as a learning problem and proposed a multimodal recommendation system based on both tag and visual correlation. Each modality is used to generate a ranking feature and Rankboost algorithm is applied to learn an optimal combination of these ranking features from different modalities. Liu *et al.* [10] proposed a tag ranking scheme, aiming to automatically rank tags for a given photo according to their relevance to the photo content. They estimated initial relevance scores of tags based on probability density estimation, and then performed a random walk over a tag similarity graph to refine relevance scores. Wang *et al.* [23] proposed a novel co-clustering framework, which takes the advantage of networking information between users and tags in social media, to discover these overlapping communities. They clustered edges instead of nodes to determine overlapping clusters (*i.e.*, a single user belongs to multiple social groups). Such social groups are also useful in determining aesthetic tendencies and communities [5], [6], [26]. Recent works [11], [14] exploit user context for photo tag recommendation.

III. System Overview

Figure 1 shows the system framework of the PROMPT system. We compute user tagging behavior (UTB) vectors for all users based on their past annotated photos using the bag-of-the-words model on a set of 1,540 user tags T, used in this study. We exploit UTB vectors to perform the grouping of users in the train set and compute asymmetric tag co-occurrence scores among all 1,540 user tags for each group during pre-processing. Moreover, the center of a group is computed by averaging UTB vectors of all users in that group. Similarly, we compute photo description (PD) vectors for photos using the bag-of-the-words model on tags in T. We do not consider user tags of photos to construct their PD vectors, instead we leverage tags derived from title, description, and visual tags which belong to T. PD vectors are used to determine semantically similar neighbors for photos based on the cosine similarity metric. During online processing to recommend user tags for a photo, we first compute its UTB vector, and subsequently a closest matching user group from the train set. We refer to the set of photos and tags in the selected user group as the candidate set and further use them in recommending user tags for photos using the following techniques.

A. Asymmetric Co-occurrence based Relevance Scores

As described in literature [24], tag relevance is standardized into mainly asymmetric and symmetric structures. Symmetric tag co-occurrence tends to measure how similar two tags are, *i.e.*, high symmetric tag co-occurrence score between two tags indicates that they most likely to occur together. However, asymmetric tag co-occurrence suggests relative tag co-occurrence $p(t_i/t_j)$, *i.e.*, it is interpreted as the probability of a photo being annotated with t_j given that it is already annotated with t_i. Thus, an asymmetric tag co-occurrence score is beneficial in introducing diversity to tag prediction, and it is defined as follows.

$$p_{ij} = p(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i|} \tag{1}$$

where $|t_i|$ and $|t_i \cap t_j|$ represents the number of times the tag t_i appears alone and with tag t_j, respectively.

Figure 2 describes the system framework of the tag prediction system based on asymmetric co-occurrence scores. We first determine seed tags from the textual metadata and visual tags of a given photo. Seed tags are the tags appeared in title, visual tags, and description of the photo, which belong to the set of 1,540 user tags used in this study. We add seed tags and their five most co-occurred tags with the highest asymmetric tag co-occurrence probabilities to the candidate set of the photo. For all visual tags of the photo, their confidence scores s_i are also given as the part of the YFCC100M dataset. We set confidence scores of seed tags from textual metadata as 1.0. For additional co-occurred tags in the candidate set, we compute their relevance scores r_{ij} or $r(t_i, t_j)$ as follows:

$$r_{ij} = r(t_i, t_j) = p_{ij} \times s_i \tag{2}$$

where, s_i is the confidence score of a seed tag and p_{ij} is the asymmetric tag co-occurrence probability of the tag t_j for the given seed tag t_i. This formula is justifiable because it assigns the high relevance score to t_j when the confidence of the seed tag t_i is high. In this way, we compute relevance

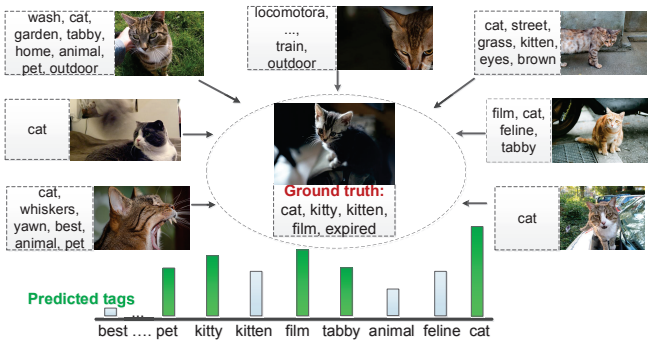

Figure 3. System framework of the tag recommendation system based on neighbor voting scores.

scores of all tags in the candidate set. Next, we aggregate all tags and merge scores of common tags. Finally, we predict top five tags with the highest relevance scores from the candidate set to the photo.

B. Neighbor Voting based Relevance Scores

Earlier works [9], [18] on computing tag relevance for photos confirm that a neighbor voting based approach is very useful in determining tag ranking. Leveraging personal and social contexts, we apply this approach for tag recommendation. Relevance scores of tags for a photo is computed in the following two steps. Firstly, N nearest neighbors of the photo are obtained from the user group of similar tagging behaviors. Next, the relevance score of a tag t for the photo is obtained as follows:

$$z(t) = vote(t) - prior(t, N) \tag{3}$$

where $z(t)$ is the tag t's final relevance score, $vote(t)$ represents the number of votes tag t gets from the N nearest neighbors of the photo. $prior(t, N)$ indicates the prior frequency of the tag t and is defined as follow:

$$prior(t, N) = N\frac{M_t}{D} \tag{4}$$

where M_t is the number of photos tagged with t, and D is the size of the train set.

C. Random Walk based Relevance Scores

Another very popular technique for tag ranking is based on random walk. Liu *et al.* [10] estimates initial relevance scores for tags based on probability density estimation, and then perform a random walk over a tag similarity graph to refine the relevance scores. We leverage the multimodal information of photos and apply this tag ranking approach for tag recommendation. Specifically, first, we determine candidate tags leveraging multimodal information such as the textual metadata (*e.g.*, title and description) and the visual content (*e.g.*, visual tags). We estimate the initial relevance scores of candidate tags adopting a probabilistic approach on co-occurrence of tags. We also use the normalized scores of tags derived from neighbor voting. Next, we refine relevance scores of tags by implementing a random walk process over a tag graph which is constructed by combining an exemplar-based approach and a concurrence-based approach to estimate the relationship among tags. The exemplar similarity φ_e is defined as follows:

$$\varphi_e = exp(-\frac{1}{N * N}\sum_{x\in\Gamma_{t_i}, y\in\Gamma_{t_j}}\frac{||x-y||^2}{\sigma^2}) \tag{5}$$

where Γ_t denotes the representative photo collection of tag t and N is the number of nearest neighbors. Moreover, σ is the radius parameter for the classical Kernel Density Estimation (KDE) [10]. Next, the concurrence similarity φ_c between tag t_i and tag t_j is defined as follows:

$$\varphi_c = exp(-d(t_i, t_j)) \tag{6}$$

where the distance $d(t_i, t_j)$ between two tags t_i and t_j is defined as follows.

$$d(t_i, t_j) = \frac{max(logf(t_i), logf(t_j)) - logf(t_i, t_j)}{logG - min(logf(t_i), logf(t_j))} \tag{7}$$

where $f(t_i)$, $f(t_j)$, and $f(t_i, t_j)$ are the numbers of photos containing tags t_i, t_j, and both t_i and t_j, respectively, in the training dataset. Moreover, G is the number of photos in the training dataset. Finally, the exemplar similarity φ_e and concurrence similarity φ_c are combined as follows:

$$\Phi_{ij} = \lambda \cdot \varphi_e + (1 - \lambda) \cdot \varphi_e \tag{8}$$

where λ belongs to [0,1]. We set it to 0.5 in our study.

We use $u_k(i)$ to denote the relevance score of node i at iteration k in a tag graph with n nodes. Thus, relevance scores of all nodes in the graph at iteration k form a column vector $u_k \equiv [u_k(i)]_{n\times 1}$. An element q_{ij} of this n×n transition matrix indicates the probability of the transition from node i to node j and it is computed as follows:

$$q_{ij} = \frac{\Phi_{ij}}{\sum_k \Phi_{ik}} \tag{9}$$

The random walk process promotes tags that have many close neighbors and weakens isolated tags. This process is formulated as follows.

$$u_k(j) = \alpha \sum_i u_{k-1}(i) q_{ij} + (1 - \alpha) w_j \tag{10}$$

where w_j is the initial score of a tag t_j and α is a weight parameter between (0, 1).

D. Fusion of Different Tag Recommendation Approaches

The final recommended tags for a given photo is determined by fusing different above-mentioned approaches. We combine candidate tags determined by asymmetric tag co-occurrence and neighbor voting schemes. Next, we initialize scores of the fused candidate tags with their normalized

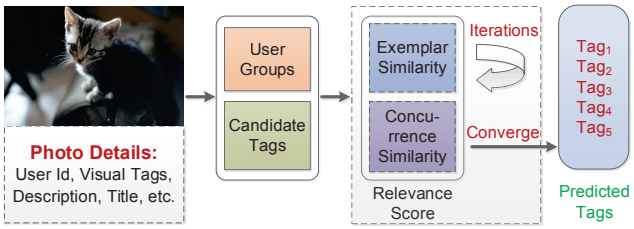

Figure 4. Architecture of the tag prediction system based on random walk.

scores from [0,1]. Further, we perform a random walk on a tag graph which has the fused candidate tags as its nodes. This tag graph is constructed by combining exemplar and concurrence similarities and useful in estimating the relationship among the tags. In this way, the random walk refines relevance scores of the fused candidate tags iteratively. Finally, our PROMPT system recommends the top five tags with the highest relevance scores to the photo, when the random walk converges.

IV. Evaluation

A. Dataset

We used the YFCC100M [22] dataset consisting of 100 million media records (approximately 99.2 million photos and 0.8 million videos) from Flickr. The reason for selecting this dataset is its volume, modalities, and metadata. For instance, each media of the dataset consists of several metadata annotations such as user tags, spatial and temporal information, and others. Moreover, all media are labelled with automatically added (visual) tags with confidence scores derived from a convolution network, which indicate the presence of a variety of concepts such as ocean, food, and scenery, say, with confidence 95%, 85%, and 92%, respectively. There are a totally 1,756 visual tags present in the YFCC100M dataset. The YFCC100M dataset has been split into 10 parts based the last digit prior to the @-symbol in their Flickr user identifier (NSID). Such split ensures that no user occurs in multiple partitions, thus avoiding dependencies between the different splits. Split 0 is used as the test set and the remaining nine splits as the train set.

For tag prediction, a specific subset of 1,540 user tags T are considered since predicting the correct tags from a virtually endless pool of possible tags is extremely challenging. Tags in T fulfill the following criteria: (i) they are valid English dictionary words, (ii) such tags do not refer to persons, dates, times or places, (iii) they appear frequently with photos in the train and test sets, and (iv) different tenses/plurals (tags) of the same word (an already added tag in T) are not considered. The train set contains all photos from the YFCC100M that have at least one tag that appeared in T and do not belong to the split 0. There are approximately 28 million photos present in the train set. The test set contains 46,700 photos from the split 0 such that each photo has at least five tags from the list of 1,540 tags. There are totally 259,149 and 7,083 unique users in the train and test sets for this study, respectively.

	Comparison	K = 1	K = 3	K = 5
Accuracy@K	Type-1	0.410	0.662	0.746
	Type-2	0.422	0.678	0.763
Precision@K	Type-1	0.410	0.315	0.251
	Type-2	0.422	0.326	0.262
Recall@K	Type-1	0.062	0.142	0.188
	Type-2	0.064	0.147	0.197

Table I
Results for the top K predicted tags.

B. Results

Recommended tags for a given photo in the test set are evaluated based on the following three metrics: (i) Precision@K, *i.e.*, proportion of the top K predicted tags that appear in user tags of the photo, (ii) Recall@K, *i.e.*, proportion of the user tags that appear in the top K predicted tags, and (iii) Accuracy@K, *i.e.*, 1 if at least one of the top K predicted tags is present in the user tags, 0 otherwise. PROMPT is tested for the following values of K: 1, 3, and 5. We implemented two baselines and proposed a few approaches to recommend personalized user tags for social media photos. In *Baseline1*, we predict the top five most frequent tags from the training set of 28 million photos to a test photo. Further, in *Baseline2*, we predict five visual tags with the highest confidence scores (already provided with the YFCC100M dataset) to a test photo. Since state-of-the-arts for tag prediction [1], [21] mostly recommend tags for photos based on input seed tags. In our PROMPT system, first, we construct a list of candidate tags using asymmetric co-occurrence, neighbor voting, probability density estimation techniques. Next, we compute tag relevance for photos through co-occurrence, neighbor voting, random walk based approaches. We further investigate the fusion of these approaches for tag recommendation.

Figures 5, 6, and 7 depicts scores@K for accuracy, precision, and recall, respectively, for different baselines and approaches. For all metrics, *Baseline1* (*i.e.*, recommending the five most frequent user tags) performs worst and the combination of all three approaches (*i.e.*, co-occurrence, neighbor voting, and random walk based tag recommendation) outperforms rest. Moreover, the performance of *Baseline2* (*i.e.*, recommending the five most confident visual tags) is second from last since it only considers the visual content of a photo for tag recommendation. Intuitively, accuracy@K and recall@K increase for all approaches when we the number of recommended tags increases from 1 to 5. Moreover, precison@K decreases for all approaches when we increase the number of recommend tags. Our PROMPT system recommends user tags with 76% accuracy, 26% precision, and 20% recall for five predicted tags on the test set with 46,700 photos from Flickr.

Table I depicts accuracy, precision, and recall scores when

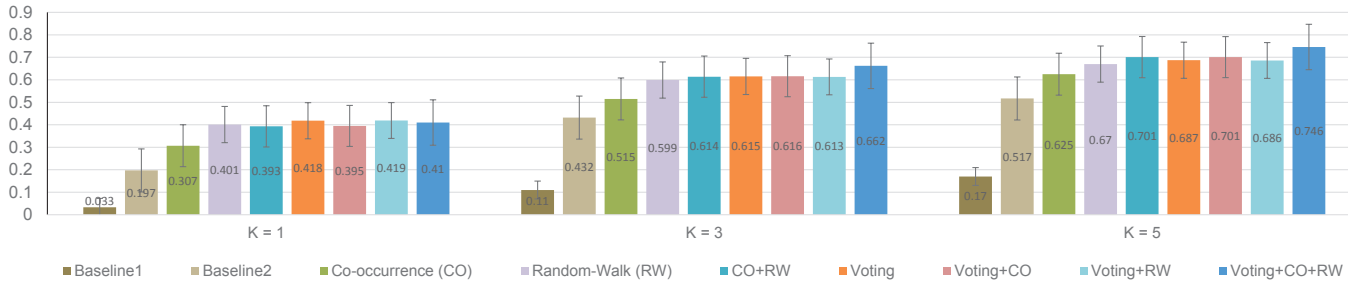

Figure 5. Accuracy@K, *i.e.*, user tag prediction accuracy for K predicted tags for different approaches.

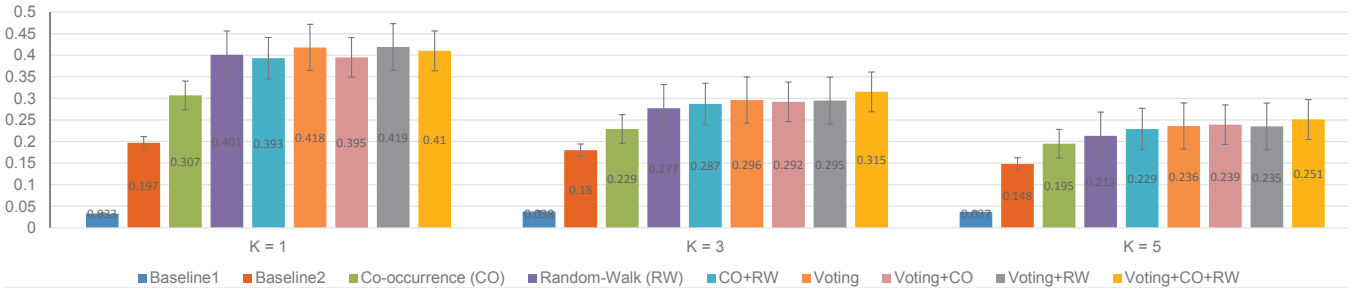

Figure 6. Precision@K, *i.e.*, the precision of tag recommendation for K recommended tags for different approaches.

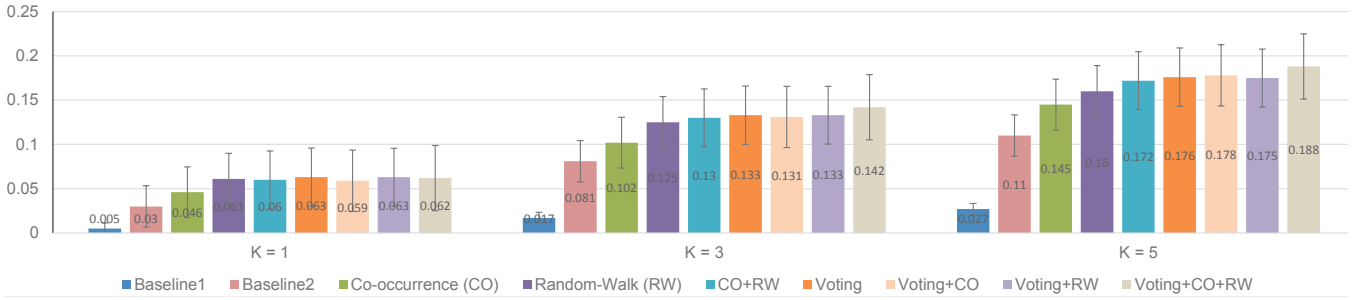

Figure 7. Recall@K, *i.e.*, recall scores for K predicted tags for different approaches.

a combination of co-occurrence, voting, and random walk is used for tag prediction. Type-1 considers a comparison as a hit if a predicted tag matches ground truth tags and Type-2 considers a comparison as a hit if either a predicted tag or its synonyms match ground truth tags. Intuitively, accuracy, precision, and recall scores are slightly improved when the Type-2 comparison is made. This is consistent with all baselines and approaches which we used in our study for tag prediction. All results reported in Figures 5, 6, and 7 correspond to the Type-1 match. Finally, Figure 8 shows the ground truth user tags and the tags recommended by our system for five sample photos in the test set.

V. Conclusions

The proposed PROMPT system is an automatic tag recommendation system. First, it determines a group of users who have similar tagging behavior as the user of a given photo. Next, we construct lists of candidate tags for different approaches based on co-occurrence and neighbor voting. Further, we compute relevance scores of candidate tags. Next, we perform a random walk process on a tag graph with candidate tags as its nodes. Relevance scores of candidate tags are used as initial scores for nodes and updated in every iteration based on exemplar and concurrent tag similarities. The random walk process iterates until it converges. Finally, we recommend the top five tags with the highest scores when the random walk process terminates. Experimental results confirm that our proposed approaches outperform baselines in personalized user tag recommendation. These approaches could be further enhanced to improve accuracy, precision, and recall in the future.

Acknowledgments

This research was supported in part by Singapores Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1415, and by JSPS KAKENHI Grant Number 16K16058.

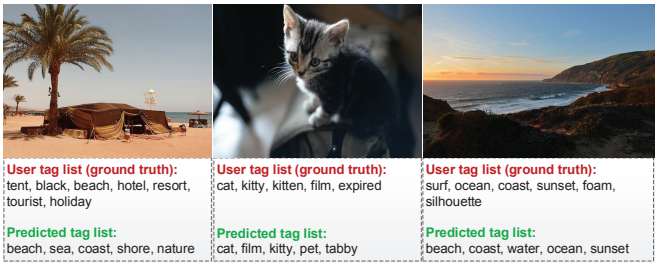

Figure 8. Demonstration of recommended and ground truth user tags.

REFERENCES

[1] A. Anderson, K. Ranghunathan, and A. Vogel. TagEz: Flickr Tag Recommendation. *In AAAI Conference on Artificial Intelligence*, 2008.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching Words and Pictures. *In Journal of Machine Learning Research*, 3(Feb):1107–1135, 2003.

[3] E. Cambria, S. Poria, A. Gelbukh, and K. Kwok. A Commonsense based API for Concept-level Sentiment Analysis. *In Workshop on Making Sense of Microposts at WWW*, pages 27–32, 2014.

[4] N. Garg and I. Weber. Personalized, interactive tag recommendation for flickr. In *ACM Conference on Recommender Systems*, pages 67–74, 2008.

[5] R. Hong, L. Zhang, and D. Tao. Unified Photo Enhancement by Discovering Aesthetic Communities from Flickr. *In IEEE Transactions on Image Processing*, 25(3):1124–1135, 2016.

[6] R. Hong, L. Zhang, C. Zhang, and R. Zimmermann. Flickr Circles: Aesthetic Tendency Discovery by Multi-view Regularized Topic Modeling. 18(8):1555–1567, 2016.

[7] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr Helps us Make Sense of the World: Context and Content in Community-contributed Media Collections. In *ACM Multimedia*, pages 631–640, 2007.

[8] J. Li and J. Z. Wang. Real-time Computerized Annotation of Pictures. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[9] X. Li, C. G. Snoek, and M. Worring. Learning Social Tag Relevance by Neighbor Voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.

[10] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag Ranking. In *ACM WWW*, pages 351–360, 2009.

[11] A. O. Nwana and T. Chen. Who Ordered This?: Exploiting Implicit User Tag Order Preferences for Personalized Image Tagging. *In arXiv preprint arXiv:1601.06439*, 2016.

[12] S. Poria, E. Cambria, and A. Gelbukh. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis. In *Empirical Methods on Natural Language Processing*, pages 2539–2544, 2015.

[13] A. Rae, B. Sigurbjörnsson, and R. van Zwol. Improving Tag Recommendation using Social Networks. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 92–99, 2010.

[14] Y. Rawat and M. S. Kankanhalli. ConTagNet: Exploiting User Context for Image Tag Recommendation. In *ACM Multimedia*, 2016.

[15] R. R. Shah. Multimodal Analysis of User-Generated Content in Support of Social Media Applications. In *ACM Int'l Conference on Multimedia Retrieval*, pages 423–426, 2016.

[16] R. R. Shah. Multimodal-based Multimedia Analysis, Retrieval, and Services in Support of Social Media Applications. In *ACM Multimedia*, pages 1425–1429, 2016.

[17] R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G. Wu. EventBuilder: Real-time Multimedia Event Summarization by Visualizing Social Media. In *ACM Multimedia*, pages 185–188, 2015.

[18] R. R. Shah, Y. Yu, S. Tang, S. Satoh, A. Verma, and R. Zimmermann. Concept-Level Multimodal Ranking of Flickr Photo Tags via Recall Based Weighting. In *MMCommons Workshop at ACM Multimedia*, 2016.

[19] R. R. Shah, Y. Yu, A. Verma, S. Tang, A. D. Shaikh, and R. Zimmermann. Leveraging Multimodal Information for Event Summarization and Concept-level Sentiment Analysis. *In Knowledge-Based Systems*, 108(Sep):102–109, 2016.

[20] R. R. Shah, Y. Yu, and R. Zimmermann. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *ACM Multimedia*, pages 607–616, 2014.

[21] B. Sigurbjörnsson and R. Van Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *ACM WWW*, pages 327–336, 2008.

[22] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The New Data in Multimedia Research. *In the Communications of the ACM*, 59(2):64–73, 2016.

[23] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering Overlapping Groups in Social Media. In *IEEE International Conference on Data Mining*, pages 569–578, 2010.

[24] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to Tag. In *ACM WWW*, pages 361–370, 2009.

[25] Y. Yin, Z. Shen, L. Zhang, and R. Zimmermann. Spatial-temporal Tag Mining for Automatic Geospatial Video Annotation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(2):29, 2015.

[26] L. Zhang and R. Zimmermann. Flickr Circles: Mining Socially-aware Aesthetic Tendency. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2015.