Peer-to-Peer Netw. Appl. (2013) 6:434–444
DOI 10.1007/s12083-013-0214-6

Tag-based personalized image ranking in event browsing

Yeqi Lu · Yao Shen · Minyi Guo

Received: 6 January 2013 / Accepted: 25 April 2013 / Published online: 21 May 2013

Abstract Many image sharing websites, e.g. Flickr, Google+, allow users to upload images as an event, and users can browse the images others uploaded as events. The fact that people usually browse only the first few images of an event then decide whether the event is what they want makes us believe that it is necessary to present those images people favor on the very first position for each event. Here we propose a new tag- based personalized image-ranking algorithm in event browsing such that it gives image higher score if it: a) is important in the event, b) matches user's preference. c) matches user's query. To this end, we first adopt a local matching model to assign images an original score based on whether this image satisfies user's query and preference. We then propose a global ranking model to take the local scores as initial values and make the salience scores iteratively smooth with respect to all images returned from the events of the query.

Keywords Tag-based image search · Personalized ranking · Event browsing · Global ranking · Local matching

1 Introduction

In many image sharing websites, like Flickr, Google+, users are allowed to upload images and group those images as events under same topics like 'Christmas Party', 'Trip to Tokyo', etc. The uploader can specify a few sentences as the description of the images, or tag the uploaded images with tags to help when other users try to search for images of a specific topic, viewers are also allowed to give comments on those images, and these comments also represent the images' contents to some extent.

For image viewers, when they search events in a website, they will receive a list of events whose images' tags, descriptions, or comments meet their query the best (Fig. 1). When the viewers get such a list of events, however, they won't eventually browse through all the images in each event, normally they only peek at the few images that are displayed first in each event and then decide whether this event is what they want to look further into. This observation gives us the idea that we shall present the images that interest the viewer the most first in each event, so as to help them to make an early decision on whether the images in those events are what they really want.

Currently, all images in an event are ranked in a one-size-fits-all style, i.e., all users receive the same ordered-list of image of the same event, and usually this is ordered simply by the capture time or upload time of the images.

However, such unified image ranking might well mislead the users to overlook the originally relevant images which they most desire to find, since different people almost always hold different criteria as for the relevance of a same image. Some relevant images may also be neglected if the user decides the event to be irrelevant to their appetite only by judging at the first few unluckily unappealing images.

Figure 1 illustrates our main scenario. The search engine returns a list of events based on the user's query, each event contains a list of images, and those images are uploaded by the same uploader and grouped under the same topic. Our algorithm will take in the user query and preference profile,

Y. Lu (✉) · Y. Shen · M. Guo
Department of Computer Science & Engineering,
Shanghai JiaoTong University, Shanghai 200240, China
e-mail: l.ainulindale@gmail.com

Y. Shen
e-mail: yshen@cs.sjtu.edu.cn

M. Guo
e-mail: guo-my@cs.sjtu.edu.cn

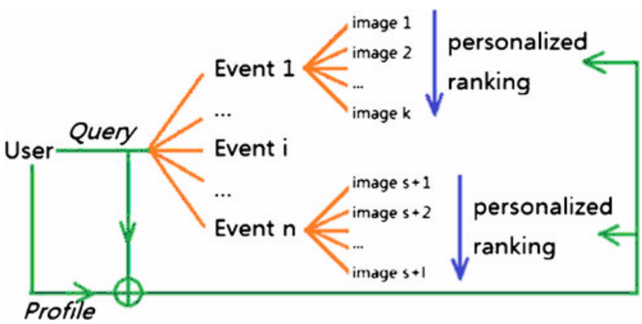

Fig. 1 Given the query, a search engine returns a list of events, each event contains a list images, our algorithm tries to rank the images within each event based on the knowledge of user's query and preference profile. *Orange Lines*: existing search engines' function; *Green Lines*: flow of user query and preference profile; *Blue Lines*: our algorithm generates the personalized rankings of images within each event

and generate personalized rankings on images within each event, so those images that satisfies the following requirements are assigned higher scores and displayed first: a) image is important in the event; b) image is relevant to the user query; and c) image is relevant to the user preference profile.

2 Related works

2.1 Personalized searching

Personalized search has long been studied in IR community, with an ultimate goal to make search results adaptive to the interests of a specific user [1, 2]. Existing work along this line has mainly two focuses: how to represent the search preferences of users in a feasible manner (i.e., personalized profile) and how to re-rank the user-oriented pages at the top of search results (i.e., personalized re-ranking). In the following, we elaborate on these two issues.

2.1.1 Personalized profile

Previous studies mostly packaged personal preferences into a user profile represented as a list of topical terms [3–5], or further as a real-valued interest vector by imposing a weighting scheme (e.g., TFIDF or BM25) on these terms [6–9]. User profile has been constructed based on a great variety of data sources, each of which reflects user interests more or less, such as browsing history [5, 9], web directory [3, 7], and desktop file [3, 4, 7, 10]. Recently, an emergent form of tagging service, social annotation, is rapidly rising on the web 2.0, where people are encouraged to adopt keyword-like annotations to tag, collect, organize and share their favorite web pages online. Since social annotations are created, selected and owned by users, they are inherently endowed with meaningful personal information. In fact, some work [6, 11, 12] has recently noticed this personal data and begun to explore it for personalized search. We also exploit social annotations to open up a wide sky of personalization clouds [29–31].

2.1.2 Personalized re-ranking

Prior research devoted to personalized search has heavily been concerned with re-ranking the generic results returned from search engines so that a current user could easily find his wanted results at the vey top. In general, there are broadly two genres of technical means to achieve this goal: result adjustment and query expansion. The former primarily pursuits a better page ranking algorithm with user query unaltered all along; it is typically achieved by encoding personalization aspects into PageRank [13], including topic-sensitive PageRank [8], topic-distributed PageRank [14] and personalized PageRank vector [15]. The latter primarily aims at substituting a better query for the original one issued by users while page ranking algorithm remains unchanged all the time; it is typically accomplished by appending personalized terms into original queries and making search results focused on personal needs indirectly [3, 10]. Our work is orthogonal to the above two means in that we rank the photos of events adaptively carter to the users' interests with need to tamper with neither user queries nor resulting events' ranks.

2.2 Tag-based image searching

Despite all valuable efforts to explore large image databases by visual similarity between a query image and the data available, using textual keyword queries still remains a powerful means to express the information need [16–20]. User tags are used in image search as an option of relevance-based ranking, [19] proposed a relevance-based ranking scheme for social image search, aiming to automatically rank images according to their relevance to the query tag. It integrates both the visual consistency between images and the semantic correlation between tags in a unified optimization framework. In our algorithm, we expand the use of image tags to the whole bunch of words related to the image, i.e. the tags, descriptions, and comments. This makes our model more robust, especially on those cases where tags are incomplete.

2.3 Event based image browsing

Event based Image Browsing is already emerging in many image sharing websites, e.g. Flickr [21] and Google+ [22], in this scenario, given the query, the returned result is not a just list of photos, but a list of events. Each event consists of several images, uploaded by the same uploader, under the same topic. That those images within each event shares the same topic gives us the idea of using a global ranking model to smooth the salience scores generated by the local matching model.

2.4 Personalized image recommendation

A personalized reinforcement-learning tool is proposed by [23], which helps user to observe the searched images that are desirable for him/her. Their tool gathers the images of the search results and selects a sample. By interacting with user and presenting samples, the personalized tool learns the user's preferences. Personalized image recommendations like JustClick [24] enables this via exploratory search from large-scale collections of manually annotated Flickr images. They also integrate kernel principal component analysis and hyperbolic visualization according to their nonlinear visual similarities.

Our method differs from existing personalized image recommendation methods in that we rank the images within each event generated from the query, while their methods rank all the images generated from the query. That is to say, they want to pick out the images that suit the query and user profile the best, while our target is to pick out representative images for each event to help users viewing and understanding those events.

Another difference between our model and theirs is that our global ranking model focuses more on the inter- connections between images within each event. Their algorithms are similar in function to our local matching model, where such relationships between images are completely ignored when doing personalized ranking. In experiments we will show that this connection enables our global ranking model to pick out related images that don't even have the query's or the profile's words in its sentences, which proves that our model is more robust to the variety qualities of the textual information of images.

3 Personalized photo ranking

3.1 Problem description

There are several observations that lead us to our model:

- Images of the same event usually share the same topic.
- Usually images are presented with tags set by theuploader, or a few lines of description.

We are concerned with generating personalized image rankings in the event. When the event is presented to the user, those images that represents the event the most and suits the user's preference and query the most shall be ranked in the front, so the users can just read a few of the first images to decide whether images in this event is what he or she needs.

However, extracting such rankings is actually non-trivial: In an event, there might be some images that only have a small number of tags, descriptions or comments, not to mention that we further impose a personalized expectation on its selection. Thereby, an effective way is required to distinguish the minority of qualified sentences from the majority of unqualified sentences.

Here in this paper, we extract for each image the tags, descriptions, and comments, and combine them into a sentence. Each image is then solely represented by this sentence. We will not use other image features as criteria in ranking. Ignoring the photo's pixel-information might be a little too harsh, but readers will see that our model is convenient when we want to add image features to it. The fact that uploaders usually give correct descriptions and tags to their images and that viewers usually give comments related to the images' topic has made our algorithm possible.

With above challenges in mind, we cast the generation of personalized image rankings as a problem of sentence ranking. We target to single out those images at the top, based on how their sentences are relevant to the users' needs, both literally and semantically.

Our problem can be described as the following: Given a list of r events $\left\{E^{(1)}, E^{(2)}, ..., E^{(r)}\right\}$, event $E^{(i)}$ consists of a list of $n(i)$ images $\left\{I_1^{(i)}, I_2^{(i)}, ..., I_{n(i)}^{(i)}\right\}$, each image $I_k^{(i)}$ corresponds to a sentence $s_k^{(i)}$, which is a vector of words extracted from the descriptions, tagging, comments of the image. The user profile P and the user's search keywords Q is also a vector of words.

Technically, we want to assign scores to images in each event $E^{(i)}$ such that images with higher scores has sentence $s_k^{(i)}$ that are more related to P and Q, and $s_k^{(i)}$ shall somehow represent how important $I_k^{(i)}$ is in the event.

This score ought to be indicative of the fitness for an image to serve as the query result and user's preference profile. Generally, the calculation of salience score for image's sentence has two earlier considered factors:

- Importance factor (IF). An image should be salient enough in the event so that it could be representative of the content of the whole list of images as an event.
- Relevance factor (RF). An image should also be relevant to user query so that it could be indicative of whether the image hits the corresponding query semantically.

Obviously the above two factors are indispensable for selecting an image. However, it is far from enough to determine the best image only considering these factors. That is because the above two factors treat all users without discrimination. We hereby introduce a third factor for picking up an image sentence.

– Personalization factor (PF). An image should further be matched to user preferences so that it could best guide the users to make a correct relevance judgment.

Since an image's sentence normally consists of a very few segments, we design a framework to combine the three factors to determine which sentences are the best ones. Specifically, the salience score of an image with sentence s is formally measured by

$$S(s) = \alpha \cdot IF(s) + \beta \cdot RF(s) + \gamma \cdot PF(s) \tag{1}$$

These coefficients are used to control the relative contributions of the three factors for calculating $S(s)$. Although only a simple linear combination is adopted in this framework, as analyzed in [25], it actually enables a lot of fast approximate ranking algorithms and also complies with distributed computing architecture in the workflow of current web search engines. Most importantly, as will be shown, we can derive two consecutive models from this unified and general framework to pick descriptive images for events in personalized ranking.

3.2 Local matching model

As a first step, we target to take advantage of local clues to select personalized photos in the given event. In intuition, the photos' sentences are good candidates if they contain some of the terms that are present in both the user query and user profile. Following the work of [26], we employ some heuristic measures to evaluate the local presence of three factors. First, importance factor is computed by

$$IF(s) = SW^2/TW \tag{2}$$

Where SW denotes the number of significant words within s while TW is the total number of words within s. A word is significant in the event if its frequency $TF(s)$ surpasses a predefined threshold t,

$$TF(s) > t = \begin{cases} 70.1 * (25 - NS), \text{ if } NS < 25 \\ 7, \text{ if } NS \in [25, 40] \\ 7 + 0.1 * (NS - 40), \text{ if } NS > 40 \end{cases} \tag{3}$$

Where NS represents the total number of sentences of photos in the event.

Second relevance factor is calculated by

$$RF(s) = TQ^2/NQ \tag{4}$$

Where TQ stands for the number of query terms within s while NQ stands for the total number of terms within the user query. Similar to Eqs. 2 and 4 here, we measure personalization factor by

$$PF(s) = TP$$

per 518ting.516,518taesithtswe6D 0 T001 T

3. Iterate $G(t+1) = \alpha \cdot G(t) + \beta \cdot RF + \gamma \cdot PF$ until convergence, where $\alpha + \beta + \gamma = 1$ and are parameters in (0, 1) and $G(0) = IF$.

Here, IF, RF and PF represent h-dimension vectors that could be calculated using the local matching model in the former section. The algorithm is divided into three steps.

Step1 calculates a symmetric similarity matrix W for all pairs of returned sentences, Section 3.6 will introduce more about $d(s_i, s_j)$ and the construction of W.

In step 2, W is normalized into U in a symmetric manner, which is necessary to guarantee the convergence of the algorithm as will be proved soon in Section 3.5

Step 3 is the heart of the framework. An iteration process is conducted, where a sentence gradually propagates its score over the similarity graph. During this course, each sentence continuously accumulates its global score from its neighbors, and also retains its original local score. As the iteration proceeds, each sentence aggregates more and more information from one possibly having the same image decision.

In the end, the salience scores of all sentences follow a smooth distribution with respect to the underlying similarity graph and take on a local and global consistency steadily.

In fact, neighbor constraint provides a guarantee of local consistency since it makes the nearby sentences have an identical salience decision as likely as possible. In consistency initial scores for some of these contiguous sentences have a good chance of being corrected during the iteration of the global model. More significantly, the global model can inherently discover many latent images. These sentences might locate at the two ends of a smooth structure, where they are seemingly irrelevant in literal content actually relevant in semantic meanings. With the effective propagation in the global model, the scores of some sentences would progressively spread to reach those remote ones in the final stable state.

3.5 Convergence proof

In this section, we prove that the iterative algorithm in Section 3.4 could theoretically converge to a stable situation.

$$
\begin{aligned}
G(1) &= \alpha U \cdot G(0) + \beta RF + \gamma PF \\
G(2) &= \alpha \cdot (\alpha U \cdot IF + \beta RF + \gamma PF) + \beta RF + \gamma PF \\
&= (\alpha U)^2 \cdot IF + (\alpha U + I)(\beta RF + \gamma PF) \\
G(3) &= \alpha U \cdot G(2) + \beta RF + \gamma PF \\
&= (\alpha U)^3 \cdot IF + \big((\alpha U)^2 + \alpha U + I\big)(\beta RF + \gamma PF) \\
&\ldots \\
G(t+1) &= \alpha U \cdot G(t) + \beta RF + \gamma PF \\
&= (\alpha U)^{t+1} \cdot IF \\
&\quad + ((\alpha U)^t + \ldots + \alpha U + I)(\beta RF + \gamma PF) \qquad (6)
\end{aligned}
$$

Where I is an identity matrix. Note that U is a stochastic matrix and its eigenvalues fall in $(-1, 1)$, the following equation holds,

$$lim_{t\to\infty}(\alpha \cdot U)^{t+1} = 0 \qquad (7)$$

Now, we can infer that $G(t)$ converges to a closed form G^*,

$$G^* = (I - \alpha U)^{-1}(\beta RF + \gamma PF) \qquad (8)$$

It can be concluded that the closed form of G is independent of its initial value IF. But this does not mean that the global model takes no account of importance factor in the process of image scoring. To see why, we take a closer look at its convergence property. Recall that the iterative equation in Section 3.4 is

$$G(t+1) = \alpha U \cdot G(t) + \beta \cdot RF + \gamma \cdot PF \qquad (9)$$

When $RF = 0$ and $PF = 0$, we have

$$G(t+1) = \alpha U \cdot G(t) \qquad (10)$$

It is exactly an algebra eigen equation. This means that G would strictly converge to the largest eigen vector of U. In this sense, the importance of each sentence is characterized by its corresponding element in the eigen vector.

It is important to understand that, the importance factor is simply encoded by U in this way. In fact, we also note that the similar technique has been adopted in [32] to perform a task of multi-document text summarization. The success they achieved is a strong support for our work. That is, the best images' sentences for one event are now collectively elected with the voting of all resulting events returned from the current user query.

At this point, the global ranking model is totally different from the local matching model, for the latter only covers a single image to make the election.

It should be further noted that Eq. 10 could only perform an unsupervised ranking upon all involved sentences. There is no way to incorporate any further constraint during this ranking course. From this perspective, our framework in Eq. 9 essentially conducts a semi-supervised ranking for all underlying sentences by taking RF and PF as their initial values.

Furthermore, we could have two variants of Eq. 9, for two special cases. In specific, when $RF \neq 0$ and $PF = 0$, we get

$$G(t+1) = \alpha \cdot G(t) + \beta RF, (\alpha + \beta = 1) \qquad (11)$$

The above equation could select for events the image with textual information most matched with user query regardless of user preference. In this case, G would converge to

$$G^* = \gamma (I - \alpha U)^{-1} \cdot RF. \qquad (12)$$

Similarly, when $RF = 0$ and $PF \neq 0$, we obtain

$$G(t+1) = \alpha \cdot G(t) + \gamma PF, (\alpha + \gamma = 1) \quad (13)$$

The above equation could select for each events the images most matched with user preference regardless of user query. At this time, G would converge to

$$G^* = \gamma (I - \alpha U)^{-1} \cdot RF \quad (14)$$

3.6 Graph building

To implement our global ranking model, we need to build the underlying graph by defining a similarity measure between two photos' sentences. Following [32], we first adopt a popular metric, namely cosine distance for this purpose. Given that two photo sentence s_i and s_j, have been represented as two TFIDF vectors and normalized in unit length, the cosine similarity between them is defined by

$$d(s_i, s_j) = \sum_{word} s_i(word) \cdot s_j(word) \quad (15)$$

It is simply the dot product of s_i and s_j. Cosine metric is hereby calculated based on comparing their common dimensions. In other words, if the two sentences agree on no terms, their cosine distance must be 0. It would cause the similarity matrix W to be sparse in many occasions. In fact, the two sentences might be very similar if they are inclusive of semantically relevant terms. In order to solve this problem, we adopt probabilistic latent semantic analysis [33] to perform a topic modeling for all sentences, and each sentence is represented with a probabilistic coefficient vector over all topics.

Finally, the similarity of two sentences is measured based on their coefficient vectors. In this way, we obtain a dense similarity graph for building global models. More favorably, both dense graph and latent similarity help to retrieve more latent images in event browsing, since latent relevant images are now connected in such a graph.

4 Experiments

In this section, we conduct an experimental evaluation to demonstrate the performance of our proposed models.

In Section 4.1, we first build up a sky of personalization clouds for users to determine their search interests. Then, we develop a system of personalized image ranking in Section 4.2 for the purpose of sentence annotation, the annotation is mainly blind of the photo, but focus on the text information (sentence), which concerns our model more, and we perform an in-depth analysis of local matching model and global ranking model by comparing their performance on sentence ranking.

4.1 Personalized sky

The first thing we have to deal with is to provide a way to simulate the user preferences, because that piece of data can only fetch from real world systems and we have no access to that.

Here we choose to explore social annotations for this end, not only because social book-markings themselves are good source of encoding personal information, but also because they are open to the public and thus help us to obviate the need to access private information of users.

In practice, privacy issue is long a hard nut to crack in the area of personalized search, since many people are reluctant to expose their private data, such as emails, search logs and browsing history, even for research purpose.

We exploit social annotations crawled from Del.icio.us [34], a popular social tagging service. In all, we obtain 1,736,268 web pages and 269,566 annotations. We further select 9,826 annotators from this corpus with 65,080 distinct tags and 90,300 pages, since they have the most annotations than others. A straightforward method is to admit a user to select an annotator by examining an annotator's book-markings so that they have similar tastes, favorites and preferences. Unfortunately, such a simple minded approach does not work in practice, because it is usually difficult for a user to find an annotator so that the two persons happen to match well in all interests, since an annotator often has a range of interests.

To steer clear of such great inconvenience for users, we decide to cluster social annotations into a set of groups, each of which is concentrated on a coherent topic. We first use an open source tool (CLUTO, [35]) to cluster the 90,300 web pages into 300 groups, then pick out 88 groups with distinct topics.

For each cluster, we plot its associate annotations as a tag could with its most frequent 20 annotations. Finally all 88 clouds scatter in a personalization sky.

A user can select a cloud from this sky as his personal profile. He can also repeat the choice of a different cloud to start his experience of personalized image service again.

Figure 2 displays some personalization clouds for illustration. We could see that each cloud has a cohesive topic, such as database, music, food and physics, etc.

4.2 Performance evaluation

4.2.1 Data preparation

We asked several users to make manual annotation of sentences. When doing annotation we keep the users blind of the images to avoid situations where images have inadequate or improper text-information.

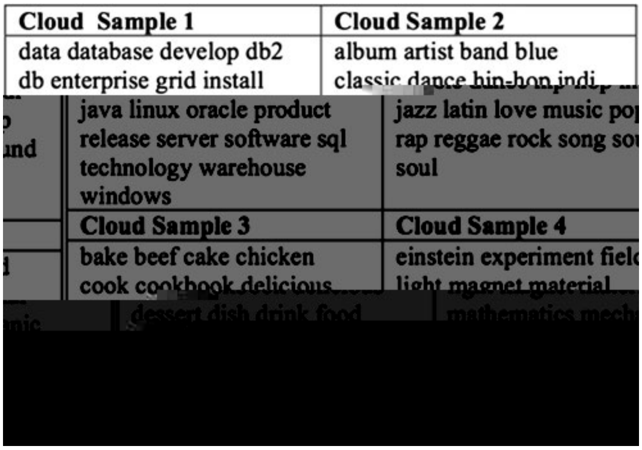

Fig. 2 Personalization sky with a variety of personalization clouds. A user could simply select a cloud as his search profile

Only with the annotated sentences, we could tune the parameters of the local and global models, compare their ranking performance and eventually apply the models in real-world system, which is used to speed up human annotation and reduce volunteers' efforts. We design the process of sentence annotation for a query as follow.

When a user triggers an annotation process, a pair of <query, cloud> is generated by the user accordingly, i.e., a user query and a cloud the user picked as his/her profile. It should be realized that the annotation is actually oriented to such a pair, other than a user query alone.

For each pair to be annotated, a total number of 10 resulting events are downloaded from Flickr [35] and returned to the user.

For an event, each scoring model returns only its first 5 images with highest salience score. The image's sentences from all models are merged before returning to the user for annotation. For the sake of fairness, for any image's sentence, the user is always kept blind to which model has generated it. For each image's sentence, the user is asked to assign a relevance score. The score has three levels, namely 2, 1 and 0. No other scores are admitted during the course of annotating a sentence.

For the sake of annotation, we make a referential criterion for the volunteers about the relevance of a sentence.

- If the image's sentence is relevant to both user query and user profile, the score ought to be 2;
- If the image's sentence is only relevant to either user query or user profile, the score ought to be 1;
- Otherwise, the score ought to be 0, as the image's sentence is relevant to neither user query nor user profile.

When doing annotation, we tried to make the decision according to our personal viewpoints. It might be the case that two users with different profile clouds annotate a small number of identical sentences for their respective query. We specially extract such sentences from all annotated ones. It turns out that different users have typically assigned them with different relevance levels.

4.2.2 Comparison ranking accuracy

By the end of paper submission, we have successfully collected a total number of 55 annotated queries, an estimated number of 430 selected events with 4,343 images (here those events that are irrelevant to the query or with inadequate images are not aggregated).

Based on these images' sentences, we investigate totally four models of photo ranking for our task of personalized, respectively Naive method (results returned by the Flickr search engine and ranked by capture timestamp and the user query), the local matching model (LocalM), the global ranking model with the sentences represented in TFIDF vectors (GR-VSM), and the global ranking model with the sentences represented in pLSI vectors (GR-pLSI).

For naive model, we simply extract the images' sentences and then send them to users for annotation. The 3 other models are developed after preprocessing each result page in following steps: 1) extracting textual content tagging, descriptions, comments from photos in photo html page body; 2) removing stop terms and common comment words through a dictionary; 3) passing all terms through a word stemmer; and 4) splitting textual content into sentences. Besides, as for GR-VSM and GR-pLSI, we further take a special measure to build their respective similarity graph. In short, we only involve those images with sentences whose IF scores (see Eq. 2) are above an expected threshold in the similarity matrix, since other images are barely qualified without a significant term in their sentences. With such a pruning scheme, the scale of similarity matrix is reduced greatly and the convergence rate of score propagation is increased greatly.

Reader can get a comprehensive example of the comparison between the three ranking methods in Fig. 3.

We evaluate the ranking performance of different models with a popular metric, namely Normalized Discounted Cumulative Gain (NDCG). NDCG is a retrieval measure devised specifically for the evaluation of ranking performance [17]. It is suited to our task of sentence ranking, where it rewards relevant sentences that are top-ranked more heavily than those ranked lower. Specifically, for a given query q, the ranked sentences are examined in a top-down order, and NDCG is given by

$$N_q = M_q \sum_{j=1}^{K} \frac{2^{r(j)} - 1}{log(1 + j)} \tag{16}$$

Where M_q is a normalization constant to ensure a perfect ordering to have an NDCG value of 1; and each r(j) an

Fig. 3 Ranking example, with the user specified keyword 'restaurant', and profile represented by a cluster of words under the topic 'food', The first line is the original ranking (rank by upload time), the second line is the 4 images with the highest ranking score by LocalM, the third line is the 4 images with the highest ranking score by GlobalM, we can see that in GlobalM, the second image is given a higher score, even though it does not have a high score in the LocalM line

integer-valued relevance label (i.e., 2, 1, or 0 in our work) of the sentence ranked at the position j.

In our experiments, we calculate for each model the final NDCG accuracy averaged over all 55 queries and for each query further averaged over all first 10 events (remember, our concern is never the ranking of the events, but the rankings of the images within each event). We are especially interested in NDCG accuracy at the first 4 positions, since the snap view of an event is seldom composed of more than 4 photos (typically 2 or 3 ones) (Table 1).

The comparison results of four models are reported in Fig. 4. We can find that:

1. LocalM is significantly better than Naive in NDCG at all positions. The minimal increase is achieved by up to 17.7 % for NDCG@2 while the maximal increase is even up to 49.0 % at NDCG@4. On the average, LocalM relatively improve NDCG by 36.8 %. It verifies that the users prefer personalized rankings for their relevance judgment in web search, and the local matching model could really retrieve many sentences by checking the presence of 3 personalization factors.
2. The global ranking model further performs much better than the local matching model. Comparing to LocalM, NDCG of GR-VSM is increased by 8.4 % averagely and that of GR-pLSI is increased by 10.4 % averagely. Overall, the average improvement is 9.4 %. It benefits from the fact that the two global models conduct an iterative propagation of salience scores among all input sentences. With this, those good candidates of photo sentences are distributed smoothly upon the underlying graph and their salience scores manifest both a local and global consistency. With such consistency, false positives (i.e., the proper sentences) and false negatives (i.e., the improper ones) are drastically reduced at the same time.
3. GR-pLSI further beats GR-VSM in terms of NDCG averagely with a relative increase by 1.9 %. It is fully owing to the fact that GR-pLSI could find many more latent photo's text materials than GR-VSM. To see this, we make a finer-grain statistical analysis on the distribution of annotated sentences with the result reported in Table 2

In Table 2, we divide all human-tagged sentences (summative information about the images, i.e. tagging and description of the uploader) into four groups: the sentences containing both cloud term and query term (shortly referred to as Both), ones containing either could term (Cloud) or query term (Query), and ones containing neither cloud term nor query term (Non). We further care about how many

Table 1 NDCG scores of different models

Model	NDCG@1	NDCG@2	NDCG@3	NDCG@4
Naive	0.6132	0.5511	0.5143	0.4844
LocalM	0.7221	0.7322	0.7543	0.7613
GR-VSM	0.7755	0.8026	0.8110	0.8237
GR-pLSI	0.7792	0.8113	0.8323	0.8544

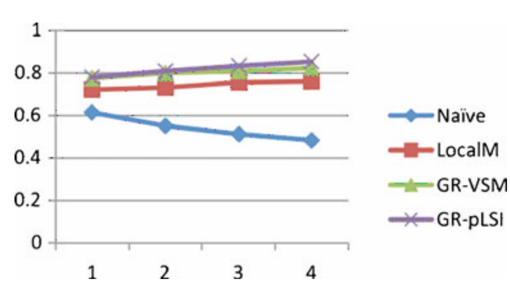

Fig. 4 NDCG comparison of 4 models for sentence ranking. This figure is a visualization of the data in Table 1

Table 2 The distribution of annotated sentences

Overall	0	1	2
Both	0.0124	0.0263	0.3538
Cloud	0.0046	0.1459	0.012
Query	0.0121	0.3139	0.0502
Non	0.0542	0.0126	0.002

sentences are respectively annotated as 0, 1, and 2 in each group.

From Table 2, we can conclude that:

1. As a whole, the four models retrieve the Both sentences the most, accounted for up to 39.25 %. It is not surprised, since Both sentences surely embody the most amount of personalization evidences. It is also observed that a considerable number of Both sentences are originated by LocalM model, where these sentences could obtain a relatively higher score by satisfying both relevance factor and personalization factor.
2. On the whole, the sentences annotated with level 2 account up to by 41.80 %. Only a very small percentage comes from Origin. In fact, most of Original photo rankings are only judged as level 1. In addition, there are practically a very small number of sentences are annotated as 0. Just for this reason, we have no need to investigate the retrieval accuracy of 4 models on Mean Average Precision (MAP), another popular ranking measure in IR domain.
3. It is important that latent image's sentence account for as much as 6.24 % in Level 2 sentences. The latent texts include Cloud, Query and Non sentences. Note that Non sentences tagged with Level 2 only account for 0.20 %. That is, most of latent photos' text contain at least one query term or at least one cloud term.
4. It is found that a great majority of latent sentences are generated from the two global models, namely GR-VSM and GR-pLSI. Further, GR-pLSI retrieves many more such sentences than GR-VSM. Specifically, the ratio of latent sentences is 17.21 % in GR-pLSI, but only 11.77 % in GR-VSM. It attributes from the fact that GR-pLSI further considers latent semantics in similarity evaluation of two sentences quite unlike GR-VSM.

4.3 Efficiency analysis

As a practical system, we need to consider the efficiency issue of our algorithm. In fact, we have considered two main points in realizing a fast service of personalized image ranking. One is convergence rate of the two global ranking models and the other is download speed of search results where there are a lot of photos.

We find that GR-VSM achieves a rapid convergence, mainly due to the fact that it is based on a sparse similarity graph. Motivated by this, we perform a pruning on the similarity matrix of GR- pLSI. Recall that there is a dense matrix in GR-pLSI. We throw away 30 % of small elements at each row (i.e., column) on this matrix, since a majority of elements with smaller weights tend to be noise. The ratio of 30 % is carefully tuned with a proper tradeoff between the efficiency and the effectiveness by repeating the trials.

Finally, on an Intel(R) Core(TM) 2 Duo CPU 2.20GHz class machine with 2.0 GB RAM, with 2M network bandwidth, our algorithm totally takes about 0.7 s to answer an offline query and 40 s to answer an online query. As seen, the great majority of response time is taken to download search results (photo events with a lot of photos). We can see 0.7 s convergence is quite fast compared to the download speed, and usually it takes less than 7 iterations before convergence.

4.4 User study

Lastly, we conduct an exploratory user study to verify the promise of personalized image ranking in image browsing. For that, we build a questionnaire with the experiment data and encourage users/colleagues to take active part in our survey. On the whole, most users realize and agree that personalized image ranking is a useful service for them to understand the contents of the events more quickly, and they would also like to see that the existing image sharing websites, such as Flickr [21] and Google+ [22], could provide such an add-on feature for them.

5 Discussion

In this section, we would like to discuss some special points about our proposed models for personalized image ranking in event browsing and the practical system implementation with a better personalization service.

Take a closer look at our global ranking framework in Eq. 9. It could be argued that such a framework actually encapsulates three progressive ranking models for select an event's images. The three-level models could be specified as

- $G(t+1) = U \cdot G(t)$
- $G(t+1) = \alpha U \cdot G(t) + \beta RF$, $(\alpha + \beta = 1)$
- $G(t+1) = \alpha U \cdot G(t) + \beta RF + \gamma PF$, $(\alpha + \beta + \gamma = 1)$

The first model could be used to locate those images that are good summary to the content of the event. The second model could be used to find those images whose sentences are better than the ones from the existing search services. The final model can be used to determine the best images as done in our work.

For simplicity, both the query issued by users and the ranks of events from search engines remains unchanged exactly. But in fact, we could have a more thorough personalization service for users. In specific, we could conduct a personalization re-ranking ranging from query level, image level to event level.

Our global ranking model could elegantly be fit for all three levels just using the same formulation. To do this, we could first build the respective G, RF and PF at a level, and then perform similarity score propagation at the corresponding level according to Eq. 9. We remain this work in the future.

6 Conclusion

Considering the fact that one hundred persons has one hundred information needs and images usually contains a variety of topical aspects explicitly or implicitly, in this work, we initiate an idea of generating personalized image ranking in event browsing.

For this goal, we propose a unified framework to determine the most proper image as the personalized representation of the event. The basic idea is that an image's textual information (sentence) is qualified with a considerable relevance to not only user query but also user profile. Under this framework, we develop two consecutive models to perform a sentence ranking for the task of personalized image ranking, one is the local matching model and the other is the global ranking model.

The local model takes the local presence of personalization terms as explicit clues and targets to select a very few salient images to constitute a resulting rank. The global model encapsulates the outputs of the local model as its initial values and makes salience scores consistently smooth with respect to all images' sentences in events returned for a query.

It is important to emphasize that our global ranking model could pick out many latent images even though they contain no keywords literally occurring in user query or user profile.

For evaluation, we build a sky of personalization clouds for users to determine their interests and implement a prototype system of personalized image for evaluation. Experimental evaluation indicates the effectiveness of our proposed models for spotting relevant images and also shows the advantages of the global model over the local model, these evidently demonstrate the prospect of personalized image ranking in real-world image event browsing applications.

Acknowledgments Our thanks to all EPCC staff and students for helping us with this paper. Our thanks to FCST-12 for giving us a chance to convey this idea to a larger group. Yao Shen is the corresponding author.

References

1. Sullivan D (2004) The older you are, the more you want personalized search. http://searchenginewatch.com/article/2064258/The-Older-You-Are-The-More-You-Want-Personalized-Search
2. Volokh E (2000) Personalization and privacy. Commun ACM 43(8):85
3. Chirita PA, Firan C, Nejdl W (2006) Summarizing local context to personalize global web search. In: Proc. of CIKM 2006, pp 287–296
4. Chirita PA, Costache S, Nejdl W, Handschuh S (2007) P-Tag: large scale automatic generation of personalized annotation tags for the web. In: Proc. of WWW 2007, pp 845–854
5. Sugiyama K, Hatano K, Yoshikawa M (2004) Adaptive web search based on user profile constructed without any effort from users. In: Proc. Of WWW 2004, pp 675–684
6. Xu SL, Bao SH, Fei B, Su Z, Yu Y (2008) Exploring folksonomy for personalized search. In: Proc. of SIGIR 2008, pp 155–162
7. Teevan J, Dumais S, Horvitz E (2005) Personalizing search via automated analysis of interests and activities. In: Proc. of SIGIR 2005, pp 449–456
8. Haveliwala T (2002) Topic-sensitive PageRank. In: Proc. of WWW 2002, pp 517–526
9. Qiu F, Cho J (2006) Automatic identification of user interest for personalized search. In: Proc. of WWW 2006, pp 727–736
10. Chirita PA, Firan CS, Nejdl W (2007) Personalized query expansion for the web. In: Proc. of SIGIR 2007, pp 7–14
11. Noll MG, Meinel C (2007) Web search personalization via social bookmarking and tagging. In: Proc. of ISWC 2007, pp 367–380
12. Russell BC, Torralba A, Murphy KP et al (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vision 77(1–3):157–173
13. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project
14. Nie L, Davison B, Qi X (2006) Topical link analysis for web search. In: Proc. of SIGIR 2006, pp 91–98
15. Jeh G, Widom J (2003) Scaling personalized web search. In: Proc. of WWW 2003, pp 271–279
16. Hentschel C, Automatic image annotation for tag-based image retrieval
17. Li W, Duan L, Xu D, Tsang IWH (2011) Text-based image retrieval using progressive multi-instance learning. In Computer vision (ICCV), 2011 IEEE international conference on IEEE, pp 2049–2055
18. Zhu G, Yan S, Ma Y (2010) Image tag refinement towards low-rank, content-tag prior and error sparsity. In Proceedings of the international conference on multimedia. ACM, pp 461–470
19. Liu D, Hua XS, Wang M, Zhang HJ (2009) Boost search relevance for tag-based social image retrieval. In Multimedia and expo, 2009. ICME 2009. IEEE international conference on IEEE, pp 1636–1639
20. Yang K, Wang M, Hua X-S, Zhang H-J (2001) Tag-based social image search: toward relevant and diverse results. In Social media modeling and computing. Springer, London, pp 25–45
21. Flickr API. http://www.flickr.com/services/api/. Accessed 30 July 2012
22. Google+ API. https://developers.google.com/+/api/. Accessed 30 July 2012

23. Shokri M, Tizhoosh HR, Kamel M (2006) Reinforcement learning for personalizing image search
24. Fan J, Keim DA, Gao Y, Luo H, Li Z (2009) JustClick: personalized image recommendation via exploratory search from large-scale Flickr image collections. IEEE Trans Circ Syst Video Technol
25. Turpin A, Tsegay Y, Hawking D, Willams HE (2007) Fast generation of result snippets in web search. In: Proc. of SIGIR 2007, pp 127–134
26. Chirita PA, Firan CS, Nejdl W (2007) Personalized query expansion for the web. In: Proc. of SIGIR, pp 7–14
27. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 2290:2319–2323
28. Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
29. Yu L, Liao X, Jin H, Jiang W (2011) Integrated buffering schemes for P2P VoD services. Peer-to-Peer Netw Appl 4(1): 63–74
30. Liao X, Jin H, Yu L (2012) A novel data replication mechanism in P2P VoD system. Future Gener Comput Syst 28(6): 930–939
31. Yang S, Jin H, Li B, Liao X, Yao H, Huang Q, To X (2009) Measuring web feature impacts in peer-to-peer file sharing systems. Comput Commun 32(12):1418–1425
32. Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22
33. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proc. of SIGIR 1999, pp 35–44
34. Del.icio.us. http://del.icio.us. Accessed 06 Apr 2009
35. CLUTO. http://glaros.dtc.umn.edu/gkhome/views/cluto/. Accessed 30 July 2012