Learning to Personalize Trending Image Search Suggestion*

Chun-Che Wu [1], Tao Mei [2], Winston H. Hsu [1], Yong Rui [2]
[1] National Taiwan University, Taipei, Taiwan
[2] Microsoft Research, Beijing, China
kenwu0721@gmail.com; {tmei, yongrui}@microsoft.com; whsu@ntu.edu.tw

ABSTRACT

Trending search suggestion is leading a new paradigm of image search, where user's exploratory search experience is facilitated with the automatic suggestion of trending queries. Existing image search engines, however, only provide general suggestions and hence cannot capture user's personal interest. In this paper, we move one step forward to investigate personalized suggestion of trending image searches according to users' search behaviors. To this end, we propose a learning-based framework including two novel components. The first component, i.e., trending-aware weight-regularized matrix factorization (TA-WRMF), is able to suggest personalized trending search queries by learning user preference from many users as well as auxiliary common searches. The second component associates the most representative and trending image with each suggested query. The personalized suggestion of image search consists of a trending textual query and its associated trending image. The combined textual-visual queries not only are trending (bursty) and personalized to user's search preference, but also provide the compelling visual aspect of these queries. We evaluate our proposed learning-based framework on a large-scale search logs with 21 million users and 41 million queries in two weeks from a commercial image search engine. The evaluations demonstrate that our system achieve about 50% gain compared with state-of-the-art in terms of query prediction accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*

General Terms

Algorithm, Experimentation.

* This work was performed when the first author was visiting Microsoft Research as a research intern.

SIGIR '14, Gold Coast, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609569.

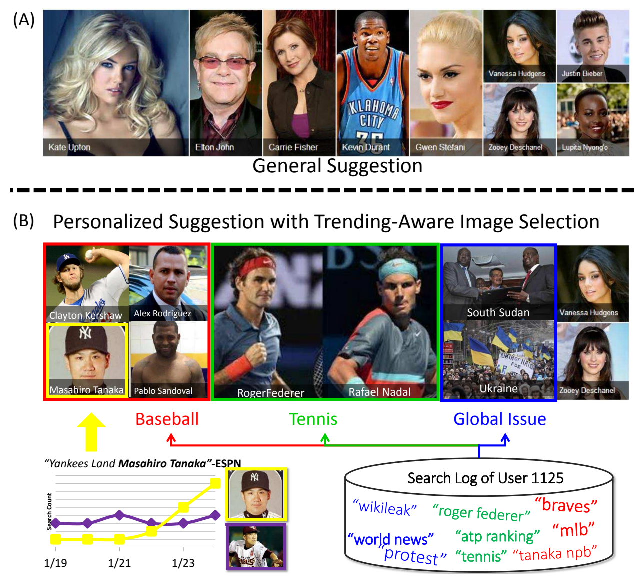

Figure 1: Trending image search suggestion: (A) most existing search engines provide general (or locality-based) trending image searches, without considering user's preference; (B) we visualize the personalized suggestion of trending image searches by combining the trending textual queries and their associated representative images. In (B), the suggested textual-visual queries not only are trending (bursty) and personalized to user's search behaviors, but also provide the compelling visual aspect of these queries. The image queries bounded with different colors correspond to different topics (interests) mined from the search history of current user #1125 (better viewed in color).

Keywords

Image search, query suggestion, personalization, matrix factorization, trending search detection.

1. INTRODUCTION

Today's Web interfaces for information seeking have predominantly focused on facilitating the "search" behavior when users are exploring the Web. Recent studies have found that the typical information retrieval activity by typing short queries into search boxes are not enough to satisfy users' search intent. This is particularly true for image search, where a significant part of queries is from recommendation (e.g., inspired from browsing a webpage, triggered by a popular social event, etc.). A new search paradigm—exploratory search—which facilitates user's searching experience with browsing (i.e., recommendation) to foster seeking and investigation, is therefore emerging.

This paper is concerned about providing image search engine users with trending queries (via a combined textual-visual form), which are highly related to users' personal interest, to further enhance their exploratory search experience (as shown in Fig. 1). Knowing the trends that most users are searching for, not only can increase user engagement by pro-actively suggesting trending news [1], but also serve for the detection of socially trending events, e.g., detecting influenza epidemics in a short latency [11].

The query log data in a search engine provide rich information of what users are interested [4, 24]. When large amount of users search for similar information, there emerges a trend, ranging from worldwide breaking news to celebrity gossips. Many search engines have already provided related services. For example, Bing Popular Now detects general trending searches [1]. To tailor the trending topics for personalization purpose, Yahoo! Trending Now [2] and Google Trends [3] suggest the trending searches based on user locality information. Google also provides search trends analytic in a set of pre-defined categories [4]. Besides, Bing serves with trending image searches [5], which can dig out different kinds of trends from image search log. Please see Figure 1(A) for an example. However, the suggested queries from most existing online services are not tailored to user preferences.

There has been extensive research on trending search. Dong *et al.* investigate the detection of trending searches [10]. Bawab *et al.* introduce the locality-aware trending search detection in Yahoo! [1]. Instead of using simple search statistics, Golbandi *et al.* propose a linear auto-regression model to shorten the latency of trending detection [12]. Meanwhile, the trending queries can be also mined from social networks [7, 28].

It can be observed that there are two unsolved issues in existing applications. First, the suggested trending queries from existing search engines are not personalized or only related to locality while ignoring individuals' search preferences [1]. However, when browsing trending queries, users are always preferring the results tailored to their personal interests. As shown in Fig. 1, user #1125 is interested in "Golden Globes" and "Baseball" related topics according to his/her search behaviors. It is therefore more reasonable to suggest related trending searches to this user. Second, besides personalized suggestion of textual queries, the selection of representative image for each trending search is able to facilitate user's understanding of the bursty events. However, this is often overlooked in existing works. Though some search engines, such as Google and Bing, have provided trending image search, the selection of representative images highly depends on editorial effort and does not consider visual aspect.

Motivated by the above observations, we propose a novel learning-based framework for personalized suggestion of trending image search by mining image search log data. The suggested textual-visual queries, each represented by the combined form of a textual query and its associated representative image, are not only socially trending due to their burst nature, but also personalized to users' search behaviors. The proposed framework consists of two components. The first component includes a novel trending-aware weight-regularized matrix factorization (TA-WRMF) algorithm which suggests the personalized trending textual queries by formulating the process as one-class collaborative filter problem. The second component selects the most representative and trending image for each textual search. Specifically, we investigate how to extract discriminative features from image search log, including popularity, visual consistency, as well as trending factor. The suggested textual-visual trending image search queries can improve user engagement and therefore foster user's exploratory search experience.

In summary, we make the following contributions:

- We have proposed a learning-based framework to personalize trending image search suggestion by mining users' click-through data. The framework consists of two novel components, i.e., TA-WRMF for personalized trending query suggestion and representative trending image selection.
- We conduct comprehensive evaluations on a large-scale search log data from a commercial search engine and validate the effectiveness of our framework.

The remaining of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the framework. Section 4 and 5 analyze the dataset and experiments, respectively, followed by the conclusions in Section 6.

2. RELATED WORK

The research related to trending image search suggestion includes trending search detection, query suggestion/auto-completion, and representative image selection.

Trending search detection. Most previous research on trending search detection depends on the statistics of search queries. Dong *et al.* detect the trending searches by the likelihood differences of query search in different time windows [10]. Bawab *et al.* extend the idea and further introduce the locality-aware likelihood to derive local trending searches [1]. Chen *et al.* propose classification-based method to distinguish *stable*, *one time burst*, *multiple time burst*, and *periodic* queries [9]. Golbandi *et al.* propose a linear auto-regression model to shorten the latency of trending search detection [12]. In addition to the search trends, other research works detect the trends from blogosphere [19], online news [8], and social media, such as Twitter [7] and Youtube [28]. To the best of our knowledge, there is few attempt for personalizing trending search suggestion with representative images.

Query suggestion and auto-completion. Query suggestion and query auto-completion are two closely related research topics. Query suggestion returns a list of ranked queries with respect to a given query, while query auto-completion returns queries with the first few letters in common with user's input. In general, there are two categories of methods. The first branch is session-based methods, where the basic idea is that the queries co-occurring in the same sessions are the candidates for suggestion [2, 14]. The second branch is based on document-click relationship [6, 23]. The query-document bipartite graph is built from search log. Two queries are assumed similar if the corresponding clicked URLs are identical [20].

However, previous research only focuses on the condition with input query and does not learn any latent topics or consider personalization. Ma *et al.* attempt to jointly learn the latent topics of users, queries, and URLs via collaborative filtering (CF) and define the query similarity on the learned latent space [22]. However, their approach needs the For personalization, Shokouhi *et al.* propose to build personal profile considering long-term search history, short-term history (previous searches in the same session), age, gender,

[1] http://www.bing.com/
[2] https://www.yahoo.com/
[3] http://www.google.com/trends/hottrends
[4] http://www.google.com/trends/explore
[5] http://www.bing.com/images

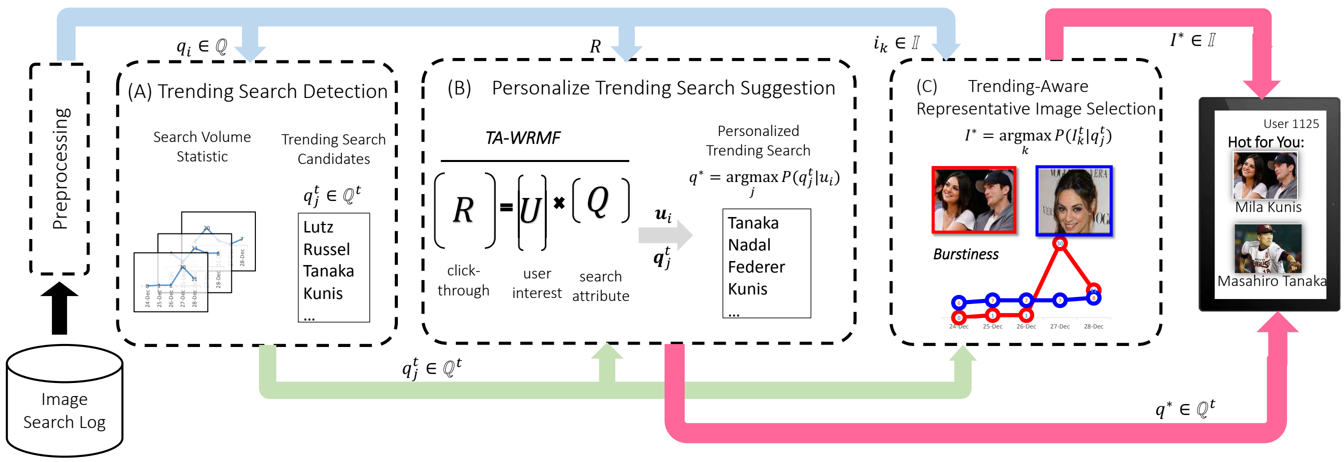

Figure 2: The framework of our proposed personalized trending image search suggestion. The system consists of three key components: (A) trending search detection, (B) personalized trending search suggestion, and (C) trending-aware representative image selection.

and region [31]. However, their approach only leverages personal search history which provides limited knowledge for understanding users.

Representative image selection. The criteria of representative image selection can be summarized as: (1) visual consistency—the selected image should be visually similar to many others [16, 33, 35]; (2) visual orthogonality—two similar images should not be included [32]; and (3) relevance—how relevant the image is to this query, which can be measured by popularity (i.e., the number of queries for this image) [16] or textual-visual similarity [33]. In addition, Kennedy *et al.* adopt a time information as criteria [16].

In summary, we move one step forward in this paper for *personalized* trending image search suggestion. Different from query suggestion/auto-completion, we do not need any query as input. For representative image selection, we extract features from image search log rather than image collection and further investigate the trending-aware time feature.

3. APPROACH

In this section, we introduce the proposed framework of learning to personalize trending search suggestion. The framework is shown in Figure 2. After preprocessing, trending image search queries are detected based on the statistics of queries from the image search log, as shown in Figure 2(A) (Section 3.1). Then we learn the user interests ($\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_i \cdots \mathbf{u}_{|\mathbb{U}|}]$) and trending search attributes ($\mathbf{Q} = [\mathbf{q}_1 \cdots \mathbf{q}_j \cdots \mathbf{q}_{|\mathbb{Q}|}]$) simultaneously via matrix factorization (MF) on the click-through matrix ($\boldsymbol{R}$), as shown in Figure 2(B). To leverage the information from common searches without sacrificing the accuracy of trending searches, we propose the trending-aware weighted regularized matrix factorization (TA-WRMF). The personalized suggestion (i.e., re-ranked list of trending searches) depends on the inner product of the learned user interests and trending query attributes (Section 3.2). Moreover, to better visualize the search trends, we select the most representative trending image for each suggested textual query, as shown in Figure 2(C). In addition to *relevance* and *visual consistency*, we propose *burstiness* as the criteria for trending-aware image selection (Section 3.3). The suggested personalized trending query is in the textual-visual form with both the textual query and its representative image. In the following sections, we will elaborate each component in details.

3.1 Trending image search detection

We start with trending search detection. To find buzzing search queries, we adopt similar strategy to [1]. First, only top frequent search queries (i.e., $10,000$ in this work) are kept and only their buzz score will be computed. The buzz score $BS(q_j)$ of the query q_j is defined by

$$BS(q_j) = \sum_{s=d-1}^{d-n} \frac{1}{d-s}\Big(P(q_j|\mathbb{Q}_d) - P(q_j|\mathbb{Q}_s)\Big), \tag{1}$$

where $P(q_j|\mathbb{Q}_d)$ is the likelihood of query q_j given the query set $\mathbb{Q}_d$ of day d. Instead of choosing max difference [1], i.e.,

$$\max_s\{P(q_j|\mathbb{Q}_d) - P(q_j|\mathbb{Q}_s)\}, \tag{2}$$

we adopt *weighted sum* to aggregate all the differences information within several days. The reason is that we prefer to detect the search with popularity rising dramatically at current moment. However, the max difference cannot distinguish this issue.

Next, we have to single out a representative search on behalf of a group of queries with similar semantics. Thus, the concept of generalized count is adopted and shown to be robust and efficient in [1]. That is, the query q_i gives a generalized count for q_j if q_i is a substring of q_j, e.g., "President Barack Obama" provides a generalized count for "Barack Obama." The buzz score of the search query q_j is then modified as in [1]:

$$BS'(q_j) = BS(q_j) \times log\big(1 + v(q_j, d) + v^*(q_i, d)\big), \tag{3}$$

where $v(q_j, d)$ is the count of q_j during day d and $v^*(q_j, d)$ is the generalized count of q_j during day d. With the help of generalized count, "Barack Obama" would be preferred, as "President Barack Obama" and "Barack Obama" have similar BS. In practice, the search queries with the top BS' can be sent to the editorial staffs to manually check the trending search quality [1].

3.2 Personalized trending search suggestion

To personalize search query suggestion for each user, it is reasonable to learn personal interest from user's search log. However, the search log data of individual users are usually sparse and thus might provide very limited personal knowledge. Leveraging users with similar search queries to expand personal interest is a feasible solution. Moreover, with the click-through data, we can assume that users are implicitly interested in the issued search queries. The non-clicked search queries might be uninterested or unknown. This scenario is similar to the one-class collaborative filtering (OCCF) problem, where only the positive data are provided and could be solved by matrix factorization techniques (MF) [25]. Here the major concern is whether the sparsity issue exists, which is a common problem to deal with search log data [3], and whether it will affect the performance. In this work, as the suggestion targets are trending searches, i.e., the searches would be issued by many users during a short period of time, there is no sparsity issue. Therefore, we can apply MF technique to solve the OCCF problem in this work.

As a result, inspired from [25], which is well-known for OCCF problem, we apply the weight-regularized matrix factorization (WRMF) as our model [6], which is formulated as follows:

$$J(\mathbf{U}, \mathbf{Q}^t) = \sum_{i,j} W_{ij}(R_{i,j} - \mathbf{u}_i^T \mathbf{q}_j^t)^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{Q}^t\|_F^2), \quad (4)$$

where $\mathbf{R}$ is the click-through matrix, $R_{ij} = 1$ indicates user u_i has issued query q_j in the training set, and (u_i, q_j) is called a positive pair. $R_{ij} = 0$ indicates user u_i has never issued query q_j in the training set, and (u_i, q_j) is called a negative pair. $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_i \cdots \mathbf{u}_{|\mathbb{U}|}]$ is the user matrix where $\mathbf{u}_i$ is the latent interest of user u_i. $\mathbf{u}_i$ is a z-dimensional vector, where z is the number of latent topics. $\mathbf{Q}^t = [\mathbf{q}_1^t \cdots \mathbf{q}_i^t \cdots \mathbf{q}_{|\mathbb{Q}^t|}^t]$ is the trending search matrix where $\mathbf{q}_j^t$ is the latent attributes of trending search q_j^t and is a z-dimensional vector. $\mathbb{Q}^t$ is the trending search set. Please note that we use superscript t to denote the detected trending image queries to differentiate from general queries. $\mathbf{Q}^t$ assigns the queries to the latent interests. $\mathbf{W}$ is the weight matrix which control the importance of positive pairs and is defined by

$$W_{ij} = \begin{cases} 1 & R_{ij} = 1 \\ W_{ij}^n & R_{ij} = 0 \end{cases} \quad (5)$$

where W_{ij}^n should be smaller than 1 because positive (u_i, q_j^t) pairs are more significant in the OCCF problem to avoid the imbalance issue. However, if only trending searches are used for user understanding, it is possible to suffer short information. The number of trending searches, i.e., $|\mathbb{Q}^t|$, is extremely small compare to the number of user $|\mathbb{U}|$. Therefore, the low-rank approximation might barely mine useful knowledge from it. The first term of the objective function $J(\mathbf{U}, \mathbf{Q}^t)$ in Equation (4) targets on modeling the two latent spaces of user and query, respectively. The similarity between two spaces are defined by inner product. The second term is to reduce the length of user and query factors, which decreases the model complexity to prevent data overfitting.

Another issue is how to use the information from both the trending and common searches (i.e., queries) as much as possible to learn the latent relationship between users and trending searches. q_i^t, $\mathbf{Q}^t$, and $\mathbb{Q}^t$ can be replaced by q_i, $\mathbf{Q}$, and $\mathbb{Q}$, respectively. $\mathbb{Q}^t = \mathbb{Q}^t \cup \mathbb{Q}^c$, where $\mathbb{Q}^c$ is the common search set, and q_j^c ($q_j^c \in \mathbb{Q}^c$) represents a common search query. Unfortunately, this setting neglects the importance of trending searches, which is the real suggestion target. Treating each common search equally to each trending searches might sacrifice the accuracy of trending searches due to the imbalance issue ($|\mathbb{Q}^c| \gg |\mathbb{Q}^t|$).

TA-WRMF. Motivated from the above observations, aiming at leveraging the information from common searches without sacrificing the accuracy of trending searches, we propose a novel *trending-aware weighted regularized matrix factorization* (TA-WRMF). The idea is similar to solve the OCCF problem, except that we apply the weighted regularization twice—once for the imbalance of positive/negative pairs and the other for the imbalance of trending/common searches. Then, we define the weighted matrix ($\mathbf{W}$) as:

$$W_{ij} = \begin{cases} W_P & R_{ij} = 1 \wedge q_j \in \mathbb{Q}^t \\ 1 & R_{i,j} = 1 \wedge q_j \in \mathbb{Q}^c \\ W_N & R_{i,j} = 0 \wedge q_j \in \mathbb{Q}^t \\ W_N \times W_{i,j}^n & R_{ij} = 0 \wedge q_j \in \mathbb{Q}^c, \end{cases} \quad (6)$$

where $W_P > 1$ is to increase the importance of positive (u_i, q_j^t) pairs, and $W_N < 1$ is to control the importance of negative (u_i, q_j) pairs. $W_{i,j}^n$ is only applied on negative (u_i, q_j^c) pairs. With such a design, the pairs containing q_j^t are more important/positive pairs. The imbalance issues are then resolved. As a result, we can not only leverage the information of common searches but also ensure the accuracy of trending searches.

Scalability and Sampling. However, one of the drawbacks of using all the queries is the scalability issue. The complexity would be $\mathcal{O}(|\mathbb{U}| \times |\mathbb{Q}| \times z \times T)$ where T is the number of optimization iterations. The time consumption would be intolerable due to enormous number of queries in real world. To reducing the complexity with competitive performance, we adopt user-oriented sampling strategy to approximate $W_{i,j}^n$. The number of negative sampled pairs N_i^{ns} is proportional to the number of positive pairs N_i^p ($N_i^p = \sum_j R_{ij}$) of a user u_i. $m = N_i^{ns}/N_i^p$ is the ratio of negative sample over positive pairs. Then, we have

$$W_{ij}^n \approx \frac{N_i^{ns}}{N_i^n} = \frac{m \times N_i^p}{N_i^n}, \quad (7)$$

where N_i^n ($N_i^n = |\mathbb{Q}| - \sum_j R_{ij}$) is the number of negative pairs of u_i. Then, the complexity becomes $\mathcal{O}\big(N^p(1+m+|\mathbb{Q}^t|)\times z\times T\big) \approx \mathcal{O}(|\mathbb{U}| \times z \times T)$, where N^p is the total number of positive (u_i, q_j) pairs in the training data [7]. To solve Eqn. (4), stochastic gradient descent (SGD) is further applied for better efficiency [17], and the iteration formulations are

$$\begin{aligned} \mathbf{u}_i &= \mathbf{u}_i + \alpha\big(-W_{ij}(\mathbf{R}_{ij} - \mathbf{u}_i^T \mathbf{q}_j)\mathbf{q}_j + \lambda \mathbf{u}_i\big), \quad (8) \\ \mathbf{q}_j &= \mathbf{q}_j + \alpha\big(-W_{ij}(\mathbf{R}_{ij} - \mathbf{u}_i^T \mathbf{q}_j)\mathbf{u}_i + \lambda \mathbf{q}_j\big), \end{aligned}$$

where α is the learning rate and λ is the regularization term [8]. At training phase, we separate part of (u_i, q_j) pairs from training data to form a validation set. When the cost function $J(\mathbf{U}, \mathbf{Q}^t)$ does not decrease for c ($c = 20$) continuous iterations on validation set, convergence is claimed. The overall optimization process is summarized in the Algorithm 1. In short, our proposed TA-WRMF adopts weighted regularization twice to solve the two imbalance issues and learns better latent structures of users and search queries.

3.3 Trending-aware representative image selection

The selected trending images are expected to facilitate the understanding of trending searches. Different from previous works mining from social image collections, we target at digging information from the commercial image search log. In this section, we investigate the widely adopted representative image selection features, i.e., *relevance* and *visual consistency* [16], as well as the features related to search popularity, i.e., *burstiness*.

[6] Many other proposed MF techniques to fit different rank-oriented optimization goals [18, 27, 30, 34] can also be adopted. But in this work, we aim to propose a unified strategy to solve the personalized trending search suggestion problem.

[7] $N^p = \sum_{i,j} R_{ij}$ and $N^p \propto |\mathbb{U}|$. m and $|\mathbb{Q}^t|$ are constants.

[8] λ in Eqn. (4) and α in Eqn. (8) are decided by cross-validation and both are set as 0.01 in the proposed TA-WRMF for all the experiments.

Algorithm 1 TA-WRMF

Input: A list of positive ($R_{ij}=1$) (*user, search*) pairs L including both trending and non-trending searches.

Output: $\mathbf{U}$, $\mathbf{Q}$

```
Initialize U, Q by uniform sampling from (−1, 1)
Append all negative (user, trending search) pairs to L
repeat
    for all (u_i, q_j) in L do
        if R(ij) == 1 then
            Randomly sample m pairs of (u_i, q_x), where (u_i, q_x) ∉ L
            Update m pairs of (u_i, q_x) by Eqn. (8)
        end if
        Update u_i, q_j by Eqn. (8)
    end for
until Converge.
```

Relevance. We assume that the *relevance* depends on the click count (c_{kj}) of an image (I_k) associated with the trending search (q_j^t). The higher c_{kj}, the more relevant I_k. Thus, the *relevance* $P_c(I_k|q_j^t)$ is decided by the likelihood of I_k given q_j^t as follows:

$$P_c(I_k|q_j^t) = \frac{c_{kj}}{\sum_l c_{lj}}. \quad (9)$$

Visual Consistency. It is important to determine the representativeness under the assumption that an image with more similar neighbors should have higher chance to be selected [16, 35]. The Bag-of-Visual-Word model on the Scale-Invariant Feature Transform (SIFT) features is applied to compute the visual similarity $Sim(I_k, I_l)$ of a pair of images (I_k, I_l) [21, 26]. Then, a random walk process is adopted to derive the visual consistency [13]. For each trending search q_j^t, each image I_k with $c_{kj} > 0$ is a node n_k and forms a image set $\mathbb{I}_j^t$. n_k has a directed edge $e_{k,l}$ to node n_l if I_l belongs to the k-Nearest Neighbor ($k = 5$) of I_k, denoted by $kNN(I_k)$. The weight of edge $e_{k,l}$ is given by $\frac{Sim(I_k,I_l)}{\sum_{I_s \in kNN(I_s)} Sim(I_k,I_s)}$. Therefore, the visual consistency $P_v(I_k|q_j^t)$ of I_k can be formulated as

$$P_v(I_k|q_j^t) = \Big(\alpha E + \frac{(1-\alpha)}{|\mathbb{I}_j^t|}\mathbb{1}^T\Big) P_v(I_k|q_j^t), \quad (10)$$

where E is the transition matrix consisting $e_{i,j}$, $\mathbb{1}$ is the vector of ones. The optimal $P_v(I_k|q_k^t)$ is the eigenvector with the largest eigenvalue.

Burstiness. The users not only care about the trending searches but also the reason why they are buzzing. As in Figure 2(C), people would issue "Mila Kunis" at the moment because they are interested in her gossip. Thus, in addition to returning the image with high relevance, it is intuitive to select the one with higher descriptivity of trending search. The image with bursty large amount of increasing clicks from the trending search is assumed with better descriptivity. Hence, we adopt the burstiness $P_b(I_k|q_j^t)$ of the image I_k given trending search q_j^t as the measurement of descriptivity, which is similar to Eqn. (1) and can be formulated as

$$P_b(I_k|q_j^t) = \sum_{s=d-1}^{d-n} \frac{1}{d-s}\Big(P(I_k|q_j^t,d) - P(I_k|q_j^t,s)\Big), \quad (11)$$

where $P(I_k|q_j^t, d)$ is the probability of image I_k being clicked given trending search q_j^t at day d. In this work, we focus on demonstrating the effectiveness of different features, especially the burst nature, i.e., *burstiness*. How to fuse the features for better performance will be covered in the future.

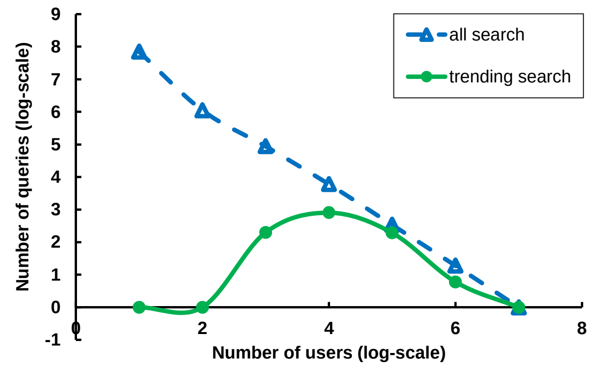

Figure 3: Distribution of users for each search query during 2012/11/01–2012/11/13. The figure aims at showing that the trending searches are issued many times and might not have cold-start issue as common searches, so TW-WRMF (CF-based method) can fit our problem.

4. QUERY ANALYSIS

We have conducted experiments (including trending search detection and personalize trending search suggestion) on a large-scale search log from a commercial image search engine. The image search logs were collected in the first two weeks of Nov. 2012. Each entry of the log contains five elements <$user$, $query$, URL, $time$, $nationality$>. The dataset contains 21 millions unique users, 41 unique millions queries, and 61 unique millions URLs. Only the search logs in US are used in our data.

Furthermore, in the raw search logs, there are many spam users who aim to promote specific searches to be detected as trends and continuously issue many queries in a short period of time. To reduce spam queries and users, we regard the users who issue more than 50 search queries in a single search session as spam users and remove their query logs. Two consecutive queries issued within a specific time threshold by a single user are considered in the same search session as defined by [5]. Here, the time threshold is set 30 min to define a single session.

Moreover, search queries with low frequency are barely considered trends and provide little information for latent topic learning. Thus, we remove the search queries below a certain threshold (i.e., 3 in this work). After spam users and low frequent searches removal, there are 15 millions unique users, 9 unique millions queries and 61 unique millions URLs in the final dataset.

One possible issue of collaborative filtering is the data sparsity (cold-start) problem. However, from Figure 3, trending searches have issued by many users, and thus might not suffer the cold-start problem. This statistic is supportive to collaborative filtering, especially our proposed TA-WRMF.

5. EXPERIMENTS

5.1 Settings

The trending search detection is based on Eqn. (1) and Eqn. (3). s is set 3, i.e., the trending strength is referred to the last three days. Searches with top 100 trending scores are claimed *trending* for evaluation. For the personalized suggestion, we will first conduct personalized suggestion on editor-labeled trending searches. Then, in the main experiments, trending searches claimed in the detection stage are directly put in the suggestion list for further personalized ranking without editor selection. There are three reasons for this

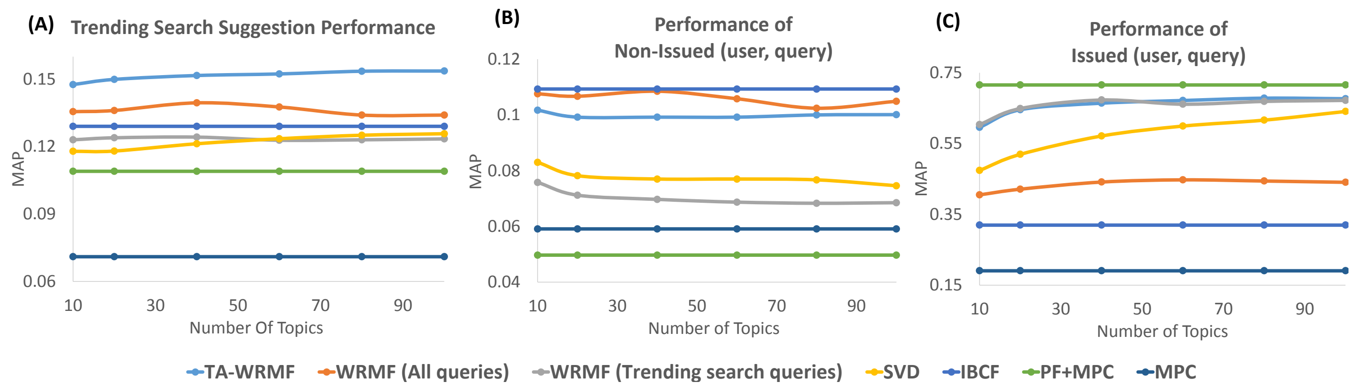

Figure 4: Performance in terms of MAP. (A) shows the overall trending search suggestion performance of all the baselines and our proposed TA-WRMF, which has the best MAP result. In (B) only the issued (u_i, q_j) pairs ($R_{i,j} = 0$) are possibly consider *relevant*, and TA-WRMF is also competitive. On the other hand, in (C), only the issued (u_i, q_j) pairs ($R_{i,j} = 1$) are possibly regarded as relevant. Our TA-WRMF approaches most closely to the upper bound achieved by $FP + MPC$ among the CF-based methods. In short, TA-WRMF has competitive performances in both conditions, and thus outperform other methods on MAP.

Table 1: Personalized suggestion evaluation protocol. For one set of dataset, we use 4-day image search log from which we detect the trending image searches and learn personal interests. Image search log of the 5-th day is used to form testing data. Then, one day is shifted to form another dataset. Such process is repeated to form more datasets.

day 1	day 2	day 3	day 4	day 5	day 6	day 7
training				testing		
	training				testing	
		training				testing

setting: 1) It is hard to manually label the trending searches during 2012 at current time; 2) It is closer to the real application scenario. For example, existing trending suggestion systems display at most 20 trending searches. It is therefore more convincing to enlarge the general trending list for further re-ranking; 3) The automatically detected trending searches also have similar properties, e.g., click-through rate, to the editor-selected. The framework should be applicable when more editor-selected trending searches are available.

Table 1 lists the personalized suggestion evaluation protocol. We simulate the suggestion evaluation based on evaluation of prediction problem. Assume the target is to suggest the trending searches of day d (e.g., day 4) to the users in day $(d + 1)$ (i.e., day 5 in the example). The users who have issued any trending searches in the day $(d + 1)$ will be considered as testing users. For each testing user, the corresponding trending searches issued by him/her in the day $(d + 1)$ are considered relevant.

Generally, the search logs from $(d - s)$ to day d (e.g., day 1 to day 4 in the example) are considered as training data. But for better experimental efficiency, in addition to testing users, only any other user who has ever issued any trending search at least generated three queries in the training data is regarded as training user. The query q_j issued by training user u_i is included as training query. After one set of experiment (day 1 to day 5), we shift one day (e.g., day 2 to day 6 in the Table 1) to have another training and testing data. Then, the evaluation process is repeated. Totally, we have 9 sets (2012/11/04–2012/11/12) in the experiments. For evaluation,

Table 2: Trending search detection results. Our weighted sum scheme outperforms the max difference [1] in terms of both MAP and Recall.

Method	MAP	Recall
max difference	0.3381	0.8778
Weighted sum	0.3482	0.9000

the mean average precision (MAP) is adopted:

$$\frac{\sum_{d=4}^{12} \sum_{u_i^t \in \mathbb{U}_d} \sum_{k=1}^{100} P(k) rel_{u_i^t}(k)}{100 \times 9 \times \sum_{l=4}^{12} |\mathbb{U}_d|}, \tag{12}$$

where $\mathbb{U}_d$ is the testing user set corresponding to the trending searches of day d, $rel_{u_i^t}(k)$ indicates whether the k-th trending search is relevant to the user u_i^t, and $P(k)$ is precision@k.

For trending search representative image selection, we focus on addressing the effectiveness of *burstiness*. Therefore, three features will be directly compared in the experiments without further fusion. The top 20 trending searches are selected from editor-labeled set during 2012/11/04–2012/11/13. The representative images are automatic generated by the corresponding features. Only the highest scored image is viewed as the representative image by each feature. Then, a user study is conducted to evaluate the results. The information why the trending searches were buzzing is manually searched by each subject and then aggregated as the reference to facilitate the subject to get familiar with the background of trending events. As a result, we invited six subjects (four males and two females) to score the *representativeness* of each image, i.e., the ability to visualize the reasoning of buzzing. Each user was required to score the selected 20 trending searches. In total, 60 images (20 queries $\times$ 3 representative images by each features) with the scores scale from 1-10 (the higher, the better) are annotated.

5.2 Baselines

Trending search detection. we compare the weighted sum detection results with maximum difference [1]. The ground truths are the editor-labeled trending searches during 2012/11/04–2012/11/12.

Trending searches suggestion. There are five baselines to compare with. Two of them are simply frequency-based methods and

Table 3: 5 nearest neighbors of trending searches in the latent space. We can see that the latent space learned from *TA-WRMF* is able to capture the semantic meanings in the real world.

Trending Search	Nearest Neighbors
katie price wedding dress train	*gwen stefani wedding dress, celine dion wedding dress, sofia coppola wedding dress, mia farrow wedding dress frank sinatra, whitney houston wedding dress bobby brown*
animals wearing glasses	*baby spectacled leaf monkey, facts about baby monkeys, dog wearing glasses, baby leaf monkey for sale, tiger wearing glasses*
lichtenstein castle	*falkenstein castle, french castles, free castle screensavers, heidelberg castle, linderhof*

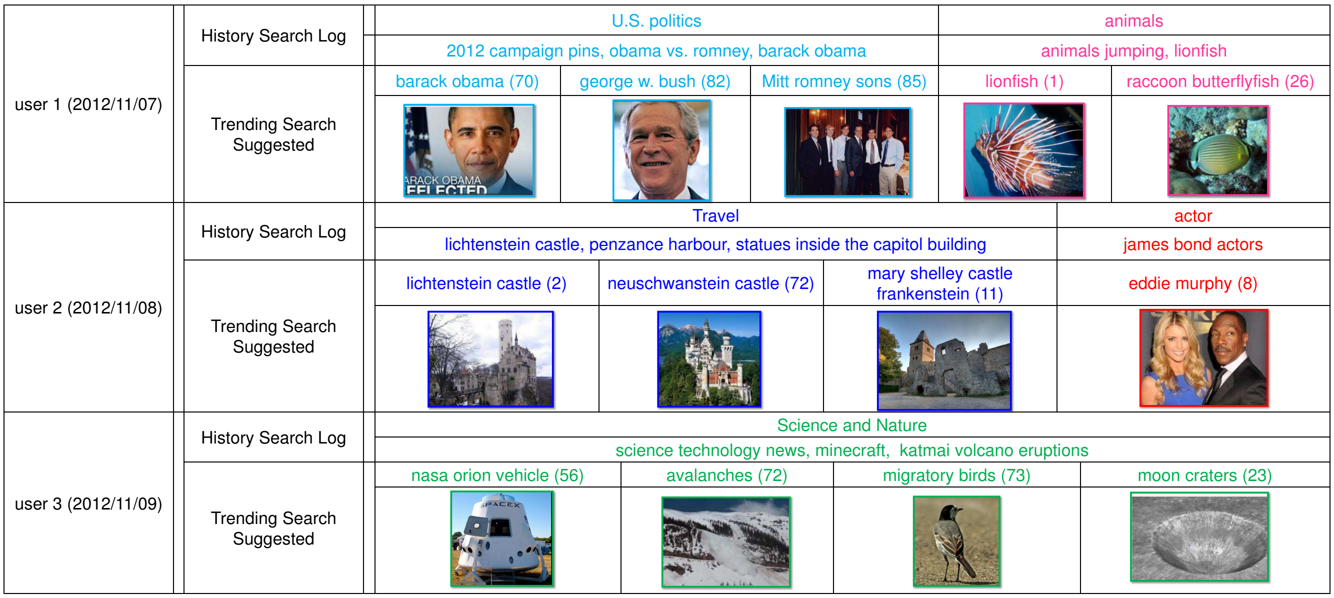

Figure 5: Examples of personalized suggestions. For each user, the top 20 in the ranked list are suggested. We select some suggested trending searches (with images) and issued queries in the search logs to show the semantic consistency between the textual and visual queries. Each color represents a topic (e.g., U.S politics) mined from the image search logs. Each number in the parenthesis following a suggested trending search means the original ranking by MPC. For example, "neuschwanstein castle" is ranked 72-th by MPC, but is promoted to top-20 by the proposed TA-WRMF.

the other three are CF-based methods. The baselines are introduced as follows.

- **Most Popular Candidates (MPC)** [2]. This is the frequency-based general suggestion without personalization. We commonly suggest the trending searches ranked by trending scores to all the users.

- **Personal Frequency (PF) + MPC**. Inspired by [31], we directly represent the personal interests by his/her past search history. That is, the search with higher frequency, or called *issued*, is ranked higher. However, the coverage of PF is not enough. So, the overall personalized suggestion list depends on the linear combination of PF and MPC.

- **Item-Based Collaborative Filtering (IBCF)** [29]. For each item (search) is normalized L_1 norm $\bar{\mathbf{R}}_{ij} = \frac{\mathbf{R}_{ij}}{\sum_i \mathbf{R}_{ij}}$, and then L_1 distance is applied. So, the suggestion score for trending search q_j^t to user u_i is measured by

$$S_{i,j} = \frac{\sum_x sim(q_j^t, q_x) \times \bar{\mathbf{R}}_{ij}}{\sum_l sim(q_j^t, q_l)} \tag{13}$$

- **Singular Vector Decomposition (SVD)**. Assume all the negative samples as zero and have the same error weight, i.e., $W_{ij} = 1$ for all positive (u_i, q_j) pairs and directly apply SVD on R to derive $\mathbf{U}$ and $\mathbf{Q}$. Note that in SVD, all queries are included in the training data.

- **Weighted Regularized Matrix Factorization (WRMF)**. WRMF is based on Eqn. (4). There are two settings for the search query set: 1) the query set only consists of trending searches ($\mathbb{Q} = \mathbb{Q}^t$), and W_{ij} depends on Eqn. (6); and 2) the query set consists all queries in the training data ($\mathbb{Q} = \mathbb{Q}^t \cup \mathbb{Q}^c$) and W_{ij} follows Eqn. (5), such that $W_{i,j}^n$ is also approximated by the sampling strategy described in Section 3.2.

5.3 Evaluation of trending image search detection

Table 2 shows the trending search detection results. First, the weight average scheme performs slightly better (on both MAP and Recall) than the max difference scheme as in [1]. The major reason is that the weighted average scheme prefers the searches with the statistical volume curves ascending in the convex shape. max difference, on the other hand, is neutral for the shape. However,

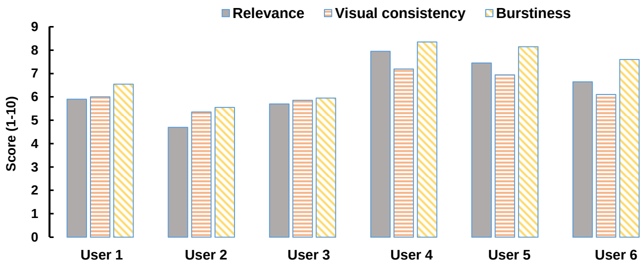

Figure 6: Evaluation of representative image selection. *Bustiness* has the best performance among all the users.

if the curve is in the concave shape, the search should be trending earlier rather than the current moment and this situation cannot be distinguished by max difference. Therefore, weighted average has slightly better performance. Besides, our approach achieves good coverage (Recall) of the trend detection. 90% editor-labeled trending searches during the period are discovered by our detection algorithm. Thus, if the personalized suggestion works well on the automatic detected trending searches, it should perform well given editor-labeled trending searches.

5.4 Evaluation of personalized trending image search suggestion

The personalized suggestion on the 20 editor-labeled trending searches improves the MAP from 0.37 (MPC) to 0.55 (TA-WRMF), which supports the intuition of personalized suggestion. Then, we focus on showing and analyzing the results on the 100 automatically detected search trends. From Figure 4 (A), we can see that TA-WRMF outperforms all other baselines with 15% and 50% relative improvements compared to CF-based methods and frequency-based method, respectively. As MPC does not use any personal information, it has the worst performance on recommendation. PF+MPC does improve a lot from MPC due to the use of personal log data. Furthermore, when we start to leverage the information from similar users, there is another improvement observed in the figure. Based on the performance from WRMF with all queries and SVD, we can see that the weighted regularized scheme does avoid imbalance issue in the OCCF problem. Moreover, leveraging more information in addition to the trending searches is helpful. Nevertheless, including more information from others queries might sacrifice the accuracy of trending searches. This should not happen because the trending searches are our suggestion targets. Therefore, the TA-WRMF increases the error weights of trending searches. The performance of TA-WRMF does confirm the observation with the improvement against others.

To further analyze the results, we separate the evaluations into two conditions: 1) for each user u_i, only the trending searches (q_j^t) not issued in the training data , i.e., $R_{i,j} = 0$, are regarded relevant candidates (called *non-issued relevant*), as results shown in Figure 4(B); and 2) conversely, for each user u_i, only the queries (q_j) issued in the training data, i.e., $R_{i,j} = 1$, are regarded relevant candidates (called *issued relevant*), as results shown in Figure 4(C). The objective is to deeply understand where the improvement comes from. Although there is an interference between these two conditions [9], the results provide much information.

[9] For example, if the ground truth contains a *issued relevant* and a *non-issued relevant*. If the *issued relevant* is ranked first, then the MAP upper bound of the second condition is only 0.5.

Figure 7: Example of trending-aware representative image selection. Four images are preferred by different methods (A-D), and (E) provides why "Don Lemon" was buzzing during the period. The Twitter feud between "Don Lemon" and "Jonah Hill" caused the burst and only the *burstiness* boosted the headshots of the two (A) to the top rank.

From Figure 4 (B)(C), we can see that if we just pull the issued searches to the top, then the performance of *issued relevant* is the best but the performance of *non-issued relevant* suffers severely, even worse than MPC (general suggestion). On the other hand, the CF-based models all outperform MPC on both conditions. This demonstrates the information from neighborhood users does help. While preferring the issued trending searches as *PF* does, the well-learned latent interests also have the ability to not only boost some *non-issued relevant* trending searches but also filter out some noisy issued trending searches. Among the CF-based models, SVD performs worse in both conditions due to imbalance of zeros. Thus, it does not fit well for the issued searches and the *non-issued relevant* are not boosted as well. WRMF (trending search only) might overfit the *issued* trending searches (q_j^t where $R_{ij} = 1$), so the *issued relevant* searches are not ranked high. WRMF (with all queries) and IBCF have similar performance. They are both good on *issued relevant* suggestion while overlooking the importance of trending searches. Hence, the accuracy of *issued relevant* trending searches are sacrificed and the overall performance is dropped. Finally, our proposed TA-WRMF achieves good performance for the *issued relevant* trending searches, and at the same time, some *non-issued relevant* trending searches are successfully ranked high. Therefore, the performance of *issued relevant* condition is also competitive [10]. In short, TA-WRMF has competitive performances under both conditions and thus has the best overall trending search suggestion performance.

[10] The slightly shy numbers might be due to the interference between two conditions.

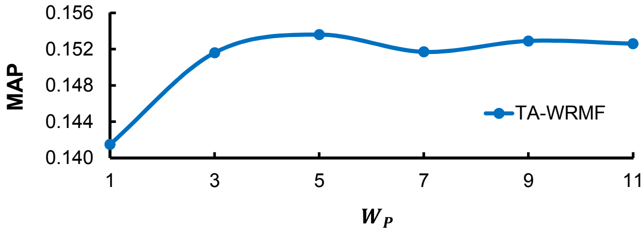

Figure 8: The sensitivity test results of W_P, i.e., the weight of positive pairs of users and trending searches, which show the effectiveness of taking more care on trending searches.

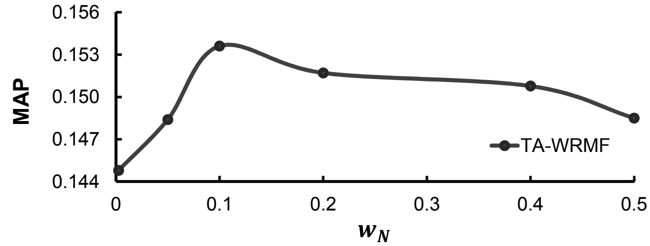

Figure 9: The sensitivity test results of W_N, i.e., the weight of negative pairs of users and search queries, which shows low W_N can save the imbalance issue.

Qualitative Analysis. Figure 5 and Table 3 provide some qualitative analysis. Table 3 shows that the latent space of search queries is quite meaningful by selecting some trending searches with their 5 nearest neighbors (NN) in the latent space. For example, "Katie price wedding dress train" is surrounded by many other wedding dress related queries. The neighbors "lichtenstein castle" are also some landmarks, especially castles. Thus, it is reasonable to expect the suggested trending searches are semantically consistent to user interest.

Figure 5 shows more trending searches suggestions (top 20). First, we can find some semantic consistency between personal search logs and suggested results. For example, user #2 concerns more about travel, related trending search, e.g., "neuschwanstein castle," is ranked very high. The *issued relevant* searches also achieve good performance, e.g., "Barack Obama" and "lichtenstein castle" for user #1 and user #2. Some *issued relevant* trending searches are boosted according to the personal interest as some science related trending searches is suggested to the user #3 who has generated plenty of science related queries.

5.5 Evaluation of trending-aware representative image selection

Figure 6 shows the evaluation results of representative image selection by user study. All the invited subjects agree with the effectiveness of *burstiness*, which achieves the best performance among all features. Figure 7 shows the example of representative image selection. During the testing time (Nov. 2012), "Don Lemon," an anchor of CNN, was a trending search target as he had a Twitter feud with "Jonah Hill," a Hollywood star, as shown in Figure 7(E). The *relevance* feature would pick the image with high popularity in the past, but does not concern the burst nature, i.e., what crowd really desires to see during the period, as shown in Figure 7(B). The *visual consistency* feature only considers the visual information, and thus cannot capture the character as well, as shown in Figure 7(C). On the other hand, the *burstiness* feature is able to detect recency, and rank the related images to the top more successfully, as shown in Figure 7(A).

5.6 Sensitivity test

We discuss the parameters in personalized trending search suggestion, including W_P, W_N ($m = 1$) in Eqn. (6). The sensitivity test results can be seen in Figure 8 and 9. Figure 8 shows the results of W_P, i.e., the importance of positive pairs of user and trending searches. The MAP reaches maximum when $W_P = 5$. Figure 9 shows the sensitivity test results of W_N, i.e., the weight of negative pairs of user and search. Because of the imbalance issue, smaller W_N should be better as shown in the figure. $W_N = 0.1$ reaches the best MAP. However, when we keep lowering W_N, the model would overfit the positive pairs of user and search, and thus the performance would be degraded.

6. CONCLUSIONS

To conclude, we have come up with a learning-based personalized trending image search suggestion framework, composed of two stages. First, we propose the trending-aware weight regularized matrix factorization (TA-WRMF) to learn the user interest by neighbor users with the help of auxiliary items, i.e., common searches. Then, different features, including *burstiness* are investigated for representative image selection of trending search. By the large-scale real commercial image search log, we demonstrate the effectiveness of our personalized suggestion model and trending-aware image selection features.

In the future, we will improve the personalized trending image search suggestion system in two-fold. First, we will leverage more information source to detect trending searches. For example, it is discovered that social media is able to shorten the trend detection latency to about 4.5 hours [15]). Second, we are trying to perform deeper understanding of the user. Long-term user preference (e.g., extending the training data to monthly scale) is aimed to be discovered to compensate the insufficiency of short-term interest learned. Besides, it is also desired to mine and aggregate other information such as user attributes or social networks for the cold-start users.

7. REFERENCES

[1] Z. Al Bawab, G. H. Mills, and J.-F. Crespo. Finding trending local topics in search queries for personalization of a recommendation system. In *Proceedings of SIGKDD*, pages 397–405, 2012.

[2] Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *Proceedings of WWW*, pages 107–116, 2011.

[3] R. Baraglia, F. Cacheda, V. Carneiro, D. Fernandez, V. Formoso, R. Perego, and F. Silvestri. Search shortcuts: A new approach to the recommendation of queries. In *Proceedings of ACM Recommender Systems*, pages 77–84, 2009.

[4] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of SIGIR*, pages 185–194, 2012.

[5] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: Model and applications. In *Proceedings of CIKM*, pages 609–618, 2008.

[6] I. Bordino, C. Castillo, D. Donato, and A. Gionis. Query similarity by projecting the query-flow graph. In *Proceedings of SIGIR*, pages 515–522, 2010.

[7] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of KDD Workshop on Multimedia Data Mining*, 2010.

[8] K.-Y. Chen, L. Luesukprasert, and S.-C. T. Chou. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans. on Knowl. and Data Eng.*, 19(8):1016–1025, 2007.

[9] Z. Chen, H. Yang, J. Ma, J. Lei, and H. Gao. Time-based query classification and its application for page rank. *Journal of Computational Information Systems*, 2011.

[10] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proceedings of WSDM*, pages 11–20, 2013.

[11] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

[12] N. G. Golbandi, L. K. Katzir, Y. K. Koren, and R. L. Lempel. Expediting search trend detection via prediction of query counts. In *Proceedings of WSDM*, pages 295–304, 2013.

[13] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of ACM Multimedia*, pages 971–980, 2007.

[14] C.-K. Huang, L.-F. Chien, and Y.-J. Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. *J. Am. Soc. Inf. Sci. Technol.*, 54(7):638–649, May 2003.

[15] S. R. Kairam, M. R. Morris, J. Teevan, D. J. Liebling, and S. T. Dumais. Towards supporting search over trending events with social media. In *Proceedings of ICWSM*, 2013.

[16] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of WWW*, pages 297–306, 2008.

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

[18] Y. Koren and J. Sill. Ordrec: An ordinal model for predicting personalized item rating distributions. In *Proceedings of ACM Recommender Systems*, pages 117–124, 2011.

[19] Y. Lee, H.-Y. Jung, W. Song, and J.-H. Lee. Mining the blogosphere for top news stories identification. In *Proceedings of SIGIR*, pages 395–402, 2010.

[20] Z. Liao, D. Jiang, E. Chen, J. Pei, H. Cao, and H. Li. Mining concept sequences from large-scale search logs for context-aware query suggestion. *ACM Trans. on Intell. Syst. Technol.*, 3(1):17, 2011.

[21] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pages 1150–1157, 1999.

[22] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceedings of CIKM*, 2008.

[23] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of CIKM*, pages 469–478, 2008.

[24] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia Search Reranking: A Literature Survey. *ACM Computing Surveys*, 46(3), Sept. 2014.

[25] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *Proceedings of ICDM*, pages 502–511, 2008.

[26] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of CVPR*, 2007.

[27] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of UAI*, pages 452–461, 2009.

[28] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Trans. on Multimedia*, 15(6):1255–1267, 2013.

[29] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of WWW*, pages 285–295, 2001.

[30] Y. Shi, M. Larson, and A. Hanjalic. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of ACM Recommender Systems*, pages 269–272, 2010.

[31] M. Shokouhi. Learning to personalize query auto-completion. In *Proceedings of SIGIR*, pages 103–112, 2013.

[32] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz. Scene reconstruction and visualization from community photo collections. *Proceedings of the IEEE*, 98(8):1370–1390, 2010.

[33] X.-J. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui. Towards indexing representative images on the web. In *Proceedings of ACM Multimedia*, pages 1229–1238, 2012.

[34] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of SIGIR*, 2009.

[35] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of ACM Multimedia*, pages 15–24, 2009.