Rescue Tail Queries: Learning to Image Search Re-rank via Click-wise Multimodal Fusion

Xiaopeng Yang †, Tao Mei ‡, Yongdong Zhang †
† Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing 100190, P. R. China
‡ Microsoft Research, Beijing 100080, P. R. China
yangxiaopeng@ict.ac.cn; tmei@microsoft.com; zhyd@ict.ac.cn

ABSTRACT

Image search engines have achieved good performance for head (popular) queries by leveraging text information and user click data. However, there still remain a large number of tail (rare) queries with relatively unsatisfying search results, which are often overlooked in existing research. Image search for these tail queries therefore provides a grand challenge for research communities. Most existing re-ranking approaches, though effective for head queries, cannot be extended to tail. The assumption of these approaches that *the re-ranked list should not go far away from the initial ranked list* is not applicable to the tail queries. The challenge, thus, relies on how to leverage the possibly unsatisfying initial ranked results and the very limited click data to solve the search intent gap of tail queries.

To deal with this challenge, we propose to mine relevant information from the very few click data by leveraging click-wise-based image pairs and query-dependent multimodal fusion. Specifically, we hypothesize that *images with more clicks are more relevant to the given query than the ones with no or relatively less clicks* and *the effects of different visual modalities to re-rank images are query-dependent*. We therefore propose a novel query-dependent learning to re-rank approach for tail queries, called "click-wise multimodal fusion." The approach can not only effectively expand training data by learning relevant information from the constructed click-wise-based image pairs, but also fully explore the effects of multiple visual modalities by adaptively predicting the query-dependent fusion weights. The experiments conducted on a real-world dataset with 100 tail queries show that our proposed approach can significantly improve initial search results by 10.88% and 9.12% in terms of NDCG@5 and NDCG@10, respectively, and outperform several existing re-ranking approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Performance

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654900.

(a) "kim kardashian" (b) "kim and kanye's baby"

Figure 1: Examples of the initial image search results of a head query ("kim kardashian") and a tail query ("kim and kanye's baby"). The red rectangles mark irrelevant images [best viewed in color]. Existing commercial search engines achieve very limited image search performance for tail queries.

Keywords

Image search; search re-ranking; tail queries; click-through data; multimodal fusion.

1. INTRODUCTION

As Web 2.0 and social networks developed, billions of images have been contributed and shared in the media-centered communities. The increasing number of images has posed a grand challenge to image search. Most existing commercial search engines adopt keyword-based technique for image search. However, the textual information of web images is often noisy and even unavailable. In order to boost the performance of web image search and overcome the "semantic gap" (the gap between the low-level features and high-level semantics) and "intent gap" (the gap between the representation of users' query/demand and the real intent of the users), many image search re-ranking approaches have been proposed in recent years [17].

However, most re-ranking methods target at improving search performance of head (popular) queries, while the tail (rare) queries have often been overlooked. In this paper, we focus on image search re-ranking for tail queries for the following reasons. First, head queries are frequently issued, which are mostly about celebrities, movies and so on. The corresponding image source material, textual information, and user clicks are generally substantial. By applying existing effective re-ranking methods, consequently, the search performance of head queries is usually satisfying. In contrast, the search performance of tail queries is poor due to the lack of enough source information and user clicks. As a result, most existing re-ranking approaches, with the assumption that *the re-ranked list should be close to the initial ranked list* or *the top ranked images from the initial ranked list suppose relevant to the given query*, cannot be extended to solve the problems of tail queries. Second, since query frequencies follow such a power-law distribution that there are large numbers of tail queries in query logs

[1], search performance of those queries would potentially affect the popularity and relevance of search engines significantly in fierce market competition. For example, Fig. 1 shows the initial search results of head query "kim kardashian" and tail query "kim and kanye's baby," respectively. We can see that the top 10 search results of head query "kim kardashian" are absolutely satisfying, while, for tail query "kim and kanye's baby," there are only two relevant images that satisfy user search intent. To summarize, we believe that appropriately handling tail queries is a newly emerging direction in image search.

To address these challenges for tail queries, we seek to leverage multiple visual features and pair-wise click-through data simultaneously. In general, most image search re-ranking approaches treat different modalities (features), such as shape, color and texture, independently. However, for different queries, discriminative modalities may have distinct effects. For instance, for the queries like "heart" and "sun," color feature may be more useful, while for the queries such as "buildings," texture feature will be more effective. The same situation happens to tail queries, for example, shape feature may be more helpful for "kim and kanye's baby" shown in Fig. 1(b) to detect the "baby." Although several approaches, which learn and predict fusion weights in a linear or non-linear way to combine multiple modalities, have proved effective [22][27], the problem of how to fuse multiple modalities adaptively and in a query-dependent way still remains. Conversely, since users browse image thumbnails before selecting the images to click, we believe that the click data can reflect users' search intent as "implicit" relevance feedback [9][26]. Intuitively, images with more clicks are more relevant to the given query than the ones with no or less clicks. However, clicked images always hold a low percentage in the search results to a given tail query, even less than 10%. It is difficult to learn the similarity among them to facilitate re-ranking for all the images. How to mine useful information from such limited click-through data is a problem that most existing re-ranking work does not take into consideration.

To address the above issues, we propose a novel query-dependent learning to re-rank approach, called "click-wise multimodal fusion," to improve the search performance for tail queries. In our paper, we define "click-wise" as the difference value of clicks for pair-wise images. Our key assumptions are based on two aspects: 1) the effects of different visual modalities to re-rank images are query-dependent, and 2) images with more clicks are more relevant than the ones with no or relatively less clicks in response to a given query. In click-wise multimodal fusion, we extract multiple visual features from all the images, and select image pairs containing click-wise information, no matter clicked-clicked image pairs or clicked-unclicked ones. On the one hand, click-wise multimodal fusion can learn and predict the fusion weights of multiple modalities adaptively and query-dependently. On the other hand, it can make full use of click-wise-based image pairs to correctly re-rank the clicked images with larger clicks higher through penalizing the misclassified pairs with different click-wise information.

Note that our re-ranking approach is general, yet, it is applicable to tail queries owing to the use of click-wise information. Specifically, as click-through data can be viewed as the footprints of user search behavior, we mine click-wise information adequately to guide image search re-ranking regardless of the initial ranked list. For a given tail query, even the number of clicked images is relatively small, such as 10 or less, we can still get enough training data by detecting the click-wise-based pairs. Moreover, the clicked images of tail queries are usually uncertain and diverse resulting in the difficulty to learn from the "similarity" among them, while using our approach, we can learn from the "dissimilarity" via click-wise information to re-rank images.

The contributions of this paper can be summarized as follows.

- We have investigated the image search re-ranking for tail queries which is often overlooked by most previous research. To the best of our knowledge, this represents one of the first attempts for formally studying the problem of image search re-ranking for tail queries.
- We propose a novel learning to re-rank approach, called "click-wise multimodal fusion," which can not only adaptively learn the fusion weights of multiple modalities in a query-dependent way, but also leverage image pairs with click-wise information to facilitate image search re-ranking.
- We conduct experiments on a one-week real-world dataset consisting of 2,682,666 queries and 20,165,208 image URLs collected from a commercial search engine. The evaluation validates that our approach is able to significantly improve search performance for tail queries, while maintaining slightly better performance for head queries.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 analyzes a one-week query log and details the definition of tail queries. Section 4 introduces our proposed query-dependent re-ranking approach, called "click-wise multimodal fusion." Section 5 reports the experimental results on the performance of our re-ranking approach. Section 6 draws the conclusion.

2. RELATED WORK

Our work is mainly related to image search re-ranking and search with click-through data. In this section, we first present image search re-ranking along two directions, i.e., recurrent pattern mining and multimodal visual feature fusion. Then, we introduce the work on search with click-through data, especially for image search.

2.1 Image Search Re-ranking

According to how many visual features are leveraged and explored, we categorize visual search re-ranking into two major directions, i.e., recurrent pattern mining and multimodal visual feature fusion. Most of them are developed based on the hypotheses that 1) the re-ranked results should not change too much from the initial ranked results or the top ranked images from the initial ranked list suppose relevant to the given query, and 2) visually similar images should be close in the re-ranked list.

Recurrent pattern mining seeks to mine recurrent patterns from relevant images to improve the re-ranking performance. For instance, Hsu *et al.* formulate re-ranking as a random walk problem along the context graph, where video stories are represented as nodes and the edges between them are weighted by contextual similarities [7]. Yan *et al.* propose to re-rank using a binary classifier where the top-ranked (bottom-ranked) documents from the initial ranked results are chosen as pseudo-positive (pseudo-negative) samples, which is the so-called pseudo-relevance feedback (PRF) [23]. Compared with exploiting the initial search results without any external knowledge, Yang and Hanjalic leverage query examples and formulate learning to re-rank as an optimization function by minimizing the distance between the re-ranked list and the initial one, while maximizing the coherence of similar ranked images in terms of the visual features [24]. Similarly, crowd-sourced knowledge, e.g., multiple initial ranked results from various search engines [15] and the suggested queries augmented from the image collection on the Web [28], is mined to find relevant visual patterns. In order to further satisfy users' search intent, many researchers

Table 1: Query frequency distribution for one-week query log.

query frequency	number of queries	region in Fig. 2
$(100, 3251]$	767	HEAD
$(1, 100]$	310,274	TAIL-A
1	2,371,625	TAIL-B

involve user interaction, such as human labeling and feedback, to guide the re-ranking process. Hauptmann *et al.* propose to employ active learning technique to exploit both the human bandwidth and machine capability for video search re-ranking [6]. For reducing the efforts of users' labeling, Tian *et al.* propose a sample selection strategy based on images' structural information, and then use a discriminative dimension reduction algorithm to capture user intent in the visual feature space [21].

Multimodal visual feature fusion aims to fuse different modalities in a unified way and make them function well accordingly. To deal with this problem, one natural way is to concentrate multiple features into a long feature vector and then use this joint modality to perform specific task. Alternatively, we can fuse the re-ranking results produced by applying modalities separately in a re-ranking algorithm. These two methods are the so-called "early fusion" and "late fusion" separately [20]. Even though the early and late fusion approaches are advantageous compared with ones using an individual modality, yet, they still suffer from "curse of dimensionality" and incapability of determining proper fusion weights for different modalities, respectively. Thus, in order to learn appropriate fusion weights of multiple modalities, Snoek *et al.* propose to assign weights heuristically and manually based on the type of query, such as text-, concept- and visual-oriented queries [19]. This "rule-based fusion," though easy to implement, may degrade the retrieval performance due to the wrong weights assigned by users. To address this issue, Kennedy *et al.* recommend to use "query-class-dependent fusion" [13]. They first classify each user query into categories learnt by cluster algorithms, and then aggregate retrieval results with the help of query-class associated weights. Nevertheless, it is difficult to categorize a user query into a specific class accurately due to its complicated semantic meanings. Therefore, "adaptive fusion" is introduced to learn query-dependent fusion weights for multiple modalities [22][25].

Overall, though proved effective for head queries, the above mentioned approaches cannot be extended to tail queries, mainly because the initial ranked results of tail queries are always so unsatisfying that they cannot support the assumptions. Moreover, click-through data, which can be viewed as an indicator of image relevance, are mostly not taken into consideration in those methods.

2.2 Search with Click-through Data

Click-through data have been studied in web search for a few years [3][4][5][12]. Compared with web document search, where users can only browse a two-line snippet in response to a given query, in image search, users browse image thumbnails before selecting the images to click. Thus, it is much more convincing that the decision to click is likely dependent on the relevance of an image. In recent years, mining click-through data from query logs to facilitate image search has attracted some researchers' attention [8][9][26]. For instance, Jain *et al.* employ Gaussian Process regression to predict the normalized click count for each image, and combine it with the original ranking score for re-ranking [9]. In [26], Yang *et al.* leverage click-through data and detect recurrent visual patterns of images simultaneously to boost the performance of image retrieval. Based on the assumption that clicked images

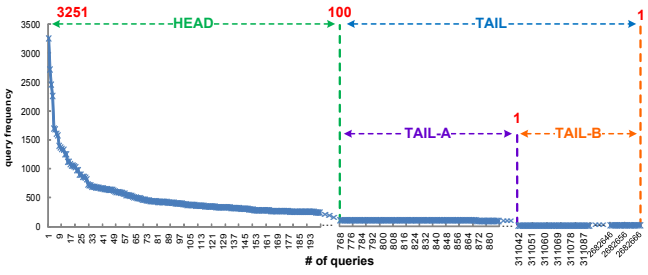


Figure 2: The power-law distribution for the query frequencies of 2,682,666 queries. Note that regions are divided by *dotted* lines [best viewed in color].

are highly correlated with relevant ones to the given query, it is inspiring to use click-through data, which are readily available and freely accessible from query logs, to guide image search instead of human intervention. As the above work presents, the click counts of images have been fully explored, yet, for tail queries, there are not enough clicked images, let alone the images' click counts. In order to increase data volume containing click information, we attempt to use click-wise-based image pairs, i.e., image pairs with the difference value of clicks. Then, our concerns are: 1) what is the influence of *the difference value of clicks*? 2) can *the difference value of clicks* be leveraged to improve search performance of images, especially for tail queries?

3. QUERY LOGS AND TAIL QUERIES

Since we focus on improving search performance of tail queries, we first analyze query logs, and then give the general definition of tail queries. We have collected query logs from a commercial image search engine for one week in Oct. 2013. The query logs are represented as plain text files that contain a line for each HTTP request satisfied by the Web server. For each record, the following fields are used in our data collection:

<Query, ClickedURL, ClickCount, Thumbnail>

where the *ClickedURL* and *ClickCount* represent the URL and the number of clicks on this URL when users submit the *Query*, respectively. *Thumbnail* denotes the corresponding image information on the *ClickedURL*.

In order to study the effect of click-through data, we use all the queries in the log with at least one click. There are 2,682,666 queries and 20,165,208 image URLs in total. Generally, a tail query is identified based on its query frequency. As Table 1 shows, there are over 88% queries with one query frequency, indicating they are issued only once by users during one week. In contrast, the number of queries with frequency larger than 100 is 767, accounting for less than 0.03% of all queries. We sort all the queries according to their frequencies in a descending order and draw Fig. 2 which shows that the query frequency distribution follows the power laws. Empirically, we believe that queries issued more than 100 times during a week can be viewed as popular (head) queries. Thus, based on the query frequency of 100, we divide the entire query frequency distribution into "HEAD" region and "TAIL" region shown in Fig. 2. Moreover, since the number of queries issued only once is extremely large, we further divide "TAIL" region into "TAIL-A" (including queries issued between 1 and 100 times) and "TAIL-B" (including queries issued only once). In our work, we mainly focus on region "TAIL-A" and "TAIL-B" consisting of tail queries.

Following the general concept based on query frequencies, we give the definition of tail queries below.

DEFINITION 1. A $Query$ is represented as a record:

<*QueryText, QueryFrequency*>

where $QueryText$ stands for the textual information of $Query$, $QueryFrequency$ indicates the query frequency during a certain time period. A tail query is a query with $QueryFrequency$ no more than T_q, where T_q is the threshold which equals to a certain query frequency for identifying tail queries.

Since the time period of query logs varies, the threshold T_q can be defined accordingly. In our work, we set T_q to 100 for a one-week query log. Taking query "kim and kanye's baby" shown in Fig. 1(b) as an example, its query frequency equals to 29 which is less than $T_q = 100$, then we can define it as a typical tail query. Concerning queries with clicks, i.e. the queries in the log with at least one click, the definition of tail queries is different from the general one as follows.

DEFINITION 2. A $Query$ with clicks is represented as a record:

<*QueryText, QueryFrequency, ClickedImages*>

where $ClickedImages$ indicates the number of clicked images in response to $Query$. A tail query with clicks is a query with *QueryFrequency* no more than T_q, and *ClickedImages* no more than T_c, where T_q and T_c are the thresholds of *QueryFrequency* and *ClickedImages*, respectively.

Similarly, the thresholds of T_q and T_c can be set depending on the time period of query logs and the requirement of click information. Compared with general definition of tail queries given in Def. 1, we add $ClickedImages$ for tail queries with clicks mainly for the following reason. A dispersive distribution of clicked images equals with diverse search intent of users. By narrowing down the range of the number of clicked images, we can get more centralized user intent via a certain number of clicked images. In our work, we set T_q and T_c all to 100 for a one-week query log. For example, as the query frequency and the number of clicked images of query "kim and kanye's baby" is 29 and 22 respectively, we can use it as a representative tail query with clicks.

4. LEARNING TO RE-RANK: CLICK-WISE MULTIMODAL FUSION

In this section, we introduce our proposed novel re-ranking approach, called click-wise multimodal fusion (CWMF). First, we present traditional ranking SVM (Supported Vector Machine), which is a commonly used "pair-wise" learning to rank algorithm. We do not introduce other learning to rank algorithms, such as "point-wise" and "list-wise," because in our work we try to use click-wise information between two images, which more matches the form of "pair-wise." Then, we extend ranking SVM to a learning to re-rank paradigm and detail our re-ranking approach.

In the following sections, we use q to denote the issued query, x to denote an image returned by a search engine, y to denote the relevance of image x, and c to denote the click count of image x.

4.1 Ranking SVM

Ranking SVM, in which the preference relations between instances are used, can be viewed as a special case of SVM [11]. Suppose we are given a set of training query-image-label triples (q_i, x_i, y_i), where $q_i \in \mathcal{Q}$ ($\mathcal{Q}$ is the set of queries), $x_i \in \mathcal{R}^n$ (n is the dimension of image's feature), $y_i \in \mathcal{R}$, $i = 1, 2, ..., N$ (N is the number of triples). Label y_i annotates the relevance of x_i in response to q_i, then we can define the set of image preference pairs as

$$\bar{\mathcal{P}} \triangleq \{(i,j) | q_i = q_j, y_i > y_j\}, \tag{1}$$

where $i, j = 1, 2, ..., N$. We use $\bar{P}$ to denote the size of set $\bar{\mathcal{P}}$, i.e., $\bar{P} \triangleq |\bar{\mathcal{P}}|$.

Then, ranking SVM can be formulated as learning for classification on the preference pairs shown as Eqn. (2).

$$\begin{aligned} \min_{\omega,\xi} \quad & F(\omega, \xi) = \tfrac{1}{2}\omega^T\omega + C\sum\nolimits_{(i,j)\in\bar{\mathcal{P}}} \xi_{ij} \\ \text{s.t.} \quad & \omega^T\big(\Phi(x_i) - \Phi(x_j)\big) \geq 1 - \xi_{ij} \\ & \xi_{ij} \geq 0, \quad \forall (i,j) \in \bar{\mathcal{P}}, \end{aligned} \tag{2}$$

where ω is a weight vector that is adjusted by learning, C is the regularization parameter and $C > 0$, Φ is a mapping onto instance feature that describes the match between q_i and x_i, ξ_{ij} is a L1 loss term. Overall, the first term of Eqn. (2) is called regularization term, and the second one is the Hinge Loss term.

Note that the optimization is equivalent to that of SVM as a quadratic optimization problem. Assume that ω^* is the optimal solution of Eqn. (2) and $f(x)$ is a ranking function, then we can leverage ω^* to calculate the ranking score of image x as follows

$$f(x) = {\omega^*}^T \Phi(x). \tag{3}$$

4.2 Click-wise Multimodal Fusion

4.2.1 Overview

Since the initial search results of tail queries are usually unsatisfying and the corresponding click data are very limited, most existing re-ranking approaches cannot be extended to tail queries. The reason lies on the basic assumption of re-ranking, i.e., *the re-ranked list should not go far away from the initial ranked list*. To address these issues, we propose to mine relevant information from the very few click data by leveraging click-wise-based image pairs (image pairs with click-wise information) and query-dependent multimodal fusion. Note that in our paper we define "click-wise information" as the difference value of clicks. Then, we develop our re-ranking approach based on the following two assumptions:

- images with more clicks are more relevant to the given query than the ones with no or relatively less clicks, and
- the effects of different visual modalities to re-rank images are query-dependent.

In order to leverage the click-wise information from click-through data, we formulate image search re-ranking as a "pair-wise" learning to re-rank problem based on ranking SVM. However, on one hand, traditional ranking SVM cannot deal with the problem of multimodal fusion resulting in the impossibility to leverage query-dependent effects of different modalities. On the other hand, without considering the difference value of clicks, ranking SVM treats the image pairs with different click-wise information equally [2]. Nevertheless, it is not desirable that the ranking model tends to be close to the clicked-clicked instance pairs, especially of which the click-wise information is relatively small. On the opposite, the ranking model should be in proximity to the clicked-unclicked instance pairs containing large click-wise information. To summarize, in order to adequately learn the multi-modality influence and adjust the ranking SVM training bias, we extend ranking SVM to make it capable of learning query-dependent multi-modality fusion weights and penalizing the misclassified click-wise-based instance pairs.

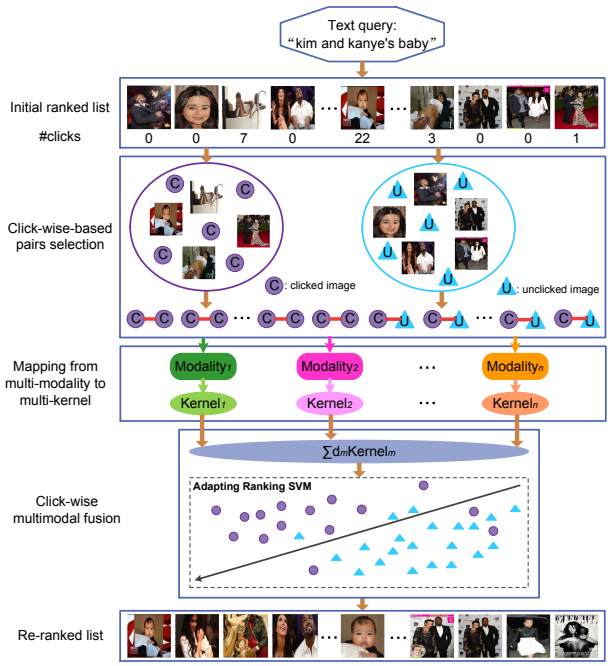


Figure 3: The overview of the proposed click-wise multimodal fusion re-ranking [best viewed in color]. Note that "♯clicks" denote the corresponding click count of each image. "Click-wise-based pairs selection" chooses image pairs with click-wise information. "Mapping from multi-modality to multi-kernel" projects different modalities into their corresponding kernel space. "Click-wise multimodal fusion" integrates multiple kernel learning with a click-based adapting ranking SVM.

The overview of the proposed query-dependent image search re-ranking approach, called "click-wise multimodal fusion," is shown in Fig. 3. Given a tail query, such as "kim and kanye's baby," to mine relevant information from click data, we first identify the set of clicked images and the set of unclicked ones from the initial ranked list based on the click counts of images (♯clicks). Since clicked images of tail queries are very limited, we select two images with different click counts as a pair to expend training data. In other words, we choose click-wise-based image pairs not only between the set of clicked images and the set of unclicked ones, but also from the interior of the set of clicked images. Then, in order to facilitate re-ranking by exploring the query-dependent effects of multiple visual features, we transform multiple visual modalities into the same dimension by mapping them to their corresponding kernel space. Finally, in our proposed query-dependent learning to re-rank approach, click-wise multimodal fusion can not only adaptively learn the linear fusion weight d_m for each modality by leveraging multiple kernel learning algorithm, but also directly output the re-ranks of images based on a click-based adapting ranking SVM.

Compared with ranking SVM, the advantages of click-wise multimodal fusion are twofold. First, we revise the Hinge Loss function by setting various losses for misclassification of image pairs between different click-wise-based pairs (pairs with discriminative difference value of clicks). To reduce errors on determining ranks of images with more clicks, the loss function heavily penalizes errors with regard to click-wise-based pairs with large difference value of clicks. This modification can be viewed as a click-based adapting ranking SVM. Second, we integrate a multiple kernel learning algorithm with the click-based adapting ranking SVM into a uniform framework. By automatically learning and predicting the query-dependent fusion weights for multiple features, we can explore how consistent (contradictory) modalities could incorporate (compromise) with each other.

4.2.2 *Formulation*

For a given query q, we represent the image instances as a set of query-instance-clicks triples (q, x_i, c_i), where $q \in \mathcal{Q}$, $x_i \in \mathcal{R}^n$, $c_i \in \mathcal{R}$, $i = 1, 2, ..., l$ (l is the number of images to be re-ranked). Suppose there are M modalities of images in total. Based on modality m, image x_i can be expressed as $x_{m,i}$ ($m = 1, 2, ..., M$). For lightening notations, we specify $\sum_m$ to represent the summation from the 1^{st} modality to the M^{th} one, and use $\phi_m(x_i)$ to represent the mapping onto feature m that describes the match between query q_i and image $x_{m,i}$. Based on the click-wise information, we define the set of click-wise-based instance pairs as

$$\mathcal{P} \triangleq \{(i,j) | (c_i - c_j) \geq \delta\}, \tag{4}$$

where δ is the threshold controlling the selection of click-wise-based pairs and $\delta > 0$. Accordingly, the difference value of clicks between x_i and x_j is expressed as c_{ij}, which equals to $|c_i - c_j|$. We also use P to denote the size of set $\mathcal{P}$, i.e., $P \triangleq |\mathcal{P}|$.

We define a new loss function which integrates multiple kernel learning with a click-based adapting ranking SVM as follows.

$$\begin{aligned} \min_{d,\omega,\xi} \quad & L(d,\omega,\xi) = \tfrac{1}{2}\textstyle\sum_m \frac{1}{d_m}\omega_m^{\mathrm{T}}\omega_m + C\sum_{(i,j)\in\mathcal{P}} \xi_{ij}\lambda_{ij} \\ \text{s.t.} \quad & \textstyle\sum_m \omega_m^{\mathrm{T}}\big(\phi_m(x_i) - \phi_m(x_j)\big) \geq 1 - \xi_{ij} \\ & \xi_{ij} \geq 0, \quad \forall (i,j) \in \mathcal{P} \\ & \lambda_{ij} = \exp\left\{\tfrac{c_{ij}}{2\gamma^2}\right\} \\ & \textstyle\sum_m d_m = 1, \quad d_m \geq 0, \quad \forall m, \end{aligned} \tag{5}$$

where d denotes the fusion weight vector for M modalities, λ denotes the adapting penalty parameter, γ denotes the average difference value of clicks among all click-wise-based pairs for query q, i.e., $\gamma = \frac{\sum_{(i,j)\in\mathcal{P}} c_{ij}}{P}$, where $i \neq j$ and $i, j = 1, 2, ..., l$. Note that in Eqn. (5) we define that when $d_m = 0$, ω_m has to be a zero vector so as to yield a finite objective value [18].

4.2.3 *Optimization*

Due to the introduction of the last constraint, it is difficult to directly solve Eqn. (5) by its corresponding dual problem. Thus, we deform Eqn. (5) as a constrained optimization problem with regard to d as follows.

$$\min_d J(d), \quad \text{such that } \sum_m d_m = 1,\ d_m \geq 0, \tag{6}$$

where

$$J(d) = \begin{cases} \min\limits_{\omega,\xi} \left\{ \tfrac{1}{2}\textstyle\sum_m \frac{1}{d_m}\omega_m^{\mathrm{T}}\omega_m + C\sum_{(i,j)\in\mathcal{P}} \xi_{ij}\lambda_{ij} \right\} \\ \text{s.t.} \ \textstyle\sum_m \omega_m^{\mathrm{T}}\big(\phi_m(x_i) - \phi_m(x_j)\big) \geq 1 - \xi_{ij} \\ \quad \xi_{ij} \geq 0, \quad \forall (i,j) \in \mathcal{P} \\ \quad \lambda_{ij} = \exp\left\{\tfrac{c_{ij}}{2\gamma^2}\right\}. \end{cases} \tag{7}$$

To solve the above problem, we first write the Lagrange function of $J(d)$ as

$$L_p = \frac{1}{2}\sum_m \frac{1}{d_m}\omega_m^{\mathrm{T}}\omega_m + C\sum_P \xi_{ij}\lambda_{ij} - \sum_P \alpha_{ij} W - \sum_P u_{ij}\xi_{ij}, \tag{8}$$

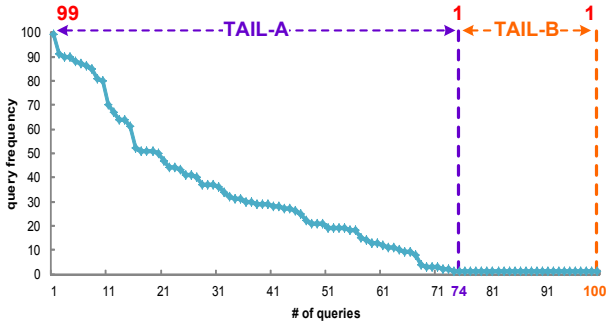


Figure 4: Query frequency distribution for the 100 tail queries dataset.

where α_{ij} and u_{ij} are the Lagrange multipliers, $\sum_P$ is the abbreviation for $\sum_{(i,j)\in\mathcal{P}}$ and

$$W = \sum\nolimits_m \omega_m^{\mathrm{T}}\big(\phi_m(x_i) - \phi_m(x_j)\big) - (1 - \xi_{ij}). \quad (9)$$

After setting the respective derivatives with respect to ω_m and ξ_{ij} to zero, we can obtain the Lagrange dual function below.

$$\begin{aligned} \max_{\alpha} \quad & L_D(\alpha) = -\tfrac{1}{2}\alpha^{\mathrm{T}} G\alpha + e^{\mathrm{T}}\alpha \\ \text{s.t.} \quad & 0 \le \alpha_{ij} \le C\lambda_{ij},\ \forall (i,j) \in \mathcal{P}, \end{aligned} \quad (10)$$

where $\alpha \in \mathcal{R}^P$ is indexed by click-wise-based pairs in set $\mathcal{P}$, $e \in \mathcal{R}^P$ is a vector of ones, and G is a P by P symmetric matrix. Specifically,

$$G_{(i,j),(u,v)} = \sum\nolimits_m d_m \Phi_{m,ij}^T \Phi_{m,uv}, \quad \forall (i,j),(u,v) \in \mathcal{P}, \quad (11)$$

where $\Phi_{m,ij} \triangleq \Phi_m(x_i) - \Phi_m(x_j)$.

Clearly, the dual problem of (7) is similar to that of ranking SVM. There are two differences lying on: 1) G in (10) is a weighted mapping from M modality, while G in ranking SVM is based on a single modality; 2) the upper bounds of α vary according to various click-wise information from pairs, i.e., pairs with larger difference value of clicks get a larger upper bounds, in contrast, the upper bounds of α in ranking SVM are the same.

In order to facilitate the calculation of fusion weight vector d and also avoid the situation that mapping Φ is infinite dimensional, we assign a kernel function $K(\cdot,\cdot)$, such as linear kernel, RBF kernel, etc., to each modality accordingly. Then, G can be indicated as

$$G = \sum\nolimits_m d_m A \tilde{G}_m A^{\mathrm{T}} \quad (12)$$

where $\tilde{G}_m \in \mathcal{R}^{l\times l}$, with $\tilde{G}_m(i,j) = K_m(x_i,x_j) = \Phi_m(x_i)^{\mathrm{T}}\Phi_m(x_j)$ and $A \in \mathcal{R}^{P\times l}$ is a particular matrix with the following format

$$A = \begin{array}{c} \\ \vdots \\ (i,j) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} \ldots & & i & \ldots & & j & \ldots \end{array} \\ \begin{bmatrix} 0\ldots0 & & +1 & 0\ldots0 & & -1 & 0\ldots0 \end{bmatrix} \end{array} \quad (13)$$

meaning that if $(i,j) \in \mathcal{P}$, the i^{th} entry of the corresponding row in A is 1, the j^{th} entry is -1, and other entries are all zeros. The reason of introducing matrix A is to save the computation cost of G from $O(l^4)$ to $O(l^2)$.

Then, by treating the weighted mapping G as a unified single one, the standard ranking SVM algorithms [11][14] can be leveraged to directly solve this problem. Once the $J(d)$ is solved, we compute the gradient of $J(d)$ with respect to d_m, and then use a reduced gradient algorithm proposed in [18] to update the fusion weight vector d for M modalities. The updating scheme is $d \leftarrow d + \theta D$, where θ is the step size which can be determined by line search and D is the descent direction

$$D_m = \begin{cases} 0 & \text{if } d_m = 0 \text{ and } \frac{\partial J}{\partial d_m} - \frac{\partial J}{\partial d_\mu} > 0 \\ \frac{\partial J}{\partial d_\mu} - \frac{\partial J}{\partial d_m} & \text{if } d_m > 0 \text{ and } m \neq \mu \\ \sum\limits_{v\neq\mu, d_v>0} \left(\frac{\partial J}{\partial d_v} - \frac{\partial J}{\partial d_\mu}\right) & \text{for } m = \mu, \end{cases} \quad (14)$$

where μ is the index of the largest component of vector d.

Finally, when the terminated criterion is met, such as the duality gap, the KKT condition, the variation of d between two consecutive steps or simply a maximal number of iterations, the re-ranking score of image x can be represented as

$$\begin{aligned} f(x) &= \textstyle\sum_m \omega_m^{*\,\mathrm{T}} \Phi_m(x) \\ &= \textstyle\sum_m d_m^* \sum_{(i,j)\in\mathcal{P}} \alpha_{ij}^* K_m(x_i - x_j, x), \end{aligned} \quad (15)$$

where ω_m^*, d_m^* and α_{ij}^* are the optimal values.

Table 2: Statistic property values of 100 tail queries.

	QueryFrequency	*ClickedImages*	*MaxClicks*
Max. value	99	69	31
Min. value	1	1	1
Avg. value	27	23.21	9.98

5. EXPERIMENTS

In this section, we describe our experimental settings and present the experimental results. To adequately validate the effectiveness of our approach, we first compare our approach with those methods lacking of multimodal fusion and several existing re-ranking methods, respectively. Then, we analyze the sensitivity of parameters used in our approach and the complexity of our approach. Finally, we give some detailed re-ranking examples of tail queries.

5.1 Experimental Settings

To facilitate evaluation and compare our proposed approach with other ones, we randomly select 100 tail queries from the one-week query log mentioned in Sect. 3 according to the settings $T_q = 100$ and $T_c = 100$, in other words, these queries belong to region "TAIL" defined in Fig. 2. Figure 4 shows the query frequency distribution of these 100 tail queries with clicks. We can find that query frequencies of these 100 tail queries follow a nearly power-law distribution as Fig. 2, indicating that these queries have certain representativeness. Among the 100 tail queries, there are 73 and 27 queries belonging to region "TAIL-A" and "TAIL-B" as shown in Fig. 2, respectively. Note that there are almost triple queries in "TAIL-A" than the ones in "TAIL-B," since we assign a larger selection weight for queries in "TAIL-A." The reason originates from the fact that the click data of queries in "TAIL-B" are provided by an individual user, which are possibly biased compared with click data aggregated by users. Moreover, there are various types of queries, such as people ("pewdiepie," and "josh freeman girlfriend"), object ("first computer," and "corner wall waterfall"), concept ("internet safety," and "small caribbean house plans"), scenery ("sea sparkle," and "east coast of the united states from space") and event ("jodie foster dating," and "elizabeth smart wedding"). Note that some typical tail queries belonging to these semantic categories are shown in the bracket next to their name severally.

Since source materials of tail queries are usually limited and the images after the top 100 results are typically irrelevant, we

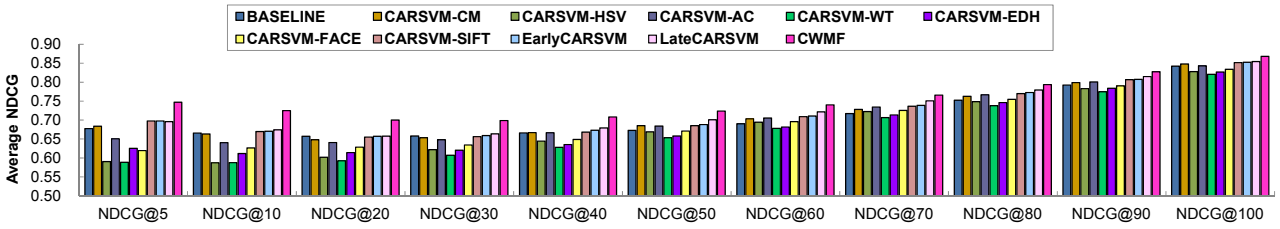


Figure 5: Comparison of re-ranking approaches using different modalities in terms of NDCG

use the top 100 images from the initial search results to perform re-ranking, which means that the number of images $l = 100$. Among the 10,000 query-image-clicks triples, the property values defined in Def. 2 are demonstrated in Table 2. The last column $MaxClicks$ recorded in Table 2 denotes the maximum click count among all the $ClickCount$ for a given query.

For these images, we extract seven features, i.e., $M = 7$, including: 225-dimensional block-wise color moments, 64-dimensional HSV color histogram, 144-dimensional color autocorrelogram, 128-dimensional wavelet texture, 75-dimensional edge distribution histogram, 7-dimensional face features, and 2000-dimensional scale-invariant feature transform (SIFT) descriptor. It is noteworthy that in our experiments SIFT descriptor [16] with a Difference of Gaussian (DoG) interest point detector is employed to extract visual patterns of images. Then, K-means is further used to cluster the similar patches into "visual words," and Bag-of-Word (BoW) is used to represent each image. Empirically, the number of visual words is set to 2,000.

In the following parts, we use CARSVM-CM, CARSVM-HSV, CARSVM-AC, CARSVM-WT, CARSVM-EDH, CARSVM-FACE, and CARSVM-SIFT to denote the methods that only use the seven modalities based on the click-based adapting ranking SVM (CARSVM), respectively. Note that CARSVM is also proposed in this paper.

In our experiments, each query-image pair is labeled carefully by annotators on a scale of 0 to 2: 0–"irrelevant," 1–"fair," and 2–"relevant." Before labeling, we first let annotators figure out the meaning of the issued tail query and check some related web documents to determine user search intent. By understanding user intent as much as possible, the relevance scores of images are more convincing relatively. We adopt the Normalized Discounted Cumulative Gain (NDCG) [10] to measure performance, which is widely used in information retrieval when there are more than two relevance levels. Given a ranked list, the NDCG at the depth p is defined as

$$\mathrm{NDCG@}p = Z_p \sum_{i=1}^{p} \frac{2^{r^i} - 1}{\log(1+i)}, \tag{16}$$

where r^i is relevance score of the i^{th} image, and Z_p is a normalization constant to guarantee that a perfect ranking's $\mathrm{NDCG@}p$ is equal to 1.

5.2 Evaluations of Re-ranking

We first compare our proposed approach, called click-wise multimodal fusion (CWMF), with methods that use only an individual modality, method "early fusion" and method "late fusion" to verify the effectiveness of multimodal fusion in our approach. Methods that use only an individual modality are all based on the click-based adapting ranking SVM (CARSVM), which means the click-wise-based pairs selection parameter δ and the Hinge Loss adapting parameter λ are all concerned. The "early fusion" refers to the method that concentrates all the seven modalities into a long vector and leverages it to perform click-wise learning to re-rank. In the "late fusion," we leverage seven click-based adapting ranking SVM classifiers, each of which uses the data of one of the seven features respectively, and we linearly fuse the re-ranking results of these classifiers, in which the fusion weights are tuned for maximum performance. For simplifying notations, we represent "early fusion" and "late fusion" based on the click-based adapting ranking SVM as EarlyCARSVM and LateCARSVM severally.

For our proposed method CWMF, we first assign a kernel function to each modality accordingly. In our experiments, linear kernel is designated to each modality for its efficiency, i.e., $K_m(x_i, x_j) = x_{m,i}^{\mathrm{T}} x_{m,j}$, where $m = 1, 2, ..., M$. Then, after tuning the parameters C and δ to obtain the optimal performance, we set $C = 0.5$ and $\delta = 5$ for our dataset. Detailed analysis about parameter C and δ will be introduced in Sect. 5.3. Moreover, we take the duality gap, which is equal to 0.01, as the stopping criterion.

Figure 5 illustrates the average measurements from NDCG@5 to NDCG@100 using the above methods. Here we also illustrate the NDCG measurements of the initial ranked lists and regard them as BASELINE results. From Fig. 5, we can find that using only an individual modality can hardly obtain the maximum improvement compared with the baseline. The performance of different single modalities, some of which even degrades to a certain extent, varies at different depths of NDCG. This phenomenon is understandable, because the effect of an individual modality depends on different queries and it is hard to determine its usefulness to a given query. But the "uncertain" single modality is still useful by integrating it with other features to work together. Combining multiple modalities, the NDCG values of EarlyCARSVM and LateCARSVM are relatively larger than those of methods using an individual modality, which demonstrate the effectiveness of the introduction of multiple modalities. Nevertheless, our proposed approach click-wise multimodal fusion achieves the maximum improvements and obviously outperforms other methods at different depths of NDCG. Thus, we can draw a conclusion that the multimodal fusion scheme of our approach can help determine the modality importance and fuse multiple modalities adaptively to obtain ideal performance.

Then, we compare our proposed re-ranking approach with several existing ones, where the parameters are optimized to achieve the best possible performance, including:

- Random walk (RW) [7]. A representative self-re-ranking method which conducts random walk on an image graph where nodes are images and edges are weighted by image visual similarities.

- Pseudo-relevance feedback (PRF) [23]. PRF performs re-ranking as a classification problem assuming that top-ranked results are more relevant than the bottom-ranked results.

- Multimodal graph-based re-ranking (MGR) [22]. An effective re-ranking method which leverages the initial ranked list and explores the effects of multiple modalities in a multi-graph-based learning scheme.

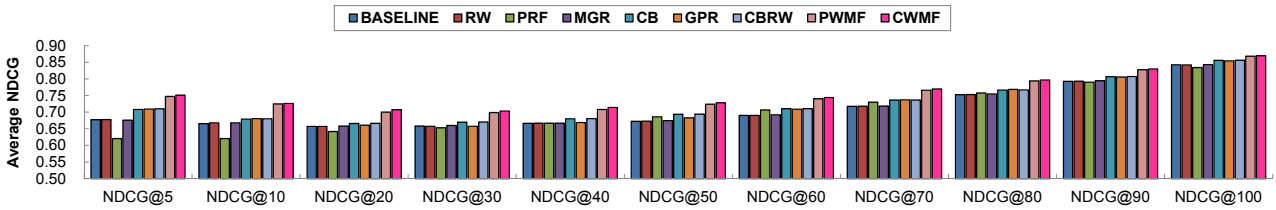


Figure 6: Comparison of re-ranking approaches in terms of NDCG.

- Click-boosting (CB). CB performs re-ranking by leveraging click-through data only, namely, CB re-ranks images according to their click counts in a descending order.
- Gaussian process regression (GPR) [9]. GPR first detects clicked images and performs dimensionality reduction on all the visual features. Then, a Gaussian process regressor is trained on the clicked images to predict the click counts of all images. The final re-ranking scores are calculated based on the predicted click counts and the initial ranking scores.
- Click-boosting random walk (CBRW) [26]. A two-step re-ranking method which is a combination of using click-through data and detecting visual recurrent patterns for image search re-ranking.
- Pair-wise multimodal fusion (PWMF). PWMF is the same as CWMF in dealing with multiple modalities and the selection of click-wise-based image pairs. The only difference is that the Hinge Loss adapting parameter λ is not concerned in PWMF, indicating that PWMF is based on the original ranking SVM with multimodal fusion. Note that PWMF is a deformation of CWMF and also proposed in this paper.

Figure 6 shows the overall performance of different re-ranking approaches in our tail queries dataset. On the whole, our proposed click-wise multimodal fusion (CWMF) outperforms other methods, and the improvements are consistent and stable at different depths of NDCG. Using our re-ranking approach the NDCG values are improved significantly. For example, the NDCG@5 is improved by 10.88% from the baseline of 0.6775 to 0.7512, and NDCG@10 is boosted by 9.12% from 0.6658 to 0.7265 on the entire dataset.

It is noted that the performance of PRF has a serious degradation from NDCG@5 and NDCG@10, respectively. This is mainly because the assumption behind PRF, i.e., top ranked images are typically relevant in response to a given query, is not suitable for tail queries whose baseline results are often much lower than head queries'. Similarly, the re-ranking scores of images by RW and MGR are partially related to the initial ranked lists. Even by tuning the tradeoff parameters, they still obtain weak improvements over the baseline results, which demonstrates the difficulty for tail queries to detect recurrent patterns via "similarity" mining. Compared with RW, PRF and MGR, we can see that CB performs better than BASELINE at all NDCG levels, indicating that the click-through data of tail queries can provide helpful information on user feedback for image re-ranking. Leveraging click counts of images, GPR and CBRW obtain better performance than CB's, however, the improvements are inconspicuous. The reasons, for GPR, possibly originate from the lack of training clicked images for Gaussian process regression and the reliance on the initial ranking scores. CBRW is due to the same reason as RW's. The fact that PWMF and CWMF achieve significant improvements proves the highly usefulness of click-wise information from image pairs. Compared with CWMF, PWMF does not take the problem of punishing the misclassified image pairs into consideration, i.e., there is no Hinge loss adapting parameter λ in PWMF. As we can see from Fig. 6, PWMF gets a comparatively poor performance, which exactly demonstrates the availability and superiority of the click-based adapting ranking SVM used in CWMF.

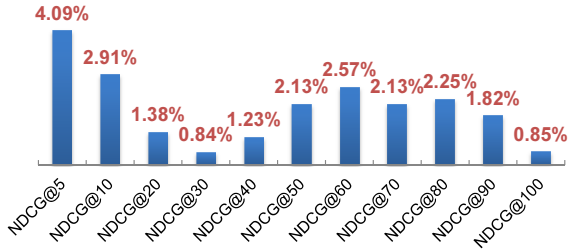


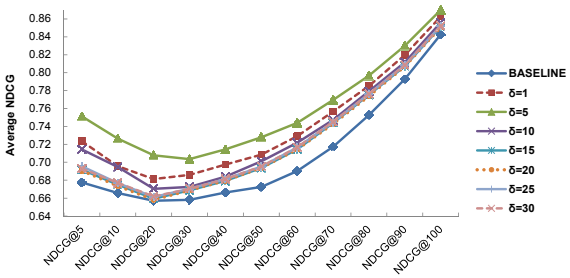
Figure 7: Improvements using click-wise multimodal fusion compared with the baseline in terms of NDCG for 50 head queries.

Overall, by simultaneously leveraging multimodal fusion and click-based adaption, our proposed approach CWMF can significantly improve the performance of image search for tail queries. To the best of our knowledge, this is the first attempt to use the difference values of clicks (click-wise information) and multiple modalities at the same time for image search re-ranking of tail queries.

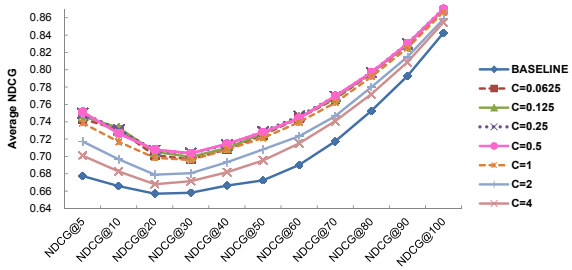
As click-wise multimodal fusion re-ranking proved significantly effective for tail queries, it is worth noting that CWMF is general and can be applied to head queries. We randomly select 50 queries with more than 100 query frequency in the one-week query log, which means that these queries belong to the "HEAD" region in Fig. 2. Due to the space limitation, for the 50 head queries, we only demonstrate the improvements using our proposed approach compared with the baseline results from NDCG@5 to NDCG@100 in Fig. 7. The improvements can be observed at all levels of NDCG, though the improvements are comparatively small because of the good initial search performance of head queries already obtained by search engines.

5.3 Parameter Analysis

We adopt five-fold cross validation to test the sensitivity of parameters δ and C in our proposed approach. We first set C to 0.5 and vary δ from 1 to 30 with an interval of 5. Figure 8(a) demonstrates the performance curve with various δ. We can find that though higher than the baseline's, the NDCG degrades when $\delta = 1$ or as δ gets larger compared with $\delta = 5$. It is understandable that when $\delta = 1$ our approach cannot obtain the best performance, because it is hard to tell the relevance relation between two images from a pair with only one click difference. For the larger δ, since the click counts of images are comparatively small for tail queries, δ should not go too far to select the possible click-wise-based pairs. If there is no pair matching the filter criterion, in our experiments, we relax the filter criterion and define the set of click-wise-based pairs as $\mathcal{P} \triangleq \{(i,j)|(c_i - c_j) > 0\}$. However, the Hinge Loss adapting parameter λ is not concerned for pairs in this $\mathcal{P}$. Thus,



(a) Re-ranking performance variation with various δ.



(b) Re-ranking performance variation with various C.

Figure 8: The sensitivity of re-ranking parameters.

the NDCG drops relatively due to the lack of click-wise-based pairs or the influence of λ. We then set δ to 5 and vary C from 0.0625 (2^{-3}) to 32 (2^5). Figure 8(b) demonstrates the performance curve with various C. We can see that the performance of our approach will not significantly degrade when C varies in a certain range.

5.4 Complexity Analysis

Since our proposed CWMF is query-dependent and computed by an approximate optimal solution through an iterative strategy, its complexity is mainly determined by the number of click-wise-based pairs, the ranking SVM solver, the reduced gradient algorithm and the iterations controlled by the terminated criterion. Take the 100 tail queries used in our experiments for example, i.e., each query corresponds to 100 images to be re-ranked by seven modalities, our approach takes 1.20 seconds on average for a given query on a regular PC (Intel quad-core 3.30GHz CPU and 8GB RAM) to complete the entire re-ranking process. To obtain satisfying results for tail queries, there is a tradeoff between the performance and the computation time using our approach. We will seek to reduce the time complexity in our consequent work.

5.5 Examples of Tail Query Re-ranking

To further verify the effectiveness of our proposed approach, in this section, we show some specific examples of tail queries. As mentioned in Sect. 5.2, most existing re-ranking approaches cannot be scaled well to tail queries, thus, in Fig. 9 we show the initial ranked list and the re-ranked list using our proposed click-wise multimodal fusion re-ranking of three typical tail queries severally.

For instance, "pewdiepie" is an online alias of a Swedish video game commentator on YouTube. Game fans may issue this query for checking on what "pewdiepie" looks like, however, we can find that there are only three "real" faces from the top 10 images returned by a search engine. Through mining the click-wise information and multiple modalities, our re-ranking approach achieves excellently satisfying results. It is worth mentioning that all the top 10 images in the re-ranked list of "pewdiepie" are portraits of himself and comply with user search intent.

6. CONCLUSIONS

In the paper, we have studied the problem of image search re-ranking for tail queries, which is often overlooked in the research communities. We have proposed a novel re-ranking approach based on the assumptions that *images with more clicks are more relevant to the given query than the ones with no or relatively less clicks* and *the effects of different visual modalities to re-rank images are query-dependent*. Our proposed approach can not only fully explore the effects of multiple visual modalities by adaptively predicting the query-dependent fusion weights, but also effectively expand training data by learning relevant information from click-wise-based image pairs. The experiments conducted on a real-world dataset demonstrate that different modalities, though the local feature SIFT is always assigned a high fusion weight, can incorporate with each other to improve search performance in a query-dependent way. For instance, a proper linear fusion of face feature and SIFT learnt by our approach is proved effective for improving the search performance of queries about characters, such as "pewdiepie" ($d_{FACE} = 0.1849, d_{SIFT} = 0.8151$). As mentioned in Sect. 5.1, there are two types of tail queries: one is from region "TAIL-A" in Fig. 2 and the other is from "TAIL-B." Using our proposed approach, the NDCG@5 of queries from "TAIL-A" is improved by 10.96% compared with the initial search results, and the corresponding improvement of queries from "TAIL-B" is 10.64%. This further shows the superiority of relevant information mined from click-wise-based image pairs used in our approach.

Intuitively, the click data provided by an individual user are believed to be possibly biased, yet from the experimental results of queries in "TAIL-B," they can be used as a useful indicator of relevance feedback. Thus, we believe that using click-through data for personal search is an interesting research topic in the future.

7. ACKNOWLEDGMENTS

This work was performed when the first author Xiaopeng Yang was visiting Microsoft Research as a research intern. In addition, this work was partially supported by National High Technology Research and Development Program of China (2014AA015202), the National Key Technology Research and Development Program of China (2012BAH39B02), the National Natural Science Foundation of China (61173054, 61172153), the Beijing New Star Project on Science & Technology (2007B071).

8. REFERENCES

[1] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. *ACM SIGKDD*, pages 76–85, 2007.

[2] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting ranking SVM to document retrieval. *ACM SIGIR*, pages 186–193, 2006.

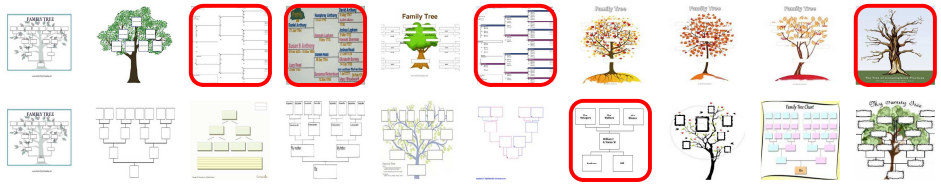
[3] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. *Advances in Neural Information Processing Systems*, 2007.

[4] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. *WWW*, pages 1–10, 2009.

[5] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. *ACM WSDM*, pages 181–190, 2010.

[6] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen. Extreme video retrieval: Joint maximization of human and computer performance. *ACM Multimedia*, pages 385–393, 2006.

(a) "pewdiepie" (81, 57)



(b) "blank family tree" (8, 20)

(c) "old engine" (1, 28)

Figure 9: The initial ranked list and the re-ranked list using our proposed approach for specific tail queries. First row of each tail query corresponds to the initial ranked list, and the second row corresponds to the re-ranked list using click-wise multimodal fusion. The red rectangles mark irrelevant images. The brackets next to the query show the corresponding query frequency and the number of clicked images during one week, respectively [best viewed in color].

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

[8] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. *ACM Multimedia*, pages 243–252, 2013.

[9] V. Jain and M. Varma. Learning to rerank: Query-dependent image reranking using click data. *WWW*, pages 277–286, 2011.

[10] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. *ACM SIGIR*, pages 41–48, 2000.

[11] T. Joachims. Optimizing search engines using clickthrough data. *ACM SIGKDD*, pages 133–142, 2002.

[12] T. Joachims, L. Granka, and B. Pan. Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR*, pages 154–161, 2005.

[13] L. S. Kennedy, A. P. Natsev, and S. Chang. Automatic discovery of query-class-dependent models for multimodal search. *ACM Multimedia*, pages 882–891, 2005.

[14] T.-M. Kuo, C.-P. Lee, and C.-J. Lin. Large-scale kernel rankSVM. *SIAM International Conference on Data Mining*, pages 812–820, 2014.

[15] Y. Liu, T. Mei, and X.-S. Hua. CrowdReranking: Exploring multiple search engines for visual search reranking. *ACM SIGIR*, pages 500–507, 2009.

[16] D. G. Lowe. Object recognition from local scale-invariant features. *IEEE ICCV*, pages 1150–1157, 1999.

[17] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(38), 2014.

[18] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[19] C. G. M. Snoek, K. van de Sande, O. de Rooij, and *et al.* The mediamill TRECVID 2008 semantic video search engine. *NIST TRECVID Workshop*, 2008.

[20] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. *ACM Multimedia*, pages 399–402, 2005.

[21] X. Tian, D. Tao, X.-S. Hua, and X. Wu. Active reranking for web image search. *IEEE Trans. on Image Processing*, 19(3):805–820, 2010.

[22] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu. Multimodal graph-based reranking for web image search. *IEEE Trans. on Image Processing*, 21(11):4649–4661, 2012.

[23] R. Yan and A. Hauptmann. Query expansion using probabilistic local feedback with application to multimedia retrieval. *ACM CIKM*, pages 361–370, 2007.

[24] L. Yang and A. Hanjalic. Supervised reranking for web image search. *ACM Multimedia*, pages 183–192, 2010.

[25] X. Yang, Y. Zhang, T. Yao, C.-W. Ngo, and T. Mei. Click-boosting multi-modality graph-based reranking for image search. *Multimedia Systems*, May 2014.

[26] X. Yang, Y. Zhang, T. Yao, Z.-J. Zha, and C.-W. Ngo. Click-boosting random walk for image search reranking. *ACM ICIMCS*, pages 1–6, 2013.

[27] T. Yao, C.-W. Ngo, and T. Mei. Circular reranking for visual search. *IEEE Trans. on Image Processing*, 22(4):1644–1655, 2013.

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