Ranking Web Objects from Multiple Communities

Le Chen*
Le.Chen@idiap.ch

Lei Zhang
leizhang@microsoft.com

Feng Jing
fengjing@microsoft.com

Ke-Feng Deng
kefengdeng@hotmail.com

Wei-Ying Ma
wyma@microsoft.com

Microsoft Research Asia
5F, Sigma Center, No. 49, Zhichun Road
Haidian District, Beijing, 100080, P R China

ABSTRACT

Vertical search is a promising direction as it leverages domain-specific knowledge and can provide more precise information for users. In this paper, we study the Web object-ranking problem, one of the key issues in building a vertical search engine. More specifically, we focus on this problem in cases when objects lack relationships between different Web communities, and take high-quality photo search as the test bed for this investigation. We proposed two score fusion methods that can automatically integrate as many Web communities (Web forums) with rating information as possible. The proposed fusion methods leverage the hidden links discovered by a duplicate photo detection algorithm, and aims at minimizing score differences of duplicate photos in different forums. Both intermediate results and user studies show the proposed fusion methods are practical and efficient solutions to Web object ranking in cases we have described. Though the experiments were conducted on high-quality photo ranking, the proposed algorithms are also applicable to other ranking problems, such as movie ranking and music ranking.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; G.2.2 [**Discrete Mathematics**]: Graph Theory; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Algorithms, Experimentation

Keywords

Web objects, image search, ranking

1. INTRODUCTION

Despite numerous refinements and optimizations, general purpose search engines still fail to find relevant results for many queries. As a new trend, vertical search has shown promise because it can leverage domain-specific knowledge and is more effective in connecting users with the information they want. There are many vertical search engines, including some for paper search (e.g. *Libra* [21], *Citeseer* [7] and *Google Scholar* [4]), product search (e.g. Froogle [5]), movie search [6], image search [1, 8], video search [6], local search [2], as well as news search [3]. We believe the vertical search engine trend will continue to grow.

Essentially, building vertical search engines includes data crawling, information extraction, object identification and integration, and object-level Web information retrieval (or Web object ranking) [20], among which ranking is one of the most important factors. This is because it deals with the core problem of how to combine and rank objects coming from multiple communities.

Although object-level ranking has been well studied in building vertical search engines, there are still some kinds of vertical domains in which objects cannot be effectively ranked. For example, algorithms that evolved from PageRank [22], PopRank [21] and LinkFusion [27] were proposed to rank objects coming from multiple communities, but can only work on well-defined graphs of heterogeneous data. "Well-defined" means that like objects (e.g. authors in paper search) can be identified in multiple communities (e.g. conferences). This allows heterogeneous objects to be well linked to form a graph through leveraging all the relationships (e.g. cited-by, authored-by and published-by) among the multiple communities.

However, this assumption does not always stand for some domains. High-quality photo search, movie search and news search are exceptions. For example, a photograph forum

*Le Chen did this work at Microsoft Research Asia as a visiting scholar.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

website usually includes three kinds of objects: photos, authors and reviewers. Yet different photo forums seem to lack any relationships, as there are no cited-by relationships. This makes it difficult to judge whether two authors cited are the same author, or two photos are indeed identical photos. Consequently, although each photo has a rating score in a forum, it is non-trivial to rank photos coming from different photo forums. Similar problems also exist in movie search and news search. Although two movie titles can be identified as the same one by title and director in different movie discussion groups, it is non-trivial to combine rating scores from different discussion groups and rank movies effectively. We call such non-trivial object relationship in which identification is difficult, *incomplete relationships*.

Other related work includes rank aggregation for the Web [13, 14], and learning algorithm for rank, such as RankBoost [15], RankSVM [17, 19], and RankNet [12]. We will contrast differences of these methods with the proposed methods after we have described the problem and our methods.

We will specifically focus on Web object-ranking problem in cases that lack object relationships or have with incomplete object relationships, and take high-quality photo search as the test bed for this investigation. In the following, we will introduce rationale for building high-quality photo search.

1.1 High-Quality Photo Search

In the past ten years, the Internet has grown to become an incredible resource, allowing users to easily access a huge number of images. However, compared to the more than 1 billion images indexed by commercial search engines, actual queries submitted to image search engines are relatively minor, and occupy only 8-10 percent of total image and text queries submitted to commercial search engines [24]. This is partially because user requirements for image search are far less than those for general text search. On the other hand, current commercial search engines still cannot well meet various user requirements, because there is no effective and practical solution to understand image content.

To better understand user needs in image search, we conducted a query log analysis based on a commercial search engine. The result shows that more than 20% of image search queries are related to nature and places and daily life categories. Users apparently are interested in enjoying high-quality photos or searching for beautiful images of locations or other kinds. However, such user needs are not well supported by current image search engines because of the difficulty of the quality assessment problem.

Ideally, the most critical part of a search engine – the ranking function – can be simplified as consisting of two key factors: relevance and quality. For the relevance factor, search in current commercial image search engines provide most returned images that are quite relevant to queries, except for some ambiguity. However, as to quality factor, there is still no way to give an optimal rank to an image. Though content-based image quality assessment has been investigated over many years [23, 25, 26], it is still far from ready to provide a realistic quality measure in the immediate future.

Seemingly, it really looks pessimistic to build an image search engine that can fulfill the potentially large requirement of enjoying high-quality photos. Various proliferating Web communities, however, notices us that people today have created and shared a lot of high-quality photos on the Web on virtually any topics, which provide a rich source for building a better image search engine.

In general, photos from various photo forums are of higher quality than personal photos, and are also much more appealing to public users than personal photos. In addition, photos uploaded to photo forums generally require rich metadata about title, camera setting, category and description to be provide by photographers. These metadata are actually the most precise descriptions for photos and undoubtedly can be indexed to help search engines find relevant results. More important, there are volunteer users in Web communities actively providing valuable ratings for these photos. The rating information is generally of great value in solving the photo quality ranking problem.

Motivated by such observations, we have been attempting to build a vertical photo search engine by extracting rich metadata and integrating information form various photo Web forums. In this paper, we specifically focus on how to rank photos from multiple Web forums.

Intuitively, the rating scores from different photo forums can be empirically normalized based on the number of photos and the number of users in each forum. However, such a straightforward approach usually requires large manual effort in both tedious parameter tuning and subjective results evaluation, which makes it impractical when there are tens or hundreds of photo forums to combine. To address this problem, we seek to build relationships/links between different photo forums. That is, we first adopt an efficient algorithm to find duplicate photos which can be considered as hidden links connecting multiple forums. We then formulate the ranking challenge as an optimization problem, which eventually results in an optimal ranking function.

1.2 Main Contributions and Organization.

The main contributions of this paper are:

1. We have proposed and built a vertical image search engine by leveraging rich metadata from various photo forum Web sites to meet user requirements of searching for and enjoying high-quality photos, which is impossible in traditional image search engines.

2. We have proposed two kinds of Web object-ranking algorithms for photos with incomplete relationships, which can automatically and efficiently integrate as many as possible Web communities with rating information and achieves an equal qualitative result compared with the manually tuned fusion scheme.

The rest of this paper is organized as follows. In Section 2, we present in detail the proposed solutions to the ranking problem, including how to find hidden links between different forums, normalize rating scores, obtain the optimal ranking function, and contrast our methods with some other related research. In Section 3, we describe the experimental setting and experiments and user studies conducted to evaluate our algorithm. Our conclusion and a discussion of future work is in Section 4.

It is worth noting that although we treat vertical photo search as the test bed in this paper, the proposed ranking algorithm can also be applied to rank other content that includes video clips, poems, short stories, drawings, sculptures, music, and so on.

2. ALGORITHM

2.1 Overview

The difficulty of integrating multiple Web forums is in their different rating systems, where there are generally two kinds of freedom. The first kind of freedom is the rating interval or rating scale including the minimal and maximal ratings for each Web object. For example, some forums use a 5-point rating scale whereas other forums use 3-point or 10-point rating scales. It seems easy to fix this freedom, but detailed analysis of the data and experiments show that it is a non-trivial problem.

The second kind of freedom is the varying rating criteria found in different Web forums. That is, the same score does not mean the same quality in different forums. Intuitively, if we can detect same photographers or same photographs, we can build relationships between any two photo forums and therefore can standardize the rating criterion by score normalization and transformation. Fortunately, we find that quite a number of duplicate photographs exist in various Web photo forums. This fact is reasonable when considering that photographers sometimes submit a photo to more than one forum to obtain critiques or in hopes of widespread publicity. In this work, we adopt an efficient duplicate photo detection algorithm [10] to find these photos.

The proposed methods below are based on the following considerations. Faced with the need to overcome a ranking problem, a standardized rating criterion rather than a reasonable rating criterion is needed. Therefore, we can take a large scale forum as the reference forum, and align other forums by taking into account duplicate Web objects (duplicate photos in this work). Ideally, the scores of duplicate photos should be equal even though they are in different forums. Yet we can deem that scores in different forums – except for the reference forum – can vary in a parametric space. This can be determined by minimizing the objective function defined by the sum of squares of the score differences. By formulating the ranking problem as an optimization problem that attempts to make the scores of duplicate photos in non-reference forums as close as possible to those in the reference forum, we can effectively solve the ranking problem.

For convenience, the following notations are employed. S_{ki} and $\bar{S}_{ki}$ denote the total score and mean score of ith Web object (photo) in the kth Web site, respectively. The total score refers to the sum of the various rating scores (e.g., novelty rating and aesthetic rating), and the mean score refers to the mean of the various rating scores. Suppose there are a total of K Web sites. We further use

$$\{S_i^{kl} | i = 1, ..., I_{kl}; k, l = 1, ..., K; k \neq l\}$$

to denote the set of scores for Web objects (photos) in kth Web forums that are duplicate with the lth Web forums, where I_{kl} is the total number of duplicate Web objects between these two Web sites. In general, score fusion can be seen as the procedure of finding K transforms

$$\psi_k(S_{ki}) \quad = \quad \widetilde{S}_{ki}, \quad k = 1, ..., K$$

such that $\widetilde{S}_{ki}$ can be used to rank Web objects from different Web sites. The objective function described in the above paragraph can then be formulated as

$$\min_{\{\psi_k | k=2,...,K\}} \quad \sum_{k=2}^{K} \sum_{i=1}^{I_{k1}} \bar{w}_i^k \left(S_i^{1k} - \psi_k(S_i^{k1}) \right)^2 \qquad (1)$$

where we use $k = 1$ as the reference forum and thus $\psi_1(S_{1i}) = S_{1i}$. $\bar{w}_i^k (\geq 0)$ is the weight coefficient that can be set heuristically according to the numbers of voters (reviewers or commenters) in both the reference forum and the non-reference forum. The more reviewers, the more popular the photo is and the larger the corresponding weight $\bar{w}_i^k$ should be. In this work, we do not inspect the problem of how to choose $\bar{w}_i^k$ and simply set them to one. But we believe the proper use of $\bar{w}_i^k$, which leverages more information, can significantly improve the results.

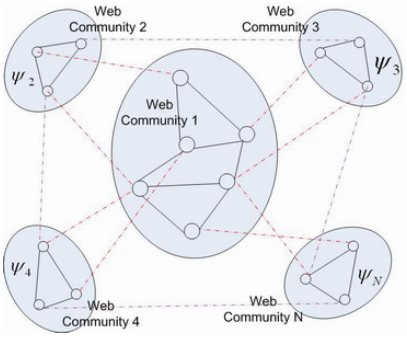


Figure 1: Web community integration. Each Web community forms a subgraph, and all communities are linked together by some hidden links (dashed lines).

Figure 1 illustrates the aforementioned idea. The Web Community 1 is the reference community. The dashed lines are links indicating that the two linked Web objects are actually the same. The proposed algorithm will try to find the best $\psi_k (k = 2, ..., K)$, which has certain parametric forms according to certain models. So as to minimize the cost function defined in Eq. 1, the summation is taken on all the red dashed lines.

We will first discuss the score normalization methods in Section 2.2, which serves as the basis for the following work. Before we describe the proposed ranking algorithms, we first introduce a manually tuned method in Section 2.3, which is laborious and even impractical when the number of communities become large. In Section 2.4, we will briefly explain how to precisely find duplicate photos between Web forums. Then we will describe the two proposed methods: Linear fusion and Non-linear fusion, and a performance measure for result evaluation in Section 2.5. Finally, in Section 2.6 we will discuss the relationship of the proposed methods with some other related work.

2.2 Score Normalization

Since different Web (photo) forums on the Web usually have different rating criteria, it is necessary to normalize them before applying different kinds of fusion methods. In addition, as there are many kinds of ratings, such as ratings for novelty, ratings for aesthetics etc, it is reasonable to choose a common one — total score or average score — that can always be extracted in any Web forum or calculated by corresponding ratings. This allows the normaliza-

tion method on the total score or average score to be viewed as an impartial rating method between different Web forums.

It is straightforward to normalize average scores by linearly transforming them to a fixed interval. We call this kind of score as *Scaled Mean Score*. The difficulty, however, of using this normalization method is that, if there are only a few users rating an object, say a photo in a photo forum, the average score for the object is likely to be spammed or skewed.

Total score can avoid such drawbacks that contain more information such as a Web object's quality and popularity. The problem is thus how to normalize total scores in different Web forums. The simplest way may be normalization by the maximal and minimal scores. The drawback of this normalization method is it is non robust, or in other words, it is sensitive to outliers.

To make the normalization insensitive to unusual data, we propose the *Mode-90% Percentile* normalization method. Here, the mode score represents the total score that has been assigned to more photos than any other total score. And The high percentile score (e.g.,90%) represents the total score for which the high percentile of images have a lower total score. This normalization method utilizes the mode and 90% percentile as two reference points to align two rating systems, which makes the distributions of total scores in different forums more consistent. The underlying assumption, for example in different photo forums, is that even the qualities of top photos in different forums may vary greatly and be less dependent on the forum quality, the distribution of photos of middle-level quality (from mode to 90% percentile) should be almost of the same quality up to the freedom which reflects the rating criterion (strictness) of Web forums. Photos of this middle-level in a Web forum usually occupy more than 70 % of total photos in that forum.

We will give more detailed analysis of the scores in Section 3.2.

2.3 Manual Fusion

The Web movie forum, IMDB [16], proposed to use a Bayesian-ranking function to normalize rating scores within one community. Motivated by this ranking function, we propose this manual fusion method: For the kth Web site, we use the following formula

$$\widetilde{S}_{ki} = \alpha_k \cdot \left(\frac{n_k \cdot \bar{S}_{ki}}{n_k + n_k^*} + \frac{n_k^* \cdot S_k^*}{n_k + n_k^*} \right) \quad (2)$$

to rank photos, where n_k is the number of votes and n_k^*, S_k^* and α_k are three parameters. This ranking function first takes a balance between the original mean score $\bar{S}_{ki}$ and a reference score S_k^* to get a weighted mean score which may be more reliable than $\bar{S}_{ki}$. Then the weighted mean score is scaled by α_k to get the final score $\widetilde{S_{ki}}$.

For n Web communities, there are then about $3n$ parameters in $\{(\alpha_k, n_k^*, S_k^*)|k = 1, ..., n\}$ to tune. Though this method can achieves pretty good results after careful and thorough manual tuning on these parameters, when n becomes increasingly large, say there are tens or hundreds of Web communities crawled and indexed, this method will become more and more laborious and will eventually become impractical. It is therefore desirable to find an effective fusion method whose parameters can be automatically determined.

2.4 Duplicate Photo Detection

We use Dedup [10], an efficient and effective duplicate image detection algorithm, to find duplicate photos between any two photo forums. This algorithm uses hash function to map a high dimensional feature to a 32 bits hash code (see below for how to construct the hash code). Its computational complexity to find all the duplicate images among n images is about $O(n \log n)$. The low-level visual feature for each photo is extracted on $k \times k$ regular grids. Based on all features extracted from the image database, a PCA model is built. The visual features are then transformed to a relatively low-dimensional and zero mean PCA space, or 29 dimensions in our system. Then the hash code for each photo is built as follows: each dimension is transformed to one, if the value in this dimension is greater than 0, and 0 otherwise. Photos in the same bucket are deemed potential duplicates and are further filtered by a threshold in terms of Euclidean similarity in the visual feature space.

Figure 2 illustrates the hashing procedure, where visual features — mean gray values — are extracted on both 6×6 and 7×7 grids. The 85-dimensional features are transformed to a 32-dimensional vector, and the hash code is generated according to the signs.

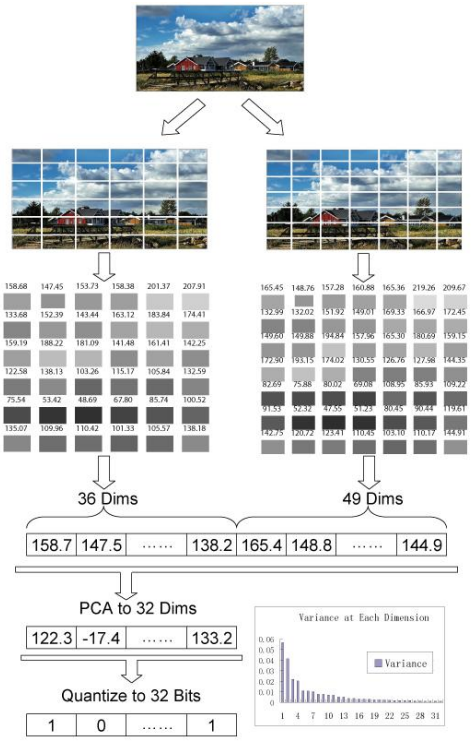


Figure 2: Hashing procedure for duplicate photo dectection

2.5 Score Fusion

In this section, we will present two solutions on score fusion based on different parametric form assumptions of ψ_k in Eq. 1.

2.5.1 *Linear Fusion by Duplicate Photos*

Intuitively, the most straightforward way to factor out the uncertainties caused by the different criterion is to scale, rel-

ative to a given center, the total scores of each unreferenced Web photo forum with respect to the reference forum. More strictly, we assume ψ_k has the following form

$$\psi_k(S_{ki}) = \alpha_k S_{ki} + t_k, \quad k = 2, ..., K \tag{3}$$
$$\psi_1(S_{1i}) = S_{1i} \tag{4}$$

which means that the scores of $k(\neq 1)$th forum should be scaled by α_k relative to the center $\frac{t_k}{1-\alpha_k}$ as shown in Figure 3.

Then, if we substitute above ψ_k to Eq. 1, we get the following objective function,

$$\min_{\{\alpha_k, t_k | k=2,...,K\}} \sum_{k=2}^{K} \sum_{i=1}^{I_{k1}} \bar{w}_i^k \left[S_i^{1k} - \alpha_k S_i^{k1} - t_k \right]^2. \tag{5}$$

By solving the following set of functions,

$$\begin{cases} \frac{\partial f}{\partial \alpha_k} = 0 \\ \frac{\partial f}{\partial t_k} = 0 \end{cases}, \quad k = 1, ..., K$$

where f is the objective function defined in Eq. 5, we get the closed form solution as:

$$\begin{pmatrix} \alpha_k \\ t_k \end{pmatrix} = A_k^{-1} L_k \tag{6}$$

where

$$A_k = \begin{pmatrix} \sum_i \bar{w}_i (S_i^{k1})^2 & \sum_i \bar{w}_i S_i^{k1} \\ \sum_i \bar{w}_i S_i^{k1} & \sum_i \bar{w}_i \end{pmatrix} \tag{7}$$

$$L_k = \begin{pmatrix} \sum_i \bar{w}_i S_i^{1k} S_i^{k1} \\ \sum_i \bar{w}_i S_i^{1k} \end{pmatrix} \tag{8}$$

and $k = 2, ..., K$.

This is a linear fusion method. It enjoys simplicity and excellent performance in the following experiments.

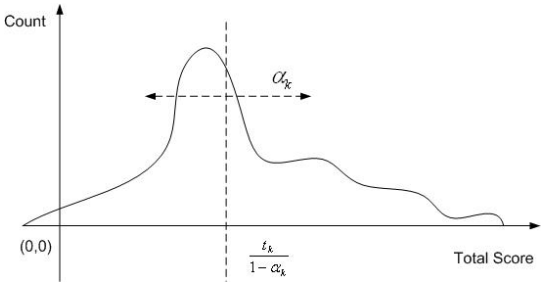


Figure 3: Linear Fusion method

2.5.2 Nonlinear Fusion by Duplicate Photos

Sometimes we want a method which can adjust scores on intervals with two endpoints unchanged. As illustrated in Figure 4, the method can tune scores between $[C_0, C_1]$ while leaving scores C_0 and C_1 unchanged. This kind of fusion method is then much finer than the linear ones and contains many more parameters to tune and expect to further improve the results.

Here, we propose a nonlinear fusion solution to satisfy such constraints. First, we introduce a transform:

$$\eta_{c_0,c_1,\alpha}(x) = \begin{cases} \left(\frac{x-c_0}{c_1-c_0}\right)^{\alpha} (c_1 - c_0) + c_0, & \text{if } x \in (c_0, c_1] \\ x & \text{otherwise} \end{cases}$$

where $\alpha > 0$. This transform satisfies that for $x \in [c_0, c_1]$, $\eta_{c_0,c_1,\alpha}(x) \in [c_0, c_1]$ with $\eta_{c_0,c_1,\alpha}(c_0) = c_0$ and $\eta_{c_0,c_1,\alpha}(c_1) = c_1$. Then we can utilize this nonlinear transform to adjust the scores in certain interval, say $(M, T]$,

$$\psi_k(S_{ki}) = \eta_{M,T,\alpha}(S_{ki}) \quad . \tag{9}$$

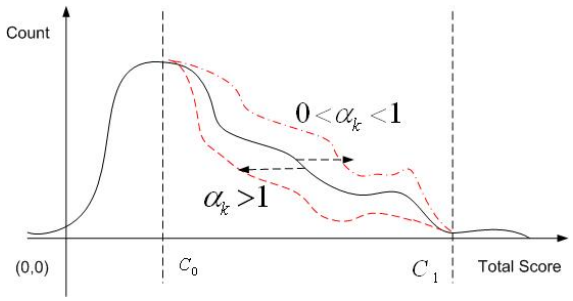


Figure 4: Nonlinear Fusion method. We intent to finely adjust the shape of the curves in each segment.

Even there is no closed-form solution for the following optimization problem,

$$\min_{\{\alpha_k | k \in [2,K]\}} \sum_{k=2}^{K} \sum_{i=1}^{I_{k1}} \bar{w}_i^k \left[S_i^{1k} - \eta_{M,T,\alpha}(S_{ki}) \right]^2$$

it is not hard to get the numeric one. Under the same assumptions made in Section 2.2, we can use this method to adjust scores of the middle-level (from the mode point to the 90 % percentile).

This more complicated non-linear fusion method is expected to achieve better results than the linear one. However, difficulties in evaluating the rank results block us from tuning these parameters extensively. The current experiments in Section 3.5 do not reveal any advantages over the simple linear model.

2.5.3 Performance Measure of the Fusion Results

Since our objective function is to make the scores of the same Web objects (e.g. duplicate photos) between a non-reference forum and the reference forum as close as possible, it is natural to investigate how close they become to each other and how the scores of the same Web objects change between the two non-reference forums before and after score fusion.

Taken Figure 1 as an example, the proposed algorithms minimize the score differences of the same Web objects in two Web forums: the reference forum (the Web Community 1) and a non-reference forum, which corresponds to minimizing the objective function on the *red* dashed (hidden) links. After the optimization, we must ask what happens to the score differences of the same Web objects in two non-reference forums? Or, in other words, whether the scores of two objects linked by the *green* dashed (hidden) links become more consistent?

We therefore define the following performance measure — δ measure — to quantify the changes for scores of the same Web objects in different Web forums as

$$\delta_{kl} = \text{Sim}(\mathbf{S}^{lk}, \mathbf{S}^{kl}) - \text{Sim}(\mathbf{S}_*^{lk}, \mathbf{S}_*^{kl}) \tag{10}$$

where $\mathbf{S}^{kl} = (S_1^{kl}, ..., S_{I_{kl}}^{kl})^T$, $\mathbf{S}_*^{kl} = (\widetilde{S}_1^{kl}, ..., \widetilde{S}_{I_{kl}}^{kl})^T$ and

$$\text{Sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||} .$$

$\delta_{kl} > 0$ means after score fusion, scores on the same Web objects between kth and lth Web forum become more consistent, which is what we expect. On the contrary, if $\delta_{kl} < 0$, those scores become more inconsistent.

Although we cannot rely on this measure to evaluate our final fusion results as ranking photos by their popularity and qualities is such a subjective process that every person can have its own results, it can help us understand the intermediate ranking results and provide insights into the final performances of different ranking methods.

2.6 Contrasts with Other Related Work

We have already mentioned the differences of the proposed methods with the traditional methods, such as PageRank [22], PopRank [21], and LinkFusion [27] algorithms in Section 1. Here, we discuss some other related works.

The current problem can also be viewed as a rank aggregation one [13, 14] as we deal with the problem of how to combine several rank lists. However, there are fundamental differences between them. First of all, unlike the Web pages, which can be easily and accurately detected as the same pages, detecting the same photos in different Web forums is a non-trivial work, and can only be implemented by some delicate algorithms while with certain precision and recall. Second, the numbers of the duplicate photos from different Web forums are small relative to the whole photo sets (see Table 1). In another words, the top K rank lists of different Web forums are almost disjointed for a given query. Under this condition, both the algorithms proposed in [13] and their measurements — Kendall tau distance or Spearman footrule distance — will degenerate to some trivial cases.

Another category of rank fusion (aggregation) methods is based on machine learning algorithms, such as RankSVM [17, 19], RankBoost [15], and RankNet [12]. All of these methods entail some labelled datasets to train a model. In current settings, it is difficult or even impossible to get these datasets labelled as to their level of professionalism or popularity, since the photos are too vague and subjective to rank. Instead, the problem here is how to combine several ordered sub lists to form a total order list.

3. EXPERIMENTS

In this section, we carry out our research on high-quality photo search. We first briefly introduce the newly proposed vertical image search engine — EnjoyPhoto in section 3.1. Then we focus on how to rank photos from different Web forums. In order to do so, we first normalize the scores (ratings) for photos from different multiple Web forums in section 3.2. Then we try to find duplicate photos in section 3.3. Some intermediate results are discussed using δ measure in section 3.4. Finally a set of user studies is carried out carefully to justify our proposed method in section 3.5.

3.1 EnjoyPhoto: high-quality Photo Search Engine

In order to meet user requirement of enjoying high-quality photos, we propose and build a high-quality photo search engine — EnjoyPhoto, which accounts for the following three key issues: 1. how to crawl and index photos, 2. how to determine the qualities of each photo and 3. how to display the search results in order to make the search process enjoyable. For a given text based query, this system ranks the photos based on certain combination of relevance of the photo to this query (Issue 1) and the quality of the photo (Issue 2), and finally displays them in an enjoyable manner (Issue 3).

As for Issue 3, we devise the interface of the system deliberately in order to smooth the users' process of enjoying high-quality photos. Techniques, such as Fisheye and slides show, are utilized in current system. Figure 5 shows the interface. We will not talk more about this issue as it is not an emphasis of this paper.

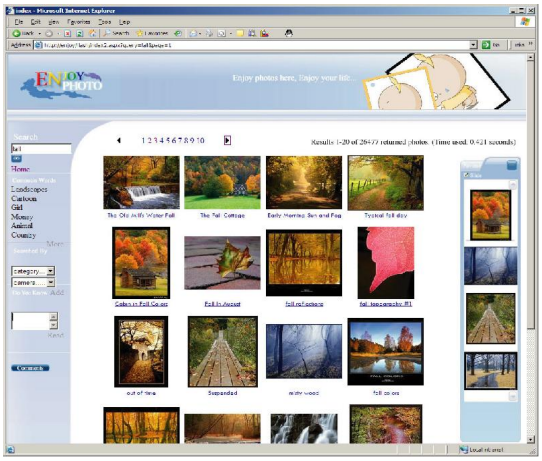


Figure 5: EnjoyPhoto: an enjoyable high-quality photo search engine, where 26,477 records are returned for the query "fall" in about 0.421 seconds

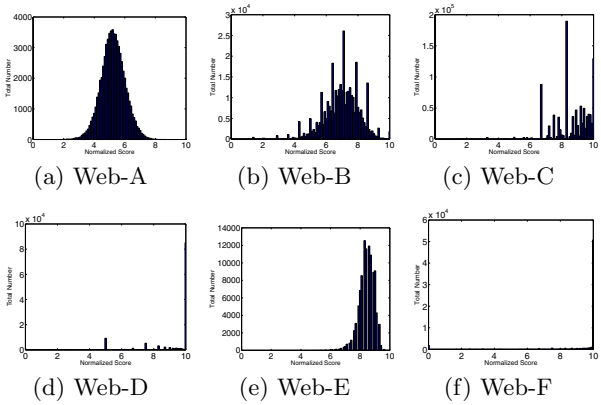
As for Issue 1, we extracted from a commercial search engine a subset of photos coming from various photo forums all over the world, and explicitly parsed the Web pages containing these photos. The number of photos in the data collection is about 2.5 million. After the parsing, each photo was associated with its title, category, description, camera setting, EXIF data [1] (when available for digital images), location (when available in some photo forums), and many kinds of ratings. All these metadata are generally precise descriptions or annotations for the image content, which are then indexed by general text-based search technologies [9, 18, 11]. In current system, the ranking function was specifically tuned to emphasize title, categorization, and rating information.

Issue 2 is essentially dealt with in the following sections which derive the quality of photos by analyzing ratings provided by various Web photo forums. Here we chose six photo forums to study the ranking problem and denote them as Web-A, Web-B, Web-C, Web-D, Web-E and Web-F.

3.2 Photo Score Normalization

Detailed analysis of different score normalization methods are analyzed in this section. In this analysis, the zero

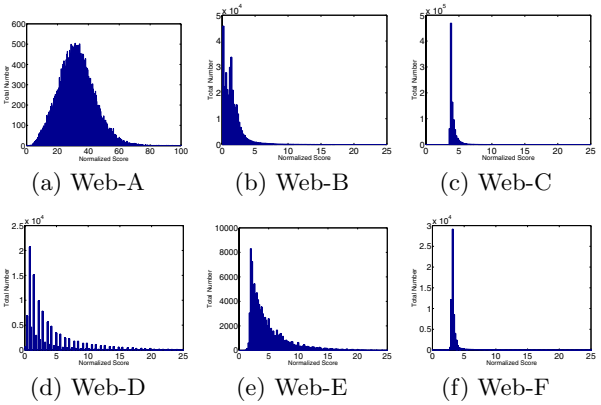
[1]Digital cameras save JPEG (.jpg) files with EXIF (Exchangeable Image File) data. Camera settings and scene information are recorded by the camera into the image file. `www.digicamhelp.com/what-is-exif/`



(a) Web-A (b) Web-B (c) Web-C

(d) Web-D (e) Web-E (f) Web-F

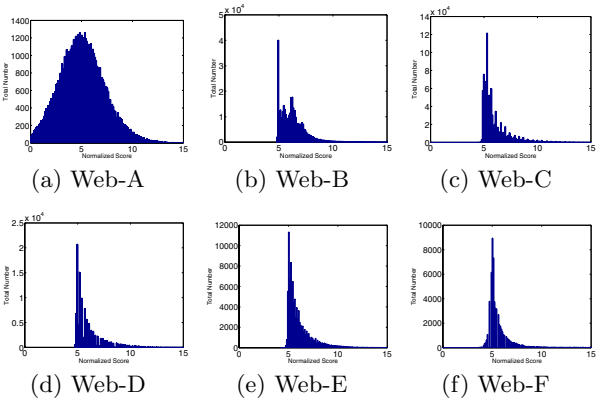
Figure 6: Distributions of mean scores normalized to $[0, 10]$



(a) Web-A (b) Web-B (c) Web-C

(d) Web-D (e) Web-E (f) Web-F

Figure 7: Maxmin Normalization



(a) Web-A (b) Web-B (c) Web-C

(d) Web-D (e) Web-E (f) Web-F

Figure 8: Mode-90% Percentile Normalization

scores that usually occupy about than 30% of the total number of photos for some Web forums are not currently taken into account. How to utilize these photos is left for future explorations.

In Figure 6, we list the distributions of the mean score, which is transformed to a fixed interval $[0, 10]$. The distributions of the average scores of these Web forums look quite different. Distributions in Figure 6(a), 6(b), and 6(e) looks like Gaussian distributions, while those in Figure 6(d) and 6(f) are dominated by the top score. The reason of these eccentric distributions for Web-D and Web-F lies in their coarse rating systems. In fact, Web-D and Web-F use 2 or 3 point rating scales whereas other Web forums use 7 or 14 point rating scales. Therefore, it will be problematic if we directly use these averaged scores. Furthermore the average score is very likely to be spammed, if there are only a few users rating a photo.

Figure 7 shows the total score normalization method by maximal and minimal scores, which is one of our base line system. All the total scores of a given Web forum are normalized to $[0, 100]$ according to the maximal score and minimal score of corresponding Web forum. We notice that total score distribution of Web-A in Figure 7(a) has two larger tails than all the others. To show the shape of the distributions more clearly, we only show the distributions on $[0, 25]$ in Figure 7(b),7(c),7(d),7(e), and 7(f).

Figure 8 shows the Mode-90% Percentile normalization method, where the modes of the six distributions are normalized to 5 and the 90% percentile to 8. We can see that this normalization method makes the distributions of total scores in different forums more consistent. The two proposed algorithms are all based on these normalization methods.

3.3 Duplicate photo detection

Targeting at computational efficiency, the Dedup algorithm may lose some recall rate, but can achieve a high precision rate. We also focus on finding precise hidden links rather than all hidden links. Figure 9 shows some duplicate detection examples. The results are shown in Table 1 and verify that large numbers of duplicate photos exist in any two Web forums even with the strict condition for Dedup where we chose first 29 bits as the hash code. Since there are only a few parameters to estimate in the proposed fusion methods, the numbers of duplicate photos shown Table 1 are sufficient to determine these parameters. The last table column lists the total number of photos in the corresponding Web forums.

3.4 δ Measure

The parameters of the proposed linear and nonlinear algorithms are calculated using the duplicate data shown in Table 1, where the Web-C is chosen as the reference Web forum since it shares the most duplicate photos with other forums.

Table 2 and 3 show the δ measure on the linear model and nonlinear model. As δ_{kl} is symmetric and $\delta_{kk} = 0$, we only show the upper triangular part. The NaN values in both tables lie in that no duplicate photos have been detected by the Dedup algorithm as reported in Table 1.

The linear model guarantees that the δ measures related

Table 1: Number of duplicate photos between each pair of Web forums

	A	B	C	D	E	F	Scale
A	0	316	1,386	178	302	0	130k
B	316	0	14,708	909	8,023	348	675k
C	1,386	14,708	0	1,508	19,271	1,083	1,003k
D	178	909	1,508	0	1,084	21	155k
E	302	8,023	19,271	1,084	0	98	448k
F	0	348	1,083	21	98	0	122k

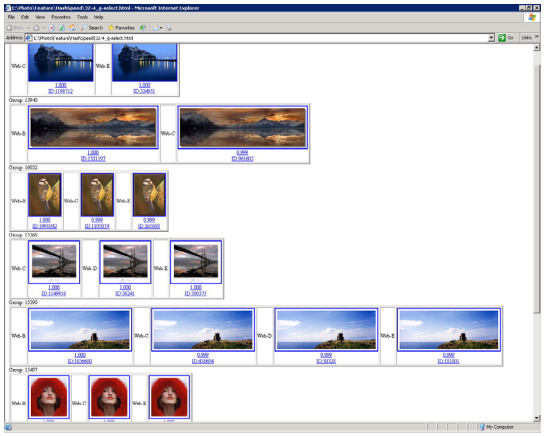


Figure 9: Some results of duplicate photo detection

Table 2: δ measure on the linear model.

	Web-B	Web-C	Web-D	Web-E	Web-F
Web-A	0.0659	0.0911	0.0956	0.0928	NaN
Web-B	–	0.0672	0.0578	0.0791	0.4618
Web-C	–	–	0.0105	0.0070	0.2220
Web-D	–	–	–	0.0566	0.0232
Web-E	–	–	–	–	0.6525

to the reference community should be no less than 0 theoretically. It is indeed the case (see the underlined numbers in Table 2). But this model can not guarantee that the δ measures on the non-reference communities can also be no less than 0, as the normalization steps are based on duplicate photos between the reference community and a non-reference community. Results shows that all the numbers in the δ measure are greater than 0 (see all the non-underlined numbers in Table 2), which indicates that it is probable that this model will give optimal results.

On the contrary, the nonlinear model does not guarantee that δ measures related to the reference community should be no less than 0, as not all duplicate photos between the two Web forums can be used when optimizing this model. In fact, the duplicate photos that lie in different intervals will not be used in this model. It is these specific duplicate photos that make the δ measure negative. As a result, there are both negative and positive items in Table 3, but overall the number of positive ones are greater than negative ones (9:5), that indicates the model may be better than the "normalization only" method (see next subsection) which has an all-zero δ measure, and worse than the linear model.

3.5 User Study

Because it is hard to find an objective criterion to evaluate which ranking function is better, we chose to employ user studies for subjective evaluations. Ten subjects were invited to participate in the user study. They were recruited from nearby universities. As search engines of both text search and image search are familiar to university students, there was no prerequisite criterion for choosing students.

We conducted user studies using Internet Explorer 6.0 on Windows XP with 17-inch LCD monitors set at 1,280 pixels by 1,024 pixels in 32-bit color. Data was recorded with server logs and paper-based surveys after each task.

Table 3: δ measure on the nonlinear model.

	Web-B	Web-C	Web-D	Web-E	Web-F
Web-A	0.0559	0.0054	-0.0185	-0.0054	NaN
Web-B	–	-0.0162	-0.0345	-0.0301	0.0466
Web-C	–	–	0.0136	0.0071	0.1264
Web-D	–	–	–	0.0032	0.0143
Web-E	–	–	–	–	0.214

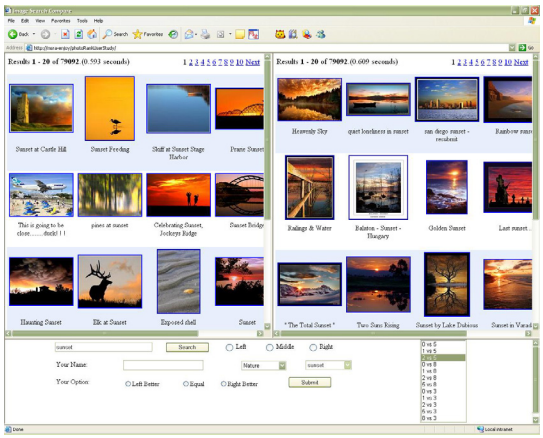


Figure 10: User study interface

We specifically device an interface for user study as shown in Figure 10. For each pair of fusion methods, participants were encouraged to try any query they wished. For those without specific ideas, two combo boxes (category list and query list) were listed on the bottom panel, where the top 1,000 image search queries from a commercial search engine were provided. After a participant submitted a query, the system randomly selected the left or right frame to display each of the two ranking results. The participant were then required to judge which ranking result was better of the two ranking results, or whether the two ranking results were of equal quality, and submit the judgment by choosing the corresponding radio button and clicking the "Submit" button.

For example, in Figure 10, query "sunset" is submitted to the system. Then, 79,092 photos were returned and ranked by the Minmax fusion method in the left frame and linear fusion method in the right frame. A participant then compares the two ranking results (without knowing the ranking methods) and submits his/her feedback by choosing answers in the "Your option."

Table 4: Results of user study

	Norm.Only	Manually	Linear
Linear	29:13:10	14:22:15	—
Nonlinear	29:15:9	12:27:12	6:4:45

Table 4 shows the experimental results, where "Linear" denotes the linear fusion method, "Nonlinear" denotes the non linear fusion method, "Norm. Only" means Maxmin normalization method, "Manually" means the manually tuned method. The three numbers in each item, say 29:13:10, mean that 29 judgments prefer the linear fusion results, 10

judgments prefer the normalization only method, and 13 judgments consider these two methods as equivalent.

We conduct the ANOVA analysis, and obtain the following conclusions:

1. Both the linear and nonlinear methods are significantly better than the "Norm. Only" method with respective P-values $0.00165(< 0.05)$ and $0.00073(<< 0.05)$. This result is consistent with the δ-measure evaluation result. The "Norm. Only" method assumes that the top 10% photos in different forums are of the same quality. However, this assumption does not stand in general. For example, a top 10% photo in a top tier photo forum is generally of higher quality than a top 10% photo in a second-tier photo forum. This is similar to that, those top 10% students in a top-tier university and those in a second-tier university are generally of different quality. Both linear and nonlinear fusion methods acknowledge the existence of such differences and aim at quantizing the differences. Therefore, they perform better than the "Norm. Only" method.

2. The linear fusion method is significantly better than the nonlinear one with P-value 1.195×10^{-10}. This result is rather surprising as this more complicated ranking method is expected to tune the ranking more finely than the linear one. The main reason for this result may be that it is difficult to find the best intervals where the nonlinear tuning should be carried out and yet simply the middle part of the Mode-90% Percentile Normalization method was chosen. The time-consuming and subjective evaluation methods — user studies — blocked us extensively tuning these parameters.

3. The proposed linear and nonlinear methods perform almost the same with or slightly better than the manually tuned method. Given that the linear/nonlinear fusion methods are fully automatic approaches, they are considered practical and efficient solutions when more communities (e.g. dozens of communities) need to be integrated.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the Web object-ranking problem in the cases of lacking object relationships where traditional ranking algorithms are no longer valid, and took high-quality photo search as the test bed for this investigation. We have built a vertical high-quality photo search engine, and proposed score fusion methods which can automatically integrate as many data sources (Web forums) as possible. The proposed fusion methods leverage the hidden links discovered by duplicate photo detection algorithm, and minimize score differences of duplicate photos in different forums. Both the intermediate results and the user studies show that the proposed fusion methods are a practical and efficient solution to Web object ranking in the aforesaid relationships. Though the experiments were conducted on high-quality photo ranking, the proposed algorithms are also applicable to other kinds of Web objects including video clips, poems, short stories, music, drawings, sculptures, and so on.

Current system is far from being perfect. In order to make this system more effective, more delicate analysis for the vertical domain (e.g., Web photo forums) are needed. The following points, for example, may improve the searching results and will be our future work: 1. more subtle analysis and then utilization of different kinds of ratings (e.g., novelty ratings, aesthetic ratings); 2. differentiating various communities who may have different interests and preferences or even distinct culture understandings; 3. incorporating more useful information, including photographers' and reviewers' information, to model the photos in a heterogeneous data space instead of the current homogeneous one. We will further utilize collaborative filtering to recommend relevant high-quality photos to browsers.

One open problem is whether we can find an objective and efficient criterion for evaluating the ranking results, instead of employing subjective and inefficient user studies, which blocked us from trying more ranking algorithms and tuning parameters in one algorithm.

5. ACKNOWLEDGMENTS

We thank Bin Wang and Zhi Wei Li for providing Dedup codes to detect duplicate photos; Zhen Li for helping us design the interface of EnjoyPhoto; Ming Jing Li, Longbin Chen, Changhu Wang, Yuanhao Chen, and Li Zhuang etc. for useful discussions. Special thanks go to Dwight Daniels for helping us revise the language of this paper.

6. REFERENCES

[1] Google image search. `http://images.google.com`.
[2] Google local search. `http://local.google.com/`.
[3] Google news search. `http://news.google.com`.
[4] Google paper search. `http://Scholar.google.com`.
[5] Google product search. `http://froogle.google.com`.
[6] Google video search. `http://video.google.com`.
[7] Scientific literature digital library. `http://citeseer.ist.psu.edu`.
[8] Yahoo image search. `http://images.yahoo.com`.
[9] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley, 1999.
[10] W. Bin, L. Zhiwei, L. Ming Jing, and M. Wei-Ying. Large-scale duplicate detection for web image search. In *Proceedings of the International Conference on Multimedia and Expo*, page 353, 2006.
[11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks*, volume 30, pages 107–117, 1998.
[12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89 – 96, 2005.
[13] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings 10th International Conference on World Wide Web*, pages 613 – 622, Hong-Kong, 2001.
[14] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134 – 160, 2003.
[15] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences.

Journal of Machine Learning Research, 4(1):933–969(37), 2004.

[16] IMDB. Formula for calculating the top rated 250 titles in imdb. `http://www.imdb.com/chart/top`.

[17] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133 – 142, 2002.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[19] R. Nallapati. Discriminative models for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64 – 71, 2004.

[20] Z. Nie, Y. Ma, J.-R. Wen, and W.-Y. Ma. Object-level web information retrieval. In *Technical Report of Microsoft Research*, volume MSR-TR-2005-11, 2005.

[21] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: Bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, pages 567 – 574, Chiba, Japan, 2005.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Technical report*, Stanford Digital Libraries, 1998.

[23] A. Savakis, S. Etz, and A. Loui. Evaluation of image appeal in consumer photography. In *SPIE Human Vision and Electronic Imaging*, pages 111–120, 2000.

[24] D. Sullivan. Hitwise search engine ratings. *Search Engine Watch Articles*, `http://searchenginewatch.com/reports/article.php/3099931`, August 23, 2005.

[25] S. Susstrunk and S. Winkler. Color image quality on the internet. In *IS&T/SPIE Electronic Imaging 2004: Internet Imaging V*, volume 5304, pages 118–131, 2004.

[26] H. Tong, M. Li, Z. H.J., J. He, and Z. C.S. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia (PCM)*, pages 198–205, 2004.

[27] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox. Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *Proceedings of the 13th international conference on World Wide Web*, pages 319 – 327, 2004.