Oracle In Image Search: A Content-Based Approach To Performance Prediction

Liqiang Nie, Meng Wang, Zheng-Jun Zha, Tat-Seng Chua
School of Computing, National University of Singapore

This paper studies a novel problem in image search. Given a text query and the image ranking list returned by an image search system, we propose an approach to automatically predict the search performance. We demonstrate that, in order to estimate the mathematical expectations of average precision (AP) and normalized discounted cumulative gain (NDCG), we only need to predict the relevance probability of each image. We accomplish the task with a query-adaptive graph-based learning based on the images’ ranking order and visual content. We validate our approach with a large-scale dataset containing the image search results of 1,165 queries from 4 popular image search engines. Empirical studies demonstrate that our approach is able to generate predictions that are highly correlated with the real search performance. Based on the proposed image search performance prediction scheme, we introduce three applications: image metasearch, multilingual image search and Boolean image search. Comprehensive experiments are conducted to validate our approach.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Index
General Terms: Algorithms, Experimentation, Human Factors
Additional Key Words and Phrases: Image search, search performance prediction, graph-based learning.

1. INTRODUCTION

The proliferation of digital cameras and the advances of networks and compression techniques have led to the overwhelming amount of image contents available online. It is predicted that the amount of digital images captured worldwide in 2011 will be over 60 billion according to IT Facts’ report1. It is also reported that there are over 2.5 billion photos uploaded to Facebook each month2. In the past two decades, intensive efforts have been dedicated to many research topics related to image search, such as content-based image retrieval [Smeulders et al. 2000][Lew 2006], image annotation [Li et al. 2006][Li and Wang 2003][Monay and Gatica-Perez 2003] and image ranking [Hsu et al. 2007][Jing and Baluja 2008]. On the other hand, online image search engines are also evolving rapidly. For example,


Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
© 20xx ACM 0000-0000/20xx/0000-0001 $5.00

ACM Journal Name, Vol. x, No. x, xx 20xx, Pages 1–077.
there are only 250 million images indexed by Google in 2001. But the number quickly turns to 1.1 billion in 2005, and now it is above 10 billion.\footnote{Google history. See: http://www.google.com/corporate/history.html}

Query performance prediction is an important research topic in information retrieval [Cronen-Townsend et al. 2002; Hauff et al. 2009; Hauff et al. 2008] and it is of great importance to many applications, such as metasearch, query suggestion and query expansion. Typically, there are two approaches, one is by directly analyzing the queries and the data collection and the other is by analyzing search results. They are usually called pre-search and post-search approaches, respectively. Many different methods have been developed for text-domain search and more details will be introduced in the next section. However, there is a lack of research on performance prediction for multimedia search based on text queries. Several pre-search methods can be applied to the multimedia search problem, however in comparison with post-search approach, pre-search methods usually can hardly achieve satisfactory performance as there is no information about search results. Clearly, the post-search methods cannot be directly applied to multimedia search as the text query and image search results belong to different media types. One way is to replace each multimedia entity, such as image or video, with its textual description that may contain title, ALT and surrounding text information. This is actually a commonly-adopted approach for indexing multimedia entities in commercial search engines. However, there is a large gap between the text description and the content of multimedia entity, and thus the conventional post-search methods will also not work well on multimedia search.

In this paper, we propose an image search performance prediction approach by exploring the visual content of search results. Given a text query and an image ranking list without any relevance labeling, our task is to predict the search performance under certain evaluation metrics, such as average precision (AP) or normalized discounted cumulative gain (NDCG). We first analyze AP and NDCG and then derive that we only need to estimate the images’ relevance probabilities to compute their mathematical expectations. We propose a query-adaptive graph-based learning approach to accomplish the task, where different image representations are used according to query types determined by a query classification component. The framework is illustrated in Figure 1. By conducting experiments on the data collected with more than 1,000 queries from four search engines, we demonstrate that our performance predictions can achieve a strong correlation with their real values, i.e., the values estimated with the ground truths of search results. Our approach is able to greatly save manual efforts, as generally estimating search performance needs a labor-intensive and tedious process to label images’ relevance. We also introduce three applications based on our approach, namely, image metasearch, multilingual image search, and Boolean image search.

The contributions of this work can be summarized as follows:

1. We propose a scheme for automatically predicting the performance of web image search. To our knowledge, this is the first work on performance prediction for multimedia search. We will show that our approach is able to obtain performance predictions that are reasonably correlated with their real performances.

2. We demonstrate that, with the search performance prediction scheme, the
performance of the three applications, namely, image metasearch, multilingual image search, and Boolean image search can be dramatically improved. This verifies the usefulness of the proposed scheme.

(3) We conduct experiments on a large dataset comprising the web image search results of 1,165 queries from 4 search engines, with the relevance of each image is manually labeled. This compares favorably to many existing works related to image search that conduct experiments on only small datasets.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. In Section 3, we introduce our image search performance prediction approach. In Section 4, we introduce the dataset and our empirical verification. The three applications and corresponding experiments are provided in Section 5, followed by our concluding remarks in Section 6.

2. RELATED WORK

Query performance prediction (also known as query difficulty prediction) is a challenging problem in information retrieval that has attracted great research interest. The task is defined as predicting the retrieval effectiveness of a query given a search system and a collection of documents. It has been demonstrated to be useful in many applications, such as query suggestion, query expansion and meta-search. The existing methods can mainly be divided into two categories, namely, pre-search approach and post-search approach.

Pre-search methods [Hauff et al. 2008] predict the performance of a query without considering search results. They are usually built based on the characteristics of the collection or query terms. He and Ounis [2004] and Zhao et al. [2008] have exploited the collection statistics such as inverse term and document frequencies. It is observed that good retrieval performance usually correlates positively with high variation of inverse document frequency of query terms. External sources such as WordNet can be employed to measure the relationship between query terms. The query terms that always appear in the same or similar senses can be considered
unambiguous and thus they are expected to perform well [Mothe and Tanguy 2005; Patwardhan 2006; Banerjee and Pedersen 2003]. These methods are usually very fast since they actually do not need to perform the search process, but the performance may not be satisfactory as search results are not taken into consideration. Post-search methods are employed after obtaining ranked search results. They are able to leverage more information clues, such as the cohesiveness of the search results, and thus can perform better. A representative work is [Cronen-Townsend et al. 2002], which proposes a Clarity Score. It measures the divergence between the query and collection language model, where a high divergence suggests a well-performing query. Following that, Zhou and Croft [2007] propose techniques to predict query performance in web search environments where document collections are significantly more heterogeneous. Several post-search methods, such as [Carmel et al. 2006; Amati et al. 2004; Hauff et al. 2008], also consider factors like the variability of similarity scores.

Despite the great success achieved, existing query performance prediction research mainly focuses on text search, whereas the literature regarding multimedia search is extremely sparse. Li et al. [2011] propose a query difficulty estimation method for content-based image search, but in their study the query is image instead of textual term. Xing et al. [2010] may be the most related work to ours. They explore the query difficulty prediction for contextual image retrieval. But it belongs to pre-retrieval approach, and, as previously analyzed, the post-retrieval approach can be more effective by exploring search results. Our approach should be the first post-search scheme for image search performance prediction. Our work well complements the work [Xing et al. 2010], and our approach can also be integrated with other query analysis based pre-search algorithms.

3. PERFORMANCE PREDICTION FOR WEB IMAGE SEARCH

If the relevance labels of search results are available, search performance can be easily estimated for different evaluation metrics. But our task is to predict the image search performance without the relevance labels. Here we first analyze two popular evaluation metrics, namely, AP and NDCG. We derive that, in order to estimate the mathematical expectations of AP and NDCG, we only need to predict the relevance probabilities of search results. Therefore, we adopt a query-adaptive graph-based learning approach to learn the relevance probability of each image.

3.1 Probabilistic Analysis of AP and NDCG

We first analyze the AP and NDCG from a probabilistic perspective. It is well known that many performance evaluation metrics support different scales of relevance, such as NDCG. To simplify our analysis, here we employ binary relevance option, that is, each image is merely judged to be relevant or irrelevant without considering more relevance grades.

Given a collection of images \( \mathcal{D} = \{x_1, x_2, \ldots, x_n\} \), denote by \( rel(x_i) \) the binary relevance label of \( x_i \) with respect to the given query, i.e., \( rel(x_i) = 1 \) if \( x_i \) is relevant and otherwise \( rel(x_i) = 0 \). Denote by \( \tau \) an ordering of the images and let \( \tau(i) \) be the image at the rank \( i \) position (lower number indicates higher rank).
The average precision measure [Wang et al. 2010] is defined as,

\[
AP = \frac{1}{R} \sum_{i=1}^{n} rel(\tau(i)) \sum_{j=1}^{i} \frac{rel(\tau(j))}{i}
\]

(1)

where \( R \) is the number of relevant images.

NDCG is defined as

\[
NDCG = \frac{DCG}{IDCG} = \frac{1}{IDCG} \sum_{i=1}^{n} \frac{2^{rel(\tau(i))} - 1}{\log_2(i + 1)}
\]

(2)

where \( IDCG \) is the normalization factor that is chosen to guarantee that the perfect ranking list’s NDCG is 1. Since in web search users focus on top results, the AP and NDCG (especially NDCG) are usually estimated only for the top results. They are usually named truncated AP or NDCG. Considering the truncated measure at depth \( T \), we can assume that the number of relevant images is greater than \( T \).

Therefore, the truncated AP can be estimated as

\[
AP@T = \frac{1}{T} \sum_{i=1}^{T} \frac{rel(\tau(i)) \sum_{j=1}^{i} \frac{rel(\tau(j))}{i}}
\]

(3)

and the truncated NDCG becomes

\[
NDCG@T = \frac{DCG@T}{IDCG@T} = \frac{1}{IDCG@T} \sum_{i=1}^{T} \frac{2^{rel(\tau(i))} - 1}{\log_2(i + 1)}
\]

(4)

where

\[
IDCG@T = \sum_{i=1}^{T} \frac{1}{\log_2(i + 1)}
\]

(5)

Now we analyze the mathematical expectations of AP and NDCG. Let \( y(x_i) \) denote the relevance probability of \( x_i \), i.e., \( Pr(rel(x_i) = 1) = y(x_i) \). Assume that the relevance of two different images is completely independent, then we derive the mathematical expectation of AP@T as

\[
E[AP@T] = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{i} \frac{E[rel(\tau(i))rel(\tau(j))]}{i}
\]

(6)

\[
= \frac{1}{T} \sum_{i=1}^{T} \frac{1}{i} \left\{ E[rel(\tau(i))^2] + \sum_{j=1}^{i-1} E[rel(\tau(i))rel(\tau(j))] \right\}
\]

\[
= \frac{1}{T} \sum_{i=1}^{T} \frac{1}{i} \left\{ y(\tau(i)) + \sum_{j=1}^{i-1} y(\tau(i))y(\tau(j)) \right\}
\]

ACM Journal Name, Vol. x, No. x, xx 20xx.
Analogously, we can derive the mathematical expectation of NDCG@T:

\[ E[\text{NDCG}@T] = \frac{1}{IDCG@T} \sum_{i=1}^{T} \frac{E[2^{rel(i)}] - 1}{\log_2(i + 1)} \]

\[ = \frac{1}{IDCG@T} \sum_{i=1}^{T} \frac{y(\tau(i))}{\log_2(i + 1)} \]

Therefore, in order to compute the mathematical expectations of AP and NDCG, we only need to estimate the relevance probability of each image.

3.2 Relevance Analysis with Query-Adaptive Graph-Based Learning

In order to estimate the relevance probability of each image, we utilize the ranking order obtained from search system and the images’ visual content. Our approach is as follows. We first estimate the ranking-based relevance probabilities according to the ranking order of images. A query-adaptive graph-based learning method is then employed to estimate the relevance probabilities of images. In the method, the query will be classified to be person-related or non-person-related according to the image search results. Different image representations are used for different query types (that is why we call the method query-adaptive).

3.2.1 Ranking-Based Relevance Analysis. We first estimate a ranking-based relevance probability for each image. We let \( \bar{y}_i \) denote the relevance probability of \( x_i \), which indicates the probability estimated from the ranking order of \( x_i \). We investigate the relationship between \( \bar{y}_i \) and the position \( \tau_i \) with a large number of queries. Actually, we can define

\[ \bar{y}_i = E_{q \in Q}[\hat{y}(q, \tau_i)] \]  

where \( Q \) denotes the whole query space, \( E_{q \in Q} \) means the expectation over the query set \( Q \), and \( \hat{y}(q, \tau_i) \) indicates the relevance ground truth of the \( i \)-th search result for query \( q \). Therefore, the most intuitive approach is to estimate \( \bar{y}_i \) by averaging \( \hat{y}(q, \tau_i) \) over a large query set.

Here we use 400 training queries to estimate \( \bar{y}_i \). The relevance score of each search result is manually labeled to be 0 or 1 if the image is irrelevant or relevant to the query. Figure 2 (a) to (d) show the averaged relevance score curves for the 400 queries with respect to the ranking position for Google, Bing, Yahoo, and Flickr, respectively. However, although we can see that the curves tend to decrease when ranking position increases, they are not smooth enough. There are fluctuations, and this is not consistent with the prior knowledge that the expected relevance scores should be decreasing with respect to ranking positions. This can be attributed to the fact that the queries used to estimate the initial relevance queries are still insufficient. Here we smooth the curves with a parametric approach. We assume \( \bar{y}_i = a \log(i) + b \), where \( a \) and \( b \) are two parameters, and we then fit this function with the points. In this way, we can estimate the parameters \( a \) and \( b \) with mean squared loss criterion. Figure 2 also shows the fitted curves from which we can see that they reasonably preserve the original information.
3.2.2 Query-Adaptive Graph-Based Learning. The query-adaptive graph-based learning is formulated based on two assumptions:

(1) The relevance probability function is continuous and smooth in visual space. That means the relevance probabilities of visually similar images should be close.

(2) The probabilities should be close to the ranking-based relevance probabilities.

A graph is constructed based on the search results of a query, where vertices are the images and edges weights indicate the pairwise similarities. We first introduce some notations. We use $W$ to denote the similarity matrix and $W_{ij}$, its $(i,j)$-th element, indicates the similarity of $x_i$ and $x_j$. Typically, it is estimated as

$$W_{ij} = \exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right)$$

where $\sigma$ is a radius parameter.

Let $d_{ii}$ denote the sum of the $i$-th row of $W$, i.e., $d_{ii} = \sum_j W_{ij}$. Then, the
A graph-based learning approach can be written as
\[
\min_y \frac{1}{2} \sum_{i,j} W_{ij} \left( \frac{y_i}{d_{ii}} - \frac{y_j}{d_{jj}} \right)^2 + \lambda \sum_i \frac{1}{d_{ii}} (y_i - \bar{y}_i)^2
\]  
(10)

where \(\lambda\) is a weighting parameter and \(y_i\) is the relevance probability of \(x_i\) that we want to estimate. We can see that the smoothness assumption is embedded in the first term of above equation, which enforces the relevance probabilities of visually similar images to be close. The second term reflects the second assumption, i.e., the probabilities we estimate should be close to the ranking-based probabilities.

We use \(D\) to denote a diagonal matrix, with \(d_{ii}\) to be its \((i; i)\)-th element; and let \(g\) denote \(\begin{bmatrix} y_{d_1} / d_{11}, y_{d_2} / d_{22}, \ldots, y_{d_n} / d_{nn} \end{bmatrix}^T\). Thus, Eq.10 can be rewritten as,
\[
\min_g g^T (D - W) g + \lambda (g - D^{-1} \bar{y})^T D (g - D^{-1} \bar{y})
\]  
(11)

It can be derived that
\[
y = \frac{1}{1 + \lambda} WD^{-1} \bar{y} + \frac{\lambda}{1 + \lambda} \bar{y}
\]  
(12)

We can iterate the above equation and the convergence can be proven.

From the above equation we can see that, if an image has many visually close images in the set, its relevance probability will be high. It is consistent with intuition. For example, if an image has many near-duplicates in top results, most likely it should be a relevant one. Up to now we have introduced the graph-based learning approach, but a remaining problem is image representation. The most straightforward approach is to extract some fixed features from each image. However, we noted that in image search, a large part of the queries is about person. For example, among the 1165 queries in our dataset, about 20% of them are person-related. Clearly, if a query is person-related, it is more reasonable to use facial features instead of global visual features as our target is to get images about the specific person.

Here we regard the judgement of whether a query is person-related as a classification task. We accomplish the classification by extracting several clues from image search results. For each image in the ranking list, we perform face detection and then extract the 7-dimensional features, including the size of the largest face area, the number of faces, the ratio of the largest face size and the second largest face size, and the position of the largest face (the position is described by the up-left and bottom-right points of the bounding box and thus there are 4-dimensional features). We average the 7-dimensional features of the top \(T\) search results and it forms the features for query classification. We learn a classification model based on the 400 * 4 training queries and it is used to discriminate person-related and non-person-related queries.

For each image in the search results of the person-related queries, we extract Local Binary Pattern (LBP) features [Ahonen et al. 2004] from the largest face to represent it. If the query is non-person-related, we extract several global features, including bag-of-visual-words, block-wise color moments, wavelet texture, and edge direction histogram. Therefore, as shown in Figure 3, given a query and its search results, we first perform query classification to decide whether it is person-related.
or not. We then employ different image representations according to the query classification result and perform graph-based learning accordingly.

3.2.3 Discussion. From the two assumptions used in our approach and the formulation in Eq. (10), we can see that our approach is actually closely related to image search re-ranking, a topic that has received several research interests in recent years [Tian et al. 2008; Yan et al. 2003; Natsev et al. 2005; Hsu et al. 2007; Liu and Mei 2011; Yang and Hanjalic 2010]. Re-ranking aims to adjust the ranking order of search results such that more relevant results can be prioritized. There are two typical approaches for image search re-ranking, one is pseudo relevance feedback and the other is graph-based re-ranking. In graph-based re-ranking, the formulation is usually developed based on two assumptions: (1) the ranking positions of visually similar images should be close; and (2) the ranking orders before and after re-ranking should not change too much. Therefore, the formulations of several re-ranking methods are very close to ours. For example, the two terms in the regularization scheme in [Hsu et al. 2007] are in the same form with our approach, and the difference only lies on the definition of initial relevance scores and image representations. Several other methods, such as those in [Tian et al. 2008] and [Jing and Baluja 2008], are also closely related to our approach. But our ranking-based relevance analysis and query classification components have not been investigated in the conventional re-ranking methods before. In most re-ranking methods, fixed image representations are used and the initial relevance scores are usually set based on several heuristics, such as letting $\bar{y}_i = 1 - \frac{1}{n}$ or $\bar{y}_i = n - i$. In the next section, we will empirically demonstrate the effectiveness of the ranking-based relevance analysis and query classification components.

From the above introduction, we can see that the computational cost of our approach mainly comes from the following three parts: (1) feature extraction (including face detection); (2) query classification; and (3) graph-based learning. For query classification, we use only 7-dimensional features and 1600 training samples, and the speed is very fast. For graph-based learning, it can be analyzed that the computational cost scales as $O(dT^2 + T^3)$, where $d$ is the dimension of features and $T$ is the number of images considered. Since we only use the top results, $T$ is usually small (its value is 140 in our experiments), and thus the computational cost is low.
cost is also low. In our experiments, the process can be finished in 0.2 second if we do not take the feature extraction part into account (Pentium4 3.0G and 2G memory). The feature extraction part is the most computationally expensive step. But actually many search engines host several pre-computed visual features for the indexed images in order to enable several services, such as re-ranking and visual search\(^4\). Therefore, if several providers want to build services based on the performance prediction approach, they can skip the feature extraction step.

4. EVALUATION

4.1 Experimental Settings

We use the 1,165 queries listed in [Li et al. 2009]. As introduced in [Li et al. 2009], the queries are actually selected from a large query log of a commercial search engine and they are the most frequent ones. For each query, we collect the top 140 results from the following four image search engines: Google, Bing, Yahoo, and Flickr. In this way, we have collected about 0.7 million images. The relevance ground truth of each image is manually labeled. Five human labelers were involved in the process. For every image, each labeler assigns a score of 0 or 1. Here 0 and 1 means irrelevant and relevant, respectively. Since there are five labelers, we perform a voting to establish the final relevance level of each image. Since there are several ambiguous queries, we perform a study on the queries before the manual labeling process, which is similar to the process in [Wang et al. 2010]. Each query was assigned a description. So, ambiguous queries will have more than one descriptions. For example, “apple” may refer to fruit, computer and cell phone. In our work, images that are consistent to different descriptions are all regarded as relevant. We randomly split the 1,165 queries into two parts: one for training and parameter tuning that contains 400 queries and the other for testing that contains 765 queries. Since we have four search engines, the total numbers of training and testing queries are 1600 and 3060, respectively. We perform face detection for each image and extract the following features:

1. 7-dimensional features about the facial characteristics of the image, including the size of the largest face area, the number of faces, the ratio of the largest face size and the second largest face size, and the position of the largest face. If there is no face detected in an image, all the 7-dimensional features are set to 0.

2. 256-dimensional LBP features [Ahonen et al. 2004] extracted from the largest face region. If there is no face detected in the image, the features are set to 0.

3. 1000-dimensional bag-of-visual-words. Difference of Gaussians is used to detect keypoints in each image and then SIFT descriptors are extracted. By building a visual codebook of size 1000 based on K-means, we obtain a 1000-dimensional bag-of-visual-words histogram for each image.

4. 225-dimensional block-wise color moments based on 5-by-5 fixed partition of the image, 128-dimensional wavelet texture, and 75-D edge direction histogram.

The first set of features are used for query classification. The second contains facial features, that is, the image representations for person-related query. The


ACM Journal Name, Vol. x, No. x, xx 20xx.
Table I. The statistics of the person-related and non-person-related classes in the 1600 training and 3060 testing queries. We can see that about 19\% of the queries are person-related.

<table>
<thead>
<tr>
<th>Class</th>
<th>Person-Related</th>
<th>Non-Person-Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Queries</td>
<td>283</td>
<td>1317</td>
</tr>
<tr>
<td>Testing Queries</td>
<td>604</td>
<td>2456</td>
</tr>
</tbody>
</table>

Table II. The confusion matrix of classification results. The classification accuracy is 95.98\%.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Person-Related</th>
<th>Non-Person-Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person-Related</td>
<td>0.86</td>
<td>0.01</td>
</tr>
<tr>
<td>Non-Person-Related</td>
<td>0.14</td>
<td>0.99</td>
</tr>
</tbody>
</table>

third and fourth sets are global features, which are used for image representation if the query is classified as non-person-related.

For each query, we predict the expected AP and NDCG and their actual values estimated using the relevance ground truths. For query classification, i.e., the component that judges whether a query is person-related or non-person-related, we learn a SVM model with RBF kernel based on the 1600 training queries and the parameters are turned by 10-fold cross-validation. For relevance estimation, the parameters $\sigma$ and $\lambda$ are tuned to the values that maximize the correlation of predicted $AP@140$ and real $AP@140$ measurements of the 1600 training queries.

The whole dataset has been put on the web for public access\(^5\).

4.1.1 On Query Classification. We introduce the ground truth establishment for the query classification task first. Since our target is to use facial features for person-related queries, we categorize the queries that are about a specific person as the person-related class. Table I illustrates the statistics of the person-related and non-person-related classes in the 1600 training and 3060 testing queries. We can see that about 19\% of the queries are person-related. Table II illustrates the query classification results on the testing queries. We can see that our approach achieves fairly good performance. The classification accuracy is 95.58\%. The mis-classification results mainly come from several queries that are about people but not a specific person, such as “drill team” and “doctor”. There are many large faces appeared in the images and this leads to mis-classification cases.

4.1.2 On Image Search Performance Prediction. Figure 4 (a) to (f) illustrate comparison of the predicted and real $AP@140$, $NDCG@5$, $NDCG@10$, $NDCG@20$, $NDCG@50$, $NDCG@100$ of the testing queries, respectively\(^6\). We can observe the reasonable correlation between the predicted and the real performance measure-


\(^6\)Here we have varied the truncated depth for NDCG because NDCG measure usually focuses more on top results.
Fig. 4. The predicted performance and the real values of the 3060 queries under the evaluation metrics of (a) AP@140; (b) NDCG@5; (c) NDCG@10; (d) NDCG@20; (e) NDCG@50; and (f) NDCG@100.

mental from the figures.

To quantitatively evaluate our approach, we employ two measures. The first measure is linear correlation. That is, we compute the linear correlation of the predicted AP or NDCG and their real values based on the 3060 testing queries. The second measure is better-worse prediction accuracy. It is defined as follows. We generate all the query pairs from the 3060 queries and then we predict which one is better in the pair (we remove the pairs that are with the same performance). We estimate the prediction accuracy using our image search performance estimation approach. We employ this measure because, in comparison with optimizing linear correlation, accurately predicting which ranking list is better can be more useful for several applications, such as metasearch, multilingual search and Boolean search introduced in the next section.

4.2 Experimental results

We compare our proposed approach with the following three methods:

(1) Using only global features (denoted as “Global Feature”). In this method, we do not classify whether a query is person-related or non-person-related and we use the 1,428 global features (bag-of-visual-words, color moments, texture and edge direction histogram) in all cases.

(2) Heuristic initial relevance score setting (denoted as “Heuristic Initialization”). In this method, we heuristically set the initial relevance score at $i$-th position as $1 - \frac{i}{n}$. That is, $y_{i} = 1 - \frac{i}{n}$.

(3) Result number based approach (denoted as “Search Number”). We assume that the number of search results is able to reflect search performance. The rationality relies on the fact that, for simple queries, good performance is usually achieved and meanwhile the numbers of search results are also great.

The comparison of our approach with the first two methods will validate the ef-
Table III. The linear correlation comparison of the three different methods with different performance measures, including AP@140, NDCG@5, NDCG@10, NDCG@20, NDCG@50, and NDCG@100. The best results are marked in bold.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Metric</th>
<th>AP@140</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>NDCG@20</th>
<th>NDCG@50</th>
<th>NDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Feature</td>
<td></td>
<td>0.627</td>
<td>0.401</td>
<td>0.462</td>
<td>0.518</td>
<td>0.579</td>
<td>0.596</td>
</tr>
<tr>
<td>Heuristic Initialization</td>
<td></td>
<td>0.568</td>
<td>0.344</td>
<td>0.402</td>
<td>0.457</td>
<td>0.518</td>
<td>0.554</td>
</tr>
<tr>
<td>Search Number</td>
<td></td>
<td>0.061</td>
<td>0.007</td>
<td>0.043</td>
<td>0.043</td>
<td>0.044</td>
<td>0.5</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td></td>
<td>0.653</td>
<td>0.422</td>
<td>0.486</td>
<td>0.542</td>
<td>0.601</td>
<td>0.621</td>
</tr>
</tbody>
</table>

Table IV. The better-worse prediction accuracy comparison of the three different methods with different performance measures, including AP@140, NDCG@5, NDCG@10, NDCG@20, NDCG@50, and NDCG@100. The best results are marked in bold.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Metric</th>
<th>AP@140</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>NDCG@20</th>
<th>NDCG@50</th>
<th>NDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Feature</td>
<td></td>
<td>0.745</td>
<td>0.607</td>
<td>0.628</td>
<td>0.648</td>
<td>0.687</td>
<td>0.711</td>
</tr>
<tr>
<td>Heuristic Initialization</td>
<td></td>
<td>0.694</td>
<td>0.593</td>
<td>0.609</td>
<td>0.624</td>
<td>0.65</td>
<td>0.674</td>
</tr>
<tr>
<td>Search Number</td>
<td></td>
<td>0.566</td>
<td>0.611</td>
<td>0.614</td>
<td>0.579</td>
<td>0.572</td>
<td>0.572</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td></td>
<td>0.766</td>
<td>0.611</td>
<td>0.633</td>
<td>0.662</td>
<td>0.716</td>
<td>0.739</td>
</tr>
</tbody>
</table>

4.3 Discussion

In this work, we only consider search relevance, but actually diversity is also an important aspect for search performance. Our task is actually to approximate a given performance evaluation measure. Most widely-used performance evaluation metrics, such as AP and NDCG, focus on relevance. That is why our approach takes no account of diversity. But there also exists performance evaluation metrics that consider diversity, such as the Average Diverse Precision (ADP) in [Wang et al. 2010]. We can also extend our approach to estimate the measurements of these performance metrics. Actually we can adopt a similar approach of Section 3.1 to perform a probabilistic analysis of ADP such that it can be estimated based...
on relevance scores. Then, diversity will be taken into account. If we apply such extended estimations to different applications such as metasearch (the applications will be introduced in the next section), the results that are more diverse will be favored.

Another noteworthy issue is that we have used facial information in image search results to classify person-related and non-person-related queries. Intuitively, we can also choose to match a query to a celebrity list to accomplish the task. We do not apply this method because it is not easy to find a complete list and it will also be difficult to keep the listed updated in time. But we may investigate the combination of our approach and the list-based method. We leave it to future work.

5. APPLICATIONS

In this section, we introduce three potential application scenarios of image search performance prediction: image metasearch, multilingual image search, and Boolean image search.

5.1 Image Metasearch

5.1.1 Application Scenario. Metasearch refers to the technique that integrates the search results from multiple search systems. In the past few years, extensive efforts have been dedicated to metasearch and most of them focus on source engine selection and multiple engine fusion [Manoj and Elizabeth 2008]. For example, MetaCrawler [Selberg and Etzioni 1995], one of the earliest metasearch engines, employ a linear combination scheme to integrate the results from different search engines. Smith and Hurson [2003] propose methods to select the best search engine for a given query. However, metasearch has been rarely touched in multimedia domain. Benitez et al. [1998] develop a content-based metasearch for images on the web. But it mainly focuses on the query by example scenario and relevance feedback is involved. Kennedy et al. provide a discussion on multimodal and metasearch in [Liu and Mei 2011]. Here we build two web image metasearch techniques based on our image search performance prediction scheme:

(1) Search engine selection. It is the most straightforward metasearch scenario. For a given query, we collect image search results from different search engines. The image search performance is then predicted for each search engine and we simply select the one with the best predicted performance.

(2) Search engine fusion. In this approach, we merge the search results from different search engines instead of selecting one from them. We adopt an adaptive linear fusion method. Note that in our image search performance prediction algorithm, we have estimated the relevance probability of each image. Denote the relevance probability of $x_i$ from the $k$-th search engine as $y_i^{(k)}$. We weight this value with the predicted performance of each search engine and then linearly fuse them. It can be written as

$$ r_i = \sum_{k=1}^{K} \alpha_k p_k y_i^{(k)} $$  

(13)

where $p_k$ is the predicted performance for the $k$-th search engine under certain performance evaluation metric, such as AP and NDCG, and $\alpha_k$ is the weight for the
Fig. 5. Image metasearch performance comparison of different methods. We can see that the “Source Selection” and “Fusion” methods, which are built based on the proposed search performance prediction approach, outperform the other approaches.

$k$-th search engine which satisfies $\sum_{k=1}^{K} \alpha_k = 1$. The final ranking list is generated with the relevance scores $r_i$ ranking in descending order. The weights $\alpha_k$ are tuned to their optimal values on the 400 training queries.

5.1.2 Experiments. We denote the search engine selection and search engine merge methods introduced above as “Source Selection” and “Fusion”. We test the metasearch performance on the 675 queries and 4 image search engines, i.e., Google, Bing, Yahoo and Flickr. For each search engine, we consider only the top 140 search results. Therefore, only the images that simultaneously appear in more than one ranking lists have multiple $y^{(k)}_i$ greater than 0. This is reasonable since, if an image appears in the top results of multiple engines, it should be prioritized.

We compare our methods with the following approaches:

1) Using individual search engines, i.e., Google, Bing, Yahoo and Flickr.
2) Search engine fusion without performance prediction (denoted as “Naive Fusion”). The formulation can be written as

$$r_i = \sum_{k=1}^{K} \alpha_k y^{(k)}_i$$

This is actually the classical score-based rank aggregation approach. Comparing Eq.(13) and Eq.(14), we can see that the only difference is that, in our “Fusion” method, we have integrated the performance prediction of different image search engines.

We first adopt the predicted NDCG@100 for $p_k$. The performance comparison of different methods are illustrated in Figure 5. We demonstrate the average NDCG measures for evaluating metasearch. First we compare the performance of “Source
Selection” with the four individual search engines. We can clearly see that the performance of “Source Selection” significantly outperforms the individual search engines. This further confirms the effectiveness of our image search performance prediction approach. The superiority of “Fusion” over individual search engines is also obvious. In addition, the proposed “Fusion” method clearly outperforms the “Naive Fusion” approach. This demonstrates that incorporating the performance prediction of search engines into their fusion is important.

We then change the performance metric for $p_k$ and demonstrate the metasearch performance variation of different methods in Figure 6. Note that actually only the performance of “Source Selection” and “Fusion” will vary, as the other methods do not rely on search performance prediction. Here we fix the performance evaluation metric for metasearch to NDCG@20. We can see that the “Source Selection” and “Fusion” methods are not very sensitive to the metric of performance prediction metrics and they consistently outperform the other approaches.

Figure 7 illustrates the top results obtained by different methods for an example query “bird of prey” for comparison (NDCG@100 is used as the performance evaluation metric for the “Source Selection” and “Fusion” methods).

5.2 Multilingual Image Search

5.2.1 Application Scenario. Multilingual search enables the access of documents in various different languages [Abusalah et al. 2005]. Typically, there are three components in multilingual search: query translation, monolingual search and result fusion. Most of the existing works focus on the fusion process. Powell et al. [2000] propose a normalized-score fusion method, which maps the scores into the same
Fig. 7. Comparison of the top search results obtained by different metasearch methods for the query “bird of prey”: (a) results retrieved from Google; (b) results retrieved from Yahoo; (c) results retrieved from Bing; (d) images retrieved from Flickr; (e) results returned by naive fusion; (f) results returned by the performance prediction based source selection method; (g) results returned by the performance prediction based fusion method.

scale for a reasonable comparison. Si and Callan [2005] propose a semi-supervised fusion solution for the distributed multilingual search problem. However, the study on multilingual multimedia search is sparse. WordNet is used to reduce the ambiguity of query in multilingual image search in [Popescu 2007]. Quénot et al. [2010] propose an approach for content-based indexing and search of multilingual audiovisual documents based on the International Phonetic Alphabet. Based on our image search performance prediction scheme, we propose a fusion approach to facilitate multilingual image search approach. Given a query, we first transform it into multiple languages and get the search results of these queries. We
then fuse the results to obtain the final ranking list. For result fusion, we adopt an approach that is similar to metasearch, i.e.,

\[ r_i = \sum_{k=1}^{K} \alpha_k p_k y_i^{(k)} \]  

where \( k \) denotes the \( k \)-th language and \( K \) is the number of considered languages.

5.2.2 Experiments. We conduct experiments with 15 queries, including black cat, sows and piglets, horse riding chebi, shanxi sandwich, Louvre, Mount Fuji with snow, Milano Politecnico logo, American flag flying, Hu Jintao shook hands with Obama, Junichi Hamada, fishing, fitness, bat, candle, and chanel. These queries are collected from several image search frequent users. We ask the users to propose a set of queries for multilingual image search that they are interested in and we then select the above 15 queries considering both their coverage and diversity. For each query, we convert it to five other languages using Google Translate, including Japanese, Chinese, French, Germany and Italian. We then get the top 140 search results from Google image search engine for each query. Therefore, the value of \( K \) in Eq.(15) equals 6. The relevance of each image is manually labeled. Similar to the experiments for metasearch, we compare our multilingual image search method with another naive approach that does not incorporate the image search performance prediction, i.e., \( p_k \) is removed in Eq.(15). The two methods are indicated as “Fusion” and “Naive Fusion”, respectively. In addition, we also compare our approach with the search performance of using different individual languages. Since for this application we do not have enough queries for training, we simply set the parameter \( \alpha_k \) to 1/6.

Similar to the experiments for metasearch, we first adopt the predicted NDCG@100 for \( p_k \) and compare the multilingual search performance of different methods in Figure 8. We then change the performance metric for \( p_k \) and demonstrate the multi-
lingual search performance variation in Figure 9. We can also see that the “Fusion” method consistently outperforms the “Naive Fusion” approach. This demonstrates the effectiveness of incorporating the performance prediction into multilingual image search. We can also observe that our fusion approach is not sensitive to the metric for performance prediction and it consistently outperforms the other approaches.

Figure 10 illustrates the top results obtained by different methods for an example query “Hu Jintao shook hands with Obama” for comparison (NDCG@100 is used as the performance evaluation metric for the “Fusion” method).

5.3 Boolean Image Search

5.3.1 Application Scenario. Boolean model is a classical information retrieval model [Manning et al. 2008]. In this model, query is represented with a Boolean expression, that is, several terms concatenated with “AND”, “OR” or “NOT”. However, many large-scale commercial systems do not support Boolean model. Actually when we issue queries that contain multiple terms concatenated with “or”, the conjunction “or” will be neglected and the relationship of the query terms becomes “and”. Google provides an advanced search option that allows users to provide up to 3 alternative query terms in the form of “term1 OR term2 OR term3”, but there is no such option for multimedia search.

Here we build a Boolean image search approach that supports multiple alternative query terms concatenated with “or”. We first perform search with each query term and then fuse the results. For result fusion, we adopt the approach similar to

\[http://www.google.com/advanced_search.\]
Fig. 10. Comparison of the top results obtained by different multilingual image search approaches for the query “Hu Jintao shook hands with Obama”: (a) only English; (b) only Japanese; (c) only Chinese; (d) only French; (e) only Germany; (f) only Italian; (g) results obtained by naive fusion; (h) results obtained by the proposed performance prediction based approach.
Fig. 11. Boolean image search performance comparison of different methods. We can see that the “Fusion” method, which is built based on our search performance prediction approach, outperforms the other approaches.

5.3.2 Experiments. We conduct experiments with 15 queries that contain multiple terms concatenated with “or”, including \textit{black cat or yellow cat, hog or a wild ox, Flickr or Yahoo, football or basketball or golf, airplane in the sky or airplane on the ground, boy in shot jogging or boy in red skirt dancing, president of us or chairman of china, people walking on the moon or people jogging on the beach, cat eating cheeseburger or dog eating bone, dad with twins or mom with twins, and Obama eating fried chicken or Bush eating hamburger}. These queries are collected in a similar way with the above multilingual image search experiment. All the experiments are conducted with Google image search engine. The top 140 results for each query term are collected. The relevance of each image is labeled. Note that in our Boolean search scenario, an image will be labeled as relevant if it is relevant to one of the alternative query term.

Similar to the experiments for metasearch, we compare our Boolean image search method with another naive approach that does not incorporate the image search performance prediction, i.e., \( p_k \) is removed in Eq.(16). The two methods are denoted as “Fusion” and “Naive Fusion”, respectively. In addition, we also compare our approach with a baseline method that directly issues a whole query on Google image search. Also due to the reason that we do not have enough queries for training, we simply set the parameter \( \alpha_k \) to \( 1/K \).

Similar to the experiments for metasearch, we first adopt the predicted NDCG@100 for \( p_k \) and compare the Boolean search performance of different methods in Figure 11. We then change the performance metric for \( p_k \) and demonstrate the Boolean

Here \( k \) denotes the \( k \)-th term in query and \( K \) is the number of alternative terms.

\[
r_i = \sum_{k=1}^{K} \alpha_k p_k y_i^{(k)}
\]

(16)
search performance variation in Figure 12. We can see that the performance of the baseline method is fairly poor. This is because that the search engine will not parse the queries that contain “or”, and for several complicated queries there is even no result return. The “Fusion” approach remarkably outperforms the “Naive Fusion” approach, and this demonstrates the effectiveness of incorporating the performance prediction into the Boolean image search. We can also see that our fusion approach is not sensitive to the metric used in performance prediction and it consistently outperforms the other approaches.

Figure 13 illustrates the top results obtained by different methods for an example query “black cat or yellow cat” for comparison (NDCG@100 is used as the performance evaluation metric for the “Fusion” method).

6. CONCLUSION AND FUTURE WORK

This paper investigates a novel problem in image search, that is, the automatic performance prediction for a ranking list returned by a search system. By analyzing NDCG and AP, we derived that we only need to predict the probabilities of images’ relevance for estimating the mathematical expectations of AP and NDCG. We proposed a query-adaptive graph-based learning approach to estimate the relevance probability of each image to a given query. Experiments demonstrate that our approach is able to achieve predictions that are highly correlated with the real image search performance. Finally, we introduced three applications based on our predicted query performance, namely, image metasearch, multilingual image search and Boolean image search. Comprehensive experiments demonstrate the effectiveness of our approaches.

We would like to mention that, although in this work we focus on mining images’ content for accomplishing search performance prediction, pre-search methods that directly analyze queries’ characteristics can also be integrated. In addition, other
Fig. 13. Comparison of the top results obtained by different Boolean search methods for the query “black cat or yellow cat”: (a) the search results returned with the whole query; (b) results based on naive fusion; (c) the results obtained by our proposed query performance prediction based approach.

information clues, such as the number of search results introduced in Section 4, can also be incorporated. For future work, we will further improve our scheme by integrating more information clues. We will also extend our method to video search performance prediction.

REFERENCES


