Image Search by Graph-based Label Propagation with Image Representation from DNN

Yingwei Pan [§], Ting Yao [†], Kuiyuan Yang [‡], Houqiang Li [§],
Chong-Wah Ngo [†], Jingdong Wang [‡], Tao Mei [‡]
[§] University of Science and Technology of China, Hefei, P. R. China
[†] City University of Hong Kong, Kowloon, Hong Kong
[‡] Microsoft Research Asia, Beijing, P. R. China
{panyw, tingyao}.ustc@gmail.com; lihq@ustc.edu.cn;
cscwngo@cityu.edu.hk; {kuyang, jingdw, tmei}@microsoft.com

ABSTRACT

Our objective is to estimate the relevance of an image to a query for image search purposes. We address two limitations of the existing image search engines in this paper. First, there is no straightforward way of bridging the gap between semantic textual queries as well as users' search intents and image visual content. Image search engines therefore primarily rely on static and textual features. Visual features are mainly used to identify potentially useful recurrent patterns or relevant training examples for complementing search by image reranking. Second, image rankers are trained on query-image pairs labeled by human experts, making the annotation intellectually expensive and time-consuming. Furthermore, the labels may be subjective when the queries are ambiguous, resulting in difficulty in predicting the search intention. We demonstrate that the aforementioned two problems can be mitigated by exploring the use of click-through data, which can be viewed as the footprints of user searching behavior, as an effective means of understanding query.

The correspondences between an image and a query are determined by whether the image was searched and clicked by users under the query in a commercial image search engine. We therefore hypothesize that the image click counts in response to a query are as their relevance indications. For each new image, our proposed graph-based label propagation algorithm employs neighborhood graph search to find the nearest neighbors on an image similarity graph built up with visual representations from deep neural networks and further aggregates their clicked queries/click counts to get the labels of the new image. We conduct experiments on MSR-Bing Grand Challenge and the results show consistent performance gain over various baselines. In addition, the proposed approach is very efficient, completing annotation of each query-image pair within just 15 milliseconds on a regular PC.

MM'13, October 21 - 25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508128.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Image search, Deep neural networks, Neighborhood graph search, Click-through Data.

1. INTRODUCTION

The rapid development of Web 2.0 technologies has led to the surge of research activities in image search. Since the query is provided as text rather than image, it is difficult to directly bring image visual features into play and build on advances in computer vision research. As a result, while visual documents are rich in image visual content and user-supplied texts, commercial image search engines to date mostly perform retrieval by employing static and textual features extracted from the image's parent web page and surrounding text, respectively. This kind of visual search approach may not always achieve satisfying results as textual information is sometimes noisy and even unavailable. Moreover, image rankers trained on query-image pairs labeled by human experts may lead to poor generalization performance due to the label noise problem and difficulty associated with understanding the user's intent.

To address the above issues, we explore user searching behavior through click-through data, which is largely available and freely accessible by search engines, for connecting image visual content with semantic textual queries as well as user's search intent. Accordingly, a novel graph-based label propagation approach is proposed. For a given query-image pair, we firstly identify approximate nearest neighbors (ANNs) of the given image from the set of previously clicked images by using neighborhood graph search. Specifically, an image neighborhood graph is constructed using image representations generated by deep neural networks, which have been proven effective in speech recognition and image classification recently. An iterated neighborhood graph search is efficiently performed to find the nearest neighbor images

with the mechanism of avoiding unnecessary neighborhood expansions and local optimum. Next, the clicked queries to these ANNs are aggregated and propagated to predict the relevance score for the given query-image pair.

Our proposed method tackles the two issues highlighted above as follows. First, the image graph is built on only visual features extracted directly from the previously clicked images. As a result, images that are visually similar to the new image in terms of measured color, texture, and edge properties are located as ANN candidates. It is worth noting that our approach does not require any textual information and any priori knowledge about the user intent.

Second, we consider exploring user click-through data to shed some light on bridging semantic gap and user intention gap for image search. In general, image rankers obtain training data by manually labelling the relevance of query-image pairs. However, it is difficult to fathom the user's intent based on the query keywords alone especially for those ambiguous queries. For example, given the query "guitar factory," experts tend to label images of guitar manufacturing base as being highly relevant. However, empirical evidence suggests that most users wish to retrieve images of a guitar with brand "guitar factory." Expert labels might therefore be erroneous. Such factors bias the training set and the ranker is learnt to be sub-optimal. Our click-based method provides a useful alternative aimed at tackling this problem directly. As in image search, users browse image thumbnails before selecting the images to click. The decision to click is likely dependent on the relevancy of an image. Therefore, click data can serve as a reliable feedback potentially useful for image search. We hypothesize that, most of the clicked images are relevant to the given query judged by the real users.

In summary, this paper makes the following contributions:

- We study the problem of image search by leveraging user click-through data.
- We propose a graph-based label propagation approach to predict query-image relevance score and the algorithm is demonstrated on a real-world user click data collected from Bing image search engine.

The remaining sections are organized as follows. Section 2 describes the related work on use of click data. Section 3 presents the graph-based label propagation method for image search, while Section 4 provides empirical evaluations, followed by the conclusions in Section 5.

2. RELATED WORK

Click-through data has been studied and analyzed widely with different Web mining techniques for improving search engines' efficacy and usability in recent years. In [1], the authors introduced another vectorial representation for the queries without considering the content information. Queries were represented as points in a high dimensional space, where each dimension corresponds to a unique URL. The weight assigned to each dimension was equal to the click frequency. Poblete *et al.* [9] proposed the query-set document model by mining frequent query patterns to represent documents rather than the content information of the documents. Li *et al.* [8] presented the use of click graphs in improving query intent classifiers. There are also several approaches that have tried to model the representation of queries or documents on the click-through bipartite. In addition, click-through data has also been used to learn the rank function. Joachims *et al.* [6] did an eye tracking experiment to observe the relationship between clicked links and the relevance of the target pages. For image search, click-through data has been found to be very reliable [3][5]. In [3], the method built a query-image click graph and performed backward random walks to determine a probability distribution over images conditioned on the given query which can be used for ranking. Jain *et al.* [5] reranked the image search results so as to promote images that are likely to be clicked to the top of the ranked list. Later in [10], an in-depth analysis of several ranking algorithms was performed on Flickr user log data to investigate the importance of many factors, including the internal and external image popularity, the overall attentions, diversity, semantic categories and visual appearance. In another work by Yao *et al.* [12], by combining click-through and video document features for deriving a latent space, the dot product of the mappings in the latent space is taken as the similarity between videos and the similarity is further applied for video tagging tasks.

3. GRAPH-BASED LABEL PROPAGATION

In this section, we first describe the image representation used in this paper, followed by the neighborhood graph search for finding ANNs on the image neighborhood graph. Then, a label propagation algorithm is proposed to further aggregate and propagate ANNs' queries/click counts to predict new query-image relevance score.

3.1 Image Representation

Recently, deep neural networks (DNN) has demonstrated its effectiveness to learn image representation and classifier simultaneously with a large number of training data. The learned image representation from DNN is close to semantics, and ever exceeds current estimate of Inferior Temporal (IT) representation performance in macaque's visual cortex [2]. Inspired by the success of DNN, we use it to generate image representation in this paper, which is a 1024-dimensional feature vector. Similar to [7], the used DNN architecture is denoted as $Image - C64 - P - N - C128 - P - N - C192 - C192 - C128 - P - F4096 - F1024 - F1000$, which contains five convolutional layers (denoted by C following the number of filters) while the last three are fully-connected layers (denoted by F following the number of neurons); the max-pooling layers (denoted by P) follow the first, second and fifth convolutional layers; local contrast normalization layers (denoted by N) follow the first and second max-pooling layers. The weights of DNN are learned on ILSVRC-2010, which is a subset of ImageNet dataset with 1.26 million training images from 1,000 categories. For an image, its representation is the neuronal responses of the layer $F1024$ by input the image into the learned DNN.

3.2 Neighborhood Graph Search

After building the image neighborhood graph on image representations, an iterated neighborhood graph search approach is exploited to locate the ANNs. We apply a recent query-driven iterated neighborhood graph search algorithm [11], to conduct the search process. The basic procedure is outlined in Algorithm 1.

GenerateInitialSolution(q, T) searches over trees T, which are constructed to index the reference images. The initial

Algorithm 1 Query-driven iterated neighborhood graph search

1: $P_0 \leftarrow$ GenerateInitialSolution(q, T)
2: $R^* \leftarrow$ LocalNGSearch(P_0, G)
3: **repeat**
4: $P' \leftarrow$ Perturbation(R^*, q, T, *history*)
5: $R^{*'} \leftarrow$ LocalNGSearch(P', G, *history*)
6: $R^* \leftarrow$ AcceptanceCriterion(R^*, $R^{*'}$)
7: **until** termination condition met

solution contains a small amount of initial NN candidates that have high probabilities to be near true NNs. Following the implementation in [11], we use kd-trees in our experiments. LocalNGSearch(P_0, G) starts from a set of seeds P_0 and searches over G by conducting neighborhood expansions in a best-first manner. Perturbation(R^*, q, T, *history*) generates new seeds from trees T according to the search history and previously selected NNs (R^*), to avoid unnecessary neighborhood expansions. LocalNGSearch(P', G, *history*) is slightly different from LocalNGSearch(P_0, G) as the search history, i.e., the NNs discovered up to the current iteration are considered in neighborhood expansion. Readers can refer to [11] for technical details.

3.3 Label Propagation

For a new query-image pair (q, I), we conduct the neighborhood graph search to get new image's Top K nearest visually similar images and aggregate their clicked queries/click counts to predict the relevance score of the new query-image. Specifically, for the new image I, let the Top K nearest neighbor in the image neighborhood graph be $I_i, 1 \leq i \leq K$. For each neighbor image I_i, let $q_i^j, 1 \leq j \leq M_i$ be the previously clicked query set to image I_i and f_i^j be the click counts of image I_i in response to query q_i^j. Then, the relevance $r(q, I)$ of query q to image I becomes

$$r(q, I) = \sum_{i=1}^{K} simi(I, I_i) \sum_{j=1}^{M_i} \frac{\left|q_i^j \cap q\right|}{\left|q_i^j \cup q\right|} \log f_i^j, \qquad (1)$$

where $simi(I, I_i)$ stands for the visual similarity between image I and its neighbor I_i which is calculated on the visual feature representation. $\left|q_i^j \cap q\right|$ and $\left|q_i^j \cup q\right|$ indicate the number of common terms between q_i^j and q, and the total unique term number of the two queries, respectively. We use the logarithm of click counts in this paper, which is verified to be effective.

The spirit of label propagation is to give a higher relevance score for query-image pair (q, I) if the visually similar neighbor images are highly clicked by queries in close semantic proximity with query q.

4. EXPERIMENT

We conduct experiments on the MSR-Bing Image Retrieval Challenge [4], which contains a training dataset and a dev dataset. Both of them are sampled from Bing user click log. In total, there are 11,701,890 distinct queries and 1,000,000 different images in training dataset, while the dev dataset consists 1,000 queries and 79,665 images.

In our experiments, we adopt the whole training dataset as our training data and the evaluations are conducted on all 1,000 queries in dev dataset.

Table 1: The NDCG@25 of different approaches with different individual visual features.

Method	WT	EDH	CM	BoW	DNN
N-Gram SVM	0.4878	0.4829	0.4873	0.4845	0.4899
GLP	0.4931	0.4866	0.4896	0.4848	0.5050

4.1 Experimental Settings

Compared Approaches: We compare the following approaches for performance evaluation:

- N-Gram SVM Modeling (N-Gram SVM). We use all the clicked images of a given query as positive samples and randomly select negative samples from the rest of the training dataset to build a SVM model for each query, and then use this model to predict the relevance of the query to a new image. In addition, in order to extend the capability of the training data to model queries that are not covered in the dataset, n-gram modeling, which attempts to model each n-gram as a "query", is used. In other words, if a query is not in the training set, but its n-grams appear in some queries of the training set, we can generate the model by linearly fusing the SVM models of these queries. In our experiments, we use bigram and unigram. The five runs using different visual features, including color moments (CM), wavelet texture (WT), edge histogram (EDH), bag of visual words (BoW), and DNN feature, are finally reported.

- Graph-based label propagation (GLP). We design five runs for our proposed graph-based approach, each based on one of the aforementioned five visual features.

Evaluation Metrics: Following the measurements in the challenge's industrial track, for each query, we use Discounted Cumulated Gain (DCG) to evaluate the performance of top 25 images. Given an image ranked list based on the score, the DCG for each query is calculated as

$$DCG@25 = 0.01757 \sum_{i=1}^{25} \frac{2^{rel_i} - 1}{\log_2^{i+1}}, \qquad (2)$$

where $rel_i = \{Excellent = 3, Good = 2, Bad = 0\}$ is the manually judged relevance for each image with respect to the query, and 0.01757 is a normalizer factor to make the score for 25 Excellent results 1. The final metric is the average of $DCG@25$ for all queries in the test set. Note that for the queries with less than 25 images in the dev dataset, we simply grade relevance score of all the supplemental positions as *Bad*. Moreover, the run time of our graph-based method is also compared and discussed.

4.2 Performance Comparison

Table 1 shows the NDCG performance of ten runs averaged over 1,000 test queries. Overall, the results across different visual features consistently indicate that GLP leads to a performance boost against N-Gram SVM modeling. Particularly, the DCG@25 performance of GLP on DNN feature can achieve 0.5050, which improves the N-Gram SVM model by 3.1%. In addition, when only evaluating the performance on the queries with more than 25 images in the dev dataset, GLP model on DNN feature (DCG@25 = 0.6339) can lead to better performance gain than N-Gram SVM model by 4.6%.

Table 2: Run time (ms) of N-Gram SVM and GLP on five different visual features. The experiments are conducted on a regular PC (Intel dual-core 3.33GHz CPU and 8 GB RAM).

Method	WT	EDH	CM	BoW	DNN
N-Gram SVM	900	890	860	6800	7500
GLP	9.9	9.8	11	15.8	14.8

This somewhat reveals the weak use of click data in training SVM model, where all the clicked images are used as positive training samples no matter how many times they have been clicked. As such, some non-relevant distracting images, which might have received only a single, or very few, clicks will also be considered as positive, thereby bringing some label noise. GLP, in contrast, is benefited from the way of taking into account the absolute click counts for relevance prediction, which should be more accurate.

Compared to global and local visual features, DNN feature achieves both the best performance by using N-Gram SVM modeling and GLP, which verifies the effectiveness of DNN feature in semantic-level similarity measurement.

Figure 1 shows the top ten images for query "portable cd player insigna" by using GLP method based on five individual features, respectively. We can easily see the results with DNN features get all excellent images, followed by WT for nine excellent images, EDH and BoW for eight excellent images, and CM for seven excellent images.

4.3 Run Time

The complexity of our method is $O(P\log^2 n+(1-P)\log^3 n)$, where n denotes the number of search seeds and P represents the probability of starting from an ideal seed. Table 2 listed the detailed run time on each feature for the two compared methods. Our approach is extremely efficient, completing relevance prediction of each image-query pair within 15 milliseconds on average. This is much faster than N-Gram SVM modeling which needs beyond 800 milliseconds.

5. CONCLUSION

In this paper, we tackled two major limitations of existing image search rankers - not incorporating visual features and learning from training data with label noise. By using user click data as a "bridge" to connect image visual content with textual queries as well as user intents, we propose a graph-based label propagation approach to tackle the two issues. The extensive experiments evaluated on 564 test queries show that our proposed GLP algorithm gave consistently better results than SVM-based method on different visual features. Moreover, an important property of GLP method is its speed, which is very efficient and can provide instant response. This aspect makes it a good candidate for online image search applications.

6. ACKNOWLEDGMENTS

This work was supported in part by a grant from the Research Grants Council of the Hong Kong SAR (CityU 119610) and Microsoft Research Asia Windows Phone Academic Program FY12-RES-OPP-107.

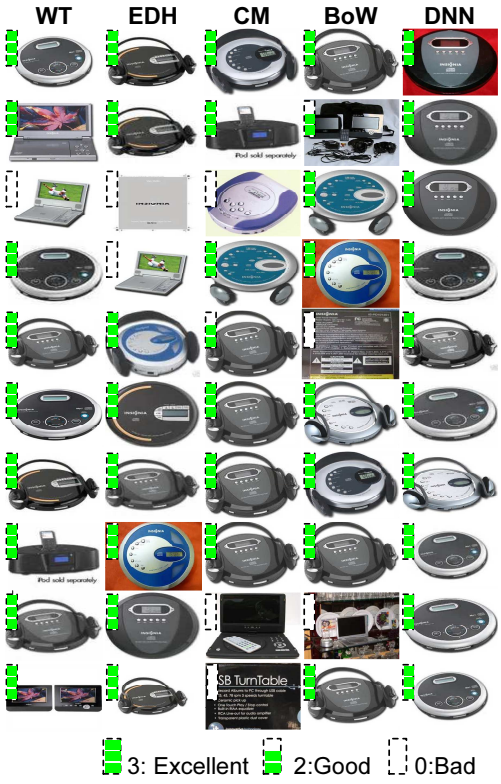

Figure 1: The exemplary list of top ten images for query "portable cd player insigna" ranked by GLP method based on five different visual features.

7. REFERENCES

[1] R. A. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD*, 2007.

[2] C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo. The neural representation benchmark and its evaluation on brain and machine. In *ICLR*, 2013.

[3] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR*, 2007.

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

[5] V. Jain and M. Varma. Learning to re-rank: Query-dependent image re-ranking using click data. In *WWW*, 2011.

[6] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. on Information Systems*, 25(2), 2007.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[8] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR*, 2008.

[9] B. Poblete and R. A. Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize web documents. In *WWW*, 2008.

[10] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. Image ranking based on user browsing behavior. In *SIGIR*, 2012.

[11] J. Wang and S. Li. Query-driven iterated neighborhood graph search for large scale indexing. In *ACM MM*, 2012.

[12] T. Yao, T. Mei, C.-W. Ngo, and S. Li. Annotation for free: Video tagging by mining user search behavior. In *ACM MM*, 2013.