Click-through-based Subspace Learning for Image Search

Yingwei Pan [†], Ting Yao [‡], Xinmei Tian [†], Houqiang Li [†], Chong-Wah Ngo [‡]
[†] University of Science and Technology of China, Hefei, China
[‡] City University of Hong Kong, Kowloon, Hong Kong
{panyw, tingyao}.ustc@gmail.com; {xinmei, lihq}@ustc.edu.cn; cscwngo@cityu.edu.hk

ABSTRACT

One of the fundamental problems in image search is to rank image documents according to a given textual query. We address two limitations of the existing image search engines in this paper. First, there is no straightforward way of comparing textual keywords with visual image content. Image search engines therefore highly depend on the surrounding texts, which are often noisy or too few to accurately describe the image content. Second, ranking functions are trained on query-image pairs labeled by human labelers, making the annotation intellectually expensive and thus cannot be scaled up.

We demonstrate that the above two fundamental challenges can be mitigated by jointly exploring the subspace learning and the use of click-through data. The former aims to create a latent subspace with the ability in comparing information from the original incomparable views (i.e., textual and visual views), while the latter explores the largely available and freely accessible click-through data (i.e., "crowdsourced" human intelligence) for understanding query. Specifically, we investigate a series of click-through-based subspace learning techniques (CSL) for image search. We conduct experiments on MSR-Bing Grand Challenge and the final evaluation performance achieves $DCG@25 = 0.47225$. Moreover, the feature dimension is significantly reduced by several orders of magnitude (e.g., from thousands to tens).

Categories and Subject Descriptors

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

General Terms

Algorithm, Experimentation.

Keywords

Image search, subspace learning, click-through data, DNN image representation.

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

http://dx.doi.org/10.1145/2647868.2656404.

1. INTRODUCTION

Keyword-based image search has received intensive research attention since the early of 1990s [8]. The significance of the topic can be partly reflected from the huge volume of published papers, particularly for addressing the problems of learning the rank or similarity functions. Despite these efforts, the fact that the queries (texts) and search targets (images) are of two different modalities (or views) has resulted in the open problem of "semantic gap." Specifically, a query in the form of textual keywords is not directly comparable with the visual content of images. The commercial search engines to date primarily reply on textual features extracted from the surrounding texts of images. 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.

Inspired by the success of subspace learning [9], this paper studies the cross-view (i.e., text to image views) search problem by learning a common latent subspace that allows direct comparison of text queries and images. Specifically, by mapping to the latent subspace, the similarity between a textual query and a visual image can be directly measured between their projections, making the information from original incomparable cross-view space comparable in the shared latent subspace.

Moreover, we consider exploring user click-through data, aiming to understand the user's intent for image search. In general, image rankers obtain training data by having human experts label 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 "gorilla hummer," experts tend to label images of animals "gorilla" and "hummer" as highly relevant. However, empirical evidence suggests that most users wish to retrieve images of a car of "gorilla hummer" type. The experts' labels might therefore be erroneous resulting in training sets with label noise and the ranker is learnt to be sub-optimal. In this work, our click-through-based learning provides an alternative to address this problem. Most image search engines display results as thumbnails. The user can browse the entire image search results before clicking on a specific image. As such, users predominantly tend to click on images that are relevant to their query. Therefore, the click data can serve as a reliable and implicit feedback for image search.

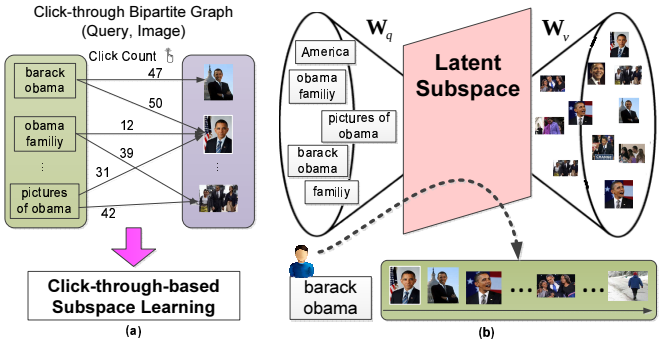

Figure 1: Click-through-based image search framework (better viewed in color). (a) Latent subspace learning between textual query and visual image based on click-through bipartite graph. (b) With the learnt mapping matrices $\mathbf{W}_q$ and $\mathbf{W}_v$, queries and images are projected into this latent subspace and then the distance in the latent subspace is directly taken as the measurement of query-image relevance. Then for each query, the images are ordered based on the relevance scores to the query.

By jointly integrating subspace learning and click-through data, this paper investigates click-through-based subspace learning approaches (CSL) for image search, as shown in Figure 1. Specifically, a bipartite graph between the user queries and images is constructed on the search logs from a commercial image search engine. An edge between a query and an image is established, if the users who issue the query clicked the image. Subspace learning aims to learn a latent subspace in the way of minimizing the distance between the mappings of query and image, or maximizing the correlation between the two views. After the optimization of subspace learning, the relevance score between a query and an image in the original space can be directly computed based on their mappings. For any query, the image search list will be returned by sorting their relevance scores to the query.

In summary, this paper makes the following contributions:

- We study the problem of keyword-based image search by jointly exploring subspace learning and the use of click-through data.
- We investigate click-through-based subspace learning methods, which aim to learn a latent subspace. By mapping to the subspace, textual queries and visual images can be directly compared.

The remaining sections are organized as follows. Section 2 presents click-through-based subspace learning methods. Section 3 provides empirical evaluations, followed by the discussions and conclusions in Section 4.

2. CLICK-THROUGH-BASED SUBSPACE LEARNING

The main goal of click-through-based subspace learning (CSL) method is to create a latent common subspace with the ability of directly comparing semantic textual query and image visual content. Four subspace learning techniques are investigated, i.e., Canonical Correlation Analysis (CCA) [2], Click-through-based Cross-view Learning (CCL) [6], Polynomial Semantic Indexing (PSI) [10], and Passive-Aggressive Model (PA) [3]. After we obtain the latent subspace, the relevance between query and image is directly measured by their mappings. The approach overview is shown in Figure 1.

In the following, we will first define the bipartite graph that naturally encodes user actions in the query log, followed by briefly presenting the four subspace learning approaches. Finally, the CSL algorithm for image search is presented.

2.1 Notation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a click-through bipartite. $\mathcal{V} = Q \cup V$ is the set of vertices, which consists of a query set Q and an image set V. $\mathcal{E}$ is the set of edges between query vertices and image vertices. The number associated with the edge represents the clicked times in the image search results of the query. Suppose there are n triads $\{q_i, v_i, c_i\}_{i=1}^n$ generated from the click-through bipartite in total, where c_i is the click counts of image v_i in response to query q_i. Let $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}^\top \in \mathbb{R}^{n \times d_q}$ and $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}^\top \in \mathbb{R}^{n \times d_v}$ denote the query and image feature matrix, where $\mathbf{q}_i$ and $\mathbf{v}_i$ are the textual and visual feature of query q_i and image v_i, and d_q and d_v are the feature dimensionality, respectively. The click matrix $\mathbf{C}$ is a diagonal $n \times n$ matrix with its diagonal elements as c_i.

2.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a classical technique, which explores the mapping matrices by maximizing the correlation between the projections in the subspace. We assume that a low-dimensional common subspace exists for the representation of the query and image. The linear mapping function can be derived from the common subspace by

$$f(\mathbf{q}_i) = \mathbf{q}_i \mathbf{W}_q, \quad \text{and} \quad f(\mathbf{v}_i) = \mathbf{v}_i \mathbf{W}_v, \tag{1}$$

where d is the dimensionality of the common subspace, and $\mathbf{W}_q \in \mathbb{R}^{d_q \times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$ are the transformation matrices that project the query textual semantics and image content into the common subspace, respectively. CCA aims to find the two linear projections making $(\mathbf{Q}\mathbf{W}_q, \mathbf{V}\mathbf{W}_v)$ maximally correlated as

$$(\mathbf{W}_q, \mathbf{W}_v) = \underset{\mathbf{W}_q, \mathbf{W}_v}{\operatorname{argmax}} \ \operatorname{corr}(\mathbf{Q}\mathbf{W}_q, \mathbf{V}\mathbf{W}_v). \tag{2}$$

Specifically, we view the click number of a query and an image as an indicator of their relevance. All the query-image pairs generated from click-through bipartite graph are used for learning the linear mapping projections.

2.3 Click-through-based Cross-view Learning

The training of Click-through-based Cross-view Learning (CCL) is performed simultaneously by minimizing the distance between query and image mappings in the latent subspace weighted by their clicks, and preserving the structure relationships between the training examples in the original feature space. In particular, the objective function of CCL is composed of two components, i.e., distance between views in the latent subspace, and the structure preservation in the original space. The overall objective function of CCL is as

$$\begin{aligned} &\min_{\mathbf{W}_q, \mathbf{W}_v} tr\big((\mathbf{Q}\mathbf{W}_q - \mathbf{V}\mathbf{W}_v)^\top \mathbf{C}(\mathbf{Q}\mathbf{W}_q - \mathbf{V}\mathbf{W}_v)\big) \\ &+\lambda\Big(\sum_{i,j=1}^{n} \mathbf{S}_{ij}^q \|\mathbf{q}_i\mathbf{W}_q - \mathbf{q}_j\mathbf{W}_q\|^2 + \sum_{i,j=1}^{n} \mathbf{S}_{ij}^v \|\mathbf{v}_i\mathbf{W}_v - \mathbf{v}_j\mathbf{W}_v\|^2\Big) \end{aligned}, \tag{3}$$

Table 1: The DCG@25 (%) of different approaches on Dev dataset.

Appr.	NGS	GLP	CCA	CCL	PSI	PA	FUS
	48.99	50.25	50.55	50.59	49.91	50.17	51.12

where λ is the tradeoff parameter, $\mathbf{S}^q \in \mathbb{R}^{n\times n}$ and $\mathbf{S}^v \in \mathbb{R}^{n\times n}$ denote the affinity matrices defined on the queries and images, respectively. The first term is the cross-view distance, while the second term represents structure preservation.

The underlying assumption of CCL is that the higher the click number, the smaller the distance between the query and the image in the latent subspace. Furthermore, the similarity between examples in the original space can be preserved in the learned latent subspace.

2.4 Polynomial Semantic Indexing

Given a query q_i and an image v_j, a polynomial ranking model with 2-degree is given by

$$f(q_i, v_j) = (\mathbf{q}_i\mathbf{W}_q)(\mathbf{v}_j\mathbf{W}_v)^T. \quad (4)$$

The training of $\mathbf{W}_q$ and $\mathbf{W}_v$ could be many forms. From our click-through data, we can easily get a set of triplets $\mathcal{T}$, where each tuple (q, v^+, v^-) consists of the query q, an image v^+ with higher click and a lower clicked image v^-. Deriving from the idea of "learning to rank" [5], it aims to optimize $\mathbf{W}_q$ and $\mathbf{W}_v$ which makes $f(q, v^+) > f(q, v^-)$, i.e., image v^+ should be ranked higher than image v^-. The margin ranking loss is employed and the optimization problem is defined as

$$minimize: \sum_{(q,v^+,v^-)\in\mathcal{T}} \max(0, 1 - f(q, v^+) + f(q, v^-)). \quad (5)$$

2.5 Passive-Aggressive Model

Similar in spirit, Passive-Aggressive model measures the match between a query q_i and an image v_j by first projecting the query into the image space. Accordingly, the ranking function is defined as

$$f(q_i, v_j) = \mathbf{q}_i\mathbf{W}_q\mathbf{v}_j^T. \quad (6)$$

For a tuple (q, v^+, v^-), the change of $\mathbf{W}_q$ is determined by whether the constrains of $f(q, v^+) > f(q, v^-)$ is verified. The optimization of $\mathbf{W}_q$ is performed by adapting the Passive-Aggressive algorithm [3]. It is worth noticing that here the subspace is set to image space and hence $\mathbf{W}_v$ is the identity matrix.

2.6 CSL Algorithm

After the optimization of $\mathbf{W}_q$ and $\mathbf{W}_v$, we can obtain the linear mapping functions defined in Eq.(1). With this, original incomparable textual query and visual image become comparable. Specifically, given a test query-image pair, we can compute the similarity or distance value between the pair as reflecting how relevant the query could be used to describe the given image. For any query, sorting by its corresponding values for all its associated images gives the retrieval ranking for these images.

3. EXPERIMENTS

We conducted experiments on the MSR-Bing Image Retrieval Challenge dataset, i.e. Clickture [4], which contains a training and a development (Dev) sets. It was collected from one year click-through data of one commercial image search engine. There are more than 11.7 millions distinct queries and 1.0 million unique images of the training set. Figure 2 shows a few exemplary images with their clicked queries and click counts in the Clickture. For example, users clicked the first image 12 times in the search results when submitting query "red wine" in total. Note that there is no surrounding text or description of images provided in the dataset.

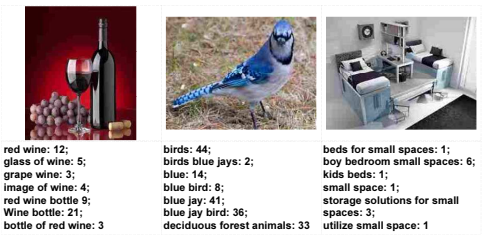

Figure 2: Examples in Clickture dataset (upper row: clicked images; lower row: search query with click times on the upper image).

In the Dev set, there are 79,926 $\langle query, image\rangle$ pairs generated from 1,000 queries, where each image to the corresponding query was manually annotated on a three point ordinal scale: Excellent, Good, and Bad. In the experiments, the training set is used for learning the latent subspace, while the Dev set is used for performance evaluation. In addition, there is an official test set for the final evaluation.

3.1 Experimental Settings

Textual and Visual Features. We take the word in queries as "word features." Words are stemmed and stop words are removed. With word features, each query is represented by a tf vector in the query space. In our experiments, we use the top 50,000 most frequent words as the word vocabulary. Visual feature derived from Convolutional Neural Networks (CNN) by using DeCAF [1] is extracted as image representation.

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

- N-Gram SVM Modeling (NGS). 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 support vector machine (SVM) model for each query, and then use this model to predict the relevance of the query to a new image. When a query is not in the training set, but its n-grams appear in some queries of the training set, we generate the model by linearly fusing the SVM models of these queries. We name this run as *NGS*.
- Graph-based Label Propagation (GLP) [7]. GLP employs neighborhood graph search to find the nearest neighbors on an image similarity graph built up with visual representations and further aggregates their clicked queries/click counts to get the labels of the new image. This run is named as *GLP*.
- Click-through-based Subspace Learning. We design five runs for CSL approaches: *CCA*, *CCL*, *PSI*, *PA*, and their linear fusion *FUS*.

Table 2: The DCG@25 (%) of our three submitted runs on test dataset.

Run	FUS	FUG	FUA
	47.225	46.404	47.441

Table 3: Run time (ms) of six different approaches. The experiments are conducted on a regular PC (Intel dual-core 2.0GHz CPU and 100 GB RAM).

Appr.	NGS	GLP	CCA	CCL	PSI	PA
	7500	14.8	4.5	1.0	1.0	1.5

Evaluation Metrics. Following the challenge's measurement, for each query, we use Discounted Cumulated Gain (DCG) to evaluate the performance of top 25 images.

3.2 Performance Comparison

Table 1 shows the DCG performance of seven runs averaged over 1,000 queries in Dev dataset. Overall, all the CSL approaches consistently lead to a performance boost against *NGS*. Particularly, the DCG@25 performance of *CCL* can achieve 0.5059, which improves *NGS* by 3.5%. As a result of linear fusion, *FUS* improves the performance up to 0.5112. Furthermore, *CCA*, *CCL* and *FUS* all exhibit better performance than *GLP*, while the performance of *PA* is slightly less than *GLP*. More importantly, by learning a low-dimensional latent subspace, the dimension of the mappings of textual query and visual image is reduced by several orders of magnitude. In our experiments, the dimensionality of latent subspace is empirically set to 50 for *CCA*, *CCL*, and *PSI*. Figure 3 shows the top ten images for query "college station texas" by using five CSL approaches, respectively.

Table 2 details the performances of our three submitted runs on test dataset. In addition to *FUS*, *FUG* is late fusion by performing GLP on several visual features including color moments, wavelet texture, histograms of oriented gradients, and CNN feature. *FUA* is average fusion of *FUS* and *FUG*. As indicated by our results, *FUS* significantly outperforms *FUG* and *FUA* achieves the best performance.

3.3 Run Time

Table 3 lists the detailed run time for each compared methods. The CSL approaches are extremely efficient, completing relevance prediction of each image-query pair within five milliseconds on average. They are much faster than *NGS* which needs beyond five seconds and *GLP* which requires about 15 milliseconds, respectively.

4. CONCLUSION

In this paper, we tackled two major limitations of existing image search rankers - highly depends on surrounding texts and learning from training data with label noise. We have investigated the issues of directly learning the cross-view distance between a textual query and an image by leveraging both click data and subspace learning techniques. The click data represent the click relations between queries and images, while the subspace learning aims to learn a latent common subspace between multiple views. The extensive experiments evaluated on 1,000 queries show that CSL approaches gave better results than SVM-based and graph-based methods. Moreover, CSL approaches have good properties on both feature dimension reduction and speed, making them good candidates for online image search applications.

Figure 3: The exemplary list of top ten images for query "college station texas" ranked by CSL approaches.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61272290), the Fundamental Research Funds for the Central Universities (No. WK2100060011), and the Shenzhen Research Institute, City University of Hong Kong.

5. REFERENCES

[1] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[2] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, pages 210–233, 2014.

[3] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *IEEE Trans. on PAMI*, 30(8):1371–1384, 2008.

[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] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[6] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui. Click-through-based cross-view learning for image search. In *SIGIR*, 2014.

[7] Y. Pan, T. Yao, K. Yang, H. Li, C.-W. Ngo, J. Wang, and T. Mei. Image search by graph-based label propagation with image representation from dnn. In *ACM MM*, 2013.

[8] Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions, and open issues. *VCIR*, 10(1):39–62, 1999.

[9] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *CoRR abs/1304.5634*, 2013.

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