France Telecom Orange Labs (Beijing) AT MSR-BING CHALLENGE ON IMAGE RETRIEVAL 2013

Lezi Wang[‡], Shusheng Cen[‡], Hongliang Bai[†], Chong Huang[‡], Nan Zhao[‡],
Bo Liu[‡],Yanchao Feng[‡],Yuan Dong[†‡]
[†]France Telecom Research & Development - Beijing, 100190, P.R.China
[‡]Beijing University of Posts and Telecommunications,100876, P.R.China
{hongliang.bai,yuan.dong}@orange.com
{wanglezi.bupt,censhusheng,huangchong661100}@gmail.com

ABSTRACT

This study addresses approaches of our team ORANGE for MSR-Bing Image Retrieval Challenge to assess the relevance on a pair of query term and image. Our approaches aim to boost the performance of web scale image retrieval, where images in the initial list (indexed by query term) can be re-ranked based on their relevances. One year BING image retrieval logs are used to develop the relevance assessment system. Several visual features are employed to describe one image. A visual similarity learning algorithm is introduced to train a weighted image similarity. Three runs are submitted for evaluation: “boostLearn”, “fast_v1”, “learn_RF”.

Categories and Subject Descriptors

I.4 [**IMAGE PROCESSING AND COMPUTER VISION**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Text Retrieval, Text Clustering, Visual Similarity Learning, Click Data

1. INTRODUCTION

The task of MSR-Bing Image Retrieval Challenge to assess the relevance between query terms and images. The system need to return a floating point score to reflect their relevance. The single score does not play any significant role. We more concern the score order of images under the same query term. The initial set of images indexed by text can be re-ranked based on the score so that the image search engine can be optimized.

Traditional solutions is training a set of classifiers for different topics and assessing the relevance by the classifier related to a specific query term. It is hard to scale up when dealing with a huge amount of concepts. The training dataset is one-year BING image retrieval logs, consisting of 11.7 million unique queries and covers a wide range of topics. Training a classifier for each topic is impractical. A retrieval-based method is proposed to handle this problem. Specifically, two kinds of schemes are presented:1)online classification scheme based on semantic information extracted from click data (“fast_v1” and “learn_RF”); 2)cluster-based query-image relevance assessment (“boostLearn”).

In the framework of online classification scheme, related images are retrieved by indexing the clicked queries in database based on test query. Therefore, textual information is mapped into visual space. The relevance of the query-image pair is scored by computing the visual similarity between the test image and images indexed by the query.

In the framework of cluster-based query-image relevance assessment, the score of a query-image pair is given through computing the visual similarity between the test image and database samples which are relevant in the text domain. Therefore, the task become a generic computer vision problem that how to assign higher similarity to images which reflect the same semantic or instance. We employed several low level visual features with different parameters, including Color-Sift(CSIFT) [1], GIST [3], LSH-RGB Historgram(LRH) [2]. Motivated by study [5], we apply a visual similarity learning algorithm to learn a weighted cosine similarity. In the evaluation stage, the primary run adopts one feature which gives the highest Discounted Cumulated Gain @25 (DCG@25) on the development dataset.

The rest of the paper is organized as follows. The system overview is described in section 2. The details of visual similarity learning are presented in section 3. The system performance and conclusion are in section 4 and 5 respectively.

2. SYSTEM PIPELINE

2.1 Framework of Online Classification Scheme

As illustrated in Fig.1, the online classification system is mainly composed of two modules: text search and visual comparison. For convenience, we denote click data as triads (Q_i, I_j, C_{ij}), where Q_i is a query term that consists of several keywords, I_j is a clicked image under the query term Q_i, and click count C_{ij} denotes how many times user click on image

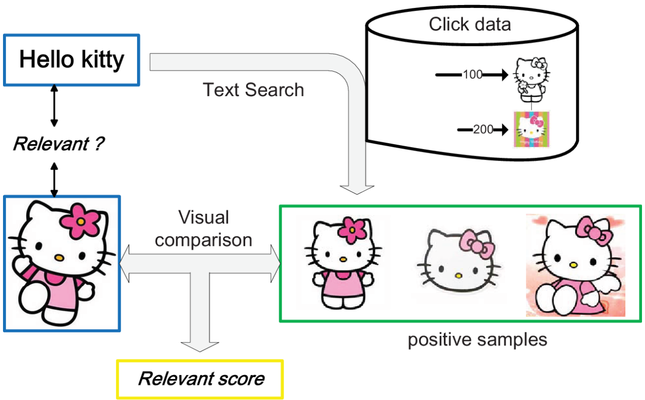

Figure 1: The system overview: online classification scheme.

I_j when using Q_i as query. Table 1 is an example of query-image logs.

Table 1: Example of Query-Image Pair in Log file

Image ID	Query Term	Click
1	border colli golden retriev mix	4
1	golden retriev mix dog	3
1	border colli golden retriev mix puppi	2
1	chow chow mix	2
1	chow mix	2

2.1.1 *Text Search*

For a given query-image pair, relevant images can be found by comparing the test query term with clicked query logs in the database. The visual pattern for a specific query can be generated from these relevant images, which is crucial for the subsequent classification.

In text retrieval stage, queries under the same image are concatenated to form a document. The click count is a good indicator for the relevance of the clicked query to the image. Higher click count indicates higher relevance. For an image I_i, we generate a document D_i and define the confidence for clicked query $Q_j \in D_i$ as

$$Confi(D_i, Q_j) = \frac{C_{ij}}{\sum_j C_{ij}} \qquad (1)$$

Input query is compared with documents to index relevant images. Measuring the similarity of a few keywords in semantic level is quite challenging. It is hard to utilize content to infer the meaning due to shortness of query terms. Fortunately, we can benefit from the completeness of big data. We can always find the exact same words between search query and relevant documents. Therefore, we just take word overlap into account in terms of efficiency and simplicity. Besides the single words, all consecutive sequences of two words in a query are also considered as a term, which are known as shingles. For example, "the great wall" will produce five terms: "the", "great", "wall", "the great", "great wall". As for shingles, heavier weight is set. The relevance $R(Q, D)$ between a query term Q and document D is defined as

$$R(Q, D) = \sigma \sum_{j \in D} S(Q, Q_j) Confi(D, Q_j) \qquad Q_j \in D \quad (2)$$

where σ is a weighting factor in proportion to the click count of the document(image) D and $S(Q, Q_j)$ indicates the similarity between two query terms Q and Q_j, given by

$$S(Q, Q_j) = \frac{V(Q) \cdot V(Q_j)}{|V(Q)| \cdot |V(Q_j)|} \qquad (3)$$

where $V(\cdot)$ is the TF_IDF vector of query. In practice, some simple preprocessing is applied to raw click queries. For example, the meaningless words like "image of" or "picture of" will be removed. And stemming is also used to group words with a similar basic meaning together.

In our implement, the open source library Lucene[1] is used to index clicked queries after preprocessing. Lucene allows us to perform fast search on the whole training set with 23 million clicked queries. A typical search can be done within one second.

2.1.2 *Visual Comparison*

To assess relevance of a query-image pair, a classifier is needed. Well-designed models like Support Vector Machine work well in task like object classification. However, a complicated model can not trained online due to expensive time cost. Nearest neighbor rule is one of the simplest classification methods in machine learning. Fairish performance can be achieved by careful parameters tuning. For this reason, a kNN scheme is used in this step.

Given a list of positive samples, the visual similarities between the test image and samples are computed. Let f_i be the feature of image I_i, Euclidean distance is used to measure the visual similarity

$$S_{visual}(I_1, I_2) = \frac{1}{||f_1 - f_2||} \qquad (4)$$

In our system, GIST [3] is used for visual similarity comparison. GIST of 960 dimensions is a global feature inspired by biological vision and shown to be useful in scene recognition. All positive samples are ranked based on visual similarity. The relevance of a query-image pair (Q, I) is scored as the sum of the top-N visual similarities:

$$Rel(Q, I) = \sigma_{I_i \in topN} S_{visual}(I, I_i) \qquad (5)$$

The value N has a great influence with predict performance. Bigger N can restrain noise, but also lower the discriminability. In our system, N is set five.

2.2 Framework of Cluster-based Query-Image Relevance Assessment

Fig. 2 shows an example of relevance assessment between the query "drone" and a test image. In off-line phase, the query logs in database are filtered by the Porter Stemming Algorithm; non-English terms and custom stop words are removed. Then images are clustered into different categories based on the text labels. One image category can be seen

[1] http://lucene.apache.org/

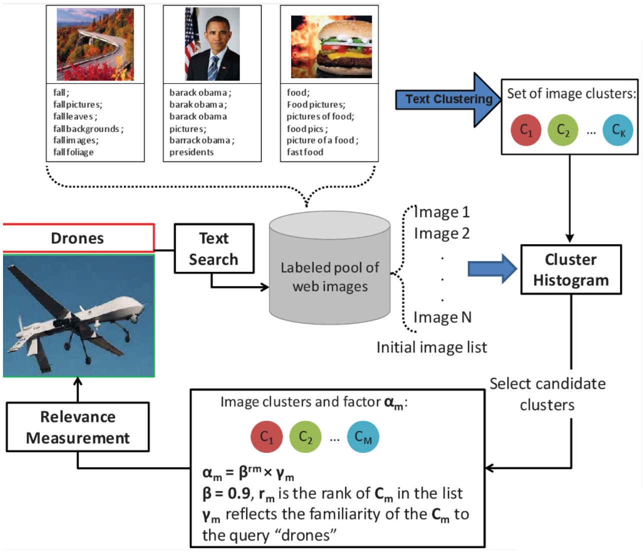

Figure 2: The system overview: cluster-based query-image relevance assessment.

as an expert that gives scores to the test images. In the experiment, near 10K image categories are generated over the training dataset. The training dataset contains 1M images and 20 query terms associate to one image on average.

In online phase, the query-image pair goes through the "text Search" module firstly (described in section 2.1.1). Several candidate images in database are indexed based on the query term "drones". The image cluster histogram is generated over the initial list to describe the clusters distribution. M maximum bins are selected as the experts to judge the relevance base on the visual similarities between the test image and images that the experts have seen. A specific weight is assigned to each expert's judgement. In addition, a visual learning algorithm is applied to train the weighted cosine similarity.

2.2.1 *Visual Features*

Three low-level image features are employed to describe the visual content: Color-Sift(CSIFT), GIST, LSH-RGB Historgram(LRH). The GIST features are extracted through the open source code[3]. The LRH feature is a combination of Uniform Local Binary Pattern (LBP) and color-histogram, introduced by Gal Chechik et. al [2].

In our run, CSIFT descriptors are extracted through [1]. Bag-Of-Words encoding algorithm is used to map a set of descriptors in one image into a sparse high-dimension TF-IDF weighting histogram. Because the vast majority of computing time is spent on calculating the nearest neighbours between the points and cluster centers, the exact k-means based nearest neighbour is computationally expensive. The Approximate K-means(AKM) [4] is adopted to speed up the computation of the assignments in each iteration by an Approximate Nearest Neighbor (ANN) method. In the experiment, the codebook sizes vary in 1K, 10K, 1M.

2.2.2 *Relevance Measurement*

The images indexed through text-based retrieval are used to generate the image cluster histogram, indicating clusters distribution over the initial list. As shown in Fig.2, m candidate clusters are selected as experts to assess the relevance on a query-image pair. For one expert, we assign it a specific weight α_m to indicate its importance

$$\alpha_m = \beta^{r_m} \cdot \gamma_m \tag{6}$$

where r_m is the rank of expert C_m in the list and β is set as 0.95 experimentally. γ_m reflects the familiarity of the expert C_m to the query, which is a ratio of the total images number in cluster C_m divided by the number of the initial list images $I_n(I_n \in C_m)$. The final relevance score $Rel(Q, I)$ between query Q and I is computed as a weighted sum

$$Rel(Q, I) = \sum_m \alpha_m \cdot S(I, C_m) \tag{7}$$

where the $S(I, C_m)$ is the visual relevance between image I and C_m, which equals the average of visual similarities between I and top-N images in C_m. N is set to 5 experimentally. We employ weighted visual similarity which is described in the next section.

3. VISUAL SIMILARITY LEARNING

Instead of using purely visual similarity measurement, we learn a weighted cosine similarity similar with [5]. The similarity $S_w(x_i, x_j)$ compares the features of images I_i, I_j, where $|x| = 1$. The S_w is defined as a weighted sum over component-wise products

$$S_w(x_i, x_j) = \sum_d w_d(x_i^d \cdot x_j^d) \tag{8}$$

If $w_d = 1$ for all d, the $S_w(x_i, x_j)$ is the simple cosine similarity. We aim at learning w which can give higher similarities to image pairs of the same categories. In training stage, we select a part of images from training and development dataset to form a set of positive pair (x_i, x_j) and negative pair (x_k, x_l). Two images of development dataset under the same query term labeled as "Both Excellent" or "Excellent-Good" are regarded as a positive pair; two images in training dataset belonging to the same category are a positive pair; we randomly select image pairs from different categories to generate negative pairs. The similarities of positive pairs $S_w(x_i, x_j)$ are set to be larger by a margin than the similarities of negative pairs $S_w(x_k, x_l)$, defined as Eq. 9

$$S_w(x_k, x_l) \leq b - 1 < b + 1 \leq S_w(x_i, x_j) \tag{9}$$

Let $X_n = x_i \cdot x_j$ be the product vector for pair $n = (i, j)$. Let $y_n = 1$ for positive pair and $y_n = -1$ for negative pair. The last inequality can be rewrrited as

$$y_n(w^T X_n - b) \geq 1 \qquad \forall n = (i, j) \tag{10}$$

which can be seen as a constrain in typical two-class support vector machine. Hence, we minimize the following objective function to get the optimum w

$$\begin{aligned} min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_n \xi_n \\ s.t.\ y_n(w^T X_n - b) \geq 1 \qquad \forall n = (i, j) \end{aligned} \tag{11}$$

4. EXPERIMENTAL RESULTS

The proposed system are evaluated on the whole development (DEV) dataset. We choose the approach of the highest DCG@25 as the primary run in the evaluation stage. We found the DEV and evaluation (EVA) dataset are very similar in terms of the data size and content. There are near 80K query-image pairs in DEV dataset, including 1000 different query terms; EVA set contains 77406 query-image pairs.

4.1 Performance of Cluster-based Query-image Relevance Assessment

Table 2 shows the DCG@25 of differen approaches conducted on the DEV dataset and time cost per query-image pair. We assign a random score on each query-image pair to simulate the initial image ranking list indexed by text only. The table shows that our approaches can boost the initial list in terms of DCG@25. In development stage, the best performance is given by 1M CSIFT. And the DCG are improved with the codebook size increasing. While the computation time cost of CSIFT is the most expensive. In evaluation stage, we select 1M CSIFT approach as our primary run "boostLearn" which achieves DCG@25 of 0.531.

Table 2: System Performance: Cluster-based Query-image Relevance Assessment

Method	DCG@25 (DEV)	Time cost
Random	0.471	-
Direct Gist	0.501	2s
weighted Gist	**0.512**	3s
Direct LRH	0.474	2s
weighted LRH	0.489	2s
CSIFT Booksize = 1k	0.472	5s
CSIFT Booksize = 10k	0.497	6s
CSIFT Booksize = 1M	**0.529**	7s

4.2 Performance of Online Classification System

Two runs are submitted to the evaluation system: fast_v1 and learn_RF. They share most of the settings mentioned above. There are only a few differences: Fast_v1 A baseline system which is the fastest version and learn_RF An improved version of baseline system. A trick is added to deal with the situation where there are not any positive images found in database. In this case, the previous test data will be treated as positive samples.

Table 3: System Performance: Online Classification Scheme

Run	DCG@25 on develop set	DCG@25 on test set
fast_v1	0.478	0.480
learn_RF	0.501	0.516

The system performances are shown in Table 3. Both develop set and test set are challenging, and our performance is better than random prediction (DCG@25 is 0.471). These results indicate that our solution is effective in solving this problem. The run learn_RF benefits from previous test data, performing better than fast_v1.

5. CONCLUSION

Two schemes are proposed to assess the relevance between query terms and images based on one year BING image retrieval logs. In online classification scheme, based on a retrieval-based framework, relevant images for a specific topic can be discovered from raw click data and the relevance between a pair of image and query text can be assessed under a k-nearest neighbor rule. In cluster-based assessment, training images are clustered into specific categories and each category can be regarded as a expert. Experts rate test images based on the images they have seen. We employ weighted cosine similarity to generate the score in image domain. In evaluation stage, our system performance can be in top among contestants. Additionally, learning a specific similarity weight over corresponding category may boost the whole system performance. And how to take advantage of the click logs to learn visual similarity is also worth to be studied.

6. REFERENCES

[1] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *Comput. Vision Image Understanding*, pages 113:48–62, 2009.

[2] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. An online algorithm for large scale image similarity learning. In *Advances in Neural Information Processing Systems*, 2009.

[3] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *International Conference on Image and Video Retrieval*. ACM, july 2009.

[4] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98: the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[5] T.Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, pages 1777–1784, June 2011.