Report on a baseline approach to the second MSR-Bing Challenge on Image Retrieval

Aleksandr Sayko
Yandex, Moscow, Russia
asayko@yandex-team.ru

Anton Slesarev
Yandex, Moscow, Russia
slesarev@yandex-team.ru

ABSTRACT

In the report we present a baseline approach for solving the problem which was set within the second MSR-Bing Challenge on Image Retrieval. Our goal is to present a simple method to calculate image's relevance to the given search query based on similarity matching with pictures from the click log. The relevance is estimated in two steps. In the first step we fetch images that were clicked for search queries that are similar to the query submitted from the click log. Within the second step we calculate the cumulative similarity of given image to the fetched pictures. Finally the cumulative similarity is returned as the requested image's relevance. Using the proposed approach we reached the value of 0.486901 of the target metric on the evaluation query set.

1. EXTRACTING CLICK-RELEVANT IMAGES

Our purpose at the first stage is to mine pictures that we could possibly use as relevant examples to a given search query from the click log. In order to find such images we build four mappings. The first index maps normalized search queries from the click log to the images that were clicked for the normalized query. We also save the total number of clicks in the index to be able to take top most clicked for a query pictures. For normalization we lemmatize the words of the query delete stop words and sort the rest lexicographically. We exploit lemmatization from the NLTK framework [2].

The second index maps words' lemmas contained in search queries to pictures that were clicked for the queries. In the index we also save the cumulative number of clicks and the number of different queries containing the lemma for that the picture was clicked. The numbers are needed to fetch top relevant images for a lemma, where relevance is estimated using heuristic rules.

The third and the fourth indices are alike with the first one but they store bigrams and respectively trigrams that are found in the click log. By n-gram we mean sorted tuple of lemmas contained in the search query.

1.1 Search query processing

A search query is processed the following way. At first we normalize the query. For each query lemma we generate a set of synsets [2]. Then we create all possible bigrams and trigrams building all combinations of elements from different sets. After that we try to retrieve 100 images from the four indices into "top 100 related pics" list. At first we retrieve top 100 most clicked images from the index build upon normalized queries. If there are less than 100 clicked images we try to fetch lacking pictures from the other indices using following heuristic rule. For each n-gram generated from the search query we fetch all clicked images from the n-gram indices. Next for each of the extracted images we estimate the click relevance. The estimated click relevance of a picture is the sum over all extraction cases(that are caused by matched n-grams) of products of the number of different queries containing the n-gram the image was clicked for with the logarithm of the total number of clicks the picture got on queries containing the n-gram and number of noun-like lemmas in the n-gram and the n-gram match type coefficient, where n-gram match type coefficients are empirically tuned. Then we sort the retrieved images according to the estimated click relevance and try to complete the "top 100 related pics" list up to 100 images with the top images from the retrieved list.

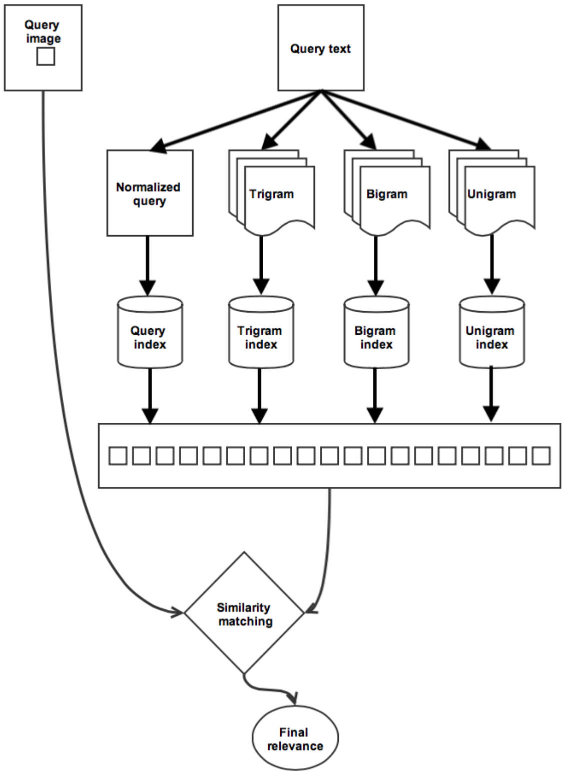
The proposed heuristic approach enables fetching of visually consistent lists of pictures for considerable part of search queries from the provided test set, which could be used for training query specific classifiers. As a proof of the concept we used the lists as relevant examples for relevance estimation. We estimated requested relevance of a given picture as sum of visual similarities with retrieved images.

2. COMPARING GIVEN IMAGE WITH CLICK-RELEVANT IMAGES

To calculate visual similarity score between two images we use a standard bag-of-words framework [5], with images represented as L_2 normalized histograms of visual words. We use a rather small vocabulary containing 32768 visual words, trained from SIFT[3] descriptors. If L_2 distance between the histograms is less than a threshold we consider the two images to be similar enough to influence the relevance. After that we consider the descriptors of any two interest points from given and retrieved images respectively as matched if they are assigned to the same visual word.

We apply RANSAC-based geometric verification to find

Figure 1: Image relevance estimation pipeline within the proposed approach.



a maximal number of pairs of matched descriptors such that the corresponding interest points from given and retrieved images can be mapped to each other using combination of scaling and translation. Also we tried more complicated transformations there were no improvements. Pairs of matched descriptors corresponding to the found transformation are called inliers. The number of inliers characterizes similarity of two images.

We can compute the final similarity score of two images S as follows:

$$S = \begin{cases} 0, & L_{2\_hist\_dist} > thr \\ I/\sqrt{N_1 \cdot N_2}, & \text{otherwise} \end{cases}$$

where I is the number of inliers, N_1 and N_2 are the numbers of descriptors in the first and the second images consequently, $L_{2\_hist\_dist}$ is the L_2-distance between the histograms of visual words and thr is manually tuned threshhold.

3. CONCLUSION

Using the described baseline approach we recieved the target metric 0.486901 on the final evaluation query set. We believe that more sophisticated machine learning techniques applied to the retrieved image lists will significantly improve the result. For instance, we believe an interesting approach is to train a SVM classifier in the space of VLAD image descriptor features [1] considering the retrieved images as positive examples and random pictures from the click log as negative ones using the margin value as requested image's relevance. Another possible approach is to exploit the ensemble of exemplar-SVMs technique that was proposed in [4].

4. REFERENCES

[1] R. Arandjelović and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[2] S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[4] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[5] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct. 2003.