Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Exploiting Text Content in Image Search by Semi-supervised Learning Techniques

Chen Shen*†, Yahui Yang§ and Bin Wang‡
*School of Software and Electronics, Peking University, Beijing 100871, China
†Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
Email: scv119@gmail.com
§School of Software and Electronics, Peking University, Beijing 100871, China
Email:yhyang@ss.pku.edu.cn
‡Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
Email: wangbin@ict.ac.cn

***Abstract*—Along with the explosive growth of the Web, Web image search has become a more and more popular application which helps users digest the large amount of online visual information. Previous research mainly exploits visual information between images while rarely uses the text information surrounding the images on the Web pages. In this paper, we consider the relevance feedback as a machine learning problem. We proposed a novel relevance feedback framework for Web image search, which exploit both text and image modalities information with semi-supervised learning techniques. In each round of relevance feedbacks, the framework trains two classifiers for the two modalities by using the feedback information collected from the user. Then, it uses the unlabeled search result to improve these two classifiers. Finally, the ranked results list produced by image and text modality classifiers are combined to get the final rank. Experiments demonstrate the promise of the proposed framework.**

***Index Terms*—Web image search, co-training, semi-supervised learning, relevance feedback**

I. Introduction

Along with the explosive growth of the Web, Web image search has become a more and more popular application which helps users digest the large amount of online visual information.

In this work, we are interested in enhancing the performance of Web image retrieval by leveraging relevance feedback algorithm. The relevance feedback algorithm[1] is a effective scheme in information retrieval, which involves users in the retrieval process so as to help search engine to understand users' information need. In each round of relevance feedback, users are asked to mark some returned images as relevant or not relevant, from which the information retrieval system can obtain additional information to achieve a better performance.

Although there exists a lot of researches on relevance feedback which provide good performance on the task of image retrieval[3][2][4], all of them focused on content-based image retrieval(CBIR) which rarely consider the text information of images. As the significant characteristic of Web images is that the Web images are embedded in Web pages and the text surrounding the Web images can be treated as descriptions of the images, we consider the text modality is also useful for the Web image retrieval. Thus, we attempt to leverage both the image and text modality in relevance feedback.

If a learning problem contains a large number of unlabeled data and multi-modal exists on the training example, it fits the co-training [5] framework. The co-training framework first trains two separate classifiers on two modalities respectively. Then the unlabeled data are classified by the two classifiers and the classified data with most confident from one classifier are used to update another classifier[5]. Inspired by co-training paradigm, we utilize the user's feedback in each feedback round to train a text modality classifier as well as a image modality classifier. The the non-feedback Web images are then classified by each classifiers to update the another classifier. At the end of each feedback round, we can combine the results of the two retrained classifier and get the final result for each image.

The rest of paper is organized as follows: Section 2 describes the background knowledge on relevance feedback and semi-supervised machine learning algorithms while Section 3 presents our proposed approaches which utilize co-training algorithms to learn from the non-feedback Web images. In section 4, the experimental results are discussed. The conclusion and discussion about potential future works are listed in the last section.

II. Related Work

In this section, we review some previous work related to our work. We first introduce the relevance feedback and its usage on image search in the following subsection.

A. Relevance Feedback in Image Search

Relevance feedback is a powerful method which has firstly used to improve the performance of traditional text-based information retrieval[1]. The main idea of relevance feedback is to ask users to label the relevant and irrelevant retrieval results in the retrieval process. In particular, the steps of relevance feedback can be listed as follows[6]:

1) : The user submits a query;

2) : The system returns the preliminary retrieval results;

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

3) : The user are asked to mark some of the retrieval results as relevant or non-relevant;

4) : The system re-ranks the retrieval results based on the user feedback;

5) : The system displays a revised set of retrieval results, and goto the step3 for next iteration.

With the annotated results from users in each feedback round, the search engine can gradually understand what information the users are seeking for.

The Rocchio algorithm[1] based on the vector space model is one of the classic algorithms for implementing the relevance feedback on text-based information retrieval, which receives significant improvement. Then, relevance feedback was transformed and introduced into content-based multimedia retrieval, mainly content-based image retrieval. The relevance feedback appeared to have attracted more attention in the area of image search–a variety of solutions have been proposed within a short period and it remains an active research topic[2][13][3][4]. Zhou et al[12] pointed out that the reason could be that more ambiguities arise when human interpreting images than words.

B. Semi-Supervised Machine Learning Algorithm

Traditional supervised machine learning algorithms use only labeled data to train the classifier, while labeled data are always hard to get. On the contrary, the unlabeled data are easier to collect, but the traditional supervised machine learning algorithms have few ways to tackle them. Accordingly, the semi-supervised machine learning becomes a new research hot spot which make use of large amount of unlabeled as well as labeled data[7].

Accounting for the advantages of the semi-supervised learning, there have been a bunch of researches conducted and methods proposed. The Expectation-Maximization algorithm propose by Dempster et al[8] is recognized as one of the basic algorithms for semi-supervised learning, which makes the identification of generative mixture models practical. Methods such as self-training, transductive support vector machines and graph-based approaches were applied and received persuasive performance[9][10][7]. Co-training[5] is based on the assumption that multi-view exists in the same samples and trains two separate classifiers on two views respectively. In each round of the co-training, the unlabeled data are classified by the two classifiers and the classified unlabeled data with most confident from one classifier are used to 'teach' anther classifier. Each classifier is then updated by the additional training examples given by the other classifier, and the round repeats. The Co-training algorithms have strong constraint that the data contain two redundant view both of which are sufficient for the correct classification[11]. As this above conditions may not be satisfied sometimes, Goldman and Zhou[11] use two learners of different types but trained by one view, and take one learner's high confidence data to teach the other learner vice versa.

III. Enhancing Relevance Feedback by Co-training (CTRF Algorithms)

As we mentioned in Section 1, the Web images have both visual contents and surrounding textual contents. In our work, we first split the Web images into image modality and text modality. Then, each iteration of our CTRF algorithm can be roughly separated into following three steps: (1) After user labels the relevance on the preliminary retrieval results, the labeled results are used to train image modality classifier and text modality classifier. (2) In the next step, the co-training framework is engaged to learn from unlabeled data, and the two classifiers are retrained. (3) Finally, the results from this two classifier are merged to get the retrieval results. We describe our algorithm in details in the following subsections.

A. Classifiers Training

After the user submit a query, the image retrieval system returns the preliminary results which contain both textual description and thumbnails. Then, users are asked to label several images according to its relevance.

Here we introduce some notations: T stands for the set of all textual descriptions of the preliminary results, and I stands for the set of all the images of the results. T_r and T_{nr} are the set of relevant and non-relevant textual descriptions respectively; I_r and I_{nr} are the set of relevant and non-relevant images respectively. As the underlying theory of relevance feedbacks is to find a query q which maximizes similarity with relevant results and minimizes similarity with non-relevant documents[6], we wish to find a query q_{opt}:

$$q_{opt} = \arg\max_{q}[sim(q, T_r) - sim(q, T_{nr})] \quad (1)$$

$$q_{opt} = \arg\max_{q}[sim(q, I_r) - sim(q, I_{nr})] \quad (2)$$

Thus, we build two classifiers based on texts and images respectively to rank a new Web images. To make the rank scores comparable, we set the rank scores from the two classifiers between -1 and 1, where positive or negative value means the learner judged the concerned image to be relevant or non-relevant, and the bigger the absolute value of the rank shows the stronger the confidence of its judgement[12]. Here, in order to avoid a complicated learning process, we use a simple model proposed by Zhou et al[12] in Equation 3 and 4.

$$L_t(x, T_r, T_{nr}) = (\sum_{y \in Tr} \frac{Sim_t(x, y)}{|T_r| + \varepsilon} - \sum_{z \in Tnr} \frac{Sim_t(x, z)}{|T_{nr}| + \varepsilon})/Z_{norm1} \quad (3)$$

$$L_i(x, I_r, I_{nr}) = (\sum_{y \in Ir} \frac{Sim_i(x, y)}{|I_r| + \varepsilon} - \sum_{z \in Inr} \frac{Sim_i(x, z)}{|I_{nr}| + \varepsilon})/Z_{norm2} \quad (4)$$

where L_i and L_t are the visual view learner and textual view learner respectively, x is the visual feature vector or textual feature vector of a Web image to be classified, Z_{norm1} and Z_{norm2} are used to normalize the result into the interval of (-1,1), ε is used to avoid the zero denominator, and Sim_i

and Sim_t are the similarity metric adopted by L_i and L_t respectively.

The similarity Sim_t between two textual feature vectors $t_1 = \{x_1, x_2, x_3...x_n\}$ and $t_2 = \{y_1, y_2, y_3...y_n\}$ is measured by the cosine metrics as shown in the following equation:

$$Sim_t(t_1, t_2) = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2}\sqrt{\sum_{k=1}^{n} y_k^2}} \quad (5)$$

The similarity Sim_i between two visual feature vectors $i_1 = \{x_1, x_2, x_3...x_n\}$ and $i_2 = \{y_1, y_2, y_3...y_n\}$ is measured by the Eucilidean distance.

$$Sim_i(i_1, i_2) = \frac{1}{(\sum_{k=1}^{n} (x_k - y_k)^2)^{\frac{1}{2}} + \xi} \quad (6)$$

where the ξ is used to avoid the zero denominator.

B. Learning from Unlabeled Data

Algorithm 1 Generalize(T_n, T_u, k)

Input:

T_n:A set of text instances whose elements are to be generalized
T_u:A set of unlabeled text instances in which neighbors are to be identified
k: Number of neighbors used in generalizing

Output: T^*:The generalized data set

$T^* \leftarrow \emptyset$
for $x \in T_n$ **do**
  $T' \leftarrow k$-nearest neighbor for x from T_u
  $T' = T' \cup x$
  $x' \leftarrow Ave(T')$
  $T^* \leftarrow T^* \cup x'$
end for

Algorithm 2 Generalize(I_n, I_u, k)

Input:

I_n:A set of image instances whose elements are to be generalized
I_u:A set of unlabeled image instances in which neighbors are to be identified
k: Number of neighbors used in generalizing

Output: I^*:The generalized data set

$I^* \leftarrow \emptyset$
for $x \in I_n$ **do**
  $I' \leftarrow k$-nearest neighbor for x from I_u.
  $I' = I' \cup x$
  $x' \leftarrow Ave(I')$
  $I^* \leftarrow I^* \cup x'$
end for

Algorithm 3 CTRF$(query, DB, L_i, L_t, k, n)$

Input:

$query$:User query
DB:Web image database
L_i: image view learner
L_t: text view learner
k:Number of neighbors used in generalizing negative examples
n:Number of Web images used in enlarging co-training examples

Output: $Result$:The Ranked Result for each Web Images

$T_r \leftarrow \emptyset; T_{nr} \leftarrow \emptyset; I_r \leftarrow \emptyset; I_{nr} \leftarrow \emptyset; U \leftarrow DB;$
while In each round of relevance feedback **do**
  Get feedbacks
  $T_r \leftarrow T_r \cup$ text descriptions of the relevant Web images
  $T_{nr} \leftarrow T_{nr} \cup$ text descriptions of the non-relevant Web images
  $I_r \leftarrow I_r \cup$ the relevant Web images
  $I_{nr} \leftarrow I_{nr} \cup$ the non-relevant Web images
  $U \leftarrow U -$Web images labeled by the user
  Train the L_t based on T_r and $Generalize_t(T_{nr}, U, k)$
  Train the L_i based on I_r and $Generalize_i(I_{nr}, U, k)$
  $Rank_t \leftarrow$ rank the Web images in U by L_t
  $Rank_i \leftarrow$ rank the Web images in U by L_i
  $T^*_{nr} \leftarrow T_{nr} \cup$ the most n non-relevant text descriptions in $Rank_t$
  $I^*_{nr} \leftarrow I_{nr} \cup$ the most n non-relevant images in $Rank_i$
  Retrain the L_t based on T_r and $Generalize_t(T^*_{nr}, U, k)$
  Retrain the L_i based on I_r and $Generalize_i(I^*_{nr}, U, k)$
  R_t rank Web images in DB by L_t
  R_i rank Web images in DB by L_i
  $Result \leftarrow Combine(R_t, R_i)$
end while

As lack of training examples for training the two learners, they are not strong to sign correct labels to the unlabeled examples. In this case, for improving the reliability, we follow Zhou et al[12]'s way to co-train the two classifier, that is, in each round of relevance feedback each learner only labels for the other learner its n most confident negative examples. This strategy is reasonable as usually most of the Web images on the Web are non-relevant for each particular query.

As mentioned in Zhou et al[12], in image retrieval the positive examples can be regarded as belonging to the same relevant class, whereas the negative examples may belong to different irrelevant classes. They considered that each negative example can be generalized by the neighboring examples. In our CTRF, for each negative textual example and visual example, we find the k-nearest neighboring unlabeled examples by the Sim_t and Sim_i metric given in the Equation 5 and 6 respectively. For both textual and visual examples, the $k+1$ examples are then averaged to derive a virtual example which is used instead of the original negative example. The pseudo-

code of the two generalize algorithms is shown in Algorithm 1 and Algorithm 2.

C. Linear Combination of the Two Ranks

In our paper, the merged rank from the L_i and L_t is calculated by the following Equation:

$$Combine(x) = \alpha L_i(x) + (1-\alpha)L_t(x) \tag{7}$$

where α is weight parameter to control the influence of the L_i and L_t.

In summary, the pseudo-code of the CTRF is presented in Algorithm 3.

IV. Experimental Results

The series of experiments are designed to compare the performance of our CTRF algorithms to the other traditional relevance feedback algorithms.

A. Experiment Set

We created a set of 10 queries including *Music, Apple, Cat, Art, Autumn, China, France, Jaguar, Halloween* and *Bush Dog*. This set contains ambiguous queries like *Jaguar* and *Apple* as well as high-level complicated queries such as *China* and *Music*. Then, we take the top 200 results together with their ranking from Google Image Retrieval as baseline for each query. Moreover, the image and the Web page of each result returned from Google are also collected as the image and text modalities. For text modality, we extract tf.idf feature to represent the text content of each web page. On the hand, color, shape and texture features are engaged to represent the images. In detail, 9-dimensional HSV moment[13] is employed to get color feature; 7-dimensional Hu moment[13] is selected to obtain the shape feature; a 16 orientations and 3 scales Gabor filter[14]s is engaged to get the 48-dimensional texture feature.

In the CTRF algorithm, the parameter n is set to be 5, that is, in each round of relevance feedback each learner get 5 most confident negative examples from another leaner. The k is set to be 10, which means we take 10-nearest neighboring unlabeled examples to generalize the original negative example. As we mentioned that the parameter α controls which classifier we trust more, we use the former five queries in the query set to determine the parameter α. We tuned α from 0 to 1 , and check the average performance of our CTRF algorithm on these 5 queries. Finally, the parameter α is set to 0.9.

Other than using Google Image Search as baseline, we introduced another two relevance feedback methods for comparison, namely text relevance feedback and image relevance feedback. In each round of relevance feedback, the text relevance feedback only uses the textual modality in users feedbacks, whereas the image relevance feedback uses the image modality.

B. Results and Analysis

The later five queries are used to conduct the experiment, including *China, France, Jaguar, Halloween* and *Bush Dog*. For each query, the precision of the compared techniques are evaluated, which is the fraction of the number of retrieved related images to that of all retrieved images. The relevance of the images to the give query is examined by human volunteer. Since few users would be patient enough to browse more than top 50 images[15], we only compare the precisions on the top 10, 20, 30, 40, and 50 retrieved Web images. Figure 1 to Figure 5 depicts the average performance of the compared techniques in five relevance feedback rounds. It can be found that in first three relevance feedback rounds, the precision of CTRF method is apparently higher than that of other method, which indicates the helpfulness of the exploitation of both the text and image information. To speak of, we found that the average precision of text relevance feedback is always lower than the baseline in each feedback round. We assume the reason for this phenomenon is that the classifier trained by text feedbacks is too weak to rank the images correctly. The parameter α which is tuned to 0.9 could also prove this assumption.

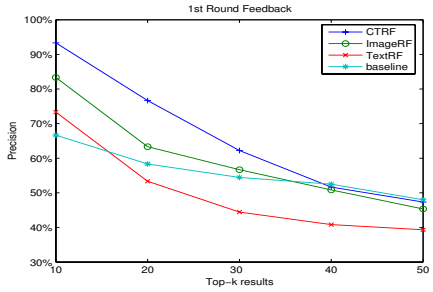

Fig. 1. First round relevance feedback

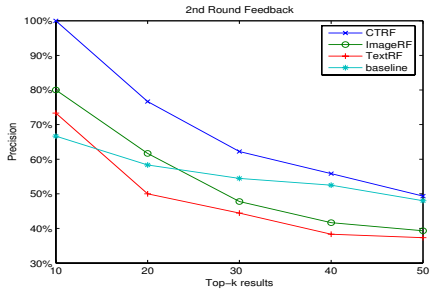

Fig. 2. Second round relevance feedback

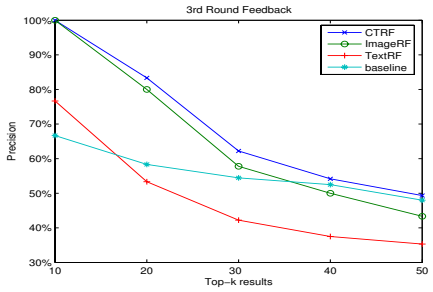

Fig. 3. Third round relevance feedback

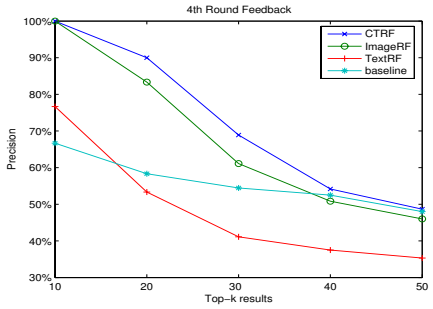

Fig. 4. Fourth round relevance feedback

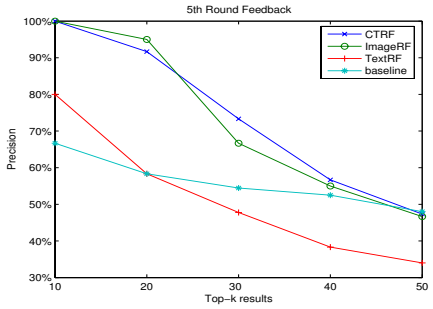

Fig. 5. Fifth round relevance feedback

V. Conclusion

In this paper, we described a method to use semi-supervised leaning to exploit both text information and unlabeled data to enhance the performance of relevance feedback in Web image retrieval. In particular, we exploit both text and image modalities information in the Web images and utilize co-training framework to learn from unlabel data in each relevance feedback round. Experiments on real-world data show that our proposed method has better performance that other traditional relevance feedback method.

One of the future research directions of this method is to explore how to set the weight parameter α which controls the influence of the two classifier for final ranking. Currently, the parameter is tuned to a fix number while we think the influences of the two classifier should be changed when users submit different queries.

Acknowledgment

This work is supported by the the Natural Science Foundation of China under grant No. 60603094, the Major State asic Research Project of China under grant No. 2007CB311103 and the National High Technology Research and Development Program of China under grant No. 2006AA010105.

References

[1] Rocchio, J. J. 1971. Relevance feedback in information retrieval. In Gerard Salton (ed.), *The SMART Retrieval System C Experiments in Automatic Document Processing,* pp. 313C323. Englewood Cliffs, NJ: Prentice-Hall. 177, 188, 301, 510

[2] Yong Rui, Thomas S. Huang, Sharad Mehrotra, and Michael Ortega, A relevance feedback architecture in content-based multimedia information retrieval systems in *Proc of IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with IEEE CVPR '97*

[3] X. Zhou and T.S. Huang, Relevance Feedback for ImageRetrieval: A Comprehensive Review, *ACM Multimedia Systems J.*, vol. 8, no. 6, pp. 536-544, 2003.

[4] J. Peng, MultiClass Relevance Feedback Content-Based Image Retrieval, *Computer Vision and Image Understanding*, vol. 90, no. 1,pp. 42-67, 2003

[5] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*. Madison, WI, 1998, 92-100.

[6] CD Manning, P Raghavan, H Schtze, Introduction to information retrieval, *Cambridge University Press* New York, NY, USA

[7] Xiaojin Zhu, Semi-Supervised Learning Literature Survey, Tech. Rep. Department of Computer Science, Madison, Wisconsin.

[8] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 1977, 39(1):1-38.

[9] Haffari, G. and Sarkar. Analysis of semi-supervised learning with the Yarowsky algorithm. *23rd Conference on Uncertainty in Artificial Intelligence*, 2007

[10] Zhang, T. and Oles, F. J. A probability analysis on the value of unlabeled data for classification problems. *Proc. 17th International Conf. on Machine Learning* (pp. 1191C1198). Morgan Kaufmann, San Francisco, CA. 2000.

[11] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In *Proceedings of the 17th International Conference on Machine Learning.* San Francisco, CA, 2000, 327-334.

[12] Zhou, Z.-H., Chen, K.-J., and Dai, H.-B. Enhancing relevance feedback inimage retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24, 219C244.

[13] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349?1380, 2000.

[14] Manjunath, B.S.; Ma, W.Y., Texture features for browsing and retrieval of image data, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.18, no.8, pp.837-842, Aug 1996

[15] Z.-H. Zhou and H.-B. Dai. Exploiting image contents in web search. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence* , pages 2928-2933, Hydrabad, India, 2007.