A Novel Active Learning Approach for SVM Based Web Image Retrieval

Jin Yuan, Xiangdong Zhou*, Hongtao Xu, Mei Wang, and Wei Wang

Department of Computing and Information Technology
Fudan University, Shanghai, China
{052021159,xdzhou,061021054,051021052,wwang1}@fudan.edu.cn

Abstract. There is a great deal of research conducted on hyperplane based query such as Support Vector Machine (SVM) in Content-based Image Retrieval(CBIR). However, the SVM-based CBIR always suffers from the problem of the imbalance of image data. Specifically, the number of negative samples (irrelevant images) is far more than that of the positive ones. To deal with this problem, we propose a new active learning approach to enhance the positive sample set in SVM-based Web image retrieval. In our method, instead of using complex parsing methods to analyze Web pages, two kinds of "lightweight" image features: the URL of the Web image and its visual features, which can be easily obtained, are applied to estimate the probability of the image being a potential positive sample. The experiments conducted on a test data set with more than 10,000 images from about 50 different Web sites demonstrate that compared with traditional methods, our approach improves the retrieval performance significantly.

1 Introduction

Statistical learning methods based on the optimal classification hyperplane, such as Support Vector Machine, are promising tools for learning complex, subjective query concepts in CBIR. It is widely understood that the major bottleneck of CBIR is the "semantic gap" between low-level visual feature representations and high-level semantic concepts of images. To reduce the gap, relevance feedback(RF) [1] was brought into CBIR to improve the retrieval performance [11,12].

Relevance feedback is a human-computer interaction process, in which users evaluate the search results. According to the evaluation, the retrieval system adjusts the classification function and returns the improved results. This process is repeated until users satisfy or abandon the search.

Traditional relevance feedback is a kind of passive learning, where the feedback samples are acquired randomly. However, randomly chosen samples often contain rare information which will increase the iterative number of relevance

* This work was partially supported by the NSF of China Grant 60403018, 973-Plan Grant 2005CB321905. Corresponding author.

H.H.S. Ip et al. (Eds.): PCM 2007, LNCS 4810, pp. 40–49, 2007.

feedback. Active learning [3,4,5] aims at actively choosing the most informative images in each feedback iteration. For example, a typical active learning method—Angle-Diversity[3], actively selects a candidate sample set which is close to the classification hyperplane as well as maintains the sample's diversity. In active learning, the performance of CBIR is suffered from the following two problems[3]: 1.*The imbalance of the positive and negative samples.* In real applications, for a certain query concept, most of the images of the database are irrelevant. Angle-diversity selects samples near hyperplane (the most uncertain) for users' feedback. However, due to the data imbalance, the probability of those samples being negative ones is higher than that as positive ones, which makes the feedback sample set containing inadequate positive samples. 2. *Visual diversity.* Due to the well-known "semantic gap", the images relevant to a certain query concept usually distribute loosely in a large scale feature space. It is often difficult to find sufficient positive sample images close to the hyperplane.

Compared to the standard image data sets, Web images are more diverse in the feature and concept space. Thus the traditional relevance feedback strategy often fails to provide sufficient positive samples. The context of Web images such as surrounding text, the Web page URLs, etc. have been exploited to improve Web images retrieval in many previous work[13,14]. However, most of the known work need to deal with the problem of "intelligently"parsing Web pages. To improve the performance of relevance feedback in Web images retrieval, we propose a new active learning approach using "lightweight"Web image features: the visual feature and image's URL, which can be obtained easily. The heuristic idea is: images with similar URLs and visual features have higher probability of similar semantics. For example, we have a positive image about concept "car" which is located in the Web set "www.acar.com.cn". Due to the "semantic gap", searching this image in the whole database will bring in a lot of negative samples. However, if we know there are some images in the result set having similar URLs with this positive sample, for instance, having the same domain name and directory path, or a similar domain name such as "www.car.com", we can "guess"such images may have similar semantic with the positive sample. Distance measure can be used to evaluate the similarity of the URLs between images. Therefore, visual features of the image sample, URL and other information can be exploited to estimate the possibility of sample as positive one. Images with the highest probability will be provided as feedback samples for the users, which compensates for the shortage of positive samples in traditional feedback strategy. In addition, in light of the returned results of SVM[2], we also propose a new ranking method. The framework of our Web image retrieval system is given in Fig.1.

2 Related Work

SVM based image retrieval and relevance feedback had been studied for decades. Chen and Huang propose one-class SVM[6] to solving the imbalance of positive and negative samples. They construct a tight hyper-sphere in the feature space to include most positive training samples. However, for some widely distributed

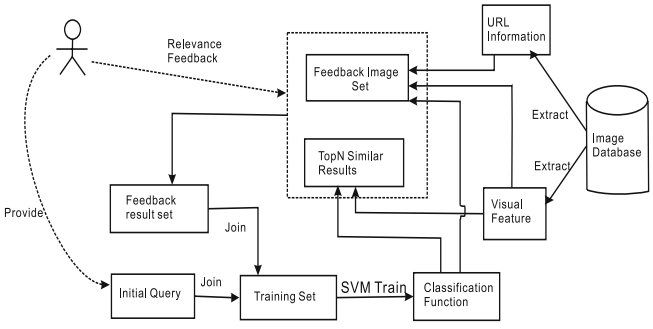

Fig. 1. The Web Image Retrieval System Framework

"concept ", how to select a proper Kernel to map those positive samples in a tight hyper-sphere is still an unsolved problem.

Gosselin and Cord [7] proposed a method to shift the hyperplane dynamically according to the ratio of the number of positive samples to negative samples learned in the last-round feedback. Due to the diversity of images in the visual space, this method results in marginal improvement especially for Web image data. Goh and Chang[9] proposed a new active learning strategy for solving the problem of scarce positive samples in feedback set.

There are a lot of work dealing with content-based Web images retrieval. Cai et al.[8] proposed a retrieval method combining visual feature, text and Web page links. It was shown that their method can improve the performance. Their Web pages preprocessing needs some "intelligent" parsing tools and suffers from parameter tuning. Quack et al.[11] proposed a Web image retrieval system *Cortina*, which combines visual feature and keywords to find relevant images. The correlations between image and keywords are estimated by data mining approach.

3 Positive Enhanced Feedback Strategy

Angle-diversity algorithm is the most commonly used active learning approach, which selects a candidate sample set close to the classification hyperplane as well as maintain the samples' diversity. It is believed that samples close to hyperplane are the most uncertain ones[3], in other words, the possibility of those samples being positive ones is near 0.5. However, compared with standard image data sets, Web image data set has more kinds of categories which means the imbalance of image data is more serious and the isolation[9] is worse. When applying angle-diversity to Web image retrieval, the improvement of retrieval performance is not significant due to lacking positive samples in feedback result set.

To deal with this problem, we propose an active learning approach based on angle-diversity by adding some potential possible positive samples to the feedback set returned to the users.

3.1 Finding High Possible Positive Sample(HPPS)

In this section, we will first give some definitions. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the set of Web images, $\mid X \mid = n$, where n is the number of instances, $Y = \{y_1, y_2, \ldots, y_n\}$ denote the set of the corresponding labels of X. y_i represents the class label of instance x_i, namely, $y_i = +1$ means x_i being a positive sample, otherwise, $y_i = -1$. $W = \{w_1, w_2, \ldots, w_m\}$ denotes the set of Web sites, $\mid W \mid = m$, where m is the number of Web sites. We assume each $x_i \in X$ is corresponded to only one element of W, and p_k denotes the number of instances in w_k. $P(x_i, +1)$ corresponds to the joint probability of x_i being positive sample. The aim of HPPS is to find the sample x_i with higher $P(x_i, +1)$ and less $P(x_i, -1)$. Let $\hat{y}_i$ denote the estimated class label of x_i, the probability $P(x_i, \hat{y}_i)$ can be given as follows:

$$\begin{aligned} p(x_i, \hat{y}_i) &= \sum_{k=1}^{m} P(\hat{y}_i | x_i, w_k) \times P(w_k | x_i) \times P(x_i) \\ &= \sum_{k=1}^{m} P(\hat{y}_i | w_k) \times P(w_k | x_i) \times P(x_i) \\ &= \sum_{k=1}^{m} P(\hat{y}_i | w_k) \times P(x_i | w_k) \times P(w_k) \\ &= \sum_{k=1}^{m} (\sum_{j=1}^{p_k} P(\hat{y}_i | x_j) \times P(x_j | w_k)) \times P(x_i | w_k) \times P(w_k) \\ &= \sum_{k=1}^{m} (\sum_{j=1}^{p_k} P(x_i, \hat{y}_i | x_j, \hat{y}_j) \times P(x_i | x_j) \times P(\hat{y}_j | x_j) \times P(x_j | w_k)) \\ &\quad \times P(x_i | w_k) \times P(w_k) \end{aligned} \tag{1}$$

The followings are the details of the computation of the above formula:

1. The probability $P(w_k)$ is related to two factors: the first is the ratio of the number of samples in Web site w_k to the total sample number n, the second is the confidence of w_k. We define the confidence W_k denoting the ability of the classification of the Web site w_k, and will discuss it in the next section. $P(w_k)$ can be calculated as follow:

$$P(w_k) = W_k \times \frac{\mid w_k \mid}{n} \tag{2}$$

2. For $P(x_i | x_j)$, sample x is drawn from a probability distribution, thus, for a new sample x_i, a non-parametric Gaussian Kernel density estimation is used to estimate the probability density $P(x_i | x_j)$, which is given by:

$$P(x_i | x_j) = (2\pi)^{-d/2} \sigma^{-d} exp^{-\frac{1}{2\sigma^2} \| x_i - x_j \|^2}, \tag{3}$$

where σ is the standard deviation, σ and d are the same for each x_i, so we ignore them in the following calculation.

3. We employ an approximate method to calculate the probability of x_j with estimated label $\hat{y}_j$, $P(\hat{y}_j|x_j)$. Given an training instance x_j, if $\hat{y}_j$ equals to the label y_j of x_j, $P(\hat{y}_j|x_j) = 1$, otherwise $P(\hat{y}_j|x_j) = 0$. In the following calculation we simply only use those instances in the training set with y_j equaling to $\hat{y}_i$ to calculate $p(x_i, \hat{y}_i)$. In this case, $P(x_j|w_k)$ corresponds to the reciprocal of the number of training data x_j with y_j equaling to $\hat{y}_i$ in w_k, which is $P(x_j|w_k) = \frac{1}{|\{x_j \mid x_j \in trainingset, x_j \in w_k, y_j = \hat{y}_i\}|}$.

4. We employ the factor α_{ij} to represent the value of $P(x_i, \hat{y}_i|x_j, \hat{y}_j)(\hat{y}_j = \hat{y}_i)$. α_{ij} can be calculated by the URLs' similarity between the instance x_i and x_j. In fact, the similarity between instances' URLs is used to measure the probability that these two instances have the same label.

5. The probability $P(x_i|w_k)$ is calculated as follows:

IF $x_i \in w_k$ $\quad P(x_i|w_k) = 1/ \mid w_k \mid$

ELSE $\quad P(x_i|w_k) = 1/ \mid n \mid$

Therefore, Eqn.1 can be rewritten as follows:

$$P(x_i, \hat{y}_i) = \sum_{k=1}^{m} (W_k \frac{\mid w_k \mid}{n} P(x_i \mid w_k) \frac{1}{p_k} \sum_{j=1, x_j \in w_k, y_j = \hat{y}_i}^{p_k} \alpha_{ij} e^{-\|x_i - x_j\|^2}) \quad (4)$$

where

$$P(x_i|w_k) = \begin{cases} \frac{1}{|w_k|}, & x_i \in w_k \\ \frac{1}{|n|}, & x_i \notin w_k \end{cases}$$

For each x_i, $f(x_i) = P(x_i, +1) - P(x_i, -1)$ is applied to estimate the probability of x_i being positive sample. Eqn.4 indicates that for a given instance x_i and a positive training data x_j, the possibility of x_i being positive sample depend on the following two factors: 1.The distance between x_i and x_j, the smaller the distance, the more possibility; 2. The URL similarity between x_i and x_j, higher similarity(leading to a large α_{ij}, higher possibility. Finally, $f(x_i)$ implies that the instance which is near positive sample set and far away from negative sample set is more possible to be a positive instance. At next section, we will discuss how to calculate α_{ij} and W_k in Eqn.4.

3.2 Web Site Weight Calculation Based on URL Similar Sample(USS)

Our method calculates W_k based on USS and users' relevance feedback. The system returns some instances which have the same Web site's URL and small distance to a positive training sample(USS), then users are prompt to label these instances. Web site weight W_k is calculated based on the feedback results. Let r_k denote the number of historic positive feedback about Web site w_k, where c_k is the number of historic negative feedback about Web site w_k. If user labels one of USS sample in w_k positive, $r_k = r_k + 1$. Otherwise, $c_k = c_k + 1$. Thus,

Algorithm 1. URL Similar Sample(USS)

Input: n,r_i,c_i {the number of feedback sample provided by USS,the number of historic positive feedback in w_i,the number of historic negative feedback in w_i}
Output: W_i,r_i,c_i{weight of Web site i}
Initialization: $w_i = r_i/(r_i + c_i)$
BEGIN
$x_p \longleftarrow Random(S_{positivetrainset})$**{**randomly select a sample from positive training set**}**
$U \longleftarrow SomeURL(x_p, S - S_{trainset})$**{**select all instances with same URL with x_p from non-training set**}**
$R \longleftarrow min(U, x_p, n)$**{**select n instances from U with smallest distance with x_p**}**
for $i = 0; i < n; i++$ **do**
  if R_i is positive by users'feedback **then**
    r_i++;
  else
    c_i++
  end if
end for
$W_i = r_i/(r_i + c_i)$
return and save W_i,r_i,c_i
END

we have the formula of Web site weight, $W_k = \frac{r_k}{(r_k+c_k)}$. The process is listed in algorithm 1:

W_i reflects the confidence weight of Web site w_i, the large value of W_i implies that Web site w_i has strong classification ability. In the practical application, some kind of Web set with great classification ability should be assigned a high confidence weight. For example, the instances in Web set "www.acar.com.cn" are almost cars, the confidence weight of this Web site is high. On the contrary, images in "www.image.taobao.com" almost belong to different categories, so this kind of Web sites should be assigned lower weight.

3.3 α_{ij} Calculation

For some Web images, their URLs can be divided into multiply layers. For example, the image "http://sina/data/cartoon/1328.jpg" has 3 layers. The image file name "1328.jpg" is not considered because it is unique in the same sub-directory. If the URLs between two images are more similar, they are more likely to be related with the same concept. We propose a method to calculate α_{ij} as follows.

We assume that $\alpha_{ij} = l_a (0 < l_a < 1)$, when x_i and x_j have not common part in their m-layer URLs, where x_j is a training instance. If the URLs of x_i and x_j are the same from the root layer to the k^{th}-layer, and from $(k+1)^{th}$-layer they are different, then we calculate α_{ij} according to Eqn.5:

$$\alpha_{ij} = l_a + (1 - l_a) * \frac{\sum_{n=0}^{k} n}{\sum_{n=0}^{m} n} \tag{5}$$

where m is the number of directory layers. We can see that when the URL of x_i's and x_j's are completely different, α_{ij} is l_a. The value l_a has important influence on the results. If l_a is too large, α_{ij} will have little effect on $f(x_i)$, thus it is not confident to find positive samples in the same URL. Otherwise, if l_a is too small, the effect on $f(x_i)$ is too great, then the samples are likely to come from the same URL. To deal with it, we propose a method based on statistics to adjust l_a dynamically. The basic idea is to record the number of images provided in HPPS. If the number of instances coming from the same URL excesses a given threshold, l_a increases; otherwise l_a decreases. The algorithm 2 is given in the followings.

Algorithm 2. l_a adjusting algorithm

Input: a,l,h {the initial l_a,the low boundary of the ratio of same URL instances number to total instance number,the high boundary of the ratio of same URL instances number to total instance number}
Output: l_a
Procedures:
maxurlnum(S_{HPPS}){the maximal instance number coming from same URL provided by HPPS in a query}
total(S_{HPPS}){the total number provided by HPPS in a query}
BEGIN
if l_a not exist **then**
  l_a=a
end if
if maxurlnum(S_{HPPS}) > $total(S_{HPPS})$*h **then**
  $l_a = 1 - (1 - (\frac{maxurlnum(S_{HPPS})}{total(S_{HPPS})} - \frac{l+h}{2})) * (1 - l_a)$
end if
if maxurlnum(S_{HPPS}) < $total(S_{HPPS})$*l **then**
  $l_a = (1 - (\frac{l+h}{2} - \frac{maxurlnum(S_{HPPS})}{total(S_{HPPS})})) * l_a$
end if
return l_a
END

4 The New Ranking Method for SVM

According to the theory of SVM, the higher the value returned by SVM classification function, the more possibility the instance is positive. However, in the content-based Web image retrieval, the performance by ranking the score from high to low is not very significative. One reason is that the positive and negative samples are not sufficient in the previous rounds of feedback, so the classification

hyperplane is not good enough to describe the characteristics of such "concept". In this case, many negative samples will obtain high SVM scores which would exclude the positive samples from the result set. Second, in content-based image retrieval, an image often contains hundreds of dimensional features, but the user can not label hundreds of instances, so the number of training data is far less than the dimension number. Thus the number of non-support vector in the training result is less than that of support vector, which is especially true for positive support vector due to the imbalance of image data. In our experiments, we found that the number of positive support vector is far more than that of positive non-support vector, which means that it is more reasonable to obtain positive samples in the local area of the positive support vector.

According to the above discussion, this paper presents a re-ranking method for the results returned by SVM. The details are given as follows:

First, we choose the samples whose SVM score value $f(x)$ is between [1-β,1+β]($\beta > 0$) and rank them according to $\mid f(x)-1 \mid$ in ascending order. These samples have higher possibility to be positive support vector. Second, we choose samples with $f(x) > 1+\beta$, and rank them according to $f(x)$ in increasing order. Finally, we select the sample with $f(x) < 1-\beta$, and sort them according to $f(x)$ in descending order. In the experiments, we set $\beta = 0.01$.

5 Experimental Results

Our experiment data set is obtained from 57 Web sites. It contains more than 11,000 images. Each image is characterized as a 528-dimension feature vector based on the color and texture features according to MPEG7. We select 23 categories as initial query categories. Because the instances are labeled manually, which is label-intensive, only 47 image queries are provided. Each initial query contains two instance, one positive and one negative. The positive instance is obtained from Google query results which is not contained in image database, while the negative one is selected by the system randomly. Six rounds feedback for each query is performed and 100 images will be returned by the system at each round.

Three experiments are conducted: the first experiment is to compare our feedback method with the traditional method(NFM) and the Angle-diversity. For the traditional method, in each feedback session, five positive images and five negative images on the top100 results are selected and returned to the retrieval system. If the number of positive images is less than 5, we increase the number of negative images until the total feedback number becomes 10. Our method is based on Angle-Diversity and HPPS+USS feedback model. The initial value of l_a is set to 0.6. l and h is set to 0.4, 0.6 respectively. The top 90 result images and 10 images obtained from Angle-Diversity+HPPS+USS are provided for users' feedback, where Angle-Diversity will provide three images, HPPS provides five, and USS provides two, respectively. Finally, five negative and five positive images are selected and return to retrieval system. In the second experiment, we compare the performance of Angle-Diversity+HPPS+USS and NFM in two different SVM

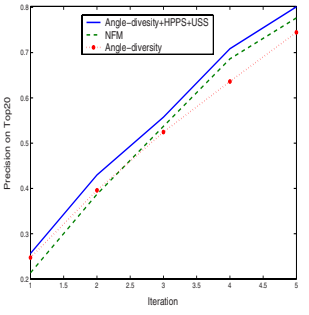

Fig. 2. Precision of three feedback strategies Top20

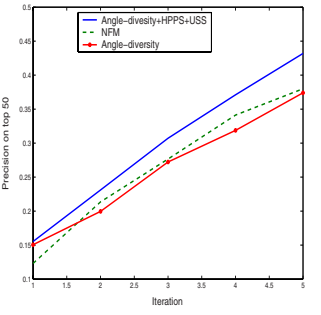

Fig. 3. Precision of three feedback strategies on Top50

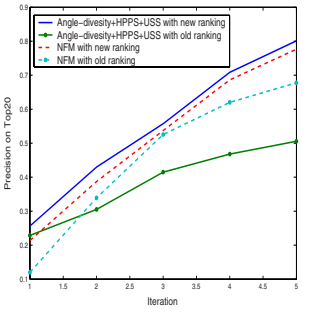

Fig. 4. Precision of Angle-diversity-+HPPS+USS and NFM with different ranking methods on Top20

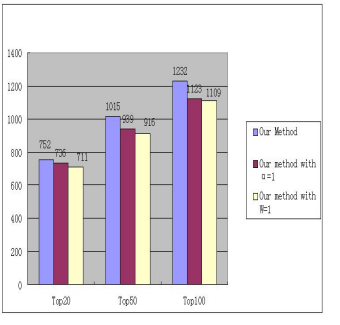

Fig. 5. The Comparison on Our Feedback Strategy with Parameters α and W

ranking methods, feedback methods are the same as the previous experiment. The third experiment evaluates the contributions of α and w to the retrieval performance based on the relevance feedback of Angle-diveristy+HPPS+USS. We compare the retrieval performance of using parameter α, W and setting these parameters as constants, $\alpha = 1$ or $W = 1$. In each of the three experiments, 47 images as provided as initial queries, 6 rounds of relevance feedback are performed. The experimental results are shown in the following figures.

From Fig.2, Fig.3, we can observe that compared with NFM and angle-diversity, our new feedback method with new SVM ranking approach can improve the precision in both top 20 and top 50 search results.

Fig.4 shows that the new ranking method in both of our new feedback method and the NFM can improve the retrieval performance significantly compared with the theory SVM ranking method.

Fig.5 shows that when we set the parameter α and W being constant, the retrieval precision is decreased, which proves that the introducing of the parameter α and W can improve the retrieval performance.

6 Conclusion

This paper presents a new active learning method and a new ranking approach for SVM based image retrieval to deal with the imbalance problem of the Web image retrieval. The experimental results show that the proposed methods improve the image retrieval performance significantly compared with the traditional methods. In the future work, the theory of our new ranking approach for SVM based Web image retrieval will be studied further.

References

1. Rui, Y., Huang, T., Ortega, M., Mehrotra, S.: Relevance feedback:A power tool in interactive content-based image retrieval. IEEE Tran. on Circuits and Systems for Video Technology 8(5) (1998)
2. Burges, C.: A Tutorial On Support Vector Machines For Pattern Recognition. Data mining and Knowledge Discovery (1998)
3. Chang, E.Y., Lai, W.-C.: Active Learning and its Scalability for Image Retrieval. In: IEEE ICME (2004)
4. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: ICML (2003)
5. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: ACM MM 2001 (2001)
6. Chen, Y., Zhou, X., Huang, T.: One-class SVM For Learning In Image Retrieval. In: IEEE ICIP 2001, Thessaloniki, Greece (2001)
7. Gosselin, P.H., Cord, M.: Active Learning Techniques for User Interactive Systems: Application to Image Retrieval, Machine Learning Techniques for Processing Multimedia Content, Bonn, Germany (2005)
8. Cai, D., Xiaofei,: Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information. In: ACM MM 2004 (2004)
9. Goh, K.S., Chang, E., Lai, W.C.: Multimodal Concept-Dependeng Active Learning for Image Retrieval. In: ACM MM 2004 (2004)
10. Quack, T., Monich, U., Thiele, L., Manjunath, B.S.: Cortina: A System for Large-scale, Content-based Web Image Retrieval. In: ACM MM 2004 (2004)
11. Jing, F., Li, M., Zhang, H.J., Zhang, B.: Support Vector Machines for Region-Based Image Retrieval. In: IEEE ICME (2003)
12. Huang, T.S., Zhou, X.S.: Image retrieval by relevance feedback:from heuristic weight adjustment to optimal learning methods. In: IEEE ICIP (2001)
13. He, X., Ma, W.Y., Zhang, H.-J.: ImageSeer:Clustering and Searching WWW Images Using Link and Page Layout Analysis, Micsoft Technical Report (2004)
14. Hua, Z., Wang, X.J., Liu, Q.: Semantic knowledge Extraction and Annotation for Web Images. In: ACM MM 2005 (2005)