A Novel Framework for Web Pages Classification

Ruiguang Hu[1] and Weiming Hu[2]

Abstract. In this paper, we propose a novel framework for classifying web pages containing images and text. Valid images are first chosen by the FOrward CompArison of Relative Sizes Sorting(FOCARSS) algorithm, and each valid image is represented by the mid-level feature vector generated by the Bag-Of-Features model. Taking these feature vectors of valid images in a web page as instances of a bag, Multi-Instance Learning is utilized to conduct the image-based web pages classification. Regarding the text information, Bag-Of-Words model is used to conduct the text-based web pages classification. Subsequently, score-level fusion schemes are used to fuse these two kinds of heterogeneous information. Experimental results on a representative dataset demonstrate that our framework can definitely take full advantage of image and text information and improve final classification performances.

Keywords: Score fusion • Multi-instance learning • Bag-of-features

1 Introduction

With the rapid development of the Internet and the extensive use of intelligent devices, such as smart phones and cameras, massive amounts of images have been emerging on the web, consequently, the content of web pages are extremely multitudinous. Additionally, web page designers prefer to utilize images to express the theme of a page. In extreme cases, a web page contains lots of images while only a few or even no text, and this kind of web pages take a increasing proportion, which can be found easily.

Regarding web pages classification, lots of algorithms have been proposed and promising classification performances have been achieved[1-5]. Unfortunately, to

[1] R. Hu (✉)
NLPR, IACAS, 95 Zhongguancun East Road, 100190, BEIJING, CHINA
e-mail: rghu@nlpr.ia.ac.cn

[2] W. Hu
NLPR, IACAS, 95 Zhongguancun East Road, 100190, BEIJING, CHINA

the best of our knowledge, most of these algorithms only take advantage of text information, no matter their web pages datasets contain lots of images or not. We argue that these algorithms, to a large extent, don't accord with the reality of the Internet. To this end, we propose a general framework for classifying web pages, which can make heavy use of both image and text information to improve final classification performance.

In our framework, information preprocessing plays a very basic role. It's commonly accepted that not all images in a web page have influence on its theme, especially those decorative images, advertising images, and hyperlink images. Getting rid of these invalid images and save those valid images can definitely improve classification performance and reduce computation complexity, memory and time consumption. The FOrward CompArison of Relative Sizes Sorting(FOCARSS) algorithm we propose can fulfil this task effectively, robustly and fast. The essence of FOCARSS is that images of large sizes should be chosen preferentially, while images of similar sizes always be evaluated delicately to determine where they are valid or not. The text of a web page that we can see in a browser are extracted by regular expression.

After preprocessing, each web page has a bag of valid images, and the number of valid images varies from zero to a large value. When conducting image-based web pages classification, Multi-Instance Learning(MIL)[6], which can measure the similarities between bags effectively, is utilized. In the MIL of our framework, each image is seen as a instance, and is represented by the mid-level feature vector of the Bag-Of-Features(BOF) model. Regarding the similarity metric between bags, the effective Multi-Instance Kernel(MIK) proposed by *Garnter etc.*[7] is used, which can be computed easily and fast. Subsequently, Support Vector Machine(SVM) is used to conduct the classification. Regarding text-based web pages classification, the well-known Bag-Of-Words(BOW) model is applied, and the vocabulary of the BOW model is selected elaborately. Analogously, SVM is used again to conduct the classification.

After conducting classification separately, image and text information are fused at the score level. In our framework, weighted sum fusion rules are utilized.

Generally, our framework can be applied to binary and multi-class classification. In the experiments, we evaluate the framework on a dataset that contains drug-related web pages and common web pages. The results demonstrate that the MIL algorithm can achieve comparable classification performance with that achieved by the BOW model. Finally, score-level fusion can fuse image and text information effectively and achieve better classification performances.

The rest of the paper is organized as follows. Section 2 introduces the information preprocessing. Section 3 presents individual classification algorithms, including MIK, BOF and BOW. Section 4 proposes the score-level fusion schemes. Section 5 conducts the experiments, and Section 6 concludes the paper.

2 Information Preprocessing

According to the contents of web pages on the network, we can generally divide them into three main categories: text-dominated web pages, exhibition-style web pages, and images-dominated web pages[8].

There are lots of text in text-dominated web pages, and probably only a few or no illustration image is in them. Old-fashioned web pages and news web pages belong to this category. Additionally, there are lots of small decorative images, advertising images, and hyperlink images in news web pages, as shown in Fig. 1(a).

Exhibition-style web pages are mainly used to exhibit an object, a painting, or a commodity, etc. There is a main image and some explanation text, the size of the main image is often much larger than those of surrounding images, as shown in Fig. 1(b).

In images-dominated web pages, lots of images are presented compactly, and the sizes of them are almost equivalent. This kind of web pages mainly exist in shopping web sites and photo-sharing web sites, only very few words are in them, as shown in Fig. 1(c).

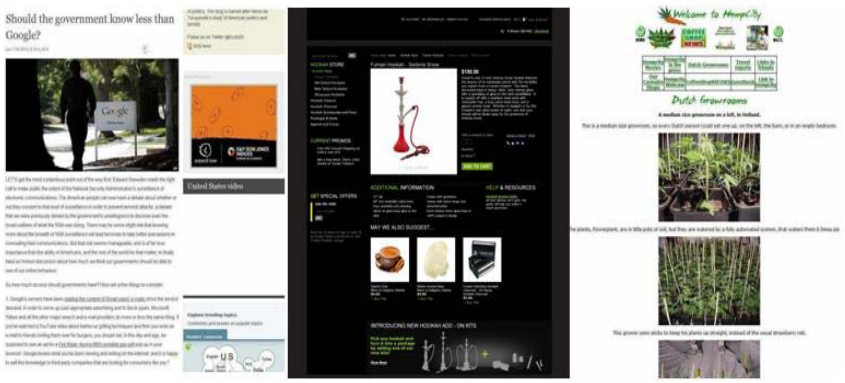

Fig.1 Three representative web pages. (a) The left one: Web page with large amount of text (b) The middle one: Web page with a dominant image (c) The right one: Web page with lots of images

In the light of deep analysis of images structure of web pages, we propose the FOrward CompArison of Relative Sizes Sorting(FOCARSS) algorithm to extract those valid images in a web page. The FOCARSS algorithm can pick out those meaningful images and get rid of those unwanted images as much as possible. Additionally, it can handle almost all kinds of web pages and is robust for the reason of using relative image sizes.

Multifarious web pages correspond to cluttered text in html files, consequently, regular expression, as a powerful text analysis tool, is utilized in our framework to extract those text that can be seen through a browser.

3 Individual Classification Algorithms

We first explain clearly the MIK, then present the BOW and BOF model together.

3.1 Multi-Instance Kernel

Actually, most of existing multi-instance learning algorithms are modifications of supervised learning algorithms by translating the objective of them from discriminating instances to discriminating bags. Multi-Instance Kernel(MIK)[7] is an excellent representative for its easy implementation and promising performance. Suppose there are bag $B_i = \{\vec{x}_{i1}, \cdots, \vec{x}_{in_i}\}$ and bag $B_j = \{\vec{x}_{j1}, \cdots, \vec{x}_{jn_j}\}$, MIK measures the similarity between bag B_i and B_j as

$$K_{MI}(B_i, B_j) = \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} K^p(\vec{x}_{ia}, \vec{x}_{jb}) \tag{3.1}$$

where $K(.,.)$ is some kind of conventional kernels defined on instances, such as linear kernel, polynomial kernel, RBF kernel etc. and p is a positive integer. To avoid numerical problems, MIK should be normalized. MIK can be applied seamlessly in SVM to conduct classification. In our framework, we set $K(.,.)$ of the MIK to the RBF kernel. Regarding RBF kernel, $K^p(.,.)$ is still a RBF kernel, consequently, p can be set to 1, and it's only needed to tune the parameters of RBF kernel to achieve the best classification performance.

In practical applications, it's well recognized that reasonable multi-instance representation is sort of more important than the selection of MIL algorithms. In our framework, the instances of a bag are the mid-level feature vectors of the BOF model.

3.2 BOW model and BOF model

Bag-Of-Words is a classical and well-known model for document representation. First, a vocabulary is selected by feature selection algorithms, then a document is

represented as a count vector of the words in the vocabulary. Finally discriminative or generative classifier is applied for classification.

Bag-Of-Features model is analogical to the BOW model. First, local features such as SIFTs are extracted, and a codebook is generated by K-means clustering or other algorithms; then all local features of an image are coded according to the codebook, and a mid-level feature vector, whose dimensionality is the same as that of the codebook, is generated. the mid-level feature vector is taken as representation of the image, i.e. a instance of a bag in our framework.

4 Score-level Fusion Schemes

Using the MIL algorithm and the BOW model, two SVMs are trained for image-based and text-based web pages classification respectively. Sigmoiding their outputs generates corresponding scores S_{im} and S_{txt}, which are then fused at the score level. Additionally, according to these scores, two ROC curves can be generated and corresponding Equal Error Rate(EER) values can be gotten, which will be used in weighted sum rules.

Generally, suppose we have K kinds of scores, then the final scores can be gotten by fusing them as follows:

$$S = \sum_{k=1}^{K} W_k S_k \tag{4.1}$$

where W_k is the weight for scores S_k. The weight can be determined in many ways such as:

$$W_k = \frac{\frac{1}{EER_k}}{\sum_{k=1}^{K} \frac{1}{EER_k}} \tag{4.2}$$

and

$$W_k = \frac{1 - 2EER_k}{\sum_{k=1}^{K} 1 - 2EER_k} \tag{4.3}$$

We refer to above-mentioned two schemes as SL-EER1 and SL-EER2. The essence of EER-weighted schemes is that, the better classification performance one classifier achieves(i.e. the lower EER value), the greater its weight is[9].

Another way to determine the weight is utilizing D-Prime[9,10]. Suppose μ_k^P and μ_k^N are the mean values of those scores of corresponding positive samples

and negative samples in the k -th classifier, σ_k^P and σ_k^N are corresponding standard deviations, then the D-Prime is

$$d_k = \frac{\mu_k^P - \mu_k^N}{\sqrt{(\sigma_k^P)^2 + (\sigma_k^N)^2}} \tag{4.4}$$

The D-Prime measures score separation and can be used to determine the fusion weight:

$$W_k = \frac{d_k}{\sum_{k=1}^{K} d_k} \tag{4.5}$$

and we refer to this scheme as SL-DP.

5 Experimental Results

We apply the novel framework to classify drug-related web pages, which contain lots of drug-taking instruments and cannabis images, as shown in Fig.1(b) and Fig.1(c). We collected 2144 drug-related web pages from about one hundred related web sites as positive samples, and 2283 normal web pages as negative samples, which cover topics of news, shopping, photo-sharing, etc. Halves of both positive and negative samples were used for training, and the reminding halves were used for testing.

A codebook of 1500 codes was generated by K-means clustering all dense-sampled RGBSIFTs[11] of valid images of training web pages, and the mid-level feature vector for each valid image was generated by hard coding and sum pooling in the BOF model. Regarding text information, 100 discriminative words were selected, and then used to form the vocabulary, which was used to generate the words count vector for each web page.

Taking those mid-level feature vectors as instances, Multi-Instance Learning was used to conduct the image-based web pages classification; According to the words count vectors, SVM[12] was used to conduct the text-based web pages classification. Subsequently, outputs of both classifiers were sigmoided as scores for score-level fusion. If there was no valid image in a web page, its score was set to 0.5, i.e. the probability of random guess.

In our experiments, ROC curves and EER values are used for presentation and explanation of experimental results. For paper space limitation, ROC curves of the image-based and text-based web pages classification are shown together with the ROC curves of weighted sum fusion schemes, as shown in Fig.2. The EER values of all ROC curves are listed in Table 1.

According to these experimental results, it's quite clear that, although text information can discriminate web pages satisfactorily, image information can achieve comparable classification performance as long as exploited appropriately. Additionally, Fusing these two heterogeneous information can improve classification performance by a large margin, which can be seen from Fig.2 and Table 1.

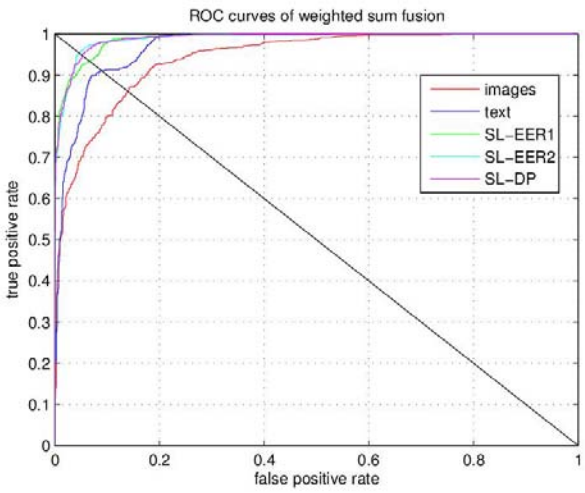

Fig.2 ROC curves of weighted sum fusion

Table 1 EER values of all classification schemes

Single source	EER(%)	Fusion scheme	EER(%)
images	14.1	SL-EER1	6.6
text	8.9	SL-EER2	4.8
		SL-DP	5.3

6 Conclusion

A novel framework for classifying web pages containing images and text has been proposed, which can exploit both image and text information thoroughly and take full advantage of them. In this framework, the effective FOCARSS algorithm is utilized to select those valid images in a web page, which are appropriately represented by BOF model. Multi-Instance Learning and BOW model are used individually to conduct respective classification, whose outputs are sigmoided and fused at the score level to achieve improved classification performance. Experimental results on a representative dataset demonstrate the effectiveness, flexibility

and robustness of the framework. In the future, video and audio information will be integrated into this framework, to accord with those more and more abundant contents of web pages.

Acknowledgments This work is partly supported by NSFC (Grant No. 60935002 of Weiming Hu and No. 61103056 of Haiqiang Zuo), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Fu, Z. M. (2008). A Comparison Study: Web Pages Categorization with Bayesian Classifiers, *10th IEEE International Conference on High Performance Computing and Communications*, Page(s): 789 - 794.
2. Bo, S. (2009). A Study on Automatic Web Pages Categorization, *IEEE International Advance Computing Conference*, Page(s): 1423 -1427.
3. Lin, Z. D.(2007). Research of Web Pages Categorization, *IEEE International Conference on Granular Computing*, Page(s): 691.
4. Zhang, R. C.(2007). Automatic Web Page Categorization using Principal Component Analysis, *40th Annual Hawaii International Conference on System Sciences*, Page(s): 73.
5. Aung, W. T.(2009). Random forest classifier for multi-category classification of web pages, *IEEE Asia-Pacific Services Computing Conference*, Page(s): 372 - 376.
6. Zhou, Z.H. (2004). Multi-instance learning: a survey, *Rapport technique*, 5.
7. Gartner, T. , Flach, P. A. , Kowalczyk, A. , & Smola. A. J. (2002). Multi-Instance Kernels. *ICML*, page(s): 179-186.
8. Hu, W. M.(2007). Recognition of Pornographic Web Pages by Classifying Texts and Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol(29), Page(s): 1019 - 1034.
9. Chia, C. , Sherkat, N., & Nolle, L. (2010). Towards a Best Linear Combination for Multi-modal Biometric Fusion, *ICPR.*
10. Nandakumar, K., Chen, Y., & Jain, A. (2008). Likelihood Ratio-Based Biometric Score Fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp.342-347.
11. van de Sande, K.E.A. , Gevers, T. & Snoek. C.G.M. (2010). Evaluation of color dscriptors for object and scene recognition, *IEEE transanction on PAMI*, Vol. 32, no. 9, pp.1582-1596.
12. Chang, C.-C. , & Lin. C.-J. (2011). LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.Accessed 20 Oct. 2013.