Web Page Classification with Heterogeneous Data Fusion

Zenglin Xu, Irwin King, and Michael R. Lyu
Department of Computer Science and Engineering
The Chinese University of of Hong Kong
Shatin, N.T., Hong Kong
{zlxu, king, lyu}@cse.cuhk.edu.hk

ABSTRACT

Web pages are more than text and they contain much contextual and structural information, e.g., the title, the meta data, the anchor text, etc., each of which can be seen as a data source or a representation. Due to the different dimensionality and different representing forms of these heterogeneous data sources, simply putting them together would not greatly enhance the classification performance. We observe that via a kernel function, different dimensions and types of data sources can be represented into a common format of kernel matrix, which can be seen as a generalized similarity measure between a pair of web pages. In this sense, a kernel learning approach is employed to fuse these heterogeneous data sources. The experimental results on a collection of the ODP database validate the advantages of the proposed method over traditional methods based on any single data source and the uniformly weighted combination of them.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; 1.5.4 [**Pattern Recognition**]: Applications—*Text processing*

General Terms

Algorithm, Experimentation

Keywords

Web page classification, Kernel combination, Data fusion

1. INTRODUCTION

With the rapid growth of web pages of the World Wide Web, web page classification has become more and more important in both research and applications. Many researchers have explored content-based and context-based data sources, such as the text, the meta data, the anchor texts, and the titles, for the web pages classification [2, 6, 5]. Although the combination of some data sources is able to improve the web pages categorization performance, there is still a lack of effective and systematic strategies of combining different data sources. In the previous work, the weights for different data sources are usually set manually and the data sources are usually put together to form a large data file. Due to the different dimensionality and different types of these heterogeneous data sources, simply putting them together would not greatly enhance the classification performance.

To solve this problem, this paper proposes a heterogeneous web page classification framework. First, these heterogeneous data sources are extracted from web pages. Via a kernel function, each data source is represented as a common format of a kernel matrix, which can be seen as the generalized similarity matrix defined among all pairs of pages. Then, the fusion of heterogeneous data sources is formulated as the combination of the kernel matrices. Exploiting Quadric Constrained Quadratic Programming (QCQP) techniques, the fusion process is reduced to a convex optimization problem. Last, the fused kernel matrix will be employed by a kernel-based classification algorithm.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. RELATED WORK

Web page classification has been widely studied in the past few years. The simplest approach for web page classification is only using the text features. However, web pages are more than texts, and they contain a lot of context and structural information, e.g., links, anchor texts, URLs, etc. [2] proposed robust statistical models by exploiting link information in a small neighborhood around documents. [1] found that the combination of the anchor text and the plain text got better accuracy. [3] found that meta data is a useful source of information and the combination of meta data with text can result in better performance. [6] claimed that the combination of the plain text, the anchor text and the title can get a large improvement on F1 measure compared with full-text.

3. WEB PAGE CLASSIFICATION WITH HETEROGENEOUS DATA FUSION

We acquire each data source by extracting one of the following features: the plain text (PT), the title (TI), the meta data (MT), the text from the inlink pages (LT) and the anchor text (AT). When the features are extracted from web pages, kernel matrices can be built.

We use θ_i to represent the weight of the i-th input source. After each data source is represent as a kernel matrix K_i, the fused kernel matrix can be represented by their linear combination, i.e., $K = \sum_{i=1}^{m} K_i$. We note the kernel matrix built on the training data points as K_i^{tr}. In the following, the kernel combination parameters θ_i and the parameters of kernel SVM are optimized at the same time.

Substitute $K^{tr} = \sum_{i=1}^{m} K_i^{tr}$ into the dual form of kernel SVM, the optimization problem can be formulated as [4]:

$$\max_{\alpha} \quad \alpha^T \mathbf{e} - \frac{1}{2}\alpha^T D(\mathbf{y})(\sum_{i=1}^{m} \theta_i K_i^{tr}) D(\mathbf{y})\alpha \qquad (1)$$

$$s.t. \quad trace(\sum_{i=1}^{m} \theta_i K_i^{tr}) = \delta,$$

$$\alpha^T \mathbf{y} = 0, 0 \le \alpha_i \le C, 1 \le i \le n,$$

where δ is a constant and $D(\mathbf{y})$ is the diagonal matrix of $\mathbf{y}$. In this paper, the weighting vector θ is constrained to be nonnegative. Thus the nonnegative combination of K_i will assure the positive semidefiniteness of K, since K_i is positive semidefinite.

It can be proved that the above problem can be transformed as follows

$$\max_{\alpha,t} \quad \alpha^T \mathbf{e} - \frac{1}{2}\delta\rho \qquad (2)$$

$$s.t. \quad \delta = \theta^T \mathbf{t}, \theta \ge \mathbf{0},$$

$$\rho \ge \frac{1}{t_i}\alpha^T G(K_i^{tr})\alpha, \quad 1 \le i \le m,$$

$$\alpha^T \mathbf{y} = 0, 0 \le \alpha_j \le C, j = 1, \ldots, n,$$

where $G(K_i^{tr}) = D(\mathbf{y})K_i^{tr}D(\mathbf{y})$, $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ is a vector of combination coefficients, and $\mathbf{t} = \{t_1, t_2, \ldots, t_m\}$ is the trace vector of each single source kernel matrix K_i. This is a Quadratically Constrained Quadratic Program (QCQP).

Thus the discriminant function can be directly written as:

$$f(\mathbf{z}) = \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{z}), \qquad (3)$$

where $\mathbf{z}$ can be a test data point.

Remark. When all kernel matrices are formulated as linear kernel matrices, finding coefficients to combine the data sources is directly equivalent to selecting coefficients to combine the kernel matrices built on heterogeneous data sources.

4. EXPERIMENTS

In this section, experiments are conducted to compare the effects of the single data sources, the universally weighted combination of them (UW), and the combinations based on kernel learning and convex optimization (KC) on web page classification.

Data Set The Open Directory Project database (ODP) is used for web page classification, because it provides more structural information and meta information for the web page classification. Three categories (with the size of 409, 546, and 402, respectively) under the "home" directory of the ODP database are selected as the test bed.

Experimental Protocols We conduct 10 trials. In each trial, a modified 5-fold cross validation is used to evaluate the performance, i.e., 1 fold is used for training and 4 fold for test in each cross. The parameters of SVM are tuned by cross-validation. The trace of combined kernel matrix is set to n, the number of web pages. The F1 metric is employed to evaluate the prediction performance of different classifiers for each category. To balance the performance measures across multi-categories, two kinds of average measures of F1 are considered, i.e., micro-F1 and macro-F1, where micro-F1 gives equal weights to each document and macro-F1 gives equal weights to each category.

Table 1: The average F1 values of SVM classifiers using different data sources and combinations.

Data	AT	LT	MT	TI	PT	UW	KC
Consumer	11.39	54.69	42.28	47.81	27.24	49.43	**80.94**
Family	49.26	69.05	43.88	58.21	58.47	73.96	**81.66**
Garden	38.61	77.02	50.50	62.48	55.66	79.54	**87.86**
Mi-F1	34.48	67.69	45.55	56.60	48.50	68.86	**83.68**
Ma-F1	33.09	66.92	45.55	56.17	47.12	67.64	**83.48**

It can be observed from Table 1 that KC achieves the best classification performance on both the micro-F1 measure and the macro-F1 measure. More specifically, KC improves by about 14% over UW on the micro-F1 measure, and 15% on the macro-F1. This is because the combination coefficients can be optimized to deal with the situation that heterogeneous data sources differ greatly in prediction performance.

5. CONCLUSION

In this paper, a statistical web page classification approach is proposed, which incorporates heterogeneous data sources extracted from web pages and formulates them into a common format of kernel matrix. Via kernel combination and convex optimization techniques, a new kernel matrix by fusing different kernel matrices can be provided to kernel-based web page classifiers. The experimental results indicate the effectiveness of this combination technique on one collection of the ODP database.

6. ACKNOWLEDGMENTS

We thank Mr. Patrick Lau, Mr. Dou Shen, Dr. Aixin Sun for their help in crawling the web pages and discussions. The work described in this paper is fully supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E).

7. REFERENCES

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory (COLT-1998)*, pages 92–100, New York, NY, USA, 1998. ACM Press.

[2] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 307–318, New York, NY, USA, 1998. ACM Press.

[3] R. Ghani, S. Slattery, and Y. Yang. Hypertext categorization using hyperlink patterns and meta data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, pages 178–185, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[4] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[5] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. A comparison of implicit and explicit links for web page classification. In *Proceedings of the 15th international conference on World Wide Web (WWW 2006)*, pages 643–650, New York, NY, USA, 2006. ACM Press.

[6] A. Sun, E.-P. Lim, and W.-K. Ng. Web classification using support vector machine. In *Proceedings of the 4th International Workshop on Web Information and Data Management (WIDM 2002)*, pages 96–99, New York, NY, USA, 2002. ACM Press.