网页自动分类融合模型研究*

□ 张晓丹 梁冰 王丽 白海燕 吕世灵 肖晶 / 中国科学技术信息研究所 北京 100038

摘要：为了提高网页自动分类的准确率，基于信息融合的模型理论，提出了一种通用的网页自动分类模型和融合算法。该模型根据完成功能的不同分为四个层次：信息抽取层、数据预处理层、特征层和决策层，其中特征层是针对网页上不同种类的媒体信息采用不同的分类方法进行分类，并将分类结果分别输入决策层和与该特征层算法相关的其他的特征层。决策层是处理特征层的分类结果，并推导出最终的网页分类融合结果，并将该模型和算法进行了实现。实验表明，文章提出的融合模型和算法可以有效地改进网页自动分类准确率。

关键词：网页自动分类，信息融合，融合模型，信息检索
DOI：10.3772/j.issn.1673—2286.2011.08.012

1 引言

网页自动分类是将网页正确分配到给定类别中的过程，目前是信息检索领域的热点研究问题。目前常用的网页分类算法都是采用文本分类的算法，比较成熟的算法有KNN算法、SVM算法和Bayes算法等，但这些方法都是针对纯文本内容进行的分类，而忽略了网页自身的特征，如有标签数据、图像和视频等媒体数据等，都是对网页分类有价值的数据。如何有效利用网页上这些有价值的数据进行网页分类，是提高网页分类准确率的重要解决方法，也是这篇论文研究的主要问题。

基于信息融合的模型理论，提出了一个通用的网页自动分类融合模型。该模型可以处理不同的数据并对其进行处理，可以得到最终的融合分类结果。

本文的结构为：第二节提出了网页自动分类融合模型，第三节提出了网页自动分类融合算法，第四节为实验，最后为结论。

2 网页自动分类融合模型

信息融合是利用多种方法处理多源信息以得到更准确的评估结果的过程。有多种融合模型和结构，即三种融合级别和两种融合结构，分别是数字层、特征层、决策层及串行连接结构、并行连接结构。

网页自动分类可以被看成是一个将WEB网页分到正确的类别中的评估问题，基于融合模型理论，一个基于多种媒体信息的特征层融合分类模型被建立起来，如图1所示。

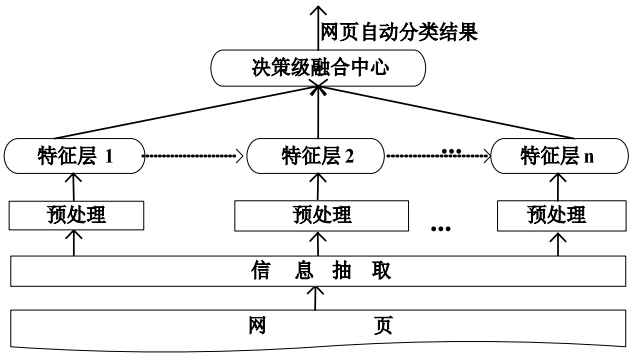

图1 网页自动分类融合模型

从图1可以看出，网页自动分类融合模型分为四层：信息抽取层、预处理层、特征层和决策层。信息抽取是指将网页中的有价值的信息抽取出来分别输入不同的预处理层。预处理层是对抽取到

* 本课题获以下项目基金资助：国家自然科学基金（基金号：60803050）；“十二五”国家科技支撑计划项目（项目编号：2011BAH10B05）；中国科学技术信息研究所预研项目（项目编号：YY—2010023）。

的不同的信息进行预处理，输入到相应的特征层。特征层针对不同的预处理信息采用不用的网页分类算法，进行分类，分类结果输入到决策层及与该特征层有关联的特征层。决策级对得到的特征层分类结果进行融合推理，得到最终的网页分类融合结果，并输出。

3 网页自动分类融合算法

文本和图像信息是网页上最常见的资源信息，也是对分类最有价值的信息，因此我们选择它们作为分类融合模型的待处理数据。为了确定特征级融合中心的分类算法，在相同的文本数据集及图像数据集的情况下，我们做实验比较了KNN、SVM、Bayes和BP网络的分类准确率。实验表明，KNN算法在采用文本数据集的情况下分类准确率优于其他算法，SVM在采用图像数据集的情况下分类准确率优于其他算法。因此，本融合模型特征层的分类算法选择KNN、SVM来分别处理从网页上抽取的文本及图像数据。D-S证据理论算法作为决策级的融合分类算法，处理特征级融合中心的KNN、SVM算法输入的分类结果。

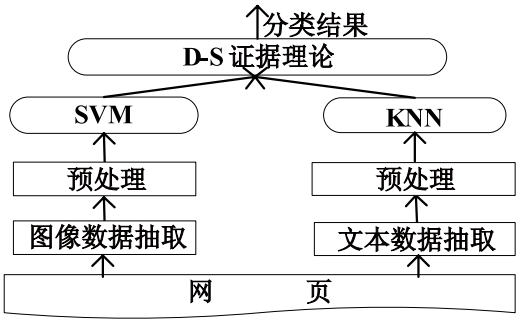

图2 网页自动分类融合算法

特征级网页自动分类融合算法见图2所示。

从图2可以看出，KNN、SVM算法作为融合模型的特征层分类算法，分别对文本和图像数据进行分类，D-S证据理论算法被用于决策层的融合分类算法，融合推导特征层的分类结果。

网页自动分类融合算法的步骤如下：

步骤1 训练文本分类器：对训练集中的网页进行信息抽取（去除广告等信息），对抽取到的文本信息进行预处理（分词、特征抽取、权重计算、向量表示等），并输入到特征层的KNN分类器中。KNN分类器确定下来。

步骤2 训练图像分类器：对训练集中的网页进行信息抽取，对抽取到的图像信息进行预处理（图像去噪、特征抽取和向量表示等），并输入到特征层的SVM分类器中。SVM分类器经训练后确定下来。

步骤3 将测试集中的网页进行信息抽取，抽取到的文本信息进行预处理，输入到KNN分类器中。输出文本信息的分类结果。

步骤4 将测试集中的网页进行信息抽取，抽取到的图像信息进行预处理，输入到SVM分类器中。输出图像信息的分类结果。

步骤5 步骤3和步骤4的结果被输入到决策层，即D-S证据理论算法中。

步骤6 最后的分类结果经D-S证据理论算法推导出并输出。

4 实验

本文提出的网页自动分类融合模型和算法是以JAVA和MYSQL实现的。为了检验该融合模型及算法的分类准确率，在相同的文本测试集和图像测试集的基础上，分别与KNN和SVM算法进行了准确率的比较。

实验中的文本数据集是从SOGOU网站上收集的。共选了6个类，包括教育、计算机、环境、交通、经济、军事。其中训练数据1049个文档，测试数据520个文档。

实验中的图像数据集是从SOGOU网站获得的。共有1700篇图像文档，每个类被分为测试数据和训练数据，比例为2：1。

表1 三种算法分类准确率及召回率比较

算法	召回率	准确率	F1
KNN	80. 3%	90. 2%	84. 8%
SVM	76. 4%	92. 6%	83. 7%
融合方法	85. 6%	95. 2%	90. 7%

三种算法的分类效率比较如表1所示。

在表1中，我们可以看到在相同的数据集的情况下，本文提出的融合算法的准确率、召回率及F1比其

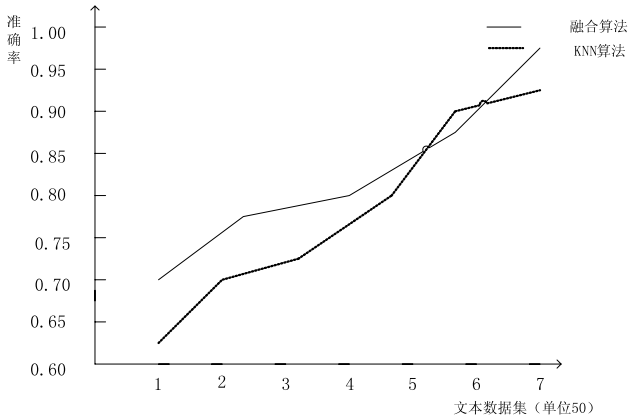

图3 相同文本数据集下的融合算法及KNN算法分类准确率比较

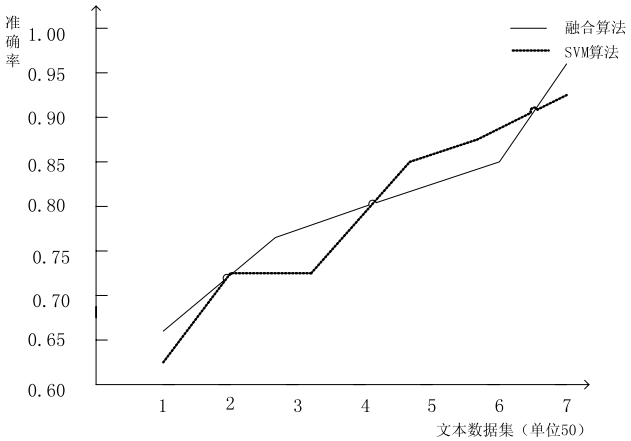

图4 相同图像数据集下的融合算法及SVM算法分类准确率比较

他两种算法高。

为了进一步说明融合算法与其他两种算法的比较，分别采用不同的数据集进行了比较；即采用相同的文本数据集，对融合算法与KNN算法进行比较，采用相同的图像数据集，对融合算法及SVM算法进行比较。如图3、图4所示。

从图3可以看出，在相同的文本数据集的基础上，本文提出的融合算法的准确率的曲线优于KNN算法的曲线。当横坐标为5.5和6.4时，两种算法的准确率相同。随着数据集的增加，准确率都是上升趋势。但是总体看来融合算法优于KNN算法。

从图4我们可以看出，在相同的图像数据集的基础上，融合算法的分类准确率优于SVM算法。当横坐标为2、4.4、6.2时，二者的准确率相同，随着数据集的增加，准确率增加。

“中信所知识管理”平台的子系统“网络科研资源获取与利用”是为了所内科研人员更方便、更好地完成自己的科研工作，满足其对互联网上的科研资源的获取和利用的需求而开发的一个系统。该系统的核心部分就是本文提出的融合模型及融合算法。目前该系统已经完成验收工作，并处于试运行阶段，已获得了良好的效果。

5 结论

随着网页上多种媒体信息的增加，如何利用这些有价值的信息来提高网页自动分类的准确率，是这篇论文的主要研究问题。根据信息融合理论的模型理论，提出了一个通用的网页自动分类融合模型。该模型对网页上的多种媒体信息进行抽取并分别进行预处理，得到的处理数据输入到特征层作分类，分类的结果分别输入到决策层，决策层经推导得到最终的融合分类结果。经实验，该模型的分类准确率优于现有的比较成熟的文本分类算法。

参考文献

[1] ZHANG B, CHEN Y, FAN W, et al. Intelligent GP fusion from multiple sources for text classification [C]//Proceedings of the 14th ACM international conference on Information and knowledgemanagement. ACM Press, Bremen, Germany, 2006:477-484.
[2] ZHANG G P. Avoiding Pitfalls in Neural Network Research. Systems, Man and Cybernetics, Part C [J]. Applications and Reviews, IEEE Transactions, 2007,37:3-16.
[3] ZHANG L, ZHU J, YAO T. An evaluation of information fusion techniques [J]. ACM Transactions on Asian Language Information Processing (TALIP), 2007(3):243-269.
[4] ZHANG Y, ZINCIR-HEYWOOD N, MILIOS, E. Narrative fusion classification for automatic key phrase extraction [C]// Proceedings of the 7th annual ACM international workshop on Web information and data management. ACM Press, Bremen, Germany, 2005:51-58.
[5] ZHENG Z, WEBB GI. Lazy Learning of Bayesian Rules [J]. Machine Learning, 2000(41):53-84.
[6] XU L Y, DU Q D. Application of neural fusion to accident forecast in hydropower station [C]// Proceedings of the Second International Conference on Information Fusion, Sunnyvale, 1999,2:1166-1171.
[7] SCHAPIRE R. E, SINGER Y, SINGHAL A. Boosting and Rocchio applied to text filtering [C]// Proceedings of SIGIR-98 21st ACM International Conference on Research and Development in Information Retrieval. ACM Press, New York, US, 1998: 215-223.
[8] SEBASTIANI F. Machine learning in automated image categorization [J]. ACM Computing Surveys, 2002(34):1-47.
[9] SHIH L K, KARGER DR. Using urls and table layout for web classification tasks [C]// Proceedings of the 13th international conference on World Wide Web. ACM Press, New York, USA, 2004: 193-202.
[10] RISH I. An empirical study of the naive Bayes classifier on image [C]// IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. T.J. Watson Research Centre, Seattle, Washington, 2001: 41-46.
[11] RILOFF E. Little words can make a big difference for text classification [C]// Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Seattle, Washington, United States, 1995:130-136.
[12] ZHANG Xiao-dan. A system of file Automated classification:China,201020200043[P].2010.

作者简介

张晓丹（1975-），研究方向为信息融合、信息挖掘。发表文章30余篇。E-mail: zhangxd@istic.ac.cn

Study on WEB Page Classification Fusion Model

Zhang Xiaodan, Liang Bing, Wang Li, Bai Haiyan, Lv Shijiong, Xiao Jing / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: For higher text classification precision, a general feature layer fusion classification model and algorithm are proposed, based on model theory of information fusion, adopting multi-information of the network for different classification, text and image information are used in the paper. The model includes two layers mainly, one is feature layer, which deals with different Media information with different classification algorithm, and inputs the classification results into the higher layer fusion centre separately. The other is decision layer, which deals with the results from the feature layer, and concludes the final classification result. The experiment expresses the fusion model can improve the text classification precision effectively.

Keywords: WEB page classification, Information fusion, Fusion model, Information retrieval