中文网页形式自动分类

董静，林鸿飞，杨志豪

（大连理工大学计算科学与工程系，大连 116024）

摘 要： 传统的网页分类大都基于内容，这种方式采用字词特征项匹配的方法，没有考虑网页的结构信息。为了充分挖掘网页的结构特征，本文提出了网页按形式分类的机制。以往关于形式分类的研究大都基于普通文本，但是网页文本不同于普通文本，它具有 URL 和 HTML 标签等对网页结构有较大影响的特征。本文从 URL 和网页的 HTML 标签中提取特征，并借用了普通文本形式分类中使用的部分特征项作为网页形式分类的特征集合，最后使用 SVM 分类器进行分类训练测试。

摘 要： 自动形式分类；特征提取；HTML 标签

Automatic Genre Classification of Chinese Web Pages

Dong Jing, Lin Hong-Fei, Yang Zhi-Hao

(Department of Computer Science and Engineering, DaLian University of Technology, DaLian 116024)

Abstract: The majority of conventional web page classification is focused on a subject. This method is generally based on common terms shared among web pages, while ignoring the structure of web pages. For mining the structure information of web pages, this paper puts forward genre classification of Chinese web pages. Most former research about genre classification is based on textual documents. While web pages are different from textual documents, which contain URL and HTML tags, which influence the page's style. This paper proposed new sets of features containing which are extracted from URL and HTML tags and which are used in the genre classification of textual documents. Finally, Support Vector Machine is used as the learning algorithm to build the classifier.

Keywords: Automatic Genre Classification; feature extraction; HTML tags

1 引言

近年来，随着因特网上的网页数量迅速膨胀，能否从海量的网页中迅速、准确的搜索用户感兴趣的信息是对网页分类技术的挑战。目前大部分的搜索引擎都采用简单的交互模式，即将结果集按相关度排序呈现给用户。这种方式实现简单，但用户通常很难从中找到自己真正想要的信息。因此如何将文档集聚类就成为必须要解决的问题，可以根据主题或其他准则，如站点地址、地理位置或标题对文档集聚类[1]。

目前，大部分关于网页分类的研究都是基于网页的主题或者内容来进行的，也实现了很多不同的网页主题聚类方法。虽然主题网页分类已经取得了一定的成果，但是这种分类方法不能完全满足用户的需求。比如，用户要

基金资助：国家自然科学基金资助项目（编号：60373095）

作者简介：董静（1983），女，山东，硕士研究生 dongjing@dl.net.cn.

查询“禽流感”，实际上他们需要的或许是关于“禽流感”的图片集，或许是新闻，或许是关于“禽流感”的常识性知识。如果按照主题聚类，返回的结果将不区分网页的这些形式，全部呈现给用户。因此，根据形式进行网页分类可以使用户方便地找到所需信息，网页形式是网页分类的另一种标准，它逐渐的引起人们的注意。

然而，如何识别、描述网页文档的形式是一项复杂而具有挑战性的工作。这是由于，目前，基于文档形式或风格的分类研究大部分针对纯文本文档，基于网页形式的分类研究很少，这样，在形式化类别的确定、分类特征项选择等方面都会存在很大的困难。另外，网页文档不同于纯文本文档，它有 URL 和丰富的 HTML 标签及链接信息，而这些信息对网页的形式有一定的限制作用。例如，若一个网页的 URL 中包含 edu.cn 且 URL 深度为 0，则这个网页就是一个学校的主页。而如何将网页的这些信息应用于形式分类也是一个难题。

本文初步尝试中文网页形式分类的可行性。首先建立自己的网页形式类别并收集相应的语料数据，然后，抽取获得特征项集合，最后用支撑向量机进行分类训练和测试。

2 网页形式分类的概述

国内外针对普通文本的自动体裁分类研究较多，也提出了很多选取特征项的方法。而针对网页文档的形式化分类研究却很少。

Karlgren 和 Cutting[2]采用结构线索和像第三人称代词的个数这样简单的线索作为特征项，在后续的工作中[3]，他们分析了被检索到和不被检索到的文档以及相关文档和不相关文档之间的关系，使用简单的统计量如句长、词长和句法复杂度如解析树的平均深度作为特征项。在 1998 年，他们又提出基于网页文本的形式类别，并构建了基于这些类别的平衡的语料库[4]，系统中使用词汇项数目，HREF 超链数目等作为特征项集，利用 C4.5 决策树方法进行分类，其分类效果没有报导。

Kessler 等人[5]指出，一些自然语言处理工具，如 POS 标记，解析和词理解的歧义性可以提高自动体裁分类的性能，这是由于每种语言都有其语法结构，且相同的词在不同的体裁中可能有不同的理解。而在网页文档中，网页形式的检测有利于从句法结构中提取指定的信息。Kessler 等人使用了四种特征项线索：句法结构线索，词汇线索，字符级线索和派生线索。

Stamatatos 等人[6]使用整个书面语中的 30 个高频词和 8 个标点符号作为特征项来进行体裁分类，这种方法不依赖于领域和语言。他们指出标点符号在识别文本体裁中发挥着重要的作用。

Lee 等人[7]研究了形式分类在信息检索中的应用。他们选用了网络中常见的 7 种形式类别：报告、社论、技术论文、评论、个人主页、常见问题与解答和产品说明。他们基于词的统计，建立了基于主题和形式的两个语料库，从中去除更依赖于主题的特征项，系统采用了贝叶斯分类方法和一般相似性方法进行了分类处理，分类精度在 50%到 80%之间。他们提出主题分类的信息能提高形式分类的性能。

国内关于文本形式分类的研究较少，且多针对普通文本进行。基于网页形式的分类尚处于起步阶段，技术还不成熟。

3 建立网页形式语料库

3.1 网页形式类别

目前还没有权威的网页形式类别体系，只有国外一些学者做了相关的研究。1998 年，Dewe 等人[4]通过调查问卷的方式，经过整理综合共得到了 11 个具体的类别：个人主页类、公共/商业主页类、交互式页面、新闻材料类、报告类、其他运行文本、常见问题解答类、链接集类、其他列表表格类、讨论类和出错信息类。

后来，Lim 等人[1]针对韩文网页，在 Dewe 等人提出的类别的基础上，添加新类并细分其中某些类别，最后共得到了 16 个形式类别：个人主页类、公共主页类、商业主页类、公告集类、链接集类、图像集类、简单表格列表类、输入页面类、新闻材料类、研究报告类、官方材料类、情报材料类、常见问题解答类、讨论类、产品说明类、其他类（非正式文本）。

本文的形式类别是在他们研究的基础上并分析中文网页的特征提出的，共有 11 个形式类别：常见问题与解

答类、产品说明类、电子商务类、个人主页类、公告集类、公共主页类、链接集类、图像集类、输入页面类、新闻材料类、知识材料类。

3.2 语料库的建立

由于目前没有公共的关于网页形式类别的语料库，因此必须自己收集语料并标以类别，为了保证选取的网页更具有一般性，我们采用以下方法来收集网页数据（网页文档只限于中文网页）：

（1）从百度搜索排行中选取 30 个查询词，在 Google 上进行检索总共获得 1080 个网页（每个查询返回 30-40 个结果），去除其中重复的网页、出错页面和禁止访问的页面，共得到 1043 个网页。

（2）人工对这些网页标以类别。

（3）平衡每个类别的网页数量。每种类别控制在 100 篇网页文档左右，对数量不足的类别用下面的方法进行补充。

最后，共得到 1196 个页面，每个类别的网页数目均在 100 个左右。

4 特征项的选取

由于网页文档包含 URL 和丰富的 HTML 标签信息，且它们对网页的形式有很大影响，我们首先从网页的这两种特殊的信息中抽取合适的特征项集。除此之外，我们还使用了文本体裁分类中使用的某些特征项。在下面的内容中，将分别介绍这些特征项集。

4.1 网页的ＵＲＬ

网页的 URL 中包含关于网页形式的重要特征。例如，若 URL 中含有 index.html 串，则此网页一定是一主页。这里，我们定义 URL 的深度为 URL 中目录的层数，如“www.dlut.edu.cn”深度为 0，而“http://gs.dlut.edu.cn/queryinfo/login.aspx”的深度为 1。此外，网页的域名、URL 的深度及 URL 中包含的某些字符串都对网页形式有一定的约束作用，例如，若文档为主页类，则它的 URL 深度一般为 0；若 URL 中包含串 faq，则该网页很可能为 FAQ 类页面。

本文中，我们统计了 URL 中出现的高频词约 100 个，为了避免特征项矩阵过于稀疏，我们提出了 URL 词汇包的概念，如，我们将 shop，mall，store，market 等定义为电子商务类的 URL 词汇包。计算时将这些词的出现次数作为一个特征项来处理。

4.2 ＨＴＭＬ标签

HTML 是一种标记语言，它定义了一系列的标记，便于浏览器解释执行。这些标记不仅设置了网页文档的格式，并且还为网页引入图像、视频、超链接等信息。显然，这些标签将影响网页的形式类别，如，含有丰富<img>标签的网页很可能就是图像集类的页面。而网页中链接的比率也将决定该网页是内容型的网页还是目录型的网页[8]，进而影响网页的形式类别。本文选用了部分 html 标签出现的比率与文字链接比和内容链接比作为特征项。

文字链接比定义（Text Link Ratio）如下：

$$TLR = \frac{TL(P)}{TL(P)+TC(P)} \times 100\% \tag{1}$$

其中，TL(P)为网页 P 中所有超链接对应的锚文本的长度之和，即，

$$TL(P) = \sum_{L\in\text{网页P中的超链接集}} \left|Anchor(L)\right| \tag{2}$$

TC(P)为网页 P 中正文的长度之和，即，

$$TC(P) = \sum_{C\in\text{网页P中的正文}} \left|Text(C)\right| \tag{3}$$

定义内容链接比（Content Link Ratio）如下：

$$CLR(P)=\frac{L(P)}{K(P)} \tag{4}$$

其中，L(P)为网页 P 中链接的数目，K(P)为页面 P 的大小（以 K 为单位）。

4.3 文本特征项

网页的形式分类与普通文本的体裁分类其基本原理是相同的，都是挖掘文档结构风格上的特征，因此在文本体裁分类中用到的特征项对网页形式分类也是适用的，基于本文提出的 11 种网页形式类别的特征，我们采用了与之相关的部分在普通文本中使用的特征项。

国际上常常把特征项根据其处理的复杂程度直观地分成两种类型：浅层特征项和深层特征项。目前对于特征项的类型概括归纳较为全面的学者是 B. Kessler[5]。他针对英文提出了四种线索：句法结构线索：如被动句数量、现在进行时数量、词性等需要对句子利用语言学工具进行解析或标注才可以得到的线索；词汇线索：如习惯用词、高频词等；字符级线索：如疑问号、冒号、分隔符等非词汇的符号线索；派生线索：即从以上线索变异得来的线索。我们在此基础上增加了适合中文结构特征的线索：格式线索，如段首序号、发文日期等；浅层分析级线索，如，平均句长、中文字数、平均段长等通过统计即可得到的线索。

系统选取的文本特征项如下所示：疑问号的频次；段首序号：即段首为“一、二、…，1、2、…，1)、2)、…”等序号的标识在网页中出现的总次数；中文字数；平均段长；新闻发文日期：如“新华网上海 11 月 27 日电”等新闻的发出日期。另外，本文针对各个形式类别提出了领域词汇的概念，对于 FAQ 类词汇，本文使用了知网中的所有疑问词语，约 100 条；知识材料类词汇，参考国内外学术期刊的常规要求，使用了段首为“摘要、关键词、引言、结束语”等的词汇；并利用从语料中统计得到的高频词来扩展各个类别的领域词汇。

4.4 数据预处理

由于我们选取的特征项的性质不同，因此会产生各种各样的量纲。单位的量纲不同，就会导致属性取值范围差别很大，所以需要数据的预处理。本文采用“极差正规化（Normalization）”来进行处理[9]。

$$x'_{ij}=\frac{x_{ij}-\min\{x_{ij}\}}{\max\limits_{1\le i\le n}\{x_{ij}\}-\min\limits_{1\le i\le n}\{x_{ij}\}} \tag{5}$$

其中，x'_{ij} 为极差正规化后的数据，x_{ij} 为第 i 个样本第 j 个特征项的属性值，n 为样本总数。

5 实验

5.1 SVM 分类器

我们采用支撑向量机作为分类器。SVM 是一种基于统计学习理论的机器学习方法。其基本思路是找到那些对分类 SVM 有较好区分能力的支持向量来最大化类间间隔，因而有较好的推广性能和较高的分类准确率。

5.2 实验结果

表 1 系统分类结果

Tab.1 the result of the system

	准确率	召回率	F1 值		准确率	召回率	F1 值
FAQ 类	91.1%	80.4%	85.4%	新闻材料类	86.7%	74.3%	80%

个人主页类	80%	31.4%	45.1%	电子商务类	98.2%	48.2%	64.7%
产品说明类	92.6%	56.3%	70%	知识材料类	92.9%	45.6%	61.2%
公共主页类	82.1%	41.8%	55.4%	输入页面类	100%	55.9%	71.7%
公告集类	90.6%	94.1%	92.3%	链接集类	93.8%	58.1%	71.8%
图像集类	94.5%	64.5%	76.7%				

在实验中，我们使用了 1186 篇网页文档作为实验语料，其中 590 篇文档作为训练数据，596 篇文档作为测试数据。分类结果采用三个指标来进行评价，即准确率（P），召回率（R）和宏观 F_1 值。其中宏观 F_1 值综合了准确率和召回率两个指标，定义如下：$F_1=\frac{P\times R\times 2}{P+R}$，系统评价结果如表 1 所示。

从表中可以看出，各个类别的准确率都比较高，而召回率略低。由于 FAQ 类、公告集类、新闻材料类的特征比较明显，宏观 F_1 值都达到了 80%以上，而个人主页类、公共商业主页类的网页由于特征不够明显，所以分类精度都较低。

分析结果，导致某些类别分类结果比较低的原因可能有如下几个方面：（1）某些类别如个人主页类形式特征不够明显；（2）语料数量偏少，导致样本训练不够充分；（3）训练数据与测试数据差异性较大，训练数据代表性不强。今后的工作中，可以通过改进以上三方面来提高精度和召回率。

6 结束语

本文通过挖掘网页的结构特征，对中文网页进行自动形式分类，是对网页形式分类一种尝试，分类结果基本令人满意，实验说明网页形式分类是可能的，有进一步研究的价值和意义。与国外网页形式分类的结果相比，选用的特征项较少，分类精度上也有了一定的提高。但还有进一步改进的地方，如扩大语料数量，寻求区分性更好的特征项等，这些都将利于精度的提高。

参考文献：

[1] Chul Su Lim, Kong Joo Lee, Gil Chang Kim. Multiple sets of features for automatic genre classification of web documents. Information Processing and Management, 2005(41):1263-1276.

[2] Jussi Karlgren，Douglass Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis[J]. Pro. Of COLING94, Kyoto, 1994.

[3] Jussi Karlgren. Stylistic Variation in an Information Retrieval Experiment. Proc. of the 2nd International Conference on New Methods in Language Processing-NeMLap, 1996.

[4] Johan Dewe, Jussi Karlgren, Ivan Bretan. Assembling a Balanced Corpus from the Internet. 11th Nordic Conference of Computational Linguistics, 1998:100-107.

[5] Kessler B, Nunberg G, Schutze H. Automatic Detection of Text Genre. Proceedings of 35th Annual Meeting of Association for Computational Linguistics and 8th Conference of European Chapter of Association for Computational Linguistics, Madrid, Spain, 1997: 32-38.

[6] Stamatatos E, Fakotakis N, Kokkinakis G. Text Genre Dtection Using Common Word Frequencies. Proceedings of 18 Int. Conference on Computational Linguistics, Luxemburg, 2001: 808-814.

[7] Yong-Bae Lee, Sung Hyon Myaeng. Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization. Proceedings of the 37th Hawaii International Conference on System Sciences –HICSS 2004: Big Island, Hawaii, USA. 2004.

[8] 高波，张忠能，查志琴. 基于文字链接比的网页分类的研究. 计算机工程与应用. 2004(27):151-153.

[9] 方鹭飞，林鸿飞，杨志豪等. 中文文本体裁的自动分类机制. 中文信息学报. 2006, 20(2):24-32.