基于URL分类库正逆向分类模型的设计实现

汤　琛，王　攀
（南京邮电大学信息网络技术研究所，江苏省南京市 210003）

摘　要　网页自动分类是Web数据挖掘中的一个重要研究方向，也是搜索引擎前期的准备工作。文章介绍了一种利用搜索引擎原理构建从网页URL到行为类别映射关系的分类系统，该系统结合爬虫原理和网页自动分类技术实现了根据网页URL来判断用户行为的类别功能。实验表明该分类系统具有较高的分类质量和较强的适应能力。
关键词　网页自动分类；用户行为分析；分类引擎

网络迅速发展到今天已有50多年历史，无论在网络通信质量和网络应用普及上都达到了空前的规模。网络在人们日常生活中扮演的角色已经逐渐从单纯的通信工具转变为生活中必不可少的一部分，网上购物、网上银行、网上医院等基于Internet业务的出现标志着人们生活模式的改变，因此各种行业都开始开拓网络商业市场。

在这种发展趋势下，信息量暴增导致了网络的信息爆炸，人们面对海量信息开始变得茫然不知所措。如何帮助网民找到他们关心的信息，帮助网络商家找到其客户群体，已经成为网络运营商目前面临的首要问题，也是运营商从“粗放式经营”向“精细化管理”转变过程中亟待解决的问题。

用户网络行为分析成为解决这一问题的关键。目前用户网络活动大多通过网络浏览器完成，因此对用户使用浏览器所产生的上网记录进行数据挖掘和行为分析可获得用户上网行为习惯，从而得知用户所关注的信息。那么如何仅从浏览网页后产生的类似URL这类数据中分析得到用户行为意向呢？目前网页URL数量庞大，要借用人工手段实现对URL对应网页的分析，虽然在准确率上有一定保证，但是效率很低且代价高昂。因此对URL自动分析、归纳网页类别已经成为快速且有效地解决这一问题的关键，也是用户行为分析中用户行为模型构建的关键。

基金资助：国家高技术研究计划“863”资助项目（2009AA01Z212）；江苏省科技成果转化专项基金（No.BA2007012）；江苏省高技术研究计划项目（No.BG2007045）

1　网页分类技术简介

文本自动分类是应信息检索领域要求产生的，随着网络的发展已演变为网页自动分类。目前网页自动分类技术的研究比较活跃，出现了很多分类器构造方法，如统计方法、机器学习方法、神经网络方法等。目前比较主流的分类方法是一种基于文档特征向量空间模型（CVSM）的机器学习方法，即利用已知类别的训练集将未知网页映射到给定的类别空间。

之后也有很多人提出对分类系统的改进，以期达到更好的分类效果，但大都把注意力放到分类算法上。算法的改进虽然在分类准确性上有一定提高，但是这种程度的提高对于能够应用到实际的高精度、高效率的分类系统来说还是远远不够的。基于向量空间的分类方法对于训练集太过依赖，分类系统的好坏主要由训练集来决定，而这种分类系统的缺点也体现在训练集的选择和生成上。

首先，该分类系统只能根据训练集中的类别对未知网页作分类处理，而训练集的生成需要相当长的时间。若有新的类别加入则需要将整个训练集重新生成，耗时耗力。

其次，网页形式多种多样，同一类别下的网页结构和形式千差万别。训练集中的网页样本固然是有限的，一旦训练集确定下来，该分类系统就只能根据训练集中的网页形式作分类处理，对训练集以外同种类别下不同结构形式的网页往往会得到错误结果。

本文考虑到基于向量空间模型分类系统的种种特点，然后结合搜索引擎原理设计并实现了基于URL分类库正逆向分类器。

2 基于URL分类库正逆向分类器设计

2.1 基本概念

定义一：能够将网页与类别关系作对应的海量关系集合称作是URL分类库。通常URL分类库是用关系表的形式来表现的，关系表的行代表某一网页，列代表网页的属性，包括网页URL、类别、网页描述等信息。

定义二：从类别到网页的映射称为正向映射，通过已知类别得到该类别网页则为正向分类。

定义三：从网页到类别的映射称为逆向映射，对未知网页判断、分析、获知类别则为逆向分类。

定义四：能够按照一定停止条件自动从Internet上下载网页，并自动将网页与给定的类别进行相关对应的算法或模型，称作分类引擎。

2.2 分类器结构

基于URL分类库的正逆向分类器体系结构如图1所示。由图1可以看出分类器的整个工作过程分为2个模块：正向分类模块和逆向分类模块。用户端通过分类系统查询URL对应网页的分类，URL分类系统将用户传递的URL放到URL分类库中进行匹配，然后返回匹配结果。若匹配到分类结果便直接返回给客户端，否则将URL传递给逆向分类模块进行预测分析，获得预测结果返回给浏览器端。

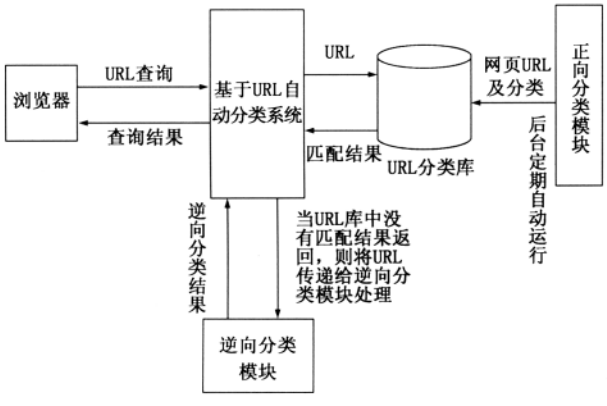

图1 分类模型框架图

2.3 正向分类模块

正向分类利用分类引擎来实现，过程简单，时效性好，负责构建URL分类库。利用Yahoo搜索引擎原理，预先给定网页分类目录和分类目录对应的种子样本，通过网页抓取器在导航网站或者借用各大搜索引擎接口（API）自动提取不同类别所关联的网页，将提取到的网页URL保存下来，根据对应类别搭建URL分类库，实现从类别到网页的映射。

如图2所示，正向分类模块需要根据不同导航网站设计不同的抓取器，每种抓取器都根据导航网页结构设计，将各导航网站上的类别体系与分类器的类别体系进行融合。对于主流搜索引擎直接利用他们的API接口提取各个类别的网页，然后对提取网页进行筛选进而保证URL分类库中数据的准确性。

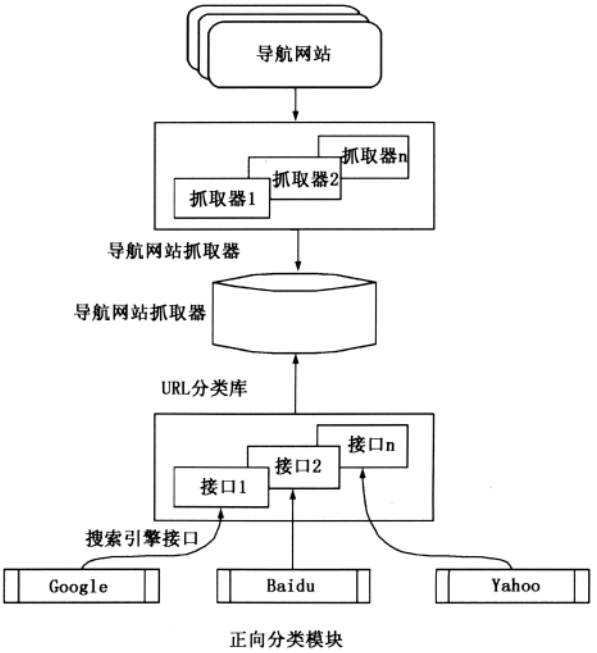

图2 正向分类结构图

正向分类模块设计特点：

a）实时性，网站内容不断丰富，导航网站的类别体系和网页结构将不定期作出更新，抓取器则需要依据网站结构变化作出相应对策以防止抓取网页错误或抓取器瘫痪。

b）分布性，由于分类模型类别体系包括五六百个类别，对一个类别的搜索一般需要花上1~2个小时。为了提高抓取网页速度，抓取器需分多进程进行和分布式地运行。

c）智能性，由于网络环境还存在着一些不稳定因素，抓取器需要有断点续传的功能，能在一次意外停止运行后继续上一次操作运行，而不是从头抓取以致造成不必要的时间浪费。

2.4 逆向分类模块

逆向分类模块能提高URL分类库覆盖率，它将URL分类库中覆盖率低的类别作为训练集来对这些类的待测网页进行分析预测获得分类。逆向分类采用中文网页分类技术中基于向量空间的文档向量比较法，该方法关键算法有：特征提取算法、网页向量

表示算法、分类算法。

特征提取在分类中起着重要作用，抽取特征项并根据特征项对类别描述的贡献计算各个特征权重大小，然后按权值排序，选取若干个评分最高的作为特征词。目前解决这一问题的算法有很多，如文档频率(DF)、信息增益(IG)、卡方统计法、互信息(MI)、开方拟和检验(CHI)等。本系统采取卡方统计法来提取特征，将网页看作普通文本，其所有文字部分的内容都用来描述网页特征。

网页特征表示算法。Salton等人提出的向量空间模型(VSM)是目前应用最多且效果较好的文本表示法之一。VSM中文本向量空间被看作是由一组正交词条向量组成的空间。TF*IDF算法从两个方面考虑了特征在文本中所起作用，即特征在文本中出现的次数要越多越重要；特征在越多文本中出现越不重要。这一方法与卡方统计法的合用将很好地解决网页向量表示。

分类算法。现有文本自动分类技术主要基于知识库方法和归纳学习法。而网页特征纷繁，类别多样，目前对于类别判断没有统一标准也就没有现存的知识库可以利用。因此对于网页分类通常都采用基于词典法的方法，即通过已知类别的训练集构造出分类模型，利用此模型将未知文档映射到给定的类别空间。而KNN算法正好符合这一思想，KNN是一种基于实力的文本分类算法，即找到未知样本X的K个最近邻，分析比较这K个近邻多属于哪一类，就把X归为哪一类。KNN也被认为是VSM理论下最好的分类算法，因此本系统将采用KNN分类算法。

预备训练集时，要考虑系统整体的时效性和准确性，不将整个类别体系作为训练集而取用URL库中网页数量少的类别作为训练集的类别体系。这类网页一般具有隐蔽性，面对的都是些特定用户，或采用的是会员制形式，直接用正向分类模块抓取到的网页数量有限，进而导致URL分类库的覆盖率不够。利用逆向分类模块的机器学习能力，仅用有限的网页训练集便能对该类别的网页作出预测分类。由于提供的类别数量少、训练样本数量规模不大，在相同数量的特征项下对于类别的描述更加具体和准确，另外训练集的生成、网页向量比较计算时所耗的时间少，这都有助于提高系统的整体分析速率和准确率。

3 系统评估指标

对于系统性能分析和评估也是网页分类研究的一个重要方面，一般从准确率和召回率两个方面来衡量分类系统的性能。

将待测数据分为4种，见表1。

表1 信息关系表

	相关	不相关
检索到	A	B
未检索到	C	D

a)A：检索到的，相关的。

b)B：检索到的，不相关的。

c)C：未检索到的，相关的。

d)D：未检索到的，不相关的。

e)召回率(recall)：检索到的相关文档数目与所有的相关文档数目的比值，简记为R。

f)准确率(precision)：检索到的相关文档数目与所有检索到的文档数目的比值，简记为P。

g)URL分类准确率(PoUC)：所有待测类别准确率的平均值为

$$P=\frac{\sum_{i=1}^{n}p_i}{n} \quad (1)$$

这两个指标来源于信息检索，通常我们希望被测数据中的相关文档被检索到的越多越好，这是追求的“召回率”，即A/(A+C)；同时我们还希望检索到的文档中相关的越多越好，不相关的越少越好，这是追求的“准确率”，即A/(A+B)。

从上面公式可以看出，“召回率”和“准确率”虽然没有必然的关系，但在大规模数据集合中，这两个指标是相互制约的。希望更多的相关文档被检索到而放宽“检索策略”，便会伴随出现一些不相关的结果，从而使准确率受到影响；而希望减少不相关文档的出现，务必要将“检索策略”定的严格一些，这样也就会使一些相关的文档不再能被检索到，从而又降低了召回率。

因此，利用F1指标综合P和R这两个指标，来对分类模型进行整体评价：

$$F_1=\frac{2\times p\times r}{p+r} \quad (2)$$

4 实验结果及分析

我们设计并实现了一个基于URL分类库的分类器。系统基于的类别体系参考《中国城市居民互联网使用及消费行为研究系列报告》中提到的网民网络

行为结构：信息获取、沟通交流、休闲娱乐、电子服务、电子商务五大结构，结合目前各大主流导航网站的分类体系将系统类别体系制定为12个一级类别，524个二级类别。利用正向分类模块，搭建了有100万条数据的URL分类库。将其中网页数量少的二级类别提取出作为逆向分类模块的训练集类别体系，其中包括网络游戏、赌博、网络电视等共20个类别。

测试集取用了10个类别，取得是用户访问比较多的类别，每个类别采用20个网页，实验结果见如表2、图3所示。

表2 试验结果

编号	类别名称	准确率	召回率	F1 值
1	体育	90.9%	90%	90.44%
2	新闻	84.9%	90%	87.33%
3	博客	84.7%	89%	86.79%
4	房产	85.2%	85%	85.42%
5	财经	88%	88%	88%
6	网络游戏	84.84%	84%	84.41%
7	网上购物	82.4%	80%	81.18%
8	博彩	82.9%	78%	80.37%
9	音乐	85.2%	87%	86.09%
10	电子	84.2%	75%	79.33%
	平均 F1 值			85.02%
	URL 分类准确率	85.32%		

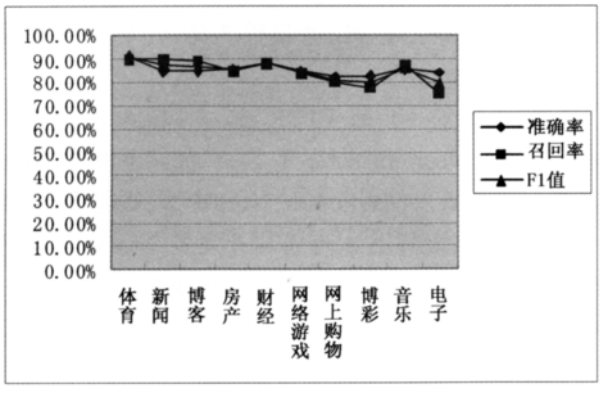

图3 数据点曲线图

从数据图3中可以看出分类器的平均F1值达到0.85，URL分类准确率达到85.32%，结果比较理想。对每一个类的分类能力都很平均，各个分类的准确率都超过80%。召回率部分，由于“电子类”的网页范围不容易界定（通常都涵盖其他类别的特征元素），所以正向分类模块搜集的网页中会将类似“电子”类的网页界定到“网上购物”类中去，从而降低“电子”类的召回率。另外如“新闻”、“体育”类别的网页特点鲜明，各大网站都有相关版面，因此容易抓取且不易混淆，因此在召回率和准确率上比其他类别高一些。“博彩”类网页由于具有一定隐蔽性，直接抓取数据量不大，URL分类库覆盖率比较低，而利用逆向分类模块进行预测在准确率上已达到82.9%，比较理想；召回率部分由于网页结构粗糙，主界面信息量少，且与“网络游戏”类网页有共同的结构特征，因此有一部分“网络游戏”测试网页被系统判断为“博彩”类，因此降低了召回率。

系统对于网络的适应力强，对网页结构内容的依赖性小。一旦有新的网页出现系统能即时补充更新，提高系统的分类覆盖率。

5 结束语

随着更多网上业务的出现，用户使用这些业务所产生的历史数据中蕴藏着巨大的商业信息。网络用户行为分析也将得到越来越多的广泛关注。网页自动分类是Web数据挖掘中一个重要的研究方向，也是搜索引擎前期的准备工作。本文介绍了一种利用搜索引擎原理构建从网页URL到行为类别的映射关系的分类系统，该系统结合爬虫原理和网页自动分类技术实现了根据网页URL判断用户行为类别的功能。实验表明，该分类系统具有较高的分类质量和较强的适应能力。用户行为分析中的用户业务识别、用户访问/使用模式分析中的关键技术和问题还有待于进一步的研究。

参考文献

1 黄萱菁，吴立德. 基于向量空间模型的文档分类系统[J]. 模式识别与人工智能,1998,11(2):147-153.
2 李晓明,闫宏飞,王继民. 搜索引擎-原理、技术与系统[M]. 北京:科学出版社,2005.
3 王 攀,张顺颐,陈雪娇. 基于DBP的Web用户行为分析关键技术[J]. 电信快报，2008(8):13-15.
4 冯是聪,张志刚,李晓明. 一种中文网页自动分类方法的实现及应用[J]. 计算机工程,2004,3(5):19-20 转108.

汤 琛(1984—)，男，硕士研究生，主要研究方向为计算机网络，网络用户行为分析。

收稿日期：2009-08-28