

## 基于链接信息的网页分类算法

秦 杰, 闫付亮, 朱海丰, 司 群, 谢 蕙

(河南工业大学 粮食信息处理教育部重点实验室, 河南 郑州 450001)

**摘 要:** 为了提高网页文本分类的准确性, 克服传统的文本分类算法易受网页中虚假、错误信息的影响, 提出一种基于链接信息的网页分类算法. 通过对 K 近邻方法的改进, 利用当前网页与其父网页的链接信息对网页实施分类. 用空间向量表示待分类网页的父链接信息, 在训练集中找到 K 篇与该网页链接信息向量最相似的网页, 计算该网页所属的类别, 通过实验与传统文本分类算法进行了对比, 验证了该方法的有效性.

**关键词:** 网页分类; 类别; K 近邻方法; 链接信息分类

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)06-0108-05

## A Webpage Classification Algorithm Based on Link Information

QIN Jie, YAN Fu-liang, ZHU Hai-feng, SI Qun, XIE Hui

(Grain Information Processing and Control Key Laboratory of Ministry of Education,  
Henan University of Technology, Zhengzhou 450001, China)

**Abstract:** To improve the performance of webpages classification system, and overcome a large number of false, erroneous information filled in the webpages affect the traditional classification algorithms, this paper presents a web page classification algorithm based on link information. Based on the K Nearest Neighbor method, the webpages are classified by the links among webpages. In this paper, the webpage currently classified is presented by the link information of vector space, and find K webpages with the highest similarity to it in the training set, then it is classified to the proper category. We compare the method to traditional classification algorithms through experiments, and the results show that it's more effective.

**Key words:** webpage classification; category; K-nearest Neighbor; link information classification

### 1 引言

在网页出现之前, 已经有大量的关于文本自动分类技术方面的研究. 随着 Web 技术的发展, 自动分类技术从文本扩展到了网页, 以及其他网络资源. 合理的分类有助于信息的组织和管理.

现有的网页分类方法基本上是基于文本分类算法的改进. 传统的文本分类算法通常分为三个处理阶段: 文本预处理阶段, 训练阶段和分类阶段. 网页信息与文本信息的主要不同在于信息的承载上, 因此, 传统的分类算法只要在文本预处理阶段稍作修改, 使其理解网络语言便可用于网络信息的分类.

现有的网页分类算法存在不足, 如过于依赖文本内容, 对网页中的虚假信息无能为力; 文本向量空间维数太大, 难在分类效率和准确率上做出取舍; 对图片、音乐无效等.

本文提出的链接信息(Link Information Categorization, LIC)分类算法, 利用网页内部的链接信息进行网页分类, 提高了网页分类的质量.

### 2 传统的文本分类方法

一些比较成熟的文本分类算法已经被应用到网页分类中, 如贝叶斯分类算法、K 近邻方法、神经网络分类算法和决策树分类算法等<sup>[1]</sup>.

收稿日期: 2011-10-26; 修回日期: 2011-12-15

基金项目: 国家自然科学基金项目(60373003); 河南工业大学校基金项目(2006BS009)

贝叶斯算法<sup>[1]</sup>适合处理不完整数据集,可以发现数据间的因果关系,但是,贝叶斯算法的计算量较大,效率较低,而且对具体问题,要合理确定许多变量的先验概率仍然是一个较困难的问题。贝叶斯算法假设待分类的特征之间是正交关系,且服从多项式独立同分布,如何判定某个实际问题是否满足这些假设,没有现成的规则,从而给实际应用带来困难。

神经网络方法<sup>[2]</sup>在小规模识别问题中比较有效,但对于如何确定网络的拓扑结构尚无可靠的理论指导,易陷入局部最优解,对于小样本的大规模模式识别问题泛化能力差。

决策树方法理论简单清晰,规则抽取能力强,适合于处理大量实例的问题。这种方法也过分依赖于样本的数量和分布,对噪音过于敏感,难于有效地处理连续属性等<sup>[3]</sup>。

K近邻方法(K Nearest Neighbor, KNN)<sup>[3]</sup>实现简单,不需要训练过程,对于未知和非正态分布可以取得较高的分类准确率,具有鲁棒性、概念清晰等优点,而且也非常便于增加新的类别,受领域影响也小。但在文本分类应用中, KNN方法也存在不足:该算法是一种 lazy-learning 算法,需要将所有样本存入计算机中,每次决策都要计算待识别样本与全部训练样本之间的距离,并进行比较,因此存储量和计算量都较大。虽然分类效果较好,但其分类时间是非线性的,当训练文本数或特征数增加时,分类时间急剧增加。

本文对 KNN 算法进行改进,把网页中的链接信息作为网页分类的一项重要参考指标,从而提高网页分类速度及准确度。

### 3 基于链接信息的网页分类算法

#### 3.1 网页组织特征

网络信息主要有四种组织形式<sup>[4]</sup>:综合性网站,专业性网站,博客,论坛。可以进一步将这些信息按照提供者的特征分为两大类:正规网站信息和网民信息。

正规网站信息通常通过综合性网站或专业性网站等形式展示。其特点:信息可靠,而且分类明确。这类网站的经营者多为政府,有实力的企业等,其结构如图1所示,综合性网站包含新闻,财经等一些内容板块,这些内容板块下面又可以分为一些更小的内容板块。

网民信息指网络用户通过互联网发布的信息,

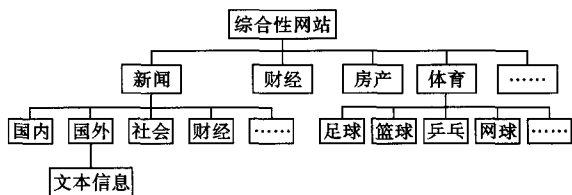


图1 综合性网站

包括博客,论坛,聊天室等等。这类信息结构性差,信息质量差异也较大,笔误、抄袭屡见不鲜,并且链接信息复杂。信息提供动机复杂,且易受信息提供者的兴趣、利益、工作等情况干扰,甚至心情都会影响发布信息的内容与质量。这些特点加大了这类信息分类的难度。但其对新知识、新信息响应速度快,容易被用户接受。相比之下,网民信息较少涉及商业利益,其内容中所引用链接信息相对较为公正。

网络信息由上述两类信息交织而成。正规网站信息本身之间就包含着较为清晰的类别关系。而网民信息类别虽然较为模糊,但是通过引用其它链接能够有效补充说明网页信息的类别。实际上,网页信息结构可以归纳为图2所示。

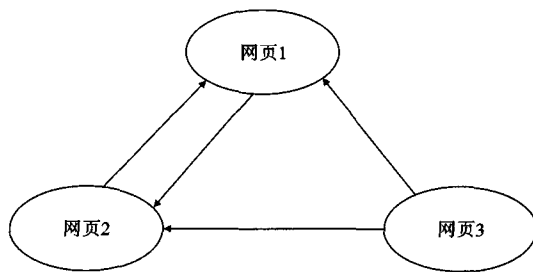


图2 网页之间的逻辑关系

图2展示了具有链接关系的网页之间的逻辑关系,网页1和网页2互相链接,同时又有网页3的链接。网页通常包含以下内容:

- (1)网页文本内容;
- (2)网页之间的链接;
- (3)用户访问模式,这从另外的角度反映了网页之间的关系。

网页信息的分类可能基于网页内容,也可能基于网页之间的链接关系。基于网页内容的分类和传统文本分类相同,只是需要在预处理阶段对网页信息进行去噪等。网页之间的链接关系反映了网页间的某种关系。链接中的网页文本对它所链接的文本起到了很好的注解作用。Craven等人利用链接中的标记文本对链接网页进行分类,取得了较好的效果<sup>[5]</sup>。网页之间的链接关系还反映了网页的重要程度,指向某一网页的链接越多,说明这篇网页越重

要,同时一篇网页被一篇重要网页链接,说明这篇网页也非常重要.因此大量的链接信息可以表征网页信息的内容重要度,相关性及结构等. Kleinberg 利用网页之间的链接关系设计提出了 HITS 算法<sup>[6]</sup>,该算法通过对网络中链接的分析,利用页面的被引用次数及其链接数目来决定不同网页的价值. Brin 和 Page 利用网页间的链接信息提出了 PageRank 算法<sup>[7]</sup>,该算法被 GOOGLE 采用.

有效利用链接信息有助于网页分类,并且具有特征维数较低、抗干扰能力强等优点.

### 3.2 网页信息来源的区分

常用的网页类型区别方法如下:

(1)通过网站内包含的网页数量判断.例如,私人网站的信息量相对较少,包含的网页数量也较为有限.

(2)通过提供信息的服务器数量或者访问速度判断.比如,政府机构或者大企业网站配置的服务器数量相对较多,访问速度也较快;而小企业、个人网站往往没有经济实力配置高性能的服务器,访问速度相对较慢.

(3)通过 URL 地址判断.由于 IP 地址匮乏等原因,绝大多数普通网络信息发布者没有自己的服务器和私人网站,而是通过其他平台完成信息的发布.谷歌、百度、搜狐、网易、腾讯、csdn 等都推出了博客、论坛等服务,URL 地址通常包含所在网站的域名,以及 blog、bbs 等关键字.

### 3.3 类别的建立与统一

类别的设定极为复杂,直接影响分类的正确性和合理性.虽然大型网站都对栏目进行了分类,但分类标准并不统一.

本文采用开放目录(Open Directory Project, ODP)<sup>[5]</sup>的组织形式,作为基于链接信息的网页分类算法的映射目标.

ODP 是互联网上最大、最广泛的人工目录.是世界各地的志愿者共同维护与建设的,是目前世界上最大的全球目录社区.目前共有 59 000 个类别,包含超过 4 500 000 个 URL 地址,其目录结构如图 3 所示.

从图 3 可以看出,ODP 分类体系具有清晰的层次关系,如 Java 是编程的子类别,编程是计算机的子类别等等.而传统的文本分类体系多为一级或两级结构,与此相比 ODP 分类更明确细致,有利于快速定位信息.

综合考虑查询速度与精度,作为折中,本文在实

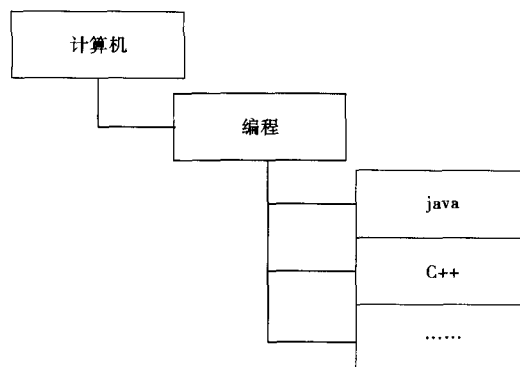


图 3 ODP 的类别结构图

际操作中选择前三层类别作为文本分类和用户兴趣映射的参照类别.

### 3.4 基于链接信息的网页分类算法设计

实际上,网页间的链接信息包含了网页内容的相关性,以及网页质量和结构方面的信息,提供了网页信息的特征.链接信息标记了网页的特征.本文提出的分类算法,就是利用了网页的这些链接信息.

基于链接信息的网页分类算法是在 KNN 算法基础上的改进<sup>[8]</sup>.首先找到在训练网页集合中与待分类网页链接信息向量最相似的 K 篇网页,根据这 K 篇网页所属的类别判定新网页所属的类别.

基于链接信息的网页分类算法与 KNN 方法的主要不同在于:后者依靠网页内容分类,前者则依赖于其父网页的引用信息分类.算法通过其它网页链接到当前页面的类别属性来决定当前页面的类别.该算法计算网页属于某个类别的重要性,对每个链入(inbound)赋与不同的权值.

算法主要步骤如下:

(1)设置待分类文档界限.网络上并非所有的网页都需要进行分类.更多的网页属于聚合网页,本身并不提供实质性的信息,主要是为了链接其它信息而存在,对这种网页进行分类毫无必要,待分类文档界限由式(1)确定.

$$\varphi = \frac{L(d_i)}{O(d_i)} \quad (1)$$

式中, $L(d_i)$ 表示文档  $d_i$  的长度, $O(d_i)$ 表示文档的链接出度,其比值  $\varphi$  用来表示其属于待分类文档集合  $D$  的可能性.当  $\varphi$  值大于某一阈值  $\alpha$  时,则认为文档  $d_i$  为待分类文档,属于待分类文档集合  $D$ .特别的,当链接出度  $O(d_i)=0$  时, $d_i$  为待分类文档.

(2)链接注释预处理.对链接注释进行分析,为其设置初始类别.获取待分类网页的所有直接父链接信息,并用向量空间  $\{(t_1, w_1), (t_2, w_2), \dots, (t_n,$

$w_n\}$ 表示,其中, $t$ 表示链接信息, $w$ 表示权重,其计算方法由式(2)确定.

$$w(t,d)=\frac{\sum_{i=1}^nq_i}{N(t)}$$

(2)

式中, $N(t)$ 表示含有关键字  $t$  的链接信息个数, $n$  为待分类网页的所有链接信息个数.  $q_i$  为待分类网页的直接父网页权重,分别为  $q_d$ 、 $q_u$ 、 $q_r$ . 如果待分类网页的域名与其父网页的域名相同,则设其权重为  $q_d$ ;否则,如果其父网页为第一类网页信息,则权重为  $q_u$ ;如果其父网页为第二类网页信息则权重为  $q_r$ ,显然  $q_d>q_u>q_r$ .

(3)链接信息文本分类

在训练集合  $d$  中选出与待分类网页  $x$  的链接信息向量最相似的  $K$  个链接信息向量,计算方法由式(3)确定.

$$\text{sim}(x,d)=\frac{x\cdot d}{\|x\|\|d\|}$$

(3)

式中, $\text{sim}(x,d)$ 为链接信息向量之间的余弦相似度.

(4)在待分类的  $k$  个邻居中,依次计算每类的权重,计算方法由式(4)确定.

$$p(x\mid c)=\sum_{i=1}^k\text{sim}(x,d_i)y(d_i,c)$$

(4)

其中, $y(d_i,c)$ 为类别属性函数,即如果  $d_i$  属于

类别  $c$ ,那么函数为 1,否则为 0.  
(5)比较每类的权重,将网页映射到权重最大的那个类别中.

4 实验结果与分析

为了验证算法的有效性,使用 java 语言,在 Java SE、Windows XP 编程环境下,Intel Pentium(R) D CPU 2.80 GHz,1 GB 内存的 PC 上实现了 KNN、支持向量机、基于链接信息的网页分类算法.

实验数据分别从综合性门户网站搜狐、网易、雅虎,以及专业网站 javaeye、csdn 上镜像得到,总计收集网页 3430 个,其中普通类别、专业类别各 9 个.

针对传统的分类算法,采用文档频度的特征提取方法对数据进行预处理,每一类选取 200 个特征,总计选取的特征维数为 1 800 维.对网页分类的性能评估采用准确率(accuracy)、查全率(recall)和 F1 评估值三种指标,对分类系统的整体性能评估采用宏平均的方法进行计算.

从经过预处理的网页中提取 20%做为测试文本集,剩余的 80%作为训练文本集,分别进行链接信息分类(LIC)、最近邻分类(KNN)、支持向量机分类(SVM)三组实验.综合性网站实验结果如表 1 所示.

表 1 综合性网站实验数据

| 类别      | 准确率     |         |         | 查全率     |         |           | F1 测试值  |         |         |
|---------|---------|---------|---------|---------|---------|-----------|---------|---------|---------|
|         | LIC     | KNN     | SVM     | LIC     | KNN     | SVM       | LIC     | KNN     | SVM     |
| 体育      | 0.994 7 | 0.999 6 | 0.999 8 | 0.769 7 | 0.156 9 | 0.786     | 0.867 9 | 0.271 2 | 0.880 1 |
| 健康      | 0.999 9 | 0.982 6 | 0.999 8 | 0.685 3 | 0.683 1 | 0.364     | 0.813 2 | 0.805 9 | 0.533 7 |
| 娱乐      | 0.996 8 | 0.995 6 | 0.999 7 | 0.676 5 | 0.522 1 | 0.395     | 0.806 0 | 0.685 0 | 0.566 3 |
| 房产      | 0.999 6 | 0.998 9 | 0.999 5 | 0.742 9 | 0.304 8 | 0.653     | 0.852 3 | 0.467 1 | 0.789 9 |
| 教育      | 0.999 7 | 0.996 3 | 0.999 6 | 0.727 8 | 0.667 2 | 0.531     | 0.842 4 | 0.799 2 | 0.693 6 |
| 汽车      | 0.999 9 | 0.985 9 | 0.999 4 | 0.738 5 | 0.465 7 | 0.558 6   | 0.849 5 | 0.632 6 | 0.716 6 |
| 电脑      | 1       | 0.995 3 | 0.222 4 | 0.689 4 | 0.819 3 | 0.949 5   | 0.816 1 | 0.898 8 | 0.360 4 |
| 科技      | 0.999 9 | 0.997 9 | 0.449 5 | 0.757 2 | 0.357 9 | 0.563 2   | 0.861 8 | 0.526 8 | 0.500 0 |
| 财经      | 0.977 3 | 0.988 9 | 0.999 8 | 0.625 9 | 0.428 6 | 0.226 9   | 0.763 1 | 0.598 0 | 0.369 9 |
| 平均准确率   |         |         | 平均查全率   |         |         | 平均 F1 测试值 |         |         |         |
| 0.996 4 |         |         | 0.489 5 |         |         | 0.631 6   |         |         | 0.601 2 |

测试结果表明,在对综合性门户网站进行分类时,LIC 分类器的准确率、查全率比其他分类器略高,其 F1 测试值较其它分类器好,但是差异并不显

著.在对专业性门户网站进行分类时,LIC 分类器准确率、查全率,以及 F1 测试值明显高于其它分类器.其原因在于传统分类算法最终是依赖文档内容

进行分类,而各种语言,特别是 SQL 语句等有极大的相似性,故而导致分类准确性、查全率的下降。

专业网站实验结果如表 2 所示。

表 2 专业网站实验数据

| 类别         | 准确率     |         |         | 查全率     |         |           | F1 测试值  |         |         |
|------------|---------|---------|---------|---------|---------|-----------|---------|---------|---------|
|            | LIC     | KNN     | SVM     | LIC     | KNN     | SVM       | LIC     | KNN     | SVM     |
| java       | 0.978 9 | 0.799 9 | 0.798 9 | 0.564 1 | 0.482 1 | 0.415 9   | 0.715 7 | 0.601 6 | 0.547 0 |
| C/C++      | 0.989 7 | 0.879 6 | 0.786 2 | 0.753 4 | 0.356 9 | 0.598 1   | 0.855 5 | 0.507 8 | 0.679 4 |
| C#         | 0.976 8 | 0.775 6 | 0.686 4 | 0.695 6 | 0.327 1 | 0.374 7   | 0.812 6 | 0.460 1 | 0.484 8 |
| python     | 0.987 6 | 0.863 9 | 0.879 5 | 0.792 3 | 0.420 8 | 0.453 2   | 0.879 2 | 0.565 9 | 0.598 2 |
| Linux      | 0.976 7 | 0.759 2 | 0.875 6 | 0.574 2 | 0.568 2 | 0.431 2   | 0.723 2 | 0.650 0 | 0.577 8 |
| Ajax       | 0.987 4 | 0.873 9 | 0.768 9 | 0.958 7 | 0.428 6 | 0.563 4   | 0.972 8 | 0.575 1 | 0.650 3 |
| Oracle     | 0.956 7 | 0.659 3 | 0.677 4 | 0.626 8 | 0.468 7 | 0.498 6   | 0.757 4 | 0.547 9 | 0.574 4 |
| MySQL      | 0.963 4 | 0.684 3 | 0.527 7 | 0.633 9 | 0.319 3 | 0.549 5   | 0.764 7 | 0.435 4 | 0.538 4 |
| SQL Server | 0.959 3 | 0.654 9 | 0.644 5 | 0.678 2 | 0.357 9 | 0.458 2   | 0.794 6 | 0.462 9 | 0.535 6 |
| 平均准确率      |         |         | 平均查全率   |         |         | 平均 F1 测试值 |         |         |         |
| 0.975 2    |         |         | 0.414 4 |         |         | 0.808 4   |         |         | 0.576 2 |

通过实验数据可以看出,与常用的 KNN 以及 SVM 分类方法相比,链接信息分类算法(LIC)具有更好的分类效果,适合 Web 的信息分类,尤其是针对专业性网站。

## 5 结束语

本文根据网页信息的特点,在 KNN 方法基础上,提出了基于链接信息的网页分类算法 LIC。与传统的分类算法相比,LIC 算法有如下优点:(1)可以有效的防止被人为加工的页面欺骗;(2)对该分类算法的主要分类依据——链接信息,稍作修改,还可以实现对非 Web 文档的分类,如视频、音乐等非文档信息的分类。

将该算法与个性化搜索引擎技术相融合,同时进一步提高网页分类的效率,是本文下一步的主要工作。

## 参考文献:

- [1] Xiaoguang Qi, Brian D Davison. Web page classification: Features and algorithms[J]. ACM Computing Surveys, 2009, 41 (2): 12: 1-12: 31.
- [2] 万中英,王明文,廖海波. 基于投影寻踪的中文网页

分类算法[J]. 中文信息学报, 2005, 29(4): 60-68.

- [3] 秦杰,王春云,谢蕙. 基于本体的元搜索引擎 4 级结果处理算法[J]. 微电子学与计算机, 2010, 27(6): 67-69.
- [4] Open Directory Project [EB/OL]. [2008-07-23]. <http://www.dmoz.org/about.html>
- [5] Craven M, Slattery S, Nigam K. First-order learning for web mining[C]// The 10th European Conference on Machine Learning. Chemnitz, 1998.
- [6] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of ACM, 1999, 46(5): 604-632.
- [7] Rungsawang A, Manaskasemsak B. Parallel adaptive technique for computing PageRank[C]// The 14th Euro-micro International Conference. Parallel, 2006: 15-17.
- [8] 朱海丰. 个性化全文搜索引擎关键技术研究[D]. 郑州: 河南工业大学, 2009.

## 作者简介:

秦杰 男, (1968—), 博士, 副教授, 硕士生导师. 研究方向为网络信息处理。

闫付亮 男, (1984—), 硕士研究生. 研究方向为信息检索。