

基于改进遗传算法的 Web 文本挖掘系统

余燕芳

(衢州广播电视大学, 浙江 衢州 324000)

摘 要: 文本分类是文本数据挖掘中一个非常重要的技术, 已经被广泛地应用于信息管理、搜索引擎、推荐系统等各个领域。现有的文本分类方法很难适用于大规模的文本数据集。为此, 提出了一种基于改进遗传算法的文本挖掘系统。提出的改进遗传算法极大地提高了文本挖掘系统的分类效率。实验结果表明, 该方法适用于大规模文本数据集; 该方法提取规则的分类正确率较高, 分类速度较快。

关键词: 文本分类; 遗传算法; 文本挖掘; 规则抽取

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2010)04-0103-03

A Novel Web Text Mining System Based on Improved Genetic Algorithm

YU Yan-fang

(Quzhou Radio and TV University, Quzhou 324000, China)

Abstract: Text classification is a very important technique in the field of text mining, and it has been widely applied to the information management, search engine, and recommendation systems. Most existing classification methods cannot be used on the occasion of classifying a large number of samples. Therefore, a novel improved genetic algorithm was proposed for Web text mining system. The performance of the proposed system has been improved largely by our improved genetic algorithm. The proposed approach can be applied to classify a large number of samples. Experimental results show that both the accuracy and the speed of categorization are high.

Key words: text classification; genetic algorithm; text mining; rule extraction

1 引言

自从有了互联网和 Web 浏览器, 人们已经能很容易地从网上访问到大量的文本和多媒体信息。在这些大量、异质的 Web 信息资源中, 蕴含着具有巨大潜在价值的知识, 人们迫切需要能够从 Web 上快速、有效地发现资源和知识的工具。因此, 我们需要研究一些新技术, 能够自动地从网上发现、抽取、过滤出用户感兴趣的信息。Web 挖掘作为数据挖掘的一个新主题, 引起了人们的极大兴趣。文本分类作为数据挖掘的一个新主题, 引起了人们的极大兴趣。目前, 普遍认为没有哪一种方法能适合于各种特点的数据。

文本分类技术出现于 20 世纪 60 年代早期^[1]。20 世纪 80 年代, 人们开始采用知识工程 (Knowledge

Engineering, KE) 的方法来实现文本的分类^[2]。KE 方法一般采用析取范式 (Disjunctive Norm Form, DNF) 为每个类别定义逻辑规则, 可以理解为一种简单的自然语言处理方法 (Natural Language Processing, NLP)。Rau 等采用复杂的词汇-语义模式来实现对自然语言资料库 (Natural Language for Data Bases, NLDB) 的分类^[3]。Jacobs 等采用统计方法来辅助构造分类规则, 进一步提高了文本分类系统的性能^[4]。采用 KE 方法实现文本分类需要手工编制规则或应用其他复杂的 NLP 技术, 难度非常大, 也非常耗时, 在很多场合显得过于低效而不实用^[5]。

到了 20 世纪 90 年代, KE 方法逐渐被机器学习 (Machine Learning, ML) 的方法所取代。机器学习就是由计算机代替人来学习关于认识世界和改造世

界的知识. ML 方法能够获得跟人类专家相媲美的分类性能, 节省人力, 因此 ML 方法比 KE 方法更有吸引力. 现有的 ML 方法大部分基于向量空间模型. 作为一种简单、有效的方法, 基于向量空间模型的方法在文本分类中引起广泛关注, 并取得了很好的分类效果.

Web 挖掘是一项综合技术, 涉及 Web、数据挖掘、计算语言学、信息学等多个领域. Web 挖掘就是利用数据挖掘技术, 自动地从网络文档以及服务中发现和抽取信息的过程. 和传统的数据挖掘相比, Web 挖掘有很多的独特之处. 首先, Web 在逻辑上是一个由文档节点和超链接构成的图, 因此 Web 挖掘所得到的模式可以是关于 Web 内容的, 也可以是关于 Web 结构的. 其次, Web 挖掘的对象是大量的、分布的、异质的 Web 文档, 它们是结构化的或半结构化的, 数据量巨大, 增长迅速, 并且具有机器难以理解的语义. 因此现有的数据挖掘工具并不完全适用于 Web 挖掘, 即使可用也需要建立在对 Web 文档进行预处理的基础之上. 这样, 开发新的 Web 挖掘技术, 以及对 Web 文档进行预处理得到关于文档的特征表示等问题, 便成为 Web 挖掘研究的重点^[6].

2 基于改进遗传算法的规则抽取方法

文中基于改进遗传算法的规则抽取方法的优化框架如图 1 所示. 该方法按照顺序覆盖的方法, 试图挖掘出一个能覆盖大多数甚至全部训练样本的分类规则列表. 该方法的执行流程可描述如下:

(1) 变量初始化. 将已发现规则列表设置为空, 同时将所有的训练样本放置到训练样本集中.

(2) 遗传算法的演化. 遗传算法的每次演化都能发现一个分类规则. 遗传算法演化完成后, 将本次演化发现的规则加入到已发现规则列表中; 同时, 将该规则所覆盖的样本从训练样本集中剔除.

(3) 终止条件. 当未覆盖样本的数目小于用户预设值时, 遗传算法停止演化.

3 中文 Web 文本挖掘系统

Web 是一个巨大的、分布广泛的、全球性的信息服务中心. Internet 的飞速发展, 使得 Web 页面以每 4 个月翻一番的速度增长. 尽管 Web 上有海量的数据, 但由于 Web 页是动态的, 内容过于复杂, 导致人们很难迅速方便地在 Web 上找到所需要的数据和信息. 如何利用文本挖掘技术实现 Web 信息的分

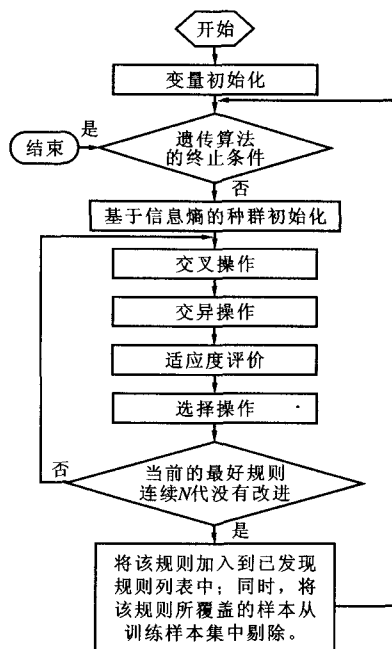


图 1 文 k 基于改进遗传算法的规则抽取方法的优化框架图

类已成为当前研究的一个热点.

文本挖掘是数据挖掘领域的一个新分支. 它首先采用文本切分技术对文本的特征进行抽取, 将文本数据转化为能描述文本内容的结构化数据; 然后利用分类、聚类和关联分析等数据挖掘技术形成结构化文本树, 并在所形成的结构中发现新的概念和获取相应的关系. 以 Web 文本文档为对象的文本挖掘被称为 Web 文本挖掘.

文中实现的中文 Web 文本挖掘系统(Web Text Miner, WTM)既是一个实验平台, 也是一个实用平台. 在该平台上, 既可以进行 Web 文本挖掘过程中的向量表示、特征抽取、分类等实验, 又可以对采集到的原始 Web 网页按照预定义的类别选择分类器进行自动分类. 作为一个实用平台, 系统在设计时考虑了对语料的无关性(即不依赖于语料集的种类, 挖掘的对象既可以是原始的 Web 网页, 也可以是普通的文本文件), 对算法的无关性(即挖掘过程的各个环节是相对独立的). 通过对各种静态数据进行设置, 在分类的各个阶段可以选择不同的算法、参数和数据文件, 同时对实验阶段得到的优化算法形成参数集, 以便在实际使用时直接利用实验的结果. 采用面向对象的程序设计思想, 我们的系统从结构上可以分为三层, 如图 2 所示^[7-8].

(1) 数据层. 将系统涉及到的各种数据和基本操作、算法封装成类, 所有的数据采用基本统一的数据结构. 对于非连续索引的需要经常查询的数据采

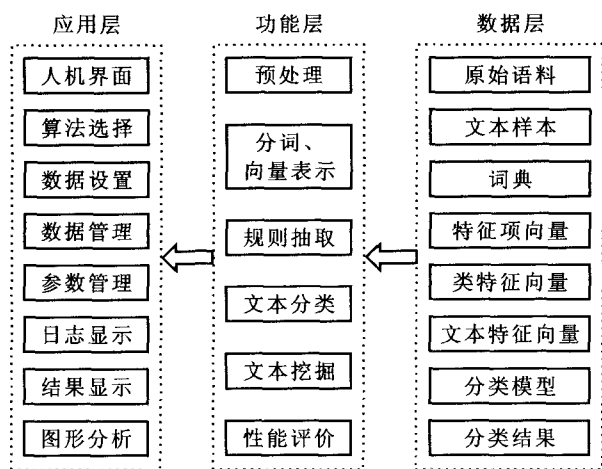


图2 中文 Web 文本挖掘系统的层次结构图

用散列表的方式,在散列表内部采用单向链表的方式以解决“冲突”问题,对于其他数据则采用结构数组的方式进行存储。一方面考虑了数据读写、查询的效率,另一方面考虑了尽量减少内存的使用。

(2) 功能层.在数据层的基础上,实现了各种功能模块。功能模块通过对数据类的操作,实现系统的基本功能,完成各种实验和自动分类过程。

(3) 应用层.完成人机界面、数据组织和参数的管理,并对各种实验和分类过程中的处理步骤进行记录,实现了过程日志的文件输出和分析结果的图形化显示。通过分析各种数据的依赖性和继承性,保证挖掘过程中数据的一致性。

4 系统性能评价

4.1 测试数据集

为了验证文中方法的效率,使用了3个标准的文本数据集来进行实验测试。

(1) 20 Newsgroups (20 NC). 20 NC 也是一个常用的文本数据集,其中收集了来自20个新闻组的近20 000篇新闻。20 NC 共包含18 828个文本。

(2) Industry Sector (IS). IS 是一个网页数据集,其网页分别来自各种不同工业经济部门的网站。这个数据集总共包含9 652个网页文本,分属于105个不同的类。

(3) 网页数据集(Web). Web 收集自 Google 的 Open Directory Project 项目。在实验中,随机选取了其中35个类,总共5 035个网页作为实验数据集。

4.2 评价方法和准则

笔者采用两种典型的测试方法来评价文中 Web 文本挖掘系统的性能。一种是训练-测试(Train and Test)方法,另一种是重交叉校验(k-fold cross

validation)方法。训练-测试方法是经典的评价方法,它将初始样本集合分成训练集和测试集两部分,训练集用来进行特征选择和分类器训练,测试集用来对分类器进行测试。而重交叉校验是将初始样本集合分成份,每次取其中的份作为训练集,剩下的一份作为测试集,然后取它们的平均值作为最终结果。将查全率和查准率作为评价文中 Web 文本挖掘系统的两个性能指标。

4.3 实验结果与分析

在处理器为奔腾 IV、主频为 2.4GHz、内存为 512MB 的个人电脑中完成这些实验测试。为了避免实验过程中的随机性,对每个实验重复进行了20次,取20次实验结果的平均值作为最终的实验结果。采用文中 Web 文本挖掘系统对3个文本数据集进行实验测试的结果如表1所示。从表1中可以看出,无论是查全率和查准率,文中的 Web 文本挖掘系统所得出的结果都比较理想。

表1 文中 Web 文本挖掘系统的性能评估结果

	训练-测试方法		k 重交叉校验方法	
	查全率	查准率	查全率	查准率
20 NC	0.93	0.89	0.95	0.92
IS	0.86	0.83	0.87	0.85
Web	0.96	0.92	0.95	0.93

5 结束语

怎样把 Web 上丰富的资源充分挖掘出来为用户所使用,是近几年研究的重点之一。文中在研究基于改进遗传算法的规则抽取方法的基础上,初步实现了一个性能较好的中文 Web 文本挖掘系统。如何将处理对象扩展到数量巨大的 Web 文本,这是下一阶段工作所要研究的内容。

本文的主要创新点:提出了一种基于改进遗传算法的规则抽取方法,初步实现了一个性能较好的中文 Web 文本挖掘系统。本文的方法和原型系统已经广泛地应用到高校文献资源的检索和推荐系统中。本项目已产生的经济效益约120万,受到了广大用户的一致好评。

参考文献:

- [1] Maron M E. Automatic indexing: an experimental inquiry[J]. Journal of the ACM, 1961, 8(3): 404-417.
- [2] Li Y H, Jain A K. Classification of text documents[J]. The Computer Journal, 1998, 41(8): 537-546.

(下转第110页)

育的工具,并为植物根系生长过程的相关研究提供帮助.

参考文献:

- [1] 刘秀海,肖伯祥,郑文刚,等. 玉米虚拟生长研究综述[J]. 玉米科学,2006,14(2):164-167.
- [2] 冯利平,高亮之,金之庆,等. 小麦发育期模拟模型研究[J]. 作物学报,1997(23):418-424.
- [3] 伍艳莲,曹卫星,汤亮,等. 基于 OpenGL 的小麦形态可视化技术[J]. 农业工程学报,2009,25(1):121-126.
- [4] 刘慧,汤亮,张文字,等. 基于模型的可视化水稻生长系统的构建与实现[J]. 农业工程学报,2009,25(9):148-154.
- [5] Penning de Vries F W T, Jansen D M, Ten Berge H F M, et al. Simulation of ecophysiological processes of growth of several annuals crops[M]. Netherlands: Pudoc. Wageningen, 1989.
- [6] Diggle A J Rootmap. A model in three-dimensional coordinates of the growth and structure of fibrous root systems[J]. Plant and Soil,1988(105):169-178.
- [7] 王功明,郭新宇,赵春江,等. 虚拟植物根系生长模型分析和比较[J]. 作物研究,2006(3):281-285.
- [8] Pages L, Jordan M O, Picard D. A simulation model of the three-dimensional architecture of the maize root system[J]. Plant and Soil, 1989(119):147-154.
- [9] 邓旭阳,周淑秋,郭新宇,等. 玉米根系几何造型研究[J]. 工程图学学报,2004(4):62-66.
- [10] 冯斌,杨培岭. 植物根系的分形及计算机模拟[J]. 中国农业大学学报,2000(5):96-99.
- [11] 张吴平,李保国. 棉花根系生长发育的虚拟研究[J]. 系统仿真学报,2006,18(1):283-286.
- [12] 杨林波,邵惠芳,章新军,等. 烟草根系研究进展[J]. 烟草科技,2002(10):45-48.
- [13] 马新明,杨娟,熊淑萍,等. 烟草根系形态发育模拟模型[J]. 中国农业科学,2005,38(12):2421-2427.
- [14] 马新明,席磊,熊淑萍,等. 大田期烟草根系构型参数的动态变化[J]. 应用生态学报,2006,17(3):373-376.
- [15] Lungley D R. The growth of root systems: a numerical computer simulation model[J]. Plant and Soil, 1973(38):145-159.

作者简介:

席磊男,(1972-),副教授.研究方向为智能系统与植物虚拟.

冀亚丽女,(1979-),硕士,讲师.研究方向为图像处理与模式识别.

汪强男,(1979-),硕士,讲师.研究方向为图像处理与模式识别.

张龙龙男,(1986-),硕士.研究方向为网络存储系统.

马新明男,(1963-),教授,博士生导师.研究方向为精准农作与信息技术.

(上接第 105 页)

- [3] Lisa F Rau, Paul S Jacobs. Creating segmented databases from free text for text retrieval[C]// proceedings of the 14th international conference on research and development in information retrieval. Chicago, Illinois: ACM Press, 1991:337-346.
- [4] Paul S Jacobs. Joining statistics with NLP for text categorization[C]//proceedings of the third conference on applied natural language processing. Trento, Italy: ACM Press, 1992:178-185.
- [5] Yang Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval [C]//Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland: ACM Press, New York, 1994:13-22.
- [6] 李士勇,梁家荣,唐志刚. 基于危险理论的 web 文本挖掘研究[J]. 微计算机信息,2007,27:170-171.
- [7] 亓洪标,李安,胡柏青. 基于 XML 的 web 文本挖掘方法与模型研究[J]. 微计算机信息,2006,26(11):196-197.
- [8] 贾志娟,胡明生. 基于网格的 Multi-Agent web 文本挖掘系统[J]. 微计算机信息,2006,26(7):266-268.

作者简介:

余燕芳女,(1976-),讲师.研究方向为计算机信息安全与智能化信息处理.