文本分类技术在海洋信息处理领域中的应用*)

徐大伟[1] 董 渊[2] 张素琴[2]

(长春大学计算机科学技术学院 长春 130022)[1] (清华大学计算机科学与技术系 北京 100084)[2]

摘 要 文本分类是数据挖掘和机器学习中非常重要的研究领域,将文本分类技术应用于海洋信息处理已经成为海洋领域研究的一个重要问题。主要研究文本分类技术在海洋信息处理领域的应用,给出了文本分类系统的关键技术设计方案,详细介绍了一种改进的 χ^2 特征提取算法以及朴素贝叶斯分类算法,实验结果具有较好的准确率和查全率,满足我国"数字海洋"信息基础建设对信息处理应用的需求。

关键词 文本分类,信息处理,数字海洋,朴素贝叶斯算法

Application of Text Classification in Marine Information Processing

XU Da-wei[1] DONG Yuan[2] ZHANG Su-qin[2]

(College of Computer Science and Technology, Changchun University, Changchun 130022, China)[1]

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)[2]

Abstract Text classification is a very important research field of data mining and machine learning, applying text classification technology to marine knowledge becomes an important point. Studied the application of text classification in marine information processing field, gave the design of text category system. Improved feature selection of χ^2 and naive bayesian categorization algorithm were in presented detail. Experiment shows good results and satisfying "Digital Marine" practical applied requiring.

Keywords Text classification, Information processing, Digital marine, Naive bayesian algorithm

1 引言

随着计算机和互联网技术的不断发展,对海量信息,特别是非结构信息的检索、过滤、管理成为一个突出的问题。海洋对于现代社会的许多领域都是至关重要的,因此对于海洋的了解也是现代社会发展的基本需求。作为数字地球的同类计划,数字海洋也成为数字计划的热点之一[1]。在北美,由海洋基金(Sea Grant)所支持的相关研究已经形成了一系列数据服务和基础数据仓库系统[1],用于各类相关的工程、生产、研究及公共服务领域。在我国,908 专项的相关工作已经在许多方面展开。

在当前科学技术飞速发展的情况下,海洋数据获取的技术和手段都得到了很大的提高,为海洋领域的发展提供了丰富的数据资料。但海洋数据信息量的急剧增长对信息管理提出了更高的要求:一方面是如何从海量海洋数据中获取规律性的知识;另一方面是如何组织这些知识以便于使用和检索。为方便信息检索,有必要先对海量的电子信息按其内容加以分类并进行索引。本文主要讨论如何将知识挖掘中文本自动分类技术应用于海洋信息处理,给出了文本分类系统的关键技术设计方案,详细介绍了一种改进的 χ^2 特征提取算法以及朴素贝叶斯分类算法。

2 文本分类技术概述

分类是数据挖掘的一种非常重要的方法。分类的概念[2]是在已有数据的基础上学会一个分类函数或构造出一个分类模型(分类器,Classifier)。该函数或模型能够在给定的分类体系下,根据文本的内容自动将文本指定到预定义的一个或多个主题类别的过程,它是自然语言处理的一个重要应用领域。

由自然语言构成的文本并不能被计算机直接处理,需要对其进行数学抽象,建立模型后进行自动分类的处理。基于机器学习的文本分类一般分为训练和分类两个阶段,在训练阶段,将一组预先分类过的文档作为训练集。然后,对训练集分析以便得出分类模式,分类模式的形成通常需要一个测试和细化的过程,在分类阶段,则利用导出的分类模式对其它文档进行分类,将文档分到最可能的类别。

3 文本分类系统的设计与实现

通过对系统功能的分析和基于对文本分类相关算法的研究,系统主要包括训练和分类过程。图 1 是系统结构框架图。

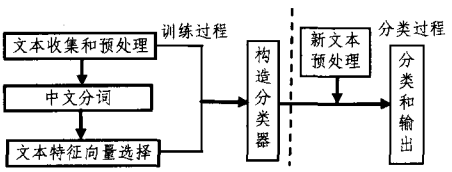

图 1 系统结构框架图

两个过程的功能如下:

(1)训练过程

训练过程包括海洋信息文本收集和预处理、中文分词、文

*)基金项目:国家自然科学基金(No. 60573017),中国'数字海洋'信息基础框架构建专项项目(908-03-01-13)。徐大伟 讲师,硕士,主要研究方向为系统软件,软件工程;董 渊 讲师,博士,研究方向为系统软件、软件工程;张素琴 教授,研究方向为程序设计语言设计与实现、编译优化。

本特征向量的提取及分类器的构建等部分。海洋信息文本收集和预处理是文本分类的基础。由于预先没有现成的可分类文档，因此需要设计和实现海洋信息文本收集器，对收集来的海洋信息进行文本预处理。中文分词是中文文本分类的重要组成部分，也是海洋信息文本分类的重要组成部分，在该过程中，需要把文本都分成中文词汇，并记录其相关的权重。

文本特征向量的提取是文本分类的关键步骤，需要把中文分词的结果中能代表分类特征的词汇提取出来，形成向量，并计算其权重，为训练做准备。

分类器的构建是对文本特征向量的训练过程。采用比较成熟的朴素贝叶斯算法来对文本特征向量进行训练，并得出可靠的训练模型来指导分类。

(2)分类过程

分类过程包括对待分类的海洋信息文本预处理及分类和输出两部分。同样，为了能够对新的海洋信息文本进行分类，首先也必须对其进行预处理，得到其特征向量的权重，然后使用已经获取的训练模型来对海洋信息文本进行分类。

海洋领域知识的文本分类系统是基于 java 并利用中间件技术进行构建的。系统采用 Bridge 与 Abstract Factory 设计模式的结合，分离模块接口与实现部分同时提高系统可扩充性。系统的核心模块包括分词模块、特征选取模块、文本分类模块。

3.1 分词模块

分词模块主要功能是从训练文本和测试文本提取词条，并过滤停用词和常用词。在分词阶段常用的方法包括 3 大类[3]：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法；常用的分词算法包括正向最大匹配法(FMM)、逆向最大匹配法(BMM)、双向匹配法(BM)、最佳匹配法(OM)等。系统采用正向最大匹配(FMM)[3]分词算法，FMM 算法分词的准确度跟分词词典的选择有很大关系，选择合适的分词词典是关键。

3.2 特征选取模块

特征选取模块的功能是处理测试文本。把测试文本经过分词得到的词条，利用特征选取算法提取出一部分数量最具分类特征的词条，供文本分类模块进行分类。常用的特征选取方法有[4]：文档频率(DF)、信息增益(IG)、互信息(MI)、χ^2 统计方法(CHI)等。

模块算法采用改进的 χ^2 统计方法，χ^2 统计方法度量词条 W 和文档类别 C 之间的相关程度，并假设 W 和 C 之间符合具有一阶自由度的 χ^2 分布。词条对于某类的 χ^2 统计值越高，它与该类之间的相关性越大，携带的类别信息也较多。χ^2 统计评估定义[4]如下：

$$\chi^2(w,c)=\frac{m[P(w,c)P(\bar{w},\bar{c})-P(w,\bar{c})P(\bar{w},c)]^2}{P(w)P(\bar{w})P(c)P(\bar{c})} \quad (1)$$

直观地看，$\chi^2(w,c)$的值越小，说明特征词条 w 关于类 c 的独立程度越高，因此选择那些 $\chi^2(w,c)$值最大的特征词条。

假设 X 是词条 w 出现在类 c 中的次数，Y 是 w 出现但不在类 c 中的次数，Z 是 w 不出现的次数，Q 是 w 和 c 都不出现的次数，N 是文档的总数，词条好坏的评估定义为：

$$\chi^2(w,c)=\frac{N\times(XQ-ZY)^2}{(X+Z)(Y+Q)(X+Y)(Z+Q)} \quad (2)$$

如果词条 w 和类 c 是相互独立的，χ^2 统计为零。在训练集中的每个词条和类之间计算每个类的 χ^2 统计，然后结合每个词条针对某个类的得分，得到如下的分数：

$$\chi^2_{avg}(w)=\sum_{i=1}^{m}P(c)\chi^2(w,c) \quad (3)$$

$$\chi^2_{max}(w)=\max_{i=1}^{m}\{\chi^2(w,c)\} \quad (4)$$

χ^2 统计得分的计算有二次复杂度，相似于互信息和信息增益。在 χ^2 统计和互信息之间主要的不同是 χ^2 是规格化评价，因而 χ^2 评估对在同类中的词是可比的。但是 χ^2 统计对于低频词来说是不可靠的。

如果对以上的 $\chi^2(w,c)$公式做个改进，将其求平方根，进一步强调特征词条 w 和类 c 之间的相关性，其中 $\sqrt{P(w)P(\bar{w})}$ 和 $\sqrt{P(c)P(\bar{c})}$ 在修改以后的公式中几乎不强调什么特征，因为对于类而言这些因素的值很小，只起到一个极小的微调作用，而$\sqrt{m}$在分子中的作用可以忽略。

将以上 3 个因素从 $\chi^2(w,c)$的平方根中去掉，将得到简化的 χ^2 统计方法：

$$\chi^2(w,c)=P(w,c)P(\bar{w},\bar{c})-P(w,\bar{c})P(\bar{w},c) \quad (5)$$

通过实验证明了简化后的 χ^2 方法比原来的 χ^2 方法要好，大幅度地减少了特征，同时也缩小了计算的工作量。

3.3 文本分类模块

分类方面，目前机器学习的文本分类方法逐渐替代了知识工程的分类方法。基于机器学习的自动分类方法有贝叶斯分类、决策树、最近邻分类和支持向量机等[5]。所构造的分类器很多，典型的有朴素贝叶斯分类器[5]，基于向量空间模型的分类器[6]和用支持向量机建立的分类器[7]等。文本分类模块的功能是实现分类算法，利用训练文本数据，为经过分词和特征选取后的测试文本确定类别。

系统采用朴素贝叶斯分类算法，构造朴素贝叶斯分类器。它的基本思想是利用特征和分类的联合概率来估计给定文档的分类概率。文档向量的分量为相应特征在该文档中出现的频度。文档属于 C 类的概率可以表示为：

$$P(C\mid Doc)=\frac{P(C)\prod P(F_j\mid C)^{TF(F_j,Doc)}}{\sum_i P(C_i)\prod_{F_i\in V}P(F_i\mid C_i)^{TF(F_i,Doc)}} \quad (6)$$

$$P(F_j\mid C)=\frac{1+TF(F_j,C)}{|V|+\sum_i TF(F_i,C)} \quad (7)$$

其中 $P(C)$为该文档属于 C 类的概率，$P(F_j\mid C)$是对在 C 类文档中特征 F_j 出现的条件概率的拉普拉斯概率估计，$TF(F_j,C)$是 C 类文档中特征 F_j 出现的频度。$|V|$为特征词典集的大小，其值等于文档表示中所包含的不同特征的总数目。$TF(F_j,Doc)$是该文档中特征 F_j 出现的频度。

朴素贝叶斯假设文本是基于特征的一元模型，即文档中特征的出现只与文档类别有关，与文档中的其它特征及文档长度无关。也就是说，特征与特征之间彼此相互独立。

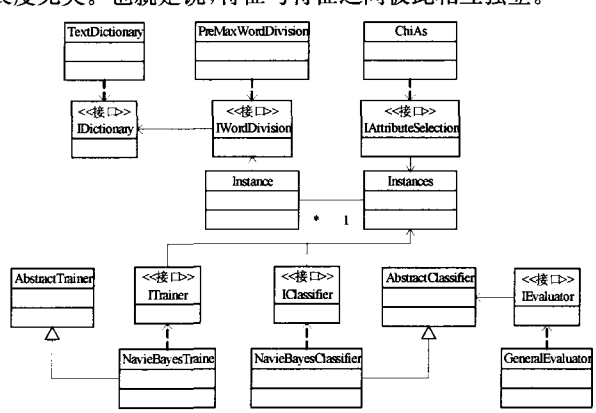

图 2 系统类图

系统类图如图 2 所示，从系统类图中也可以看到系统设计是以分词接口 IWordDivision、特征选择接口 IAttributeSe-

lection、训练接口 ITrainer、分类接口 IClassifier、评估接口 IEvaluator 及 Instance 与 Instances 两个数据类为核心，其中 Instance 代表一篇文档，Instances 代表文档集。

4 实验结果与分析

文本分类从根本上说是一个映射过程，因此评价分类系统的标志是映射的准确程度和映射的速度。评价分类效果的标准很多，本系统采用准确率[8]和查全率[8]作为评价标准。

实验过程中，把海洋类信息网站上收集来的信息分为海洋调查与观测、区域海洋学、海洋基础科学、海洋资源与开发、海洋工程、海洋环境科学、潜水医学、军事海洋学 8 个大类，其中各大类又分成众多子类别。以海洋基础科学为例，分为：海洋水文学、海洋气象学、海洋物理学、海洋化学、海洋生物学、海洋地质学、海洋地貌学、海洋地球物理学。在实验初期，我们选择海洋基础科学信息这一大类共 240 篇文档，平均每类 30 篇文档。每一类选择训练集和测试集的方法如下：将这些分类好的数据平均分成 10 份，选择其中 1 份作为开放测试集，剩余的 9 份作为训练集和封闭测试集。经测试，封闭测试准确率平均达到 86.27%，查全率平均达 86.25%，具有较好的分类效果。表 1 给出有实际意义的开放测试结果。

表 1 分类结果

类别	人工归入	机器正确归入	机器实际归入	准确率 (%)	查全率 (%)
海洋水文学	30	26	29	89.66	86.67
海洋气象学	30	27	31	87.09	90.00
海洋物理学	30	25	30	83.33	83.33
海洋化学	30	28	31	90.32	93.33
海洋生物学	30	26	31	83.87	86.67
海洋地质学	30	23	28	82.14	76.67
海洋地貌学	30	28	33	84.85	93.33
海洋地球物理学	30	24	27	88.89	80.0

结束语 本文主要讨论基于海洋信息处理的文本分类系统的设计与实现，提出系统结构，详细介绍训练和分类的步骤及所使用的相关算法，提出一种改进 χ^2 特征选取算法，最后给出实验结果和分析。从实验结果看，系统具有较好的查全率和准确率。进一步的研究工作改进是分词和分类算法，尝试几种不同的分类算法，并比较其性能，提高分类准确度，完善分类评估体系。

参 考 文 献

[1] 刘宝银，张杰. 海洋科学的前沿——"数字海洋"[J]. 地球信息科学，2000(1)：8-10

[2] 苏新宁，杨建林，江念南，等. 数据仓库和数据挖掘[M]. 北京：清华大学出版社，2006

[3] 朱珣. 中文自动分词系统的研究[D]. 武汉：华中师范大学，2004

[4] 刘丽珍，宋瀚涛. 文本分类中的特征提取[J]. 计算机工程，2004(4)：14-16

[5] Yang Yiming, Liu Xin. A Re-Examination of Text Categorization Methods[J]// 22nd Annual International SIGIR. 1999：42-49

[6] Wu T-F, Lin C-J, Weng R C. Probability estimates for multi-class classification by pair wise coupling. J. of Machine Learning Research, 2004(5)：975-1005

[7] Zhang J, Yang Y. Robustness of regularized linear classification methods in text categorization// SIGIR'03. 2003：190-197

[8] 宋枫溪，高林. 文本分类器性能评估指标[J]. 计算机工程，2004(13)：107-110

(上接第 136 页)

两个背景对应的标尺如图 3。

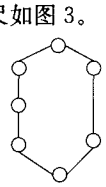

图 3 双序标尺

从标尺的同构关系可以直接得到"B 良"与关键词"80 多分"相匹配。

结束语 模式匹配不是一个刚性的构架，而是一个从实用主义出发考虑的思想体系，放宽模式匹配原有的某些约束是可以接受的，甚至是有利的。

基于概念格标尺的语义匹配是自然的，符合人们的思维方式，具有一定的理论基础，也便于计算机的实现，从而使计算机具有一定程度的智能，使之具有一定的解决实际的问题的能力。

需要进一步研究的内容和克服的困难在于如何量化多个语义相关概念间的距离，使得各个概念能较为精确地映射到一个标尺，从而使概念在语义上得到较好的匹配。

参 考 文 献

[1] Rahm E, Bernstein P A. A Survery of Approaches to Automatic Schema Matching. The International Journal on Very Large Data Bases[J], 2001, 10(4)：334-350

[2] 张治，车皓阳，施鹏飞. 模式匹配问题的描述框架与算法模型. 模式识别与人工智能[J], 2006, 12(6)：715-721

[3] Ganter B, Wille R. Formal Concept Analysis: Mathematical-Foundation[M]. New York: Springer-Verlag, 1999

[4] Dtintsch I, Gediga G. Algebraic aspects of attribute dependencies in information systems. Fundamenta Informaticae[J], 1997, 29：119-133

[5] Dtintsch I, Gediga G. Approximation operators in qualitative data analysis// de Swart H, Orlowska E, Schmidt G, et al., eds. Theory and Application of Relational Structures as Knowledge instruments[M]. Heidelberg: Springer, 2003：216-233

[6] Pagliani P. From concept lattices to approximation spaces: Algebraic structures of some spaces of partial objects. Fundamenta in Formaticae[J], 1993, 18(1)：1-25

[7] Yao Y Y. Concept lattices in rough set theory//Proceedings of 23rd International Meeting of the North American Fuzzy Information Processing Society. 2004

[8] Krantz D H, Luve R D, Suppes P, et al. Foundation of Measurement[M]. New York: Academic Press, 1971, 1, 1989, 2, 1990, 3

[9] Ganter B, Wille R. Conceptual scaling// Fred S. Roberts, ed. Applications of combinations and graph theory to the biological and social science[M]. Berlin, Heidelber, New York: Springer-Verlag, 1989：139-167

[10] Reurer K, Wille R. Complete congruence relations of concept lattice. Acta Sci. Math[M], 1987, 51：319-327