Web 文本分类技术研究现状述评

高 淑 琴
(江苏徐州师范大学图书馆 ,徐州 ,221116)

[摘要] 本文在分析国内外 Web 文本分类方法研究现状的基础上 ,对新近出现的基于群的分类方法、基于模糊 —粗糙集的文本分类模型、多分类器融合的方法、基于 RBF 网络的文本分类模型、潜在语义分类模型等新方法 ,以及 K —近邻算法和支持向量机的新发展等进行了深入探讨 ;并对 Web 文本分类过程的几个关键技术 :文本预处理、文本表示、特征降维、训练方法和分类算法进行了分析 ;最后总结了 Web 文本分类技术存在着新分类方法不断涌现、传统分类方法的进一步发展、文本、语音和图像分类技术的融合等几种发展趋势 ,以及存在着分词问题、目前还没有发现“最佳”的特征选择等研究的不足之处。
[关键词] Web 文本分类 数据挖掘 机器学习
[中图分类号] G254.11 [文献标识码] A [文章编号] 1003-2797(2008)03-0081-06

[Abstract] This article has analyzed the research present situation of domestic and foreign Web text classification method firstly , has analyzed the new methods which recently appeared , swarm-based approaches , based on the fuzzy - rough collection text classification model , the multi- sorters fusion method , based on RBF network text classification model , latent semantic classification model and so on , as well as the recent development of the K-NN and the support vector machine (SVM) method ; And has discussed the Web text classification process several essential technologies : The text pretreatment , the text expressed , the characteristic fell Uygur , the training method and the classified algorithm ; Finally summarized the development tendency and deficiency of Web text classification technology.
[Key words] Web text classification Data mining Machine learning

1 Web 文本分类方法的研究现状

在 Web 出现之前 ,人们已经对文本自动分类问题进行了大量的研究 ,形成了文档自动分类技术。随着 Web 上海量的文本信息的增加 ,文档自动分类技术的处理对象从普通的文档扩展到了 Web 文本。很显然 ,文档自动分类技术也成为 Web 文本分类技术的基础。

1.1 国外 Web 文本分类方法的研究现状

国外对文本自动分类的研究开展较早 ,50 年代末 ,H. P. LUhn 在这个领域进行了开创性的研究 ,提出了基于词频统计思想的文本自动分类方法。1960 年 ,Maron 发表了关于自动分类算法的第一篇论文 ,随后以 K. Spark , G. Salton 以及 K. S. Jones 等人为代表的众多学者也在这一领域进行了很有成效的研究工作[1] , 目前国外的文本分类研究已经从实验性阶段进入到了实用化阶段 ,并在邮件分类 ,电子会议等方面取得了广泛的应用 ,其中较为成功的有麻省理工学院为白宫开发的邮件分类系统和卡内基集团为路透社开发的 construe 系统[2] 。

1.2 国内 Web 文本分类方法的研究现状

相比于英文文本分类 ,中文文本分类的一个重要的差别在于预处理阶段 :中文文本的读取需要分词 ,不像英文文本的单词那样有空格来区分。在很长一段时间内 ,中文文本分类的研究没有公开的数据集 ,使得分类算法难以比较。现在一般采用的中文测试集有 :北京大学建立的人民日报语料库、清华大学建立的现代汉语语料库等。其实一旦经过预处理将中文文本变成了样本矢量的数据矩阵 ,那么随后的文本分类过程和英文文本分类相同 ,也就是随后的文本分类过程独立于语种。因此 , 当前的中文

[作者简介] 高淑琴 ,女 ,1968 年生 ,馆员 ,从事图书馆数字信息资源建设研究。

文本分类主要集中在如何利用中文本身的一些特征来更好地表示文本样本。国内对于文本自动分类的研究起步较晚,但从简单的查词典的方法,到后来的基于统计语言模型的分词方法,中文分词的技术已趋于成熟。

1981 年,侯汉清教授对计算机在文本分类工作中的应用作了探讨和阐述[3]。此后,我国陆续研究产生了一些文本分类系统,其中具有代表性的有上海交通大学研制的基于神经网络算法的中文自动分类系统,清华大学的自动分类系统等。同时在不同的分类算法方面也展开了广泛的研究和实现,中科院计算所的李晓黎、史忠植等人应用概念推理网进行文本分类[4],召回率达到 94. 2%,准确率达到 99. 4%。中国科技大学的范众等人在 KNN、贝叶斯和文档相似性研究的基础上提出了一个超文本协调分类器[5],正确率接近 80%,它的特点是适当的考虑了 HTML 文本中结构化信息。复旦大学和富士通研究中心的黄萱菁、吴立德等人研究了独立语种的文本分类[6],并以词汇和类别的互信息量为评分函数,考虑了单分类和多分类,最好的召回率为 88. 87%。上海交通大学的刁倩、王永成等人结合词权重和分类算法进行分类[5],基于 VSM 的封闭式测试实验中分类正确率达到 97%。

目前,一些比较成熟的经典的文本分类算法已经被应用到了 Web 文本分类中,主要包括:决策树方法(经典的决策树算法主要包括:ID3 算法、C4. 5 算法和 CART 算法等)、神经网络方法(BP 算法)、遗传算法(GABIL 系统)、贝叶斯分类、K2 近邻算法和基于案例的推理。粗糙集方法、模糊集方法和支持向量机是较新的分类方法。

1.3　文本分类方法的新发展

近年来,文本分类已成为众多领域研究者的热门研究课题,研究者们从不同的角度把越来越多的知识引入文本分类领域,推动着文本分类的不断发展,产生了许多新的方法。

(1)基于群的分类方法(swarm-based approaches)。这种方法可以看作是进化计算的一个新的分支,它模拟了生物界中蚁群、鱼群和鸟群在觅食或者逃避敌人时的行为。纵观文献中对基于群的分类方法的研究,我们将这种方法分为两类:一类是蚁群算法或者蚁群优化(ant colony optimization ACO),另一类称为 Particle Swarm Optimisers(PSO)。用蚁群优化来进行分类规则挖掘的算法称为 Ant-Miner[8-9],Ant-Miner 是将数据挖掘的概念和原理与生物界中蚁群的行为结合起来形成的新算法。PSO 是进化计算的一个新的分支,它模拟了鱼群或鸟群的行为。PSO 将群中的个体称为 Particles,整个群称为 swarm。在优化领域,PSO 可以与遗传算法相媲美。对 ACO 或者 PSO 在数据挖掘中应用的研究仍处于早期阶段,要将这些方法用到实际的大规模数据挖掘中还需要做大量的研究工作。

(2)基于模糊—粗糙集的文本分类模型。文本分类过程中由于同义词、多义词、近义词的存在导致许多类并不能完全划分开来,造成类之间的边界模糊。此外交叉学科的发展,使得类之间出现重叠,于是造成许多文本信息并非绝对属于某个类。这两种情况均会导致分类有偏差,针对上述情形,文献[10]提出利用粗糙-模糊集理论结合 KNN 方法来处理在文本分类问题中出现的这些偏差。模糊-粗糙集[11-12]理论有机的结合了模糊集理论与粗糙集理论在处理不确定信息方面的能力。粗糙集理论体现了由于属性不足引起集合中对象间的不可区分性,即由于知识的粒度而导致的粗糙性;而模糊集理论则对集合中子类边界的不清晰定义进行了模型化,反映了由于类别之间的重叠体现出的隶属边界的模糊性。它们处理的是两种不同类别的模糊和不确定性。将两者结合起来的模糊-粗糙集理论能更好地处理不完全知识。

(3)多分类器融合(fusion)的方法。实际应用的复杂性和数据的多样性往往使得单一的分类方法不够有效。因此学者们对多种分类方法的融合(fusion)进行了广泛的研究,取得了一系列研究成果。纵观文献中的研究,可以大致将多分类器的融合技术分为以下几类:投票机制(voting)、行为知识空间方法(Behavior-Knowledge Space BKS)、证据理论(Dempster-Shafer theory)、贝叶斯方法和遗传编程(genetic programming GP)。采用投票机制的方法主要有装袋(bagging[13])和推进(boosting[14])。Buhlmann P. 和 Yu B[15]对 bagging 进行了深入的分析;Hothorn T. and Lausen B. 在文献[16-17]中将 bagging 用于决策树,并在文献[18]中对 bagging 进行了发展,采用 boot strap aggregation 来融合分类器。而 Schwenk 和 Bengio 则将 boosting 用于神经网络,提出了 AdaBoost 方法[19],从而提高了神经网络的预测精度。文献[20]采用 BKS 进行分类器融合。文献[21]用证据理论将 4 个不同的分类方法(SVM,KNN,KNN Model-based approach 和 Rocchio)结合起来,形成融合的分类器。用贝叶斯方法进行分类器融合有两种情况:一种是有独立性假设

的贝叶斯方法[22],另一种是没有独立性假设的贝叶斯方法[23-24]。

(4) 基于 RBF 网络的文本分类模型。基于 RBF 网络的文本分类模型把监督方法和非监督方法相结合,通过两层映射关系对文本进行分类,首先利用非监督聚类方法根据文本本身的相似性聚出若干个簇,使得每个簇内部的相似性尽可能高而簇之间的相似性尽可能低,并由此产生第一层映射关系,即文本到簇的映射,然后通过监督学习方法构造出第二层映射关系,即簇集到目标类集合的映射[25]。然后为每一个簇定义一个相应的径向基函数(Radial Basis Function,RBF),并确定这些基函数的中心和宽度,利用这些径向基函数的线形组合来拟合训练文本,利用矩阵运算得到线性组合中的权值,在计算权值时,为了避免产生过度拟合的现象,采用了岭回归技术,即在代价函数中加入包含适当正规化参数的权值惩罚项,以保证网络输出函数具有一定平滑度。

(5) 潜在语义分类模型。潜在语义索引方法,已经被证明是对传统的向量空间技术的一种改良,可以达到消除词之间的相关性,化简文档向量的目的,然而 LSI 在降低维数的同时也会丢失一些关键信息。LSI 基于文档的词信息来构建语义空间,得到的特征空间会保留原始文档矩阵中最主要的全局信息。但在某些情况下,一些对特定类别的正确分类非常重要的特征,因为放在全局下考虑显得不重要,而在维数约减的过程中被滤掉;该情况对稀有类别尤为明显。而如果这样,稀有类的分类性能就肯定会受到影响。针对上述问题,在扩展 LSI 模型的基础上,文献[26]提出了一种新的文本分类模型:潜在语义分类模型(Latent Semantic Classification:LSC)[27-28]。与 LSI 模型类似,文献[29]希望从原始文档空间中得到一个语义空间;然而不同的是,通过第二类潜在变量的加入,把训练集文档的类别信息引入到了语义空间中。这样,就可以得到比 LSI 模型的语义空间更适合文本分类的语义空间。

(6) K—近邻算法(K-NN)的新发展:K-NN 是一种有效的分类方法,但是它有两个最大的缺陷:第一,由于要存储所有的训练实例,所以对大规模数据集进行分类是低效的;第二,K-NN 分类的效果在很大程度上依赖于 k 值选择的好坏。Gongde Guo 和 Hui Wang 等人[30-31]针对 K-NN 的两个缺陷,提出了一种新颖的 KNN 类型的分类方法,称为基于 KNN 模型的分类方法。新方法构造数据集的 KNN 模型,以此代替原数据集作为分类的基础,而且新方法中 k 值根据不同的数据集自动选择,这样减少了对 k 值的依赖,提高了分类速度和精确度。实验证明,基于 KNN 模型的方法在分类精确度上与 C5.0 和标准的 K-NN 相当。另外,针对 K-NN 方法的第一个缺陷,Nong Ye and Xiangyang Li 将聚类方法和经典的 K-NN 方法结合起来,提出了一种新颖的分类方法,称为 CCA-S[32]。CCA-S 能够处理大规模数据集,可伸缩性好,并且支持增量式学习。但 CCA-S 只能处理连续属性,而且只针对类别为两类的分类问题。如何扩展 CCA-S,以使其能够处理多类别的问题,还有待进一步研究。文献[33]将遗传算法和 KNN-Fuzzy 方法[34]结合起来,用遗传算法来寻找最优的 k 值,从而优化 KNN-Fuzzy 方法,提高了分类精确度。文献[35]基于模糊粗集理论提出了一种新的 KNN 分类方法,与传统的 NN 和 fuzzy NN 相比,新方法有更高的预测精度。

(7) 支持向量机(SVM)方法的新发展:SVM 是进行分类、聚类和时间序列分析的有效数据挖掘工具。但是,由于 SVM 的训练时间会随着数据集的增大而增加,所以在处理大规模数据集时,SVM 往往需要较长的训练时间。而实际的数据挖掘应用往往包含了数以百万计的数据,这使得 SVM 很难发挥作用。针对这个问题,文献[36-37]用选择性采样或者主动学习方法来训练 SVM,它的基本思想是从整个训练数据集中选择一小部分最有代表性的数据来最大化学习效果。与主动学习思想类似的方法还有随机采样[38-39]。这两种方法都需要对数据集进行多遍扫描。与主动学习和随机采样不同,文献[40]将层次聚类用于 SVM,以加快 SVM 对大规模数据的处理速度。文中所提出的新方法称为 Clustering-Based SVM(CB-SVM),CB-SVM 用一个层次微聚类(hierarchical micro-clustering)算法对整个数据集进行单遍扫描,为 SVM 提供带有整个数据集统计概要信息的高质量样本。CB-SVM 对于大规模数据集有很好的伸缩性,而且有较高的分类精确度。另外,为了解决实际应用中数据集大小动态变化的问题,Fung G. 和 Mangasarian O.L[41]提出了增量式 SVM,新提出的方法用于二分类问题。文献[42]在文献[43]的基础上,提出了一种有效的基于内存的增量算法,以支持多分类问题。除了以上所归纳的几种新的方法之外,还有基于粒度计算的分类方法、基于投影寻踪回归的文本模型及一些正处于探索阶段的新方法。

2　Web 文本分类中的关键技术

在对 Web 文本进行分类的过程中,包括几个关键步骤:文本预处理、文本表示、特征降维、训练方法和分类算法,这些关键技术的研究和实现对最终的分类算法都有一定程度上的影响。

2.1　Web 文本预处理

Web 文本作为一种非结构化的数据类型,其特点表现为特征空间的高维性、文本特征表示向量的稀疏性及文本主题特征表现不突出等特点。与数据库和数据仓库中的结构化数据相比,Web 文本具有有限的结构或者根本就没有结构。文本信息源的这些特征使得现有的数据挖掘技术无法直接应用于其上,因此需要对 Web 文本进行预处理,抽取其特征并用结构化的形式保存,作为文本的中间表示形式。

文本预处理即去掉一些标记,例如 HTML 中的 Tag,去除停用词、词根还原。对于中文文本而言,因为词与词之间没有明显的切分标志,所以需要分词。此外还需要进行词性标记、短语识别等。

汉语自动分词是机器翻译、文献标引、智能检索、自然语言理解与处理的基础,也是中文文本分类的一个关键的环节。自从 20 世纪 80 年代初自动分词被提出以来,有众多的专家和学者为之付出了不懈的努力,涌现了许多成功的汉语分词系统,主要有北京航空航天大学研制的 CDWS 和 CWSS 分词系统,分词速度为 200 字每秒[44]。清华大学黄昌宁、马晏等开发的 SEG 系统,分词速度为 258 字每秒,正确率为 99.3%[45]。东北大学姚天顺建立的基于规则的汉语分词系统;南京大学王启祥等人实现的 WSBN 分词系统[46]。中科院计算所研制出的汉语词法分析系统 ICTCLAS 等等[47]。

汉语自动分词系统的实现及效果依赖于分词理论与方法。目前国内分词系统所采用的或者正在研究的方法基本上分为三类:基于字符串匹配的方法、基于理解的方法和基于统计的方法。

2.2　文本表示

Web 文档的内容是用自然语言描述的,计算机很难处理其语义,为了便于计算机的处理,所以必须将文本的内容特征转化为计算机可以处理的格式。目前文本的表示模型有多种:布尔逻辑型、向量空间型(VSM)、潜在语义索引模型、概率型以及混合型等。向量空间模型是近几年来信息检索领域应用较广且效果比较好的模型。

2.3　特征降维

训练文本和待分类文本经过分词并去除停用词和高频词后,表示文本的向量空间和类别向量的维数也是相当大的,因此需要进行特征降维。是否进行特征降维对文本分类的训练时间、分类准确性都有显著的影响,而且分类器的算法和实现的复杂度都随特征空间维数的增加而增加。所以,特征集的降维操作是文本分类准确率和效率的关键。特征选择(Feature Selection)和特征抽取(Feature Extraction)是特征降维中的主要方法。

(1)特征选择(Feature Selection)。特征选择就是从特征集 $T = \{t_1, \cdots, t_s\}$ 中选择一个真子集 $T' = \{t_1, \cdots, t_{s'}\}$,满足($s' < s$)。其中,s 为原始特征集的大小,s 为选择后的特征集大小。选择的准则是经特征选择后能有效提高文本准确率。选择没有改变原始特征空间的性质,只是从原始特征空间中选择了一部分重要的特征,组成一个新的低维空间[48-49]。

文献[50]中认为文本分类中,用于特征选择的统计量大致有:特征频度(Term Frequency),文档频度(Document Frequency),特征熵(Term Entropy),互信息(Multi-Information),信息增益(Information Gain),X^2 统计量(Chi-square),特征权(Term Strength),期望交叉熵(Expected Cross Entropy),文本证据权(Weight of Evidence for Text),几率比(Odds Ratio)等。这些统计量从不同的角度度量特征对分类所起的作用。

(2)特征抽取(Feature Extraction)。特征抽取也叫特征重参数化(Feature Reparameterization)[51]。特征抽取是文本分类系统中十分关键的问题,它可降低向量空间的维数,提高系统的速度和精度,还可以防止过拟合。由于自然语言中存在大量的多义词、同义词现象,特征集无法生成一个最优的特征空间对文本内容进行描述。特征抽取是将原始特征空间进行变换,重新生成一个维数更小、各维之间更独立的特征空间。常用的特征抽取方法可以分为三类:主成分分析、潜在语义标引、非负矩阵分解。

2.4　训练方法和分类算法

Web 文本分类是一个典型的有教师的机器学习问题,一般的可分为训练和分类两个阶段。其中训练算法的工作是对训练文档集合中每篇文本对应的词表进行统计,计算出类别向量矩阵同时进行归一化,最后保存训练得到的向量表,即得到了分类知识库;分类算法(也可称为识别算法)则依据训练得到的分类知识库,并用一定的算法对待分类文本进行分类。

3 研究现状分析

通过以上分析,我们可以看出 Web 文本分类技术存在以下几种发展趋势:

一是新分类方法不断涌现,比如基于群的分类方法和基于粒度计算的分类方法。新分类方法出现得益于人工智能、机器学习、进化计算和粒度计算等领域中新技术的涌现和发展。

二是传统分类方法的进一步发展,比如支持向量机的不断改进和 KNN 方法的发展。传统分类方法的发展主要利用了机器学习、进化计算、数据挖掘、模糊集和粗糙集等理论中的原理和方法。

三是根据实际问题需要,有针对性地综合众多领域的技术,以提高分类的性能。

四是文本、语音和图像分类技术的融合,随着互联网和多媒体技术的进一步发展,文本分类技术将与图像识别、语音识别融合,比如图像文本的分类、语音文本的分类、多媒体数据库索引等。

目前国内对 Web 文本分类的研究还没有到达一个成熟的阶段,其中还存在一些有待进一步研究的问题:

(1)分词是影响文本分类的重要因素之一,分词的速度和准确率与最终的分类结果密切相关。尤其是 Web 上不断出现新词汇,对分词理论的创新和词典的构造都提出了较高的要求。就中文文档分类而言,分词是一项非常复杂的工作,分类系统一般都比较复杂和庞大,分词速度慢,且准确度不高,因此,研究无须词典支持、领域独立的文本分类系统无疑具有重要价值,这使得文档分类系统成为真正意义上的通用系统。

(2)目前还没有发现"最佳"的特征选择方法,针对中文 Web 文本分类的组织特点,需要结合特定的特征选择,因此在使用不同分类算法时如何选择最佳的特征选择方法也是我们需要深入研究的问题。

(3)由于中文文本分类起步晚和中文不同于英文的特性,目前中文 Web 文本分类还没有标准的开放的文本测试集,各研究者大多使用自己建立的文本集进行训练和测试,其分类结果没有可比性,不利于交流和提高。一般地,训练文档集应该是公认的经人工分类的语料库。国外文档研究都使用共同的测试文档库,这样就可以比较不同分类方法和系统的性能,而就中文文档分类而言,各研究者使用自己建立的训练文档库进行测试,测试结果没有可比性,这一现状应当引起国内文本处理研究者的重视。

(4)将自然语言理解和处理技术、语义 Web 概念、Agent 技术和机器翻译等技术应用于 Web 自动文本分类中,进一步解决中文文本分类的难点,提高文本分类的智能化水平。

(5)目前存在多种成熟的文本分类算法,大部分分类系统都是应用某一种分类算法,分类性能受到制约。

参考文献

1 王继成等. Web 文本挖掘技术研究. 计算机研究与发展, 2000(5)

2 王本年等. Web 智能研究现状与发展趋势. 计算机研究与发展,2005(5)

3 侯汉清. 分类法的发展趋势简论. 北京:中国人民大学出版社,1981.

4 李晓黎等. 概念推理网及其在文本分类中的应用. 计算机研究与发展,2000(9)

5,7,44 张滨. 中文文档分类技术研究. 武汉大学硕士学位论文,2004.

6 黄萱菁,吴立德. 独立于语种的文本分类方法. 2000 International Conference on Multilingual Information Processing,2000:37-43

8 Parpinelli R S, Lopes H S, Freitas A A. Data Mining with an Ant Colony Optimization Algorithm. IEEE Trans. on Evolutionary Computation, 2002, special issue on Ant Colony algorithms.

9 Parpinelli R S, Lopes H S, Freitas A A. Mining Comprehensible Rules from Data with an Ant Colony Algorithm. In: Bitten court G, Ramalho,. eds. SBIA 2002, LNAI 2507, 2002. 259-269

10 付雪峰,王明文. 基于模糊 —粗糙集的文本分类方法. 2004 年度全国搜索引擎和网上信息挖掘学术研讨会,华南理工大学学报(自然科学版),2004,32:73-76

11 Yao, Y Y. A Comparative Study of Fuzzy Sets and Rough Sets. Information Sciences,1998,109(1-4):227-242

12 Dubois D, Prade H. Putting Rough Sets and Fuzzy Sets Together. Intelligent Decision Support: Handbook of Applications and Advanced of the Rough Set Theory. Boston: Slowinski R ED, Kluwer Academic Publishers, 1992,203-222

13 Breiman L. Bagging predictors. Machine Learning, 1996, 24:123-140

14 Schapire F, Freund Y, Schapire R E. Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the thirteenth International Conference, Morgan Kaufmann, 1996. 148-156

15 Buhlmann P, Yu B. Analyzing bagging. The Annals of Statistics, 2002,30:927-961

16 Hot horn T, Lausen B. Building classifiers by bagging trees. Preprint, Friedrich Alexander University Erlangen

Nuremberg ,URL. http ://www. mathpreprints. com/

17 Hot horn T , Lausen B , Benner A. et al. Bagging survival trees ,Statistics in Medicine ,2004 ,23 :77-91

18 Hot horn T , Lausen B. Double2bagging :Combining classifiers by boot strap aggregation , Pattern Recognition , 2003 ,36 :1303-1309

19 Schwenk H , Bengio Y. Boosting neural networks. Neural Computation , 2000 ,12 (8) :1869-1887

20 Huang Y S , Suen C Y. A Met hod of Combining Multiple Experts fort he Recognition of Unconst rained Handwritten Numerals. IEEE Trans on Pat tern Analysis and Machine Intelligence ,1995 ,17 (1) :90-94

21 Bi Yaxin , et al. Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization. In :MDAI2004 , 2004. 127-138

22 Xu L , Krzyzak A , Suen C Y. Several Methods for Combining Multiple Classifiers and Their Applications in Handwritten Character Recognition. IEEE Trans on System , Man and Cybernetics ,1992 ,22 (3) :418-435

23 Kang H J , Kim K , Kim J H. Optimal Approximation of Discrete Probability Distribution with kth-2order Dependency and Its Applications to Combining Multiple Classifiers. PRL ,1997 ,18 (6) :515-523

24 Kang Hee-Joong , Doermann D. Combining Multiple Classifiers Based on Third Order Dependency. In : Proceedings of t he Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) .

25 王晓庆. 基于 RBF 的文本自动分类的研究. 南昌 :江西师范大学计算机信息工程学院 ,2003.

26 ,27 ,29 曾雪强等. 一种基于潜在语义结构的文本分类模型. 2004 年度全国搜索引擎和网上信息挖掘学术研讨会 ,华南理工大学学报 (自然科学版) ,2004 ,32 :99-102

28 Wang M , Nie J . A Latent Semantic Structure Model for Text Classification. ACM-SIGIR-2003 , Workshop on Mathematic/ Formal Methods in Information Retrieval , Toronto , Canada , 2003.

30 Guo Gongde ,et al. KNN Model-Based Approach in Classification. In :Coop IS/ DOA/ ODBASE 2003 ,2003. 986-996

31 Guo Gongde , et al. A KNN Model-Based Approach and Its Application in Text Categorization. In :CICLing2004 , 2004. 559-570

32 Ye Nong , Li Xiangyang. A machine learning algorithm based on supervised clustering and classification. In : Liu J , et al , eds. AMT 2001 , LNCS 2252 , 2001. 327-334

33 Rosa L A , Ebecken N F F. Data mining for data classification based on the KNN-Fuzzy met had supported by genetic algorithm. In : VECPAR 2002 , LNCS 2565 , 2003. 126-133

34 Keller J M , Gray M R , Givens J A J r. A Fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on Systems , Man and Cybernetics ,1985 ,SMC215 (4) :258-260

35 Bian H Y. Fuzzy Rough Nearest Neighbor Classification : an Integrated Framework. In : Proceedings of IASTED International Symposium on Artificial intelligence and Applications , 2002. 160-164

36 Schohn G , Cohn D. Less : Active learning wit h support vector machines. In : Proc. 17th Int Conf. Machine Learning ,Stanford ,CA ,2000.

37 Tong S , Koller D. Support vector machine active learning with applications to text classification. In : Proc. 17th Int Conf. Machine Learning , Stanford , CA , 2000.

38 Balczar O W J L , Dai Y. A random sampling technique for training support vector machines. In : The 2001 IEEE Int Conf. Data Mining , San Jose , CA , 2001.

39 Lee YJ , Mangasarian O L. RSVM : Reduced support vector machines. First SIAM Int Conf Data Mining , Chicago , IL , 2001.

40 Yu Hwanjo , Yang Jiong , Han Jiawei. Classifying Large DataSets Using SVMs with Hierarchical Clusters. SIGKDD'03 ,Washington , DC , USA ,Aug. 2003.

41 ,43 Fung G , Mangasarian O L. Incremental support vector machine classification. In : Grossman R , Mannila H , Motwani R ,eds. Proceedings of the Second SIAM International Conference on DataMining ,SIAM (2002) ,2002. 247-260

42 Tveit A , Hetland M L. Multicategory Incremental Proximal Support Vector Classifiers. In : Palade V , Howlett R J , J ain L C , eds. KES 2003 , LNAI 2773 , 2003. 386-92

45 张治平. Web 信息精确获取技术研究. 硕士学位论文 ,国防科学技术大学 ,2004.

46 罗强. 基于粗糙集理论的知识发现在 Web 文本挖掘上的应用研究. 硕士学位论文. 广西大学 ,2003.

47 张海燕. 基于分词的中文文本自动分类研究与实现. 硕士学位论文 ,湖南大学 ,2002.

48 David W Aha , and Richard L Bankert. A comparative evaluation of sequential feature selection algorithms. In : Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics , 1995 :1-7

49 Ron Kohavi , and George H John. Wrappers for feature subset selection. Artificial Intelligence Journal. Special Issue on Relevance , 1997 :273-324

50 陈涛 ,谢阳群. 文本分类中的特征降维方法综述. 情报学报 ,2006 (6)

51 Schutze H , Hull D A and Pedersen J O. A comparison ofclassifiers and document representations for the routing problem. In : Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR295) . 1995. 229-237

(收稿日期 :2007-11-22)