ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No.35, Dec. 2017
E-mail: xsjl@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

新闻文本自动分类技术概述

刘冬瑶,刘世杰,陈宇星,张文波,周振

(中国矿业大学(北京) 机电与信息工程学院,北京 100083)

摘要:文本分类是对文本集按照一定的分类体系或标准划分为不同的类别。该文总结了文本分类的基本流程,讨论了中文文本分类的主要特点和常用技术,指出了现今新闻文本分类存在的问题,并对中文文本分类未来的发展前景和研究方向做出展望。

关键词:文本预处理;新闻文本分类;机器学习;自然语言处理

中图分类号:TP391 文献标识码:A 文章编号:1009-3044(2017)35-0087-05

DOI:10.14004/j.cnki.ckt.2017.4078

The Research Summary of News Text Automatic Classification Technology

LIU Dong-yao, LIU Shi-jie, CHEN Yu-xing, ZHANG Wen-bo, ZHOU Zhen

(China University of Mining & Technology, Beijing 100083,China)

Abstract: The text classification is divided into different categories by the classification of the text set according to certain classification system or standard. This paper summarizes the basic flow of text classification, discusses the characteristics and key technologies of Chinese text classification, points out the existing problems of news text classification, and prospects the future development of Chinese text classification and its research direction.

Key words: text preprocessing; news text automatic classification; machine learning; NLP

1 概述

随着网络信息技术的迅速发展和传统纸媒逐渐向信息化媒体的转型,网络中有越来越多的信息积累,尤其是新闻的无纸化使得人们更倾向于在网络上搜索信息。其中大部分是以文本形式存在。文本分类则能有效解决这一问题,而传统的文本分类主要使用手工分类的途径,这种做法有着很多的弊处:首先,这样会耗费大量的人力、物力;其次,存在获得的成果与所要求的不一致的现象。效率低下的手工分类方式面临愈来愈多的困难,面对大数据更显得无从下手,为了提高分类的准确率和速度,新闻文本自动分类顺理成章地成为了发展方向。

新闻是对时事、最新消息进行了解的重要途径,新闻信息分类有助于实现新闻有序化、对新闻进行挖掘,从而引导决策等,很有意义。新闻文本分类已经有了大量的相关研究,包括分类的流程和大量的相关算法。

本文组织如下,第2节介绍了文本自动分类的三个步骤及各种分类方法,第3节介绍了新闻文本分类的应用方向和现今仍然存在的问题,第4节对新闻分类的成长发展远景及研究方向进行展望。

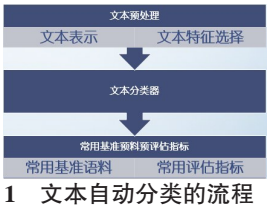

图1 文本自动分类的流程

2 文本自动分类的流程

文本自动分类一般有三个步骤组成:文本预处理、文本分类和常用基准语料预评估。图1为文本自动分类的流程。

2.1 文本预处理

文本预处理是用预先处理原始文本数据的方式,来提高学习算法的精准度、分类效果和文本弹性。

2.1.1 文本表示

一般来说,语言在现实使用中的形式是文本。现实使用中,文本是根据一定的语言衔接和语义连贯规则构成的语句系统。主要采用向量空间模型 VSM (Vector Space Model)来进行文本表示,这种模型将高维词条空间中的向量与文本逐一对应。

1970s,向量空间模型由 Salton 等提出,并应用于有名的 SMART 文本检索系统。把对内容的处理简化为向量的变化,文档间的相似程度可以通过计算向量之间的相似程度来衡量,直观易懂。多数情况下,用余弦距离来进行相似性度量。

M个无序特征项 t_i ,词根/词/短语/其他每个文档 d_j 可以用特征项向量来表示(a_{1j} , a_{2j} ,…, a_{Mj})权重计算,N个训练文档 AM*N= (a_{ij}) 文档相似度比较

收稿日期:2017-10-25
基金项目:国家大学生创新创业训练计划(C201604063)
作者简介:刘冬瑶(1995—),女,黑龙江哈尔滨人,中国矿业大学(北京)本科生,主要研究方向为自然语言处理。

1) Cosine计算，余弦计算的好处是，正好是一个介于0到1的数，如果向量一致就是1，如果正交就是0，符合相似度百分比的特性，余弦的计算方法为，向量内积/各个向量的模的乘积。

2) 内积计算，直接计算内积，计算强度低，但是误差大。

向量空间模型只是一个理论模型，不同的应用中对项的权重评价、相似度的计算有着不同的定义，在各种权重评价函数和确定相似度的办法中可以有不同的表示方法，所以模型适应能力很强，很好地应用于各种不同的系统。

VSM在计算性和操纵性上有着其独特的优势，在大数据时代，它的应用已经有了许多新的扩展，可以帮助人们检索文本、摘录文章段落语句、自动提取关键词等，还可以在数据检索方面代理信息、构建搜索引擎、实现web新闻发布等。在VSM中，用多维空间的一个点来代表文本内容，将文本用向量的形式与实数域中的数一一对应，在模式识别或者其余范畴中的成熟计算方法的辅助下，文本的可计算性和可操作性得到了很大的提高。

2.1.2 文本特征选择

文本特征选择是从原始特征中基于一些准则来选择那些最能将类别区分开的特征词。特征选择筛选出相关性较低和多余的特征并将它们删去，使处理效率大大提升。文本数据的特征选择研究的重点就是用来衡量单词重要性的评估函数，其过程就是首先根据这个评估函数来给每一个单词计算出一个重要性的值，然后根据预先设定好的阈值来选择出所有其值超过这个阈值的单词。目前，国内外常用文本特征选择方法主要有以下几种：文档频率、信息增益、互信息、X^2 统计量、期望交叉熵等。

1) 信息增益

在信息增益中，以特征能给分类程序带来的信息量来作为度量标准的，带来的信息与特征的重要程度成正比。是否有这个特征将为系统的信息量带来波动，而信息量的差值即为系统中受它影响的信息量，换言之即为熵。设有变量X，有n种取值可能，Pi为每种可能被取值的概率，则定义X的熵为

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{1}$$

换言之，X变化的可能性与其能带来的信息量成正比，也就是与熵成正比。对聚类来说，就是文档属于哪个类别的变化越多，类别的信息量就越大。所以特征T给聚类C或分类C带来的信息增益为 $IG(T)=H(C)-H(C|T)$。$H(C|T)$ 有两种可能：一种是出现特征，将其记为T，用t表示，一种是特征T不出现，用 t' 表示。所以 $H(C|T)=P(T)H(C|t)+P(t')H(C|t')$，再通过熵的计算公式求出特征与类别的信息增益公式。

信息增益的最大的不便就是只能判断特征在整个系统中的影响，而判断它在哪个类别中。

2) 文档频率

在文档频率方法中，使用特征词在一个类别中出现的文档数来表示这个特征词与该类别的相关度。更大概率通过筛选的特征词是在更多的文档中出现过的。

文档频率是最简易的特征抽取方法，由于它有基于训练语料规模的线性计算复杂度，更适合大规模的语料统计，能够极大地使效果改善。[1]

3) 互信息

互信息(Mutual Information)是基于信息熵概念上的，它是度量两个随机事件相关性的特征，广泛使用于统计语言模型中。词条(记为t)和文本类别(记为c)的互信息定义是：

$$MI(t,c)=\log\frac{P(tc)}{P(t)\times P(c)} \tag{2}$$

其中 $P(tc)\in$C且包含t的文档概率，$P(t)$ 表示包含词条t的语料中的文本的概率，$P(c)$ 表示C类文本在语料中出现的概率。根据概率，如果在分布上一个词与一个类别是在统计上独立的，即 $P(tc)=P(t)\times P(c)$，则 $MI(t,c)=0$，也就是说词t的频率无法对预测类别C产生影响。

在实际运用中，互信息表达式可近似为语料库中对应的出现频数。如果包含t且属于C的文档频数(记为X)，包含t但不属于C的文档频数(记为Y)，属于C但不包含t的文档频数(记为Z)，语料中文本总数(记为N)，则有：

$$MI(t,C)=\log\frac{X\times N}{(X+Y)\times(X+Z)} \tag{3}$$

对于属于不止一个类别的应用，算出t在每一类中的MI值，再算出对整个语料而言t的MI值：

$$MI(t)=\max_{i=1}^{m}MI(t,C_i) \tag{4}$$

互信息计算的时间复杂度与信息增益相似。互信息的缺点是评价结果受到词条频率影响较大，且之前的计算量很大。

4) 卡方统计量

卡方检验的基本方法就是对比实际值与理论值，以差值来判断它的正确性。通常先假设两个变量为独立的，然后观察实际值与理论值的偏差值，若偏差趋近于0，我们就将此误差当作可接受样本误差，是受测量条件限制或小概率发生的，所以两者独立，原假设成立；如果发生了很大的偏差，对这种误差来说，一般不可能是受测量条件限制或小概率发生的，我们就认为两者不独立，是相关的，原假设不成立。

训练语料中的文本总数(记为N)，某一特定类别(记为c)，特定的词条(记为t)，属于C类且包含t的文档频数(记为A)，不属于C类但包含t的文档频数(记为B)，属于C类但不包含t的文档频数(记为C)，既不属于C也不包含t的文档频数(记为D)。则对于C来说词条t的CHI值是：

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{5}$$

近似于处理互信息，为能够在多个类别中应用CHI统计，先计算各个词条与每一类的CHI值，再用式(6)计算对整个语料而言它的CHI值：

$$\chi^2(t)=\max_{i=1}^{m}\chi^2(t,c_i) \tag{6}$$

其中m为类别数。

词条对于某类的 χ^2 统计学数值与该类之间的相关性和其具有的类别信息成正比。

5) 期望交叉熵

$$ECE(t)=P(t)\sum_i P(c_i|t)\log\frac{P(c_i|t)}{P(c_i)} \tag{7}$$

如果词条t和类别Ci相关性越强，则 $P(c_i|t)$ 越大，如果与此同时，$P(c_i)$ 又很小，则说明该词对该类的影响大。这个量代表了文本类别的概率分布，以及文本类别在某种特征的基础上的概率分布之间的距离。

2.2 分类

2.2.1 分类方法

文本分类是依照文本内容或特征,在规定的分类系统下将待划分文本分配到一个及以上的之前定义好的分类中的方法[2]。

文本分类是一一对应的方法,将未明确的待分类文本对应到已定义的分类中,由于一篇文本可以同多个类别相关联,这个映射一般来说是一对一或一对多的映射。数学公式为:

$$f:X\rightarrow Y \quad 其中:X=(M_1,M_2,\cdots,M_n)\quad Y=(N_1,N_2,\cdots,N_m) \tag{8}$$

即:X为所有待划分的文本的集合;Y为规定的分类系统下,所有分类的集合。X可以为无限集合,而Y必须为有限集合。

分类方式一般依照基本划分方法不同而分为两种:基于机器学习的分类方法和基于规则的分类方法。

2.2.1.1 基于机器学习的分类方法

基于机器学习的分类方法是通过学习给定的训练集,从而归结出各分类的模板,从而使用模板来进行文本分类。

此方法的优点是简易可行,一般来说分类精确度较高;但它的缺点主要是:

1) 当重叠现象在各个类别中较多时(特征重复),精确度将严重下降,特别在多层分类中,特征重叠现象在子类中更为多见,因此在基本分类大体正确的时候,却发生了子类的分类精度严重下降的情况。比如说,在对金融,历史,科学技术,医疗卫生等方面的种类的检测中,显示出分类效果中体育分类的效果最好,精确度趋近于100%,这主要是因为体育类的特征与其他类的重叠很少;而医药卫生和科学技术类的精确度不理想,都低于90%,因为这两个类的特征之间重叠很多,并且与其他分类之间也有交叉。

2) 严格控制训练语料的量与质。如果训练集不全面,无法代表所在分类的特征,那么自动分类的精度将受到严重影响。对于每个分类来说,训练集最好全面覆盖该类。搜集训练集一定要保证语料准确属于所在类别,否则分类器的分类效果将受到影响。

文献[3]中提出使用机器学习分类方法会降低分类速度,因而使用了基于标题的新闻网页分类方法,然而目前的新闻信息玉石杂糅,很多新闻为博取读者眼球,尝尝文不对题,且内容真实性不高,据此分类则会给人们的生活带来不便甚至给人们带来危害。

2.2.1.1.1 朴素贝叶斯分类器

贝叶斯分类是一类分类算法的总称,这类算法均以贝叶斯定理为基础,故统称为贝叶斯分类。朴素贝叶斯算法(Naive Bayesian) 是其中应用最为广泛的分类算法之一。朴素贝叶斯分类器是一系列以假设特征之间强(朴素)独立下运用贝叶斯定理为基础的简单概率分类器。它基于一个简单的假定:属性之间在确定目标值的情况下彼此条件独立。朴素贝叶斯分类器的一个优势在于只需要根据少量的训练数据估计出必要的参数(变量的均值和方差)。

贝叶斯定理是一个与随机事件A和B的边缘概率相关的定理。[4]其中P(A|B)是在B发生的情况下A发生的可能性。

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)} \tag{9}$$

朴素贝叶斯的思想大体上是:对于待分类项来说,解出各个类别在此项出现时出现的概率,此待分类项的类别就是最大概率的分类。

朴素贝叶斯分类模型的优势是:

1) 时间复杂度、空间复杂度较低;

2) 算法逻辑清晰简便,易于理解和转化为具体程序;

3) 算法效果不易受其他因素干扰,模型健壮性良好。

在条件独立性假设的基础上,朴素贝叶斯分类器假设一个属性对指定类别的影响与其他属性无关,朴素贝叶斯分类算法的最小的误分类率是在条件独立性假设生效的情况下[5]。但朴素贝叶斯假设在实际中往往并不成立,多少影响了朴素贝叶斯分类器的分类效果。[6]

2.2.1.1.2 神经网络算法

人工神经网络(ANN),简称神经网络,是以生物神经网络的结构和功能的为原型的数学计算模型。一般来说,人工神经网络是自适应系统,可以根据外界信息来改变内部结构。在现代,ANN是统计学中的一种工具,常用于非线性数据建模,它将输入和输出间复杂的关系转化为模型,在探索数据的情况下用途甚广。

现今,神经网络的问题主要是收敛速度慢、计算量大、训练时间长和泛化能力不足[7],很多研究人员仍在不懈地研究神经网络算法,其研究目的是创新或改善神经网络的算法和性能,追求更快的收敛速度、降低陷入局部极小的概率或消除局部极小问题、提高泛化能力等。[8]

2.2.1.1.3 KNN分类方法

1968年,KNN算法由Cover和Hart提出,该算法的思路是:用经典的向量空间模型把文本内容转化为特征空间中的加权特征向量。计算检测文本与训练语料里的文本的相似程度,找出M个最相似的文本,用加权距离来判断测试文本的种类。最大权重的类别即为文本所在的类别。[9]

KNN(k-Nearest Neighbor)算法稳定性好、准确率高,但由于其时间复杂度与样本数量成正比,导致其分类速度慢,难以在大规模海量信息处理中得到有效应用。由于KNN方法主要依靠邻近的样本,但周围样本有限,因此对于类域重叠较多的待分语料来说,更适合使用KNN方法。

2.2.1.1.4 支持向量机(SVM)方法

在机器学习中,支持向量机(SVM)是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法,由Vapnik在1995年提出[10]。给定一组训练实例,每个训练实例被标记为属于两个类别中的一个或另一个,SVM训练算法建立一个将新的实例分配给两个类别之一的模型,使其成为非概率二元线性分类器。SVM模型是将实例表示为空间中的点,这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后,将新的实例映射到同一空间,并基于它们落在间隔的哪一侧来预测所属类别。

而对于非线性分类,SVM还可以有效地使用所谓的核技巧(kernel trick),把它的输入隐式映射到高维特征空间中。

如果数据未被标记,则需要进行非监督式学习,它会试着找出从数据到簇的自然聚类,并将心数据映射到这些已形成的簇。支持向量聚类[11]就是指由SVM改进的聚类算法,当数据并未或少量被标记时,支持向量聚类经常在应用中被用作分类步

骤的预处理。

2.2.1.2 基于规则的分类方法

决策树分类方法:使用决策树的主要原因是其构造过程不需要使用者学习其他专业的知识来决定参数,因此在实际应用中,为了探索新知识,更应该使用决策树。决策树算法主要是通过构造决策树来从给定样本中找出区分它们的规则,构造精度高、规模小的决策树是决策树算法的主要目的。

决策树被广泛应用是因为其原理简单易懂,分类过程中不需要人为设定参数,速度快、计算量较小,易于生成可理解的规则。但决策树难以预测连续性的字段,对于有时间顺序的数据,要进行许多预处理,类别太多时,错误增加速度很快。而且使用贪心算法的决策树方法做出的只是某种意义上的局部最优选择。[12]

2.2.2 方法的评估指标

机器学习(ML),自然语言处理(NLP),信息检索(IR)等领域,评估(Evaluation)是一个必要的 工作,而其评价指标往往有如下几点:准确率(Accuracy),精确率(Precision),召回率(Recall)。

表 1 评价标准量化表

	相关(Relevant),正类	无关(NonRelevant),负类
被检索到(Retrieved)	true positives(TP 正类判定为正类)	false positives(FP 负类判定为正类,"存伪")
未被检索到(Not Retrieved)	false negatives(FN 正类判定为负类,"去真")	true negatives(TN 负类判定为负类)

准确率(**accuracy**),其定义是:对于给定的测试数据集,分类器正确分类的样本数与总样本数之比。也就是损失函数是0-1损失时测试数据集上的准确率。Accuracy需要得到的是分类正确的数量占总数的比例。

精确率(**precision**)的公式是 $P=\frac{TP}{TP+FP}$ (12),它计算的是所有"正确被检索的item(TP)"占所有"实际被检索到的(TP+FP)"的比例。

召回率(**recall**)的公式是 $R=\frac{TP}{TP+FN}$ (13),它计算的是所有"正确被检索的item(TP)"占所有"应该检索到的item(TP+FN)"的比例。

Precision和recall在某些情况下是矛盾的。比如极端情况下,我们只搜出了一个结果,且是准确的,那么P就是100%,但是R就很低;而如果我们把所有结果都返回,那么必然R是100%,但是P很低。

F-Measure又称为F-Score,是IR(信息检索)领域的常用的一个评价标准,计算公式为:

$$F_\beta=\frac{(\beta^2+1)PR}{\beta^2P+R} \tag{10}$$

其中 β 是参数,P是精确率(Precision),R是召回率(Recall)。

当参数 $\beta=1$ 时,就是最常见的F1-Measure了:

$$F_1=\frac{2*PR}{P+R} \tag{11}$$

2.3 常用基准语料

语料库是对人们日常生活中的常用语言进行的搜集和整合,便于进行研究和实验分析。语料的选择恰当与否很大程度上影响了实验结果的正确性。

Reuters-21578是路透社发布的常用英文语料库,是一个常用于文本分类任务的测试集。它包含了21578个从1987年以来的路透社新闻文档,它们被路透社全体成员人工的分类标记。标签有五种不同的目录分类,目录总数为672个,但它们许多都仅仅出现的非常少。一些文档从属于许多不同的目录,一些只从属于一个分类,还有一些不属于任何目录。在过去的几十年里,他们做出了许多努力来使数据库升级并改进它,以使它能够被用于各种科学研究。

20 Newsgroups包含了大约20000份新闻文档,涵盖了约20中不同的新闻分类,这个数据集最早是Ken Lang建立的。现在20 Newsgroups在机器学习技术的文本应用方面的实验中成为了一个常用的数据集,例如文本分类和文本聚类。

TanCorp由谭松波整理搜集,是一个开放的中文语料库,是中国的权威语料库之一,通常用于自然语言处理等研究。语料库共有两个层次,收集14150篇文本,第一层为12个类别,第二层为60个类别。

复旦大学中文文本分类语料库共9804篇文档,分为20个类别,训练预料基本按照1:1的比例来划分。来源于复旦大学计算机信息与技术系国际数据库中心自然语言处理小组,也是目前中文文本分类实验常用的语料库。

搜狗实验室文本分类语料库是根据Sohu新闻网站收集的大量经过编辑手工整理与分类的新闻语料及相应的类别信息。其分类体系有着几十个分类节点,约为100,000个网页。目的是构建一个较大规模的标准化的中文文本分类测试平台,可用于中文文本分类和主题跟踪与检测等,是近年新兴的中文文本分类语料库。

3 中文新闻文本分类研究

中文新闻文本分类起步较晚,而且汉语语法结构繁杂,所以在技术和成果上中文新闻文本分类目前还略逊于英文新闻文本分类的研究进度。但近些年来发展仍是突飞猛进,尤其是基于中文新闻信息的相对详细的语料库的出现和发展使中文新闻文本分类变得更加便捷和容易。

3.1 新闻文本分类应用

1) 新闻网站的分类引导

新闻网站对于新闻进行分类使得读者可以更准确地找到自己感兴趣的分类并有利于关联阅读的进行,例如目前新闻页面下方的新闻推荐,可以使读者便捷地阅读到与当前阅读新闻相关的其他新闻页面。这一点有些类似传统纸媒的不同版面,在目前的新闻网站中已经非常普及。

2) 热点话题追踪

新闻的特点之一是时效性,有时一个事件的发展会经历一段相当长的时间。如果读者想要持续关注这一事件,则需要花费额外的精力每天在网络上搜索相关报道。但新闻文本分类技术可以实现对热点话题进行追踪,从而实时推送相关事件的最新进展,节省了用户耗费在信息搜集上的时间,例如,微博中的话题功能也有类似的新闻分类作用[13],但此功能由于是用户依照主观意识自主添加的,所以在准确率上非常不如意。

3) 特定类别新闻过滤和筛选

网络是一把双刃剑,既提升了信息传播的便捷性,又使有害信息的传播也变得相对容易。例如前些时间的"莆田系"医院时间、求职陷入传销组织事件等,均是由于对网络有害信息的不准确判断而酿成的悲剧。若能利用新闻文本分类将有害

信息在传播过程中筛选和拦截,将使用户接触到这些不良信息的几率大大降低。

3.2 中文新闻文本分类的问题

新闻的概括性较强,叙述时以较简洁利落的文字,在有效时间内的发布附近新近发生的、有价值的事实,能够让特定的受众获得信息。六大新闻要素5W1H(Who\Where\What\Why\When\How)中,时间、人物、地点等实体要素在大部分情况下可以表现出新闻内容中的主体对象。所以在对新闻文本进行预处理时,主要将这些词提炼出来以便后续分类的进行。

然而,新闻媒体经常以夸张标题吸引读者注意是无论中外媒体均有的通病,这导致了新闻文本常常文不对题,从而使依靠标题进行文本分类的准确率大大下降,也增加了读者的时间支出,使用户体验下降。

现有语料库的分类层次太浅,仍是依靠大类来进行文本分类,但太过详细的分类又会使新闻类别的数量指数增长,应做好新闻文本分类的准确性和类别数量之间的平衡。

4 总结和展望

文章主要介绍了在现今社会中,新闻文本自动分类的必要性和需求,重点介绍了文本分类的主要流程、基本原理和方法,介绍了中文新闻文本分类的进展,然后设想了文本分类技术在新闻领域的具体应用方向。虽然中文新闻文本分类技术在前辈学者的研究下已经有了一定的进展,但仍有许多方面需要进一步的研究和努力。

1) 新闻文本分类层次加深

将新闻文本的分类再进行细化,使新闻的分类更加准确和细致。但这会导致新闻数据的维护难度增加,并且需要计算速度提高方面的支持。

2) 新闻文本分类维度拓广

现有的新闻文本分类语料库大多是以主题进行的分类,这样的分类方向太过于单一。今后可以建立以情感[14]、应用、行业综合等不同方向的新闻文本分类语料库,以满足不同行业、不同用途的应用。

3) 新领域新闻分类

新闻的发展越来越快,承载形式从传统纸媒发展到现在的网络传媒。而新闻的类型也在不断增多,从过去的文字、图像等单一形式,到现在视频、音频等多种形式。一大批自媒体的兴起代表着视频新闻的时代已经到来,所以新闻分类已经不能拘泥于传统的文本分类,更要放眼于图像识别、语音识别以及视频中的动态图像识别等技术,甚至于新近兴起的AR、VR等技术。

4) 新闻文本分类在大数据方面的应用

网络信息的爆炸式增长,掀起了大数据的浪潮。新闻分类也得益于大数据的到来,可以进行充足的数据分析和学习。通过分析用户日常阅读的新闻的兴趣所在,从而实现新闻的个性化推荐,使新闻的受众更精准,用户体验大大提高。

参考文献:

[1] 刘依璐 . 基于机器学习的中文文本分类方法研究[D]. 西安:西安电子科技大学,2009:22-24.
[2] 庞剑锋,卜东波,白硕 . 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23-26.
[3] 钱爱兵,江岚 . 基于标题的中文新闻网页自动分类[J]. 现代图书情报技术,2008(10):59-68.
[4] 张磊 . 文本分类及分类算法研究综述[J]. 电脑知识与技术,2016,(34):225-226,232.
[5] 李旭升,郭春香,郭耀煌 . 扩展的树增强朴素贝叶斯网络信用评估模型[J]. 系统工程理论与实践,2008(6):129-136.
[6] 王国才 . 朴素贝叶斯分类器的研究与应用[D]. 重庆:重庆交通大学,2010.
[7] 杨旭华 . 神经网络及其在控制中的应用研究[D]. 杭州:浙江大学,2004.
[8] 周瑛 . 神经网络作为分类器的算法研究及在信息检索中的应用[D]. 合肥:安徽大学,2006.
[9] 卜凡军 . KNN算法的改进及其在文本分类中的应用[D]. 无锡:江南大学,2009.
[10] Boser B,I.Guyon V N.Vapnik."A training algorithm for optimal margin classifiers[C]//Fifth Annual Workshop on Computational Learning Theory, San Mateo,CA: Morgan Kaufumann.1992:144-152,139.
[11] Ben-Hur Asa, Horn David, Siegelmann Hava,et al. Support vector clustering[J]. Journal of Machine Learning Research, 2001(2):125-137.
[12] 王煜 . 基于决策树和K最近邻算法的文本分类研究[D]. 天津:天津大学,2006.
[13] 刘志远,高俊波 . 基于话题的微博多特征情感极性分类[J]. 微型机与应用,2017(16):60-62+66.
[14] 陈巧红,孙超红,贾宇波 . 文本数据观点挖掘技术综述[J]. 工业控制计算机,2017(2):94-95,102.