ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.8, No.4, February 2012

E-mail: xsjl@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

文本分类及算法综述

张征杰,王自强

(河南工业大学,河南 郑州 450001)

摘要:为了能在海量的文本中及时准确地获得有效的信息,文本分类技术受到了广泛的关注。该文概括地介绍了文本分类的一般分类过程,详细阐述文本表示、特征选取和权重的计算,并对几种典型文本分类算法的基本思想、优缺点等进行了讨论。

关键词:文本分类;文本表示;向量空间模型;特征选择;权重;分类算法

中图分类号:TP301 **文献标识码**:A **文章编号**:1009-3044(2012) 04-0825-04

在当今的信息社会,各种形式的信息都得到了极大的丰富了我们的生活,尤其随着Internet的大规模普及,网络上的信息量在飞速增长当中,如各种电子文档、电子邮件和网页充满网络上,从而造成信息杂乱。为了快速、准确、全面地找到我们所需要的信息,文本分类成为了有效组织和管理文本数据重要方式,越来越受到广泛的关注。文本分类在信息检索、信息过滤、搜索引擎、文本数据库、数字化图书馆等领域得到广泛的应用。

1 文本分类的一般过程

文本分类是一个有指导的学习过程,它根据一个已经被标注的训练文本集合,找到文本属性(特征)和文本类别之间的关系模型(分类器),然后利用这种学习得到的关系模型对新的文本进行类别判[1]。文本分类的过程总体可划分为训练和分类两部分。训练的目的是通过新本和类别之间的联系构造分类模型,使其用于分类。分类过程是跟据训练结果对未知文本进行分类,给定类别标识的过程。具体流程图如图1:

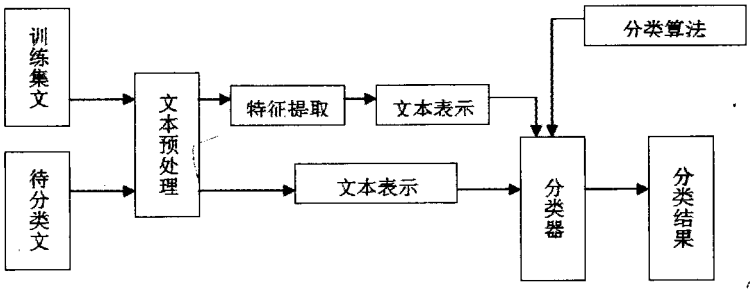

图1

2 文本预处理

文本预处理是从文本中提取关键词来表示文本的处理过程,它的主要任务是进行中文分词和去停用词。不同于英文中词与词之间是靠空格隔开,中文文本的自然语言中词与词间没有明显的切分标志,所以首先要对文本进行分词处理。中文分词方法主要有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]。

基于字符串匹配的分词方法是按照一定的策略将待分析的字符串与一个机器词典中的词条进行匹配,若从词典中找到某个字符串,则匹配成功。依据不同的扫描方向,可分为正向匹配和逆向匹配;依据不同长度优先匹配的情况,可分为最大匹配和最小匹配。

基于理解的分词方法是通过让计算机仿照人对句子的理解方式,从而达到识别词的效果。其基本思想就是在分词的同时进行句法和语义分析,利用句法信息和语义信息来处理歧义现象。

基于统计的分词方法是测试字与字相邻共现的频率,并把它作为成词的可信度评价标准。具体做法是先统计语料库中相邻共现的各个字的组合频度,计算它们的互信息。因为互信息体现了汉字之间结合关系的关联程度,当关联程度高于某一个阈值时,便认为这些字组可能构了一个词。

目前歧义词和新词是中文分词面临的最大困难所在。前者要解决自然语言理解的问题,根据上下文环境,在不同切分结果中选择最优解;后者要解决词典中未收录词(如人名、地名、机构名等)的识别[2]。

停用词通常指在各类文本中都频繁出现,因而被认为带有很少的有助于分类任何信息的代词、介词、连词等高频词。通过构造一个停用表,在特征提取过程中删除停用表中出现的特征词。

3 文本的表示

自然语言文本是非结构化的杂乱无章的数据,须将它们转换为结构化的计算机可识别处理的信息,即对文本进行形式化处理,

收稿日期:2011-12-25

本栏目责任编辑:谢媛媛

结果称为文本表示。目前通常采用的文本表示模型有概率模型、潜在语义索引模型和空间向量模型[3]。其中,向量空间模型是应用最广的文本表示模型。

向量空间模型(Vector Space Model,VSM)是Salton等人在20世纪60年代提出的,初期在信息检索领域应用,现在已成为文本分类中最广泛采用的一种文本表示。向量空间模型基于如下假设:文章中词条出现的顺序无关紧要,它们之间是相互独立的而忽略其依赖性,把文本看作一系列无序词条的集合。在该模型中,每篇文本表示为特征空间的一个向量,向量中的每一维对应于文本中的一个词条,每一个词条称为一个特征项,每一个特征词的值为该向量维对应的特征在文本集中的权值。其数学描述如下:

假设特征词集合为$T=\{t_1,t_2,t_3,t_4,\cdots t_n\}$,文本集合为$D=\{d_1,d_2,d_3,d_4,\cdots d_m\}$,文档$d_j$用一个n向量表示为$d_j=(w_{j1},w_{j2},\cdots,w_{jn})$,每一维对应特征词集合中的一个特征项,其值通过权值计算公式w_{jk} $(1\leqslant k\leqslant n)$给出。权值一般是特征词在文本集中出现频率的函数。

考虑到词语与词语之间是有语义上的联系的,图模型[4]利用图来表示文本。图中的节点表示文本中的词语,边表示词语之间的相互关系。另外,也有把概念和概念距离引入向量空间模型,从语义,概念的角度出发,以概念作为文本的特征项,建立基于概念的文本表示模型[5],解决同义词和多义词的问题而实现对向量空间模型的改进。

4 特征项的选择和特征权重

通常原始特征空间维数非常高,且存在大量冗余的特征,因此需要进行特征降维。特征选择是特征降维中的其中一类,它的基本思路:根据某种评价函数独立地对每个原始特征项进行评分,然后按分值的高低排序,从中选取若干个分值最高的特征项,或者预先设定一个阈值,把度量值小于阈值特征过滤掉,剩下的候选特征作为结果的特征子集。

文本分类中常用的特征选择方法有:文档频次、互信息量、信息增益、χ统计量(CHI)等方法[6]。

4.1 文档频率(DF:Document Frequency)

文档频率指训练集中包含该特征的文本总数。所谓包含特征的文本是指这个特征在该文本中是否出现,而忽略其出现次数。采用文档频率基于如下假设:文档频率值低于某个阈值的词条是低频词,可认为它们不包含有类别信息(不具有分类的能力),将这样的词条从原始特征空间中除去,能够降低特征空间的维数从而提高分类精度。

文档频率是最简单的特征选择技术,由于其具有相对于训练语集规模的线性计算复杂度,它能够容易地被用于大规模语料统计。但是在信息抽取研究中却通常认为DF值低的词条相对于DF值高的词条具有较多的信息量,将这些词条从特征空间中移除会降低分类器的准确率[5]。

4.2 信息增益(IG:Information Gain)

信息增益在机器学习领域被广泛使用,它通过特征词在文本中出现和不出现前后的信息量之差来推断该特征词所带的信息量。采用如下公式:

$$IG(t)=-\sum_{i=1}^{m}P(c_i)\log P(c_i)+P(t)\sum_{i=q}^{m}P(c_i|t)\log P(c_i|t)+P(\bar{t})\sum_{i=1}^{m}P(c_i|\bar{t})\log P(c_i|\bar{t})$$

其中P(t)表示样本集中包含词t的文本的概率,$P(c_i)$表示类文本在样本集中出现的概率,$P(c_i|t)$表示文本包含词t时属于c_i类的条件概率,$P(c_i|\bar{t})$表示文本不包含词t时属于c_i类的条件概率,$\bar{t}$表示样本集中不包含词t的文本的概率。

4.3 互信息(MI:Mutual Information)

互信息是信息论中的一个重要概念,它用来衡量一个消息中两个信号之间的相互依赖程度。在文本分类中,互信息是用来衡量特征词和类别之间的共现关系,其类别c_i和特征词t之间的互信息定义如下:

$$I(t,c_i)=\log\frac{p(t,c_i)}{p(t)p(c_i)}=\log\frac{p(t|c_i)}{p(t)}$$

其中$p(t,c_i)$表示特征t与类别c_i共现的概率,$p(t)$表示特征t在整个训练集中出现的文本频率,$p(c_i)$表示类别c_i在训练集中出现的概率。其$I(t,c_i)$表示特征项t与类别c_i的关联程度。它越大说明t与类别c_i的联系越紧密。

4.4 卡方统计法(CHI)

卡方统计也用于表征两个变量的相关性,与互信息相比,它同时考虑了特征在某类文本中出现和不出现时的情况。卡方值越大,它与该类的相关性就越大,携带的类别信息也就越多。

$$\chi^2(t,c_i)=\frac{p(t,c_i)p(\bar{t},\bar{c}_i)-p(\bar{t},c_i)p(t,\bar{c}_i)}{p(c_i)p(t)p(\bar{c}_i)p(\bar{t})}$$

4.5 特征权重的计算

在文本中,每一个特征项赋予一个权重,表示这一特征项在该文本中的重要程度。特征权值一般都是以特征项的频率为基础进行计算的。特征权重(term weight)的计算公式很多,假定特征t_k在文本d_j中的词频为f_{jk}、,特征权值为w_{jk},N表示文本集中的文本数,M表示所有文档的词汇量,n_k表示特征t_k在整个文档集中的出现频率,则常见的权值计算方法包括:

1)布尔权值法

如果某个词条在一篇文本中出现,则将其权值w_{jk}定义为1,否则定义为0。

2)词频权值法

词频权值法是根据特征词在文本中的出现频率来衡量其重要程度,即$w_{jk}=f_{jk}$

3)TF/IDF权值法

TF/IDF(Term Frequency/ Inverse Document Frequency)方法是应用最为广泛的一种权值法,其中TF表示特征词在某文本中的出

现频率,IDF表示特征词在整个文本集中的出现频率。文本k中词i的TF/IDF权值与其在该文本中的出现频率成正比,而与其在整个文本集中的出现频率成反比,用公式表示为:

$$w_{jk}=f_{jk}\times\log(\frac{N}{n_k})$$

4)TFC权值法

TF/IDF权值法虽然最常用,但它没有考虑文本长度对权值的影响。TFC权值法在TF/IDF方法的基础上利用文本长度对其进行规范化。

$$w_{jk}=\frac{f_{jk}\times\log(\frac{N}{n_k})}{\sqrt{\sum_{j=q}^{M}(f_{jk}\times\log(\frac{N}{n_k}))^2}}$$

5 文本分类算法

5.1 朴素贝叶斯分类算

朴素贝叶斯分类算法(Naïve Bayes)是一种典型的概率模型算法,根据贝叶斯公式作,算出文本属于某特定类别的概率。它的基本思路是计算文本属于类别的概率,该类别概率等于文本中每一个特征词属于类别的概率的综合表达式,而每个词属于该类别的概率又在一定程度上可以用这个词在该类别训练文本中出现的次数(词频信息)来粗略估计。

假定文本集中每一个样本可用一个n维特征向量$d_j=\{t_{j1},t_{j2},t_{j3},t_{j4},\cdots t_{jk}\}$表示,基于贝叶斯理论类计算待定新文本$d_j$的后验概率用$p(c_i|d_j)$表示:

$$p(c_i|d_j)=\frac{p(c_i)p(d_j|c_i)}{p(d_j)}$$

其中$p(d_j)$对计算结果与影响,因此可以不计算。贝叶斯方法的基本假设是词项之间的独立性,于是:

$$p(d_j|c_i)=p(t_{j1}\cdots t_{jk}|c_i)=\prod_{k=1}^{n}p(t_{jk}|c_i)$$

类别的先验概率$p(c_i)$和条件概率$p(t_{jk}|c_i)$在文本训练集用下面的公式来估算:

$$p(c=c_i)=\frac{n_i}{N}$$

$$p(t_{jk}|c_i)=\frac{n_{ik}+1}{n_i+r}$$

其中,n_i表示属于类c_i训练文本数目;N表示训练文本总数;n_{ik}表示类c_i中出现特征词t_k的文本数目;r表示固定参数。

朴素贝叶斯算法优点是逻辑简单,易实现,分类过程中时空开销小,算法稳定。它的不足处是它基于文本中各个特征词之间是相互独立的,其中一词的出现不受另一词的影响,但是显然不对。

5.2 Rocchio算法

Rocchio算法又称类中心最近距离判别算法,最早由Hull在1994年引进文本分类,是基于向量空间模型和最小距离的算法。它的基本思路是用简单的算术平均为每类中的训练集生成一个代表该类向量的中心向量,然后计算测试新向量与每类中心向量之间的相识度,最后判断文本属于与它最相似的类。

向量相似性的度量一般常采用:

1)夹角余弦:

$$Sim(d_i,d_j)=\cos(\theta)=\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{\sum_{k=1}^{n}w^2{}_{ik}\times\sum_{k=1}^{n}w_{jk}^2}}$$

夹角余弦表示一篇文本相对于另一篇文本的相似度。相似度越大,说明两篇文本相关程度越高,反之,相关程度越低

2)向量内积:

$$Sim(d_i,d_j)=d_i\cdot d_j=\sum_{k=1}^{n}w_{ik}w_{jk}$$

3)欧氏距离:

$$D(d_i,d_j)=\sqrt{\frac{1}{N}(\sum_{k=1}^{n}(w_{ik}-w_{jk})^2)}$$

距离越小,两篇文本的相关程度就越高,反之,相关程度越低。

在Rocchio算法中,训练过程是为了生成所有类别的中心向量,而分类阶段中,系统采用最近距离判别法把文本分配到与其最相似的类别中从而判别文本的类别。所以,如果类间距离比较大而类内距离比较小的类别分布情况,此方法能达到较好的分类效果,反之,类中心最小距离算法效果比较差。但由于其计算简单、迅速、容易实现,所以它通常用来实现衡量分类系统性能的基准系统,而很少采用这种算法解决具体的分类问题。

5.3 k最近邻算法

K最近邻算法(KNN)最初由Cover和Hart于1968年提出[7],是一种基于实例的文本分类方法,将文本转化为向量空间模型。其基本思路是在给待定新文本后,计算出训练文本集中与待定文本距离最近(最相似)的k篇文本,依据这k篇文本所属的类别判断新文本所属的类别。

可以用夹角余弦、向量内积或欧氏距离计算出K篇最相似文本。而决策规则是统计K篇训练样本中属于每一类的文本数,最多文本数的类即为待分类文本的类。但考虑到样本平衡问题时,目前应用较广的是SWF决策规则,该决策规则是对上面DVF规则的改进,根据K个近邻与待分类文本的相似度之和来加权每个近邻文本对分类的贡献,这样可以减少分布不均匀对分类器的影响。SWF决策规则数学描述:

$$SCORE(d,c_i)=\sum Sim(d,d_j)y(d_j,c_i)-b_i$$

其中, $SCORE(d,c_i)$ 为文本d属于类 c_i 的分值; $Sim(d,d_j)$ 为 d 与 d_j 之间的相似度;当 $y(d_j,c_i)$ 如果属于类别 c_i 时,则 $y(d_j,c_i)=1$,当 $y(d_j,c_i)$ 不属于类别 c_i ,则 $y(d_j,c_i)=0$,; b_i 为阈值,它可在集上通过训练来得到。

KNN的不足处之一是判断一篇新文本的类别时,需要把它与现存所用训练文本都比较一遍。另一个不足处是当样本不平衡时,即如果一个类的样本容量很大而其它类很小,可能导致输入一个新样本时,该样本的K个邻居中大容量样本占多数。

5.4 决策树

决策树(Decision Tree)基本思路是建立一个树形结构,其中每个节点表示特征,从节点引出的每个分支为在该特征上的测试输出,而每个叶节点表示类别[8]。大致需要下面几个步骤:

1)根据信息增益法在特征集中选取信息增益最高特征项作为当前节点的测试属性;

3)按测试属性(特征权重)不同取值建立分支;

3)对各子集递归进行以上两步操作建立决策树节点的分支,直到所有子集仅包含同一类别的数据为止;

4)对决策树进行剪枝,生成更紧凑的决策树。

决策树算法的核心问题是选取测试属性和决策树的剪枝。除了常用的信息增益法,选择测试属性的依据还有熵、距离度量、G统计、卡方统计和相关度等度量方法。从决策树的根节点到每个叶节点的每一条路径形成类别归属初步规则,但其中一些规则准确率较低,需要对此决策树进行剪枝。

决策树实际上是一种基于规则的分类器,其含义明确、容易理解,因此它适合采用二值形式的文本描述方法。但当文本集较大时,规则库会变得非常大和数据敏感性增强会容易造成过分适应问题。另外,在文本分类中,与其它方法相比基于规则的分类器性能相对较弱。

5.5 人工神经网络

人工神经网络(Artificial Neural Networks)是一种按照人脑的组织和活动原理而构造的一种数据驱动型非线性模型。它由神经元结构模型、网络连接模型、网络学习算法等几个要素组成,是具有某些智能功能的系统。在文本分类中,神经网络是一组连接的输入输出神经元,输入神经元代表词条,输出神经元表示文本的类别,神经元之间的连接都有相应的权值。训练阶段,通过某种算法,如正向传播算法和反向修正算法,调整权值,使得测试文本能够根据调整后的权值正确地学习。从而得到多个不同的神经网络模型,然后令一篇未知类别的文本依次经过这些神经网络模型,得到不同的输出值,通过比较这些输出值,最终确定文本的类别。

6 分类性能评估

分类器性能评估通常采用评估指标来衡量,评估指标是在测试过程中所使用的一些用来评价分类准确度的量化指标,文本分类中常用的性能评估指标有查全率又称召回率(Recall)、查准率又称准确率(Precision)和F1标准。

查全率是衡量所有实际属于某个类别的文本被分类器划分到该类别中的比率,查全率越高表明分类器在该类上可能漏掉的分类越少,它体现系统分类的完备性。数学公式如下:

$$查全率=\frac{分类的正确文本数}{应有的文本数}$$

查准率是是衡量所有被分类器划分到该类别的文本中正确文本的比率,准确率越高表明分类器在该类上出错的概率越小,它体现系统分类的准确程度。数学公式如下:

$$查准率=\frac{分类的正确文本数}{实际分类的文本数}$$

F1标准即考虑了查全率,又考虑了查准率,将两者看作同等重要。数学公式如下:

$$F1=\frac{准确率\times查全率\times2}{(准确率+查全率)}$$

7 总结

本文分析了文本分类的一般过程,详细介绍文本分类中的文本表示、特征选择和权重计算,并且讨论几种常见分类算法,最后叙述分类器性能评价。希望能给该领域感兴趣的读者一些有益的参考。

参考文献:

[1] Aas K,Eikvil L.Text Categorization:a survey[Z].Teehnical Report 941,NOwegian Computing Center,1999:90-100. (下转第841页)

字符串。

```
Conn=@"Data Source=.\SqlExpress;Initial Catalog= AddrBook;" +
"User ID=" + UserIDTBox.Text + ";" +
"Password=" + PasswordTBox.Text + ";"
```

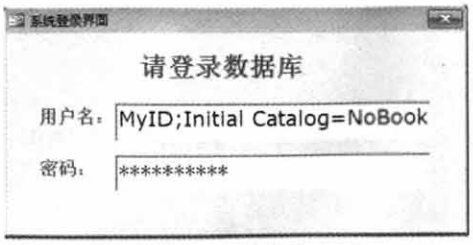

图1 输入用户名和密码

如恶意用户按照上图方式输入用户名来修改连接字符串。通过Conn和用户输入得到的连接字符串如下：

```
Data Source=.\SqlExpress;Initial Catalog=AddrBook;
User ID=MyID;Initial Catalog=NoBook;Password=MyPswd;
```

Initial Catalog被两次赋值，那么不知道连接到哪个数据库？连接字符串生成器技术可以帮助用户处理来自恶意的用户输入。

2)用连接字符串生成器防止连接字符串注入

ADO.NET 2.0为每个数据提供程序引入连接字符串生成器，提供与每个数据提供程序允许的已知键/值对相对应的方法和属性。每个类都有一个固定的同义词集合，可以将同义词转换为相应的已知键名，并执行键/值对的有效性检查，无效对会引发异常，此外，还会以一种安全方式处理插入的值。运行时构造有效连接字符串。

利用SqlConnectionStringBuilder生成SqlClient连接字符串：

```
SqlConnectionStringBuilder buld = new SqlConnectionStringBuilder();
buld.DataSource = @".\SqlExpress";
buld.InitialCatalog = "AddrBook";
buld.UserID = UserIDTBox.Text;
buld.Password = PasswordTBox.Text;
```

字符串生成器生成以下连接字符串：

```
Data Source=.\SqlExpress;Initial Catalog= AddrBook;
User ID=MyID;Initial Catalog=NoBook;Password=MyPswd
```

这会使得ADO.NET以“MyID;Initial Catalog=NoBook”为User ID来登录到SQL Server数据库而无法实现。防止了恶意连接字符串的注入。

3 小结

配置文件用于已编译的应用程序外部，使用XML存储信息，就动态属性而言配置文件是可以根据需要更改的。根据连接不同的数据源而生成的连接字符串采取不同的安全漏洞防范措施，从而保护整个系统。

参考文献：

[1] 李志强,张少华,郦雅芳.基于XML的动态用户界面实现技术[J].电脑知识与技术,2006(36).

(上接第828页)

[2] 龙树全,赵正华.中文分词算法概述[J].电脑知识与技术,2009,5(10):2605-2607.
[3] 骆昌日.基于统计方法的中文文本自动分类研究[D].武汉:华中师范大学.2004:8-11.
[4] 周昭涛,卜东波,程学旗.文本的图表示初探[J].中文信息学报,2005,19(2).
[5] 陈龙,范瑞霞,高琪.基于概念的文本表示模型[J].计算机工程与应用,2008,44(20):162-164.
[6] Yank Y.A Comparative Study on Feature Selection in 1'ext Categorization[C]//Proceeding of the Fourteenth International Conference on Machine Learning,1997:412-420.
[7] Cover T M,Hart P E.Nearest neighbor pattern classification[J].IEEE Transactions on Information Theory,1967,13(3):21-27.
[8] 古平,朱庆生.基于贝叶斯模型的文档分类及相关技术研究[D].重庆:重庆大学博士论文,2000.