适于Internet新闻文本实时分类的动态向量空间模型DVSM*)

张晓辉 李 莹 常桂然 赵 宏
(东北大学软件中心 沈阳110004)

摘 要 传统向量空间模型(VSM)特征间无关联,且不能动态增量训练,不适合主题和焦点实时变化的Internet新闻信息,为此提出了一种改进的文本实时分类模型——动态向量空间模型(DVSM)。通过对VSM的特征提取策略进行改进,提出了特征聚合和增量训练算法。通过将对分类有相同贡献的文本特征词聚合,使用它们共同的分类贡献向量特征模式作为文本特征向量的基本维;采用增量动态训练改变对分类贡献已改变的特征词在文本向量的特征模式中的位置,适应Internet新闻信息的实时特性。使用静态训练集和动态训练集进行的DVSM与传统VSM的对比实验表明,采用特征聚合和动态训练的DVSM在Internet新闻实时分类中优势效果明显优越。

关键词 动态向量空间模型,特征聚合,增量动态训练,Internet新闻分类,分类贡献向量特征模式

A Dynamic Vector Space Model for Internet News Textual Categorization

ZHANG Xiao-Hui LI Ying CHANG Gui-Ran ZHAO Hong
(Software Center, Northeastern University, Shenyang 110004)

Abstract Traditional Vector Space Model does not consider the relationship between features, and is not suitable for dynamic training. Focus on the Internet news with dynamically changing topics and focus, a Dynamic VSM (DVSM) is proposed. Multiple discriminating features with similar contribution to classification are combined into one pattern, which is used as the basic feature dimension. When new samples need to be learned, the changed discriminating features are moved between patterns with dynamic incremental training method for the real-time characteristics of Internet. Comparison experiments using static and dynamic training sets respectively show that DVSM outperforms the traditional model significantly in Internet News Real Time Categorization.

Keywords Dynamic vector space model, Feature combination, Dynamic incremental training, Internet news text categorization, Contribute pattern to classification

1 引言

自动文本分类被广泛应用于文本处理和信息检索的各个领域。通常,自动文本分类指自动将一篇文章指定到一个或几个预定义好的文本类别中[1,2]。传统分类方法包括支持向量机(SVM)、k近邻法(KNN)、神经元网络方法(NNet)、线性最小二乘方估计法(LLSF)和朴素贝叶斯方法(Naïve Bayesian)等[2]。其中大部分方法采用由Salton等人提出的向量空间模型VSM(Vector Space Model)[3,6]。该模型将文本表示为向量,作为向量空间的一个点,向量各维对应用于表征文本的各特征词,采用TF-IDF统计值作为特征词的测度,通过计算向量间的距离决定文本所属类别。目前存在多种基于向量空间模型的文本分类方法,如神经网络方法、支持向量机模型、KNN方法等,得到广泛应用。

传统向量空间模型下的静态分类方法不适应快速更新的Internet新闻,原因主要为:静态的训练集不能体现快速变化的网络信息的主题和焦点特征;未考虑特征词间的关联关系,使得距离的计算没有足够的模糊性;以及多数特征维对分类作用有限。问题集中在特征相关性和静态训练集上。

针对VSM不考虑特征间相关性的缺点,一些方法进行了改进。文[4]提出一种WAKNN加权方法,该方法在给每个特征词加权时,逐一尝试权值,直到找到最有效的为止。WAKNN方法取得了一定的效果,但同时计算代价也大幅增加。文[5]主要考虑文档间特征词属性关联与"共现"对相似度的贡献,用一个匹配系数调整两文档间的距离。其实质是对相似度的一种加权,未对特征词的选取和特征向量做任何改进。文[7]将向量空间模型与概率推理网络相结合,虽增强了关联词的模糊性,但算法需基于一定的语义分析,复杂且效率不高。

针对动态的数据,重新训练方式会造成已分类文本的再训练。现有增量训练方法主要有:文[8]中的增量SVM算法通过遗忘因子对分类面附近样本点的不同处理来提高训练速度;文[9]针对贝叶斯网络的增量训练进行了研究;文[10]提出将训练样本聚合,以样本集合来代替单个样本,但对老数据没有删除或降权处理;长时间的处理后,将出现大量类别,影响分类质量。这些方法取得了一定的成功,但并没有解决特征间的关联问题。

为解决VSM在Internet新闻数据实时分类中的特征关

*)基金项目:国家"八六三"高技术计划资助项目(863-306-ZD02-02-6)。张晓辉 博士研究生,主要研究方向为计算机网络与信息处理技术。李莹 硕士研究生,主要研究方向为计算机网络与信息处理技术。常桂然 教授,博士生导师,主要研究方向为计算机网络与信息处理技术。赵宏 教授,博士生导师,主要研究方向为分布式多媒体及其支撑环境。

联和增量训练问题，本文在传统 VSM 基础上提出一种动态向量空间模型 DVSM。该模型通过分析特征词的 CHI 分类贡献向量，将相关联的特征词聚合为一个分类贡献向量特征模式（简称为：特征模式），作为文本特征的基本维，从而实现特征关联和降维；针对增量训练集，采用增量动态训练和 hash 算法定位分类贡献已变化的词条所要改变到的目标特征模式序号，从而在不影响分类的情况下，完成分类特征维的改变。

2 动态向量空间模型 DVSM

针对 Internet 新闻文本，DVSM 对 VSM 的特征提取策略进行了改进，并提出了特征聚合和增量训练算法，下面进行详细说明。

2.1 特征提取

以特征提取这个 VSM 的关键步骤作为 DVSM 的出发点。目前提取文本特征词的方法主要有互信息、信息增益、几率比、词频法和 CHI 概率统计等[1]。研究表明，在包括 Reuters-21578和 OHSUMED 的一个测试集合中，CHI 概率统计通常比其它特征词选取方法优越[1]。其实质是：认为词与类别之间符合 χ^2 分布。χ^2 统计量充分体现了词汇和类别之间的相关性，χ^2 值越高，词和类别之间的独立性越小、相关性越强、词对这一类别的贡献越大[5]。

根据 CHI 统计量的意义，我们可以利用 χ^2 提供的信息，改进传统的 VSM 模型。由 CHI 的本质可知，每类中的 χ^2 值体现了该词条与该类别的相关程度，据此可定义特征词的关联性。我们的目标是提高动态文本数据集分类的精度，特征词间的关联性并不局限于同义词和近义词，只要对单个类别的分类贡献相同，即可认为其相关联，据此可以聚合所有相关特征词；当增量训练引起特征词对各类别贡献发生变化时，只需根据该词 χ^2 值的变化，把它移动到对应模式中，不必对整个训练集重新训练。

DVSM 不介意特征词数目，特征模式数和特征词数不成正比关系。其目标是为聚合对分类贡献相同的特征词提供原始数据，包括词条和该词的所有 χ^2 值。因此，能够在保证一定维数的同时，扩大特征词数，从而包括更多的低频词，以加强分类效果。

2.2 特征词聚合

特征词聚合是 DVSM 的一个重点。每个特征词可以得到 C（类别总数）个 CHI 值，表示为 $\chi^2(t,i)(1\leqslant i\leqslant C)$。根据该值，每个词可得到一个对不同类别的分类贡献向量。根据上一节的分析，那些有相同分类贡献向量的特征词是相关联的，对分类有相同的贡献。我们可将分类贡献相同的特征词聚合为一个特征模式。最终，可得到 m 个不同的特征模式，以特征模式替代原有的单个特征，作为文本向量的一个特征维。

将总类别设为15，即 $C=15$，则特征词 t 可得到15个 CHI 值。图1为5个词的 CHI 对各类（类别定义和名称含义见表1）的分类贡献向量图。由图可以看出，“欧洲”和“意大利”两词向量基本重合，“CPU”和“硬盘”两个词的向量也基本重合，而“软驱”、“CPU”和“硬盘”三个词几乎都只在 D 类出现，即认为它们只对 D 类有分类贡献。聚合的目的首先就是找出这些词。

为描述特征聚合算法，给出下面三个定义：

定义1 $value_t=(value_t[1],value_t[2],\cdots,value_t[C])$ 称为特征词 t 的规格化分类贡献向量。其中：$value_t[i]=\lfloor\chi^2(t,i)/5\rfloor,(i\in[1\cdot\cdot c])$ 称为特征词 t 的规格化公式。

定义2 对于特征词，若在其规格化分类贡献向量中，只有一个分量 i 的 *value* 值大于0，而其他分量的 *value* 值都等于0，即满足：

$$\begin{cases} value_t[i]>0(i\in[1\cdot\cdot C]) \\ value_t[j]=0(\forall\ j\in[1\cdot\cdot C]\wedge j\neq i), \end{cases}$$

则将满足该条件的所有特征词根据分量 i 的不同分别合并。称此规则为合并规则。

定义3 对于特征词，如果它们的规格化分类贡献向量中各个分量 $value_t[i](i\in[1\cdot\cdot C])$ 都相等，或者满足合并规则，则称这些特征词具有相同“特征模式”，并以该模式作为文本向量的一个特征维。

例如，我们对5000个特征词的分类贡献向量的分析知，单个词的最大 CHI 值变化范围是8.347877－76.56693，大部分的 CHI 值都低于5。可以应用定义1来规格化每个特征词的分类贡献向量。大约有40%的特征词 $value_t[i]$ 只有一项大于零，即只在一类中出现，可以认为这样的词只对这一类有分类贡献，则可应用定义2的合并规则。

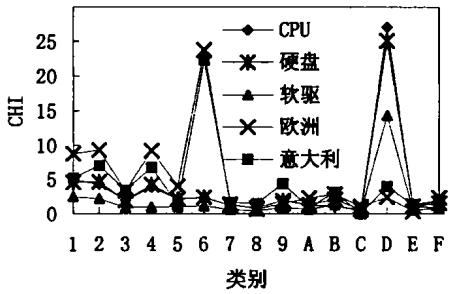


图1 分类贡献向量

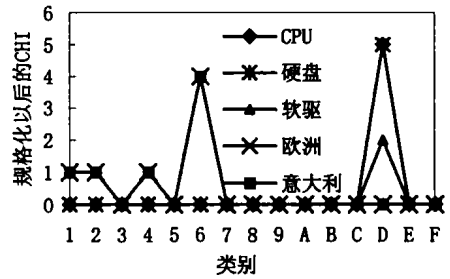


图2 规格化分类贡献向量

如图1所示，“CPU”、“硬盘”、“软驱”、“欧洲”和“意大利”五个特征词，采用定义1规格化以后的结果如图2所示。可以看出“CPU”和“硬盘”两词的向量重合了，“欧洲”和“意大利”的向量也重合了。而采用定义2的合并规则后，“软驱”与“CPU”、“硬盘”具有相同“特征模式”。以形成的特征模式作为文本向量的一个特征维，则本例中可得到两个特征维，其对应的特征词集合为“欧洲－意大利”和“CPU－硬盘－软驱”。

5000个特征词应用规格化公式和合并规则后，最终得到了776个不同的特征模式。其中最大的模式包含“致癌物质”、“免疫”、“人工受精”、“平衡状态”、“正电荷”等266个特征词，均只在大众科技类别出现；最小的模式只有一个特征词，这样的模式共有546个，即有89%的特征词都聚合在不只包含一个词的模式下。可见，大多数特征词都存在与其相关联的词。同时，模式和类别间无一一对应关系，如包含特征词“欧洲”的模式同时对应国际新闻、大陆新闻、社会新闻和国际足坛四个类。

每篇文本根据不同的特征模式构造成一个 m 维的向量。

在该文本向量中，将同一维中所有特征词的特征值(即 TF-IDF 值)相加，其和作为这一维的特征值，并对其归一化得到特征向量。

2.3 动态增量训练

DVSM 需要根据训练集的变化修改各特征维所包括的特征词，因而必须支持增量训练。

在 DVSM 中维护两张表。一张表称为词条记录表，记录特征词条和 CHI 计算公式中的各分量，及该词条对应的模式序号(该序号为该词条规格化分类贡献向量的 hash 值)。另一张表称为文本记录表，记录样本及该样本对应的词条。

动态经验训练的算法描述如下：

Step1. 对于新增的所有样本，重复执行 Step2～5；

Step2. 从文本记录表中选取最陈旧的样本 O；

Step3. 从词条记录表中修改 O 相关的每个词条 t 出现的类别 c 对应的各分量值；如果各分量值均为0，则删除该词条记录；

Step4. 在文本记录表中添加新样本 N；

Step5. 在词条记录表中修改新样本 N 中每个词条 t 在出现的类别 c 对应的各分量值；如果不存在该词条对应记录，则增加该词条记录；

Step6. 计算 Step3和 Step5中发生变化的词条的 CHI 值；并重新定位该词条对应的模式(计算 hash 值)；

Step7. 动态训练完成。

为加快算法速度，算法只对发生变化的词条处理，包括 CHI 的分量值和 CHI 值的计算等；通过 Hash 函数计算，可快速定位目标模式。同时在动态训练中，模式的数目和词条的数目会发生变化，这也反映了当前网络信息的重点变化情况。

3 数据集选取及评估方法

迄今为止，还没有一个标准的、普遍可接受的中文文本分类静态/增量训练集和测试集，因而我们使用自己的一个新闻收集系统来获得各网站上的 Internet 新闻。该系统覆盖新浪[11]、搜狐[12]、新华网[13]、齐鲁热线网[14]等400余个中文网站，一天收集新闻约20000篇。数据中存在20—24%的重复新闻，由机器自动滤除。对于初始数据集合，首先收集连续两天的新闻，去除重复新闻，前一天数据作为训练集，包含10028篇文本；第二天的作为测试集，包含11173篇文本。手工将其分成15个类别(类别定义和名称含义见表1)，由于网络中信息分布的不均衡，每类包含的文本数分布不均，57%的类别包含大于600篇的文本，13%的类别少于100篇。表1给出初始训练集和测试集文本类别分布情况。

表1 初始训练集和测试集文本类别

编号	类别名称	训练集	测试集
1	国际新闻	2286	1904
2	大陆新闻	1865	1859
3	港澳台新闻	390	308
4	社会新闻	1605	1209
5	国内足坛	471	806
6	国际足坛	567	782
7	篮排球	160	247
8	棋牌	77	78
9	综合体育报道	370	849
A	生活时尚	177	290
B	影视娱乐	725	885
C	休闲旅游	35	37
D	计算机科技	929	999
E	大众科技	128	447
F	文教卫生	244	473

增量训练部分，我们使用该新闻收集系统，每天从新浪和搜狐等些著名的网站上搜集已经分好类的新闻。每天自动添加2000篇左右的文本到训练集里，并去掉相应篇数的旧新闻。每10天进行一次手工数据分类，进行完整的结果评估。

主要采用以下几个指标评估文本分类系统：查全率(recall)、查准率(precision)[1]和 F1测试值[16]。同时使用的还有微平均(micro-average)和宏平均(macro-average)两种评估方法[2]。

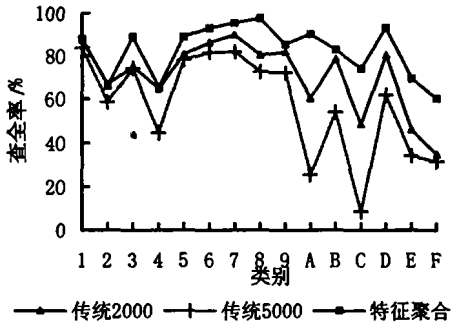
KNN 方法被证明是分类效果最好的方法之一[2]，因而我们以 KNN 为例进行实验，并设置 k 值为10。测试环境为 PIII-800，128M 内存，VC6.0。

4 实验结果

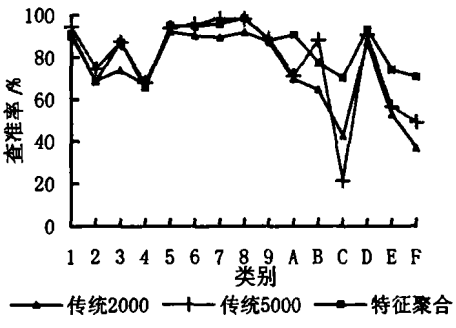
首先，对比应用特征词聚合的改进算法与传统 VSM 的静态分类效果，以及特征词数对分类的影响。表2给出了使用初始训练集和测试集的微平均查全率(miR)、微平均查准率(miP)和微平均 F1(miF1)对比情况。图3是按照类别对比的查全率和查准率。

表2 miR、miP 和 miF1对照表

项目	传统2000词/%	传统5000词/%	特征聚合5000词/%
miR	70.85	57.56	82.50
miP	74.04	78.56	83.91
miF1	70.46	66.44	83.20



(a) 按类别对比两种方法的查全率



(b) 按类别对比两种方法的查准率

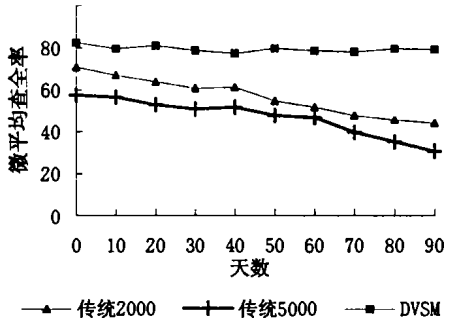
图3 静态环境两种模型性能比较

从结果中可以看出，改进方法明显优越于传统方法。而且对于传统方法来说，相对于特征词数的增加，由于不考虑特征间的关联关系，使得特征过于分散，并增加了无用特征词的数目，因而多数类别的分类查准率和查全率反而降低。

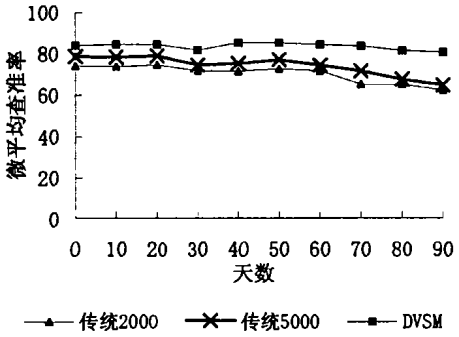
经特征聚合后，篮排球类、棋牌类、计算机科技类等类别界限较清晰的类，查准率和查全率提高比较明显。这些类别有些是由于其划分比较细，如体育类可分成若干明显内聚的小

类；有些是类别本身的区分特征非常明显，如计算机科技类。而对于大陆新闻和社会新闻等有交叉和类似内容的类别，由于区分特征不明显，则特征词聚合方法对其影响不大，甚至导致了查准率和查全率的下降，可通过选取代表性强的文章及特征项加权的方式来解决。

接下来对比动态数据情况下传统 VSM 和 DVSM 的分类效果。分别对两种模型持续测试了90天。图4给出了每10天进行一次评测的微平均查全率和查准率对比结果。



(a) 90天的 miR 对比结果



(b) 90天的 miP 对比结果

图4 动态环境两种模型性能比较

测试中，DVSM 能够保持 miR 在80%左右，miP 在83%～87%之间，miF1保持在80%以上。传统 VSM 在选择2000个特征词时，miR 下降到44%，miP 下降到62%，miF1下降到51%；选择5000个特征词时，miR 最终下降到31%，miP 下降到65%，miF1下降到42%。

静态数据情况下，DVSM 的训练时间为0.25秒/篇，比传统 VSM 的训练时间0.22秒/篇要长，分类时间基本相同(0.17秒/篇)。相对于结果的质量提高，这个时间是可以接受的。

动态数据情况下，如采用传统 VSM，由于需对所有样本重新训练，平均每次仍要约37分钟；动态经验训练只需0.35秒/篇，可以保证增量训练过程不影响正常的分类，验证了 DVSM 模型高效。

总的说来，DVSM 模型比传统 VSM 模型优越的主要原因在于：

1)聚合突出稀有词，排除了大量对分类无意义的低有效词，从而降低向量维数，减小计算开销。

2)特征模式作为文本特征的基本维，加强了特征项对文本分类的贡献，达到与概念推理网模型[15]类似的相关概念自动提取作用。

3)动态经验训练使得特征词可以随文本信息主题的改变在模式中移动位置，调整其对分类的贡献，适应 Internet 文本信息分类的实时要求。

4)只对增加的训练集合进行增量训练，对分类过程的中断影响较小。

结束语 本文提出的 DVSM 对 VSM 的特征提取策略进行了改进，并提出了特征聚合和增量训练算法，实际测试证实其有效。可以说 DVSM 是一种新思路，是在 Internet 文本分类方面的有益尝试。进一步的工作包括对特征词和聚合模式的加权处理，以优化界限不清晰类别的分类精度；搜集包含网络各个领域的更完整、更有代表性的新闻集合进行测试等。

参 考 文 献

1 Yang Y, Pedersen J P. A comparative study on feature selection in text categorization. In: Proc. of the Fourteenth Intl. Conf. on Machine Learning (ICML'97), San Francisco: Morgan Kaufmann Publishers. 1997. 412～420

2 Yang Y, Liu X. A re-examination of text categorization methods. In: Proc. 22nd Annual Intl. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999. 42～49

3 Salton G, Fox E A, Wu H. Extended Boolean information retrieval. Communications of the ACM, 1983, 26(11): 1022～1035

4 Han E H, George K, Vipin K. Text categorization using weight adjusted k-nearest neighbor classification: [Technical Report # 00-046]. Computer Science Department, University of Minnesota, 2000

5 孙丽华，张积东，李静梅. 一种改进的 KNN 方法及其在文本分类中的应用. 应用科技，2002，29(2)：25～27

6 Slaton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989

7 李宏乔，樊孝忠，李良富. 基于关键词与概念相结合的混合信息检索模型，搜索引擎与 web 挖掘进展. 高等教育出版社，2003. 41-45

8 萧嵘，王继成，孙正兴，张福炎. 一种 SVM 增量训练算法 α-ISVM. 软件学报，2001，12(12)：1818～1824

9 宫秀军，刘少辉，史忠植. 一种增量贝叶斯分类模型. 计算机学报，2002，25(6)：645～651

10 Ye Nong, Li Xiangyang. A scalable, incremental learning algorithm for classification problems. Computers & Industrial Engineering. 2002, 43: 677～692

11 http://www.sina.com.cn

12 http://www.sohu.com

13 http://www.xinhua.org

14 http://www.sdinfo.net

15 李晓黎，刘继敏，史忠植. 概念推理网及其在文本分类中的应用. 计算机研究与发展，2000，37(9)：1032～1038

16 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. 计算机应用研究，2001，18(9)：23～26