第45卷第4期 中国科学技术大学学报 Vol. 45, No. 4
2015年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2015

文章编号:0253-2778(2015)04-0314-07

基于类别信息优化的潜在语义分析分类技术

季 铎[1,2],毕 臣[2],蔡东风[2]

(1. 中国刑事警察学院网络犯罪侦查系,辽宁沈阳 110854;2. 沈阳航空航天大学知识工程中心,辽宁沈阳 110136)

摘要:潜在语义索引作为一种公认有效的矩阵降维技术,在关键词检索、文本分类等多种基于统计的机器文本学习任务中被广泛应用.基于专业文献的文本分类任务,结合严格分类体系下同类与不同类文本的特点,以专利文献分类为例,提出了一种基于类别信息优化的潜在语义分析分类技术.该方法根据分类文本各类别的特征信息,将原始文档分解为多种伪文档,强化不同分类的专属特征出现频率,进而优化构建潜在语义空间,提升模型分类性能.实验结果证明,专利文本分类任务结合该方法时,可以有效地提高分类的准确性.

关键词:潜在语义分析;特征共现;文本分类

中图分类号:TP311 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.04.009

引用格式:JI Duo, BI Chen, CAI Dongfeng. An latent semantic analysis classification technique based on optimized categorization information[J]. Journal of University of Science and Technology of China, 2015,45(4):314-320.

季铎,毕臣,蔡东风. 基于类别信息优化的潜在语义分析分类技术[J]. 中国科学技术大学学报,2015,45(4):314-320.

A latent semantic analysis classification technique based on optimized categorization information

JI Duo[1,2], BI Chen[2], CAI Dongfeng[2]

(1. *Cyber Crime Investigation Department, National Police University of China, Shenyang* 110854, *China*;
2. *Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang* 110136, *China*)

Abstract: As an effective method in the way of dimensionality reduction, latent semantic analysis(LSA) has been widely applied to many text learning missions, such as information retrieval and text categorization. Based on professional literature text classification tasks, features of text from same and different categories were analyzed under a strict classification system, patent documents classification was taken as an example, an optimized LSA classification technique was purposed based on categorization information. Utilizing features information from different category text, the technique divided original documents into a variety of fake documents, strengthens occurrence frequency of exclusive features from different categories, thus building optimized latent semantic space and improving the performance of the classification model. The experimental result shows that the method effectively improves categorization precision when applied to text categorization.

Key words: Latent Semantic Indexing; Term Co-occurrence; Text Categorization

收稿日期:2014-03-21;**修回日期**:2014-11-04

基金项目:辽宁省教育厅自然科学基金(L201120302)资助.

作者简介:季 铎(通讯作者),男,1981年生,博士/副教授.研究方向:数据挖掘. E-mail:jiduo_1@163.com

0　概述

随着网络化的推进和信息化时代的到来，信息的产生和传播速度不断提升，新的生产技术和理论方法产生速度逐年加快. 新技术、新专利浩如烟海，用户能否从海量数据中快速、准确地检索到适用于实际生产建设需求的信息，制约着企业乃至国家的经济发展和建设能力.

由于概念的表达具有多样性和随意性，同一概念可能有多种表述方法，而同一特征在不同语境下可能代表不同概念. 传统基于特征精确匹配的向量空间检索/分类模型很难为用户推送真正有效的信息.

潜在语义分析[1] (latent semantic analysis，LSA)，或称为潜在语义索引/隐含语义索引(latent semantic index，LSI)，是一种基于统计的无指导数据挖掘技术，能够在一定程度上解决传统向量空间模型难以处理的特征同义和歧义问题以及在此基础上出现的表达多样性问题，故已成为近年来国内外自然语言处理领域研究的一个重点. 如文献[2]将LSA 应用于科技文献和技术文档领域的分类任务，取得了较好的效果；文献[3]将 LSA 与支持向量机(support vector machine，SVM)结合并应用于情感分类任务，其方法被证明是可行的，具有优良的分类能力；文献[4]将 LSA 应用于电子邮件反钓鱼欺诈任务，对 E-mail 内容进行特征提取并基于可信性进行分类，取得了可喜的成果. 文献[5]提出了基于LSA 的多类分类模型，既能较好地解决文档中同义词和多义词的问题，又能解决多类属分类问题，在维度较低时分类性能稳定.

注意到现有的相关研究主要以将 LSA 与其他技术的融合应用到分类任务中为主，对 LSA 技术本身的优化较少；且对于专业文献的分类仅国外有初步研究，国内相关研究较罕见. 故针对专利文献领域分类任务，本文提出一种基于类别信息优化的潜在语义分析分类技术，与文献[6]单纯以人工语义资源作为伪文档的方法不同，该方法通过对类别信息进行特征提取和过滤，构建更为通用的伪文档并据此优化原始潜在的语义空间，进而增强潜在语义分析技术的分类性能. 实验表明，该方法能够有效地提升专利领域分类系统的性能.

1　潜在语义分析的基本方法

潜在语义分析以数学方法奇异值分解(single value decomposition，SVD)为基础. 奇异值分解将任意 $m\times n$ 阶实矩阵 $\boldsymbol{A}_{m\times n}$ 分解为三个矩阵 $\boldsymbol{U}$、$\boldsymbol{S}$、$\boldsymbol{V}$，且有

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

如图 1 所示(图中为 $m<n$ 情况)，其中 $\boldsymbol{U}$ 是 m 阶正交方阵，它的每一列称为 $\boldsymbol{A}$ 的左奇异向量. $\boldsymbol{V}$ 是 n 阶正交方阵，它的每一列是 $\boldsymbol{A}$ 的右奇异向量. $\boldsymbol{S}$ 是 $m\times n$ 阶的对角矩阵，其对角线元素称为 $\boldsymbol{A}$ 的奇异值.

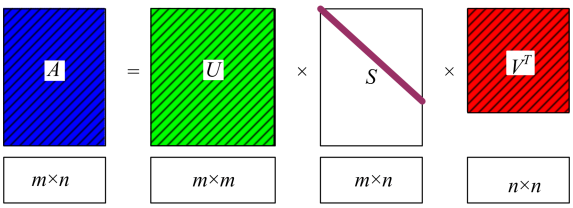

图 1　奇异值分解图示

Fig.　Singular value decomposition diagram

按照大小顺序排列，即有

$$\forall i,j,1\leqslant i\leqslant j\leqslant \min(m,n),\exists S_i\geqslant S_j \tag{2}$$

式中，$\boldsymbol{S}_i$ 为 $\boldsymbol{S}$ 对角线上第 i 个元素.

保留 $\boldsymbol{U}$、$\boldsymbol{V}$ 的前 k 个列向量和 $\boldsymbol{S}$ 中的前 k 个奇异值，分别得到 $\boldsymbol{U}_k$、$\boldsymbol{V}_k$ 和 $\boldsymbol{S}_k$，则有

$$\boldsymbol{A}_k=\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}} \tag{3}$$

式(3)称为截断的奇异值分解，如图 2 所示，且 $\boldsymbol{A}_k$ 为 $\boldsymbol{A}$ 在秩为 k 条件下最小二乘意义上的最优近似矩阵[1].

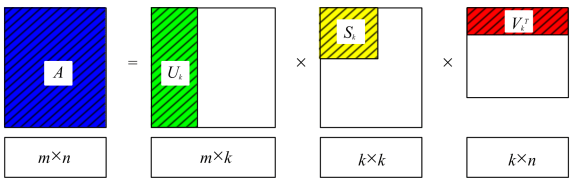

图 2　截断的奇异值分解

Fig. 2　Truncated singular value decomposition

当 $\boldsymbol{A}$ 为特征-文档矩阵时，即原始数据集含有 m 个特征、n 个文档，$\boldsymbol{A}_k$ 的每一行代表一个特征，每一列代表一个文档. 特征间相似度如下：

$$\boldsymbol{A}_k*\boldsymbol{A}_k^{\mathrm{T}}=\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}}*(\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}})^{\mathrm{T}}=$$
$$\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}}*\boldsymbol{V}_k\boldsymbol{S}_k^{\mathrm{T}}\boldsymbol{U}_k^{\mathrm{T}}=(\boldsymbol{U}_k\boldsymbol{S}_k)*(\boldsymbol{U}_k\boldsymbol{S}_k)^{\mathrm{T}} \tag{4}$$

文档间相似度如下：

$$\boldsymbol{A}_k^{\mathrm{T}}*\boldsymbol{A}_k=(\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}})^{\mathrm{T}}*\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}}=$$
$$\boldsymbol{V}_k\boldsymbol{S}_k^{\mathrm{T}}\boldsymbol{U}_k^{\mathrm{T}}*\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathrm{T}}=(\boldsymbol{V}_k\boldsymbol{S}_k)*(\boldsymbol{V}_k\boldsymbol{S}_k)^{\mathrm{T}} \tag{5}$$

故 $\boldsymbol{V}_k\boldsymbol{S}_k$ 的行向量表示原始文档在潜在语义空间中的投影向量.

许多研究者对于 LSA 理论进行分析,文献[7]指出 LSA 并非基于实体语义关系,而是通过统计方法获得了特征的共现信息,并通过数学方法进行了证明. 由此,加强合理的特征共现,优化构建潜在语义空间,有助于提升潜在语义分析模型的性能.

2 基于类别信息优化的潜在语义分析分类技术

对已分类文本进行潜在语义分析建立模型时,需要将合理的特征共现关系映射到潜在语义空间中. 其中,“合理的特征共现关系”指共现的特征对或特征集能够体现所属类别的特点、对于 query 文本的分类有指导作用. 本文的指导思想是对各类别文档进行处理,抽取具有合理共现关系的特征集,构建类别信息伪文档,将伪文档与原始文档合并处理. 伪文档中包含的大量合理特征共现信息对构建潜在语义空间起优化作用.

专利文献由于其专业性,文本中大量包含本专业领域的专业术语和相关词汇(以下简称为“专业特征”). 如农业专利多出现“肥料”“扦插”等专业词汇. 随分类层次加深,同类文档中涉及的专业特征越相近,共现程度越高. 而不同类专利之间专业特征的共现程度则较弱,如机械专利较少出现“肥料”等特征. 文献[8-9]的研究表明,在同样的设定下,采用术语作为特征的分类结果优于使用一般特征词,故使用与之作用近似的专业特征能够较好地区分专利所在的类别.

本文研究使用专利文档作为数据集. 不失一般性,对任何带有分类标签的文档集,均可对各分类利用下述方法构建类别信息伪文档并应用于潜在语义分析中. 下述方法应用对象包括但不限于专利文档.

2.1 基于全局特征分布信息的伪文档构建方法

对含有多个分类的数据集 D,且数据集中每一篇文档都都可以被唯一地分类到 n 个类(记为 C_i,$i=1,\cdots,n$). 则根据分类的不同,D 中文档可划分为 n 个文档子集 D_i,D_i 中所有文档均分类为 C_i,$i=1,\cdots,n$),$\{D_i\}i=1,\cdots,n$)是 D 的一个划分,即有

$$\forall 1\leqslant i\neq j\leqslant n, D_i\cap D_j=\Theta,\ \bigcup_{i=1,\cdots,n} D_i=D \tag{6}$$

原始数据集的每一个特征,可看作在一个或数个数据子集中出现. 特征在某一数据子集出现与否,依赖于该数据子集中包含的文档,是否出现这一特征. 根据特征在哪些分类对应的数据子集出现过,可以判断特征与数据子集的归属关系,示例文档如图 3 所示.

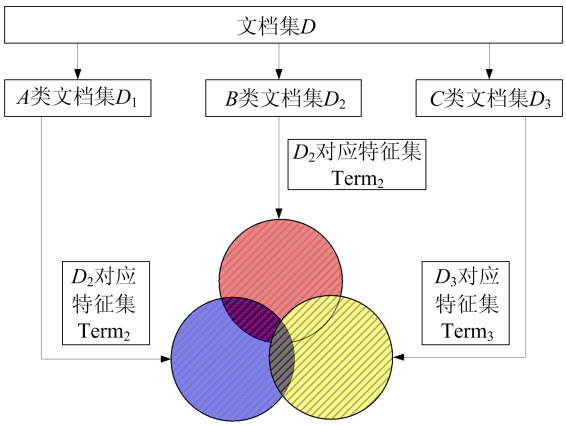

图 3 分类文档与特征关系示意图

Fig. 3 The relation of document classification and its characteristics

对每一个子集 D_i,可从中提取特征构成特征集 Term_i. 这里的特征为一元词语和二元词对,没有专利类别限制,故对不同的 i、j,Term_i 与 Term_j 之间存在特征交集. 这类特征(跨分类特征)对分类相似度计算是有一定混淆作用的. 根据下文 3.2 节相关实验,设置阈值,将出现在多个分类的特征过滤后,分类结果不升反降. 这是因为过滤过多特征后,文档与文档之间计算相似度时向量维度较低,考虑因素过少,因此需要对每一分类的专属特征权重进行增强,以在不减少特征数量的前提下淡化跨分类特征对分类计算的负面效果.

对每一分类子集 D_i,提取本分类 C_i 专属特征集 T_i 方法如下:

$$\forall i\in Z\cap[1,n], T_i=\mathrm{Term}_i-\bigcup_{\substack{j=1,\cdots,n\\ j\neq i}}\mathrm{Term}_j \tag{7}$$

用此方法提取某一类专属特征时,需要依赖其他各类文档的特征集进行特征筛选. 根据此方法提取出的特征集 T_i 即为该类文档集 D_i 的专属特征集合. 此时 T_i 的所有特征出现且仅出现在 D_i 所属分类中. 当语料文档为专利文献时,T_i 为给定 IPC 分类的专利文献集的专业特征. 提取各分类专属特征集后,将特征集附加分类标记,作为伪文档追加到原始文档集合中,如图 4 所示.

图 4 中，D 为原始的文档集；$C_i:D_i$ 表示文档集 Di 中的所有文档均包含分类标签 C_i；$C_i:T_i$ 表示该文档由分类标签 C_i 与特征集 T_i 中所有特征词构成的伪文档；全局伪文档集 GF 为所有根据本方法构建的伪文档构成的集合，其文档数应小于或等于分类数量；扩充文档集 DGF 为原始文档集 D 与全局伪文档集 GF 的合集，在 DGF 基础上进行潜在语义分析.

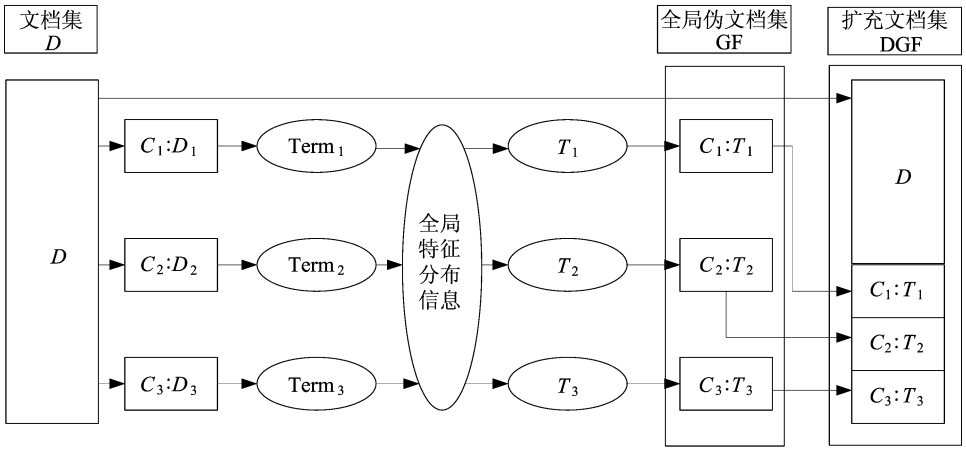

图 4　基于全局特征分布信息的扩充伪文档集构建过程图解

Fig. 4　Construct process of extended pseudo document set based on global feature distribution information

2.2　篇章伪文档构建方法

注意到全局伪文档构建方式构建出的伪文档数量较少，一个分类只能构建一篇伪文档，对于合理共现特征的加强效果较弱，引入篇章伪文档构建方法.

根据 2.1 中的示例，对于每类文档集 D_i，有

$$D_i = \{\mathrm{doc}_{ij}\}_{j=1,\cdots,n_i} \tag{8}$$

式中，doc_{ij} 为文档集 i 中第 j 篇文档，n_i 为 D_i 中包含文档的总数量.

考虑到 D_i 中每个文档 doc_{ij} 都包含特征集 Term_{ij}，Term_{ij} 为 D_i 特征集 Term_i 的子集，且有

$$\bigcup_{j=1,\cdots,n_i} \mathrm{Term}_{ij} = \mathrm{Term}_i \tag{9}$$

则对每篇文档 doc_{ij}，取其特征集 Term_{ij} 与本类专属特征集 T_i 的交集，构建基于篇章的伪文档. 构建过程如图 5 所示，仅取原始文档集的子集 D_1，设 D_1 含有两个文档.

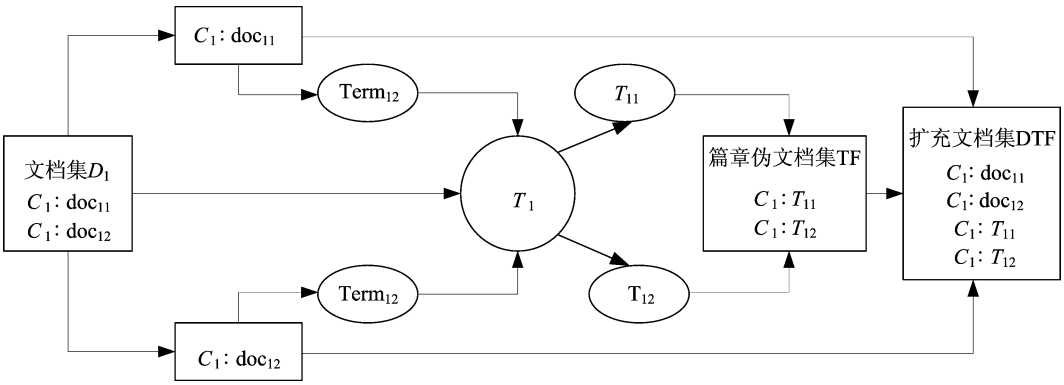

图 5　基于篇章特征分布信息的扩充伪文档集构建过程图解

Fig. 5　Construct process of extended pseado document set based on textural features distribution information

图 5 中，T_{1j} 为 Term_{1j} 与 T_1 的交集；$C_1:T_{1j}$ 为由文档分类标签 C_1 与特征词集 T_{1j} 中所有特征词构成的伪文档；篇章伪文档集 TF 为所有根据本方法构建的伪文档构成的集合，其文档数应小于或等于所有子语段的数量；扩充文档集 DTF 为原始文档集 D 与篇章伪文档集 GF 的合集，在 DTF 基础上进行潜在语义分析.

2.3　伪文档构建方法的融合

结合全局伪文档构建方法与篇章伪文档构建方法，将两种方法构建的伪文档全部加入原始文档集进行潜在语义分析处理、构建潜在语义空间，如图 6 所示.

3 实验与分析

3.1 实验基本设置

本文实验以专利分类性能为实验标准，采用 A 类专利文献 14 000 篇作为数据集（细节如表 1 所示）. 所涉及专利文献均为中文语料，按照国际专利分类标准进行分类.

表 1 中的文档在预处理前，类别分布均匀，每一分类含有 1 000 个文档. 选择均匀分布的文档原因是：根据文献[10]，不均衡数据集中的小类别对分类具有负面影响，均衡数据集的分类性能远超过不均衡的数据集.

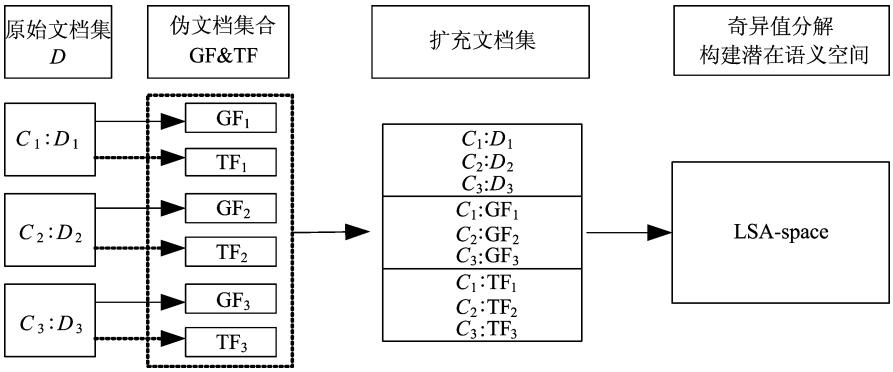

图 6 融合伪文档集构建过程图解

Fig. 6 Construct process of fusion pseudo document set

表 1 训练语料明细表

Tab. 1 Training corpus subsidiary

IPC	本类内容	数量
*A*01	农业；林业；畜牧业；狩猎；诱捕；捕鱼	1 000
*A*21	焙烤；制作或处理面团的设备；焙烤用面团	1 000
*A*22	屠宰；肉品处理；家禽或鱼的加工	1 000
*A*24	烟草；雪茄烟；纸烟；吸烟者用品	1 000
*A*41	服装	1 000
*A*42	帽类制品	1 000
*A*43	鞋类	1 000
*A*44	服饰缝纫用品；珠宝	1 000
*A*45	手携物品或旅行品	1 000
*A*46	刷类制品	1 000
*A*47	家具；家庭用的物品或设备；咖啡磨；香料磨；一般吸尘器	1 000
*A*61	医学或兽医学；卫生学	1 000
*A*62	救生；消防	1 000
*A*63	运动；游戏；娱乐活动	1 000

对将本数据集采用一元文法与二元文法相结合的特征抽取方法，过滤掉出现文档数低于 5 篇和高于 30% 的特征后，共获取 20 504 个不同有效特征. 在此基础上从数据集的每一分类中随机抽取 50 篇文档作为测试语料，其余作为训练语料.

本文采用 KNN 分类方法对专利文本进行分类. K 近邻法（K nearest neighbors，KNN）是机器学习中的一种重要的基于实例的学习方法. 对于已经标注好分类标签的训练集，给定一个测试文档，KNN 将该文档分到它 k 个距离最近的“邻居”（本文中指相似度最高的 k 个训练文档）中最多数文档所属的类别. 在本文中，k 取值为 10.

对给定测试文档 q，q 与类别 C_i 的相似度公式如下：

$$\mathrm{sim}(q,Ci)=\sum_{d\ \mathrm{in}\ Ci}\mathrm{sim}(q,d) \tag{10}$$

式中，d 为所有分类为 C_i 的训练文档，$\mathrm{sim}(q,d)$ 为 q 与文档 d 的相似度.

本文采用准确率作为评价分类性能好坏的标准. 计算公式如下：

$$\mathrm{score}=nr/n \tag{11}$$

式中，nr 为被正确分类的测试文档数量，n 为全部测试文档数量.

3.2 特征类别限制实验

本实验的主要目的是验证当一个特征同时在多个类别的文档中出现时，假设该特征对分类存在混淆效果，会降低系统性能. 基于此假设，在特征筛选时根据设置的阈值过滤掉出现在多个类别的特征，可能提升分类系统的性能. 本实验基于此假设设置，用于验证假设成立与否.

给定含有 n 个文档的数据集 D，D 具有 m 个分类 $\{C_i\}_{i=1,\cdots,m}$，从该数据集中可抽取合计含有 P 个特征的特征集 $T=\{t_i\}_{i=1,\cdots,p}$. 每个特征 t_i 都可能在多个分类的文档中出现，记为 $t_i \in \{C_j\}$. 其中 $\{C_j\}$ 为 t_i 出现过的文档所属的分类集合，其大小随特征 t_i 不同而变化. 对 $\{C_j\}$ 的大小进行限制，当 $\{C_j\}$ 的大小超过某阈值时，则认为该特征 t_i 对分类算法没有帮助，将其从 T 中移除. 实验增加此特征筛选过程后，直接构建得到潜在语义空间对专利分类任务的性能.

实验预期为进行特征筛选后 LSA 得到更合理的潜在语义空间，表现在数据上，为专利分类任务的性能得到提升. 为了避免其他优化算法的干扰，本实验仅基于传统的潜在语义分析技术进行处理，对比在不同类别数量限制下，筛选特征后的 LSA 分类性能.

表 2 为分别限制类别数量为 5、10 以及无限制时，分类准确率随 SVD 参数变化的数据表(根据实验设置，KNN 参数始终为 10).

表 2　特征类别数量限制与分类准确率相关性

Tab. 2　Feature category limit and classification accuracy

SVD	类别限制		
	无	5	10
60	0.6580 71	0.628	0.656
120	0.6708 48	0.628	0.656
180	0.6991 05	0.636	0.68
240	0.7320 92	0.628	0.648
300	0.7264 52	0.624	0.668

实际实验数据如表 2 所示，其中类别限制为 5 表示单个特征最多允许出现在 5 个类别中；类别限制为 10 表示单个特征最多允许出现在 10 个类别中；类别限制为无表示不限制特征出现的类别数量. 不符合类别限制的特征被过滤，不参与建模和后续处理. 截断奇异值分解的参数取值范围为从 60 至 300.

与实验预期相反，对整体比较而言，限制特征出现的类别阈值越小，相同 SVD 参数下专利分类的准确率越低，这一现象在 SVD 参数大于 100 后稳定出现. 分析认为，较严格的类别数量限制大幅减少了特征数量，降低了原始向量空间模型的维度，进而使映射的潜在语义空间中的特征点稀疏化. 这使得在提问式与训练数据做相似度计算时，不能很好地匹配特征，在 KNN 分类前不能根据相似度选择合适的邻居，因此会导致限制严格时分类性能降低.

实验结论：直接通过出现类别数量筛选特征是不可行的，严格的筛选策略反而降低模型分类性能. 要加强合理的特征共现现象，需要强化合理共现特征的出现频率，而不是去除导致不合理共现的特征.

3.3　多策略伪文档优化方案的比较

本实验将传统的潜在语义分析方法、基于全局特征分布的伪文档优化方法、基于篇章的伪文档优化方法、融合的伪文档优化方法进行横向比较，通过对专利分类性能的数据评估，分析不同方法的性能优劣及各方法的优势与不足.

实验预期为与传统的潜在语义方法相比，基于伪文档的优化方法应具有更好的性能；与其他伪文档优化方法相比，融合的伪文档优化方法性能最优.

实际实验数据如图 7 所示，其中“LSA”表示使用基本的潜在语义分析方法；“LSA＋ Global”表示使用基于全局特征分布的伪文档优化方法；“LSA＋Term”基于篇章的伪文档优化方法；“LSA ＋Combine”表示使用融合的伪文档优化方法.

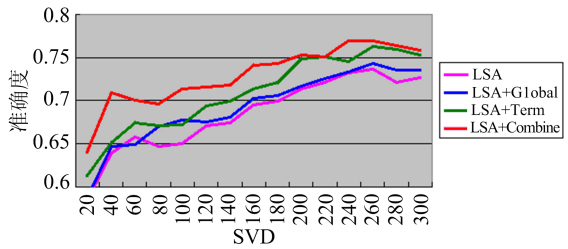

图 7　不同维度下 SVD 准确率变化趋势

Fig. 7　Change trend of SVD accuracy in different dimensions

由图 7 中的趋势曲线可知，由于基于全局和基于篇章的伪文档构建方式都是在原有数据集基础上扩充文档的，所以在 SVD 截断参数较低时会有曲线混叠现象，即优化方法的正确率反而不如基本方法；但 SVD 截断参数较高时，潜在语义空间能够更好地把握特征之间的共现关系，此时两种未融合的优化方法的效果要好于基本的潜在语义方法.

同时，由于基于全局特征分布的优化方法构建伪文档较少，虽然它能较好地提取各类的专属特征，但是伪文档数量较少导致其对原始文档集的影响较小，从而使该方法的优化幅度较低(相对基于篇章的优化方法).

两种方法融合后，由于是模型级别的融合，两种伪文档对特征的合理共现现象强化效果更为显著，

其融合效果明显高于基本的潜在语义分析方法.

4 结论

本文对潜在语义分析中的特征共现问题以及其对映射潜在语义空间合理性的影响进行了充分的分析,由强化特征的合理共现现象角度提出了基于伪文档优化的潜在语义分析技术思想.提出了多种不同的伪文档构建方式,并试验了不同方式对潜在语义模型性能的影响.实验显示单纯地特征过滤对潜在语义分析具有负面影响,而融合的伪文档构建方式对潜在语义模型的性能有较好的优化作用.

参考文献(References)

[1] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[2] Thorleuchter D, Van den Poel D. Technology classification with latent semantic indexing[J]. Expert Systems with Applications, 2013, 40(5): 1786-1795.

[3] Wang L, Wan Y. Sentiment classification of documents based on latent semantic analysis[A]// Advanced Research on Computer Education, Simulation and Modeling Communications in Computer and Information Science, Berlin: Springer, 2011, 176: 356-361.

[4] L'Huillier G, Hevia A, Weber R, et al. Latent semantic analysis and keyword extraction for phishing classification [C]// International Conference on Intelligence and Security Informatics. Vancouver, Canada: IEEE Press, 2010: 129-131.

[5] 叶浩,王明文,曾雪强. 基于潜在语义的多类文本分类模型研究[J]. 清华大学学报(自然科学版), 2005, 45(S1): 1818-1822.
Ye H, Wang M W, Zeng X Q. Automatic text mult-classification model based on latent semantic [J]. Journal of Tsinghua University (Science & Technology), 2005, 45(S1): 1818-1822.

[6] 郭东波. 基于伪文档的潜在语义索引优化技术的研究[D]. 沈阳航空工业学院, 2010.

[7] KontostathisA, Pottenger W M. A mathematical view of latent semantic indexing: Tracing term co-occurrences [R]. LU-CSE-02-006, Department of Computer Science and Engineering, Lehigh University, 2002.

[8] LarkeyL S. Some issues in the automatic classification of U. S. patents[R]. Technical Report WS-98-05, Learning for Text Categorization, 1998: 87-90.

[9] 屈鹏,王惠临. 专利文本分类的基础问题研究[J]. 现代图书情报技术, 2013, (3): 38-44.

[10] 张启蕊,张凌,董守斌,等. 训练集类别分布对文本分类的影响[J]. 清华大学学报(自然科学版), 2005, 45(S1): 1802-1805.
Zhang Q R, Wang H L Dong S B, et al. Effects of category distribution in a training set on textcategorization[J]. Journal of Tsinghua University (Science & Technology), 2005, 45(S1): 1802-1805.