doi:10.11835/j.jssn.1000-582X.2013.08.023

结合改进非负矩阵分解的模糊网页文本分类算法

贾兆红,李龙澍,朱建建
(安徽大学 计算智能与信号处理教育部重点实验室,合肥 230039)

摘 要:通过构建向量空间模型可以获得表征网页数据的词-文本权重矩阵,然而直接基于此高维矩阵进行分类学习效率较低,为此提出一种结合改进非负矩阵分解的模糊网页文本分类算法。首先,通过迭代的归一化压缩非负矩阵分解将高维的原数据映射到低维语义空间,以降低问题的复杂性。然后,将模糊逻辑引入分类模型,通过特征词与类别的模糊隶属度来生成文本的类别模糊集,以解决确定性矩阵难以判定语义模糊词所属类别的问题。实验结果表明,与其他方法相比,所提出的分类算法具有较高的分类准确度和较好的时间性能。

关键词:分类;非负矩阵分解;模糊逻辑;隶属函数
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1000-582X(2013)08-156-07

Fuzzy webpage text classification algorithm combined with improved NMF

JIA Zhaohong, *LI Longshu*, *ZHU Jianjian*
(Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230039, China)

Abstract: An item-document weight matrix representing the web pages could be generated by constructing the vector space model. Since the efficiency of direct classification through the high-dimensional matrix is relatively low, a fuzzy webpage text classification algorithm combined with improved nonnegative matrix factorization (NMF) is presented. Firstly, the original high-dimensional data are mapped into the low-dimensional semantic space via an iterative normalized compression NMF(NCMF) to reduce the complexity of the problem. Secondly, in order to solve the problem of categorizing ambiguous words by using deterministic matrices, fuzzy logic is incorporated into the classification model, where the fuzzy categorization set of the document is constructed with the fuzzy membership degree between features and categories. Comparative experiment results demonstrate the proposed classification algorithm possesses higher accuracy and better time performance.

Key words: classification; nonnegative matrix factorization; fuzzy logic; membership function

“信息爆炸”使得从网络获取有用信息遇到一些瓶颈问题,如用户就某个主题搜索特定信息,结果却找到大量不相关的网页等[1]。解决这些问题的有效方法之一是通过遍历网页,收集与某个主题相关的

收稿日期:2013-03-01
基金项目:国家自然科学基金资助项目(71171184);安徽省自然科学基金资助项目(090412054);教育部高等学校博士学科点专项科研基金资助项目(200803580024);安徽大学青年科学研究基金项目(3305044);人才科研启动项目(2303224)
作者简介:贾兆红(1976-),女,安徽大学副教授,博士,主要从事数据挖掘及商务智能研究,(E-mail) zhjia@mail.ustc.edu.cn。

文档并找出这些文档与该主题之间的可能联系,而这必定需要一个用于判定某个网页是否与特定主题相关的分类机制[2]。相对于人工分类,自动网页分类的成本更低,速度更快,因而随着网络资源的持续增长,自动分类的重要性变得越来越明显。自动网页分类是一个有监督的学习过程,通过一组加了标签的网页来训练分类器,然后用分类器为待分类的网页加上一个预定义的标签[3]。由于在海量的网页信息资源中,绝大多数信息以文本形式存在,因此自动网页文本的分类技术成为最具研究意义的网页分类热点问题之一。

近年来许多研究者将模式分类技术应用到网页文本分类问题中,于是出现了大量应用于网页文本分类的模型和学习技术。此外,还出现了一些针对中文网页分类的技术,如基于线性的分类器即向量空间模型(vector space model,VSM)[4]、神经网络模型[5]等,其中 VSM 因简单而广泛用于自动文档分类。基于 VSM 进行网页文本数据分类时,通常先将文本抽象为一个矩阵,然后针对所构造的矩阵进行分类训练。文本分类器常常采用决策树[6]、K 最近邻(K-nearest neighbor, KNN)[7]、神经网络[8]、朴素贝叶斯[9]、支持向量机[10]等机器学习方法,这些方法适用于特征数较少的分类问题,在对高维文本特征矩阵进行分类训练时通常效率较低,而网页文本数据经过抽象得到的是高维矩阵,因而上述方法直接用于求解网页分类问题时,其性能往往受到影响。针对这个问题,可通过先对高维文本特征矩阵降维,减小问题规模,降低问题复杂度,再进行分类学习的方法来提高求解效率。常用的数据降维方法有奇异值分解(singular value decomposition, SVD)[11]、NMF[12]等算法。SVD 算法将原矩阵分解为 3 个矩阵的乘积形式,NMF 算法通过简单迭代计算将原矩阵分解为 2 个矩阵的乘积,前者执行时间较长且分解过程较繁琐,而后者求解过程相对简单,分解所得的左右非负矩阵占用的存储空间相对较小且语义解释性强,因而适用于处理规模较大的问题[13]。在用 NMF 进行矩阵分解时一般需要对分解后的 2 个矩阵进行初始化,而且要通过复杂的更新规则来提高降解矩阵的稀疏度,以提高模型的区分能力。研究采用一种基于 NMF 的归一化压缩(NCMF)的矩阵分解算法[14],此方法只需初始化一个矩阵且采用的迭代更新规则更简单。

在 VSM 中常存在大量难以用"有"或"无"来明确定义其与文本相关性的特征词,而且文本中也包含一些语义多样的词汇,常使得难以明确判断文本与类别是"有关"还是"无关",因而直接对 VSM 的高维确定性矩阵进行分类通常效果也不理想。针对此问题,在所提出的分类模型中分别引入模糊相似度来计算特征词与文本及类别的相关系数,以获得更好的分类效果。

1　结合 NCMF 的模糊网页分类算法

假定 $P=\{p_1,p_2,\cdots,p_n\}$ 表示 n 个网页文档构成的文本集,$W=\{w_1,w_2,\cdots,w_m\}$ 表示特征词集,类别集合为 $C=\{c_1,c_2,\cdots,c_c\}$。训练样本集由已标注类别的文本组成,即 $T=\{<p_j,c(p_j)>\mid p_j\in P,c(p_j)\in C,1\leqslant j\leqslant n\}$,其中 $c(p_j)$ 表示网页文本 p_j 所属的主题类别。

1.1　构造向量空间模型

构建 VSM 是将网页文本数据转化为计算机可识别格式的有效方法之一,研究采用 TF-IDF 方法,该方法综合考虑各词在相应文本及其他类别主题中分别出现的次数,具有较好的区分特性[15]。这里,每个网页文本 p_j 表示为一个向量 $\boldsymbol{\Phi}_j=\{\varphi_{j1},\varphi_{j2},\cdots,\varphi_{jm}\}$,$\varphi_{ji}(w_i,p_j)$ 表示特征词 w_i 在 p_j 中的权重,其计算公式如下

$$\varphi_{ji}=\frac{f_{ji}\times\log\left(\frac{n}{n_i}\right)}{\sqrt{\sum_{i=1}^{m}\left(f_{ji}\times\log\left(\frac{n}{n_i}\right)\right)^2}},\tag{1}$$

其中:f_{ji} 表示 w_i 在 p_j 中出现的次数;n 表示网页文本总数,n_i 为含有 w_i 的网页文本数。于是基于文本集 P 可构造一个 $n\times m$ 的词-文本权重矩阵 $\boldsymbol{\Phi}(W,P)=(\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_2,\cdots,\boldsymbol{\Phi}_n)^{\mathrm{T}}$。

1.2　基于 NCMF 的特征约简

在文本分类中,不需要精确区分文本的每个细节,只需获得该文本在总体上的类别倾向,因此直接基于稀疏的 VSM 进行文本分类必然产生数据冗余并导致大量的资源消耗[16]。针对这个问题,可以将这个高维稀疏矩阵向低维紧凑语义空间进行映射,以削减空间维数,找出最具代表性的特征[17]。这里可用 NMF 算法进行特征提取,将高维词-文本矩阵 $\boldsymbol{\Phi}$ 分解为 2 个非负矩阵的乘积,即 $\boldsymbol{\Phi}=\boldsymbol{VH}$,左矩阵 $\boldsymbol{V}$ 可看作数据的压缩形式,右矩阵 $\boldsymbol{H}$ 可看作描述不同维映射关系的投影矩阵,并以 KL 散度作为相似度衡量,其目标函数如公式(2)所示

$$F(\boldsymbol{V},\boldsymbol{H})=\min_{V\geqslant 0,H\geqslant 0}\sum_i\sum_j\left(\varphi_{ij}\cdot\lg\frac{\varphi_{ij}}{(\boldsymbol{VH})_{ij}}-\varphi_{ij}+(\boldsymbol{VH})_{ij}\right),\tag{2}$$

其中:$\boldsymbol{\Phi}\in\boldsymbol{R}^{n\times m}$,$\boldsymbol{V}\in\boldsymbol{R}^{n\times r}$,$\boldsymbol{H}\in\boldsymbol{R}^{r\times m}$,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant$

n,r 为低维空间的维数,一般满足$(n+m)\ll n\times m$。由于分解前后的矩阵中仅包含非负元素,因此 $\boldsymbol{\Phi}$ 的任一行向量也可解释为 $\boldsymbol{H}$ 中所有向量(称为基向量)的加权和,即 $\boldsymbol{\Phi}_j=\sum_{k=1}^{r}\boldsymbol{V}_{jk}\times\boldsymbol{H}_j$,权重系数为 $\boldsymbol{V}$ 中对应行向量中的元素,故 $\boldsymbol{V}$ 也被称为系数矩阵。进一步分析可以发现,每个样本通过 $\boldsymbol{H}$ 被映射到一个 r 维空间中,即文本集中每个文本的 $\boldsymbol{H}_j$ 都是相同的,所以 $\boldsymbol{V}$ 中保留了所有样本的基本信息。$\boldsymbol{V}$ 和 $\boldsymbol{H}$ 分别采用如下迭代公式进行求解

$$\boldsymbol{V}_{ik}\leftarrow\boldsymbol{V}_{ik}\frac{\sum_{j=1}^{m}\dfrac{\boldsymbol{H}_{kj}\cdot\varphi_{ij}}{(\sum_{k'=1}^{r}\boldsymbol{V}_{ik'}\boldsymbol{H}_{k'j})}}{\sum_{j'=1}^{m}\boldsymbol{H}_{kj'}},\tag{3}$$

$$\boldsymbol{H}_{kj}\leftarrow\boldsymbol{H}_{kj}\frac{\sum_{i=1}^{n}\left(\dfrac{\boldsymbol{V}_{ik}\cdot\varphi_{ij}}{(\sum_{k'=1}^{r}\boldsymbol{V}_{ik'}H_{k'j})}\right)}{\sum_{i'=1}^{n}\boldsymbol{V}_{i'k}},\tag{4}$$

$\boldsymbol{V}$ 和 $\boldsymbol{H}$ 随公式(3)和(4)的迭代计算将变得越来越稀疏,然而这 2 个对称的更新规则会使它们的稀疏趋势相互制约,并最终导致其失去明显的稀疏特征,而 $\boldsymbol{V}$、$\boldsymbol{H}$ 特别是 $\boldsymbol{H}$ 的高稀疏性将有利于发现有意义的隐含特征,因而研究采用一种可有效提高 $\boldsymbol{H}$ 稀疏度的 NCMF 算法进行求解,假定 Max_iter 是预先设定的最大迭代次数,则算法过程描述如下。

Step1 按公式(5)归一化词-文本矩阵 $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi}_{ij}\leftarrow\frac{\boldsymbol{\Phi}_{ij}}{\sum_{i'=1}^{n}\boldsymbol{\Phi}_{i'j}};\tag{5}$$

Step2 初始化 $r\times m$ 维的 $\boldsymbol{H}$ 矩阵,iter=1;

Step3 分别按公式(6)、(7)更新矩阵 $\boldsymbol{V}$、$\boldsymbol{H}$

$$V_{ik}\leftarrow\frac{\sum_{j'=1}^{m}\boldsymbol{\Phi}_{ij'}\boldsymbol{H}_{kj'}}{\sum_{j=1}^{m}\boldsymbol{H}_{kj}},\tag{6}$$

$$\boldsymbol{H}_{kj}\leftarrow\sum_{i=1}^{n}\left(\frac{\boldsymbol{\Phi}_{ij}}{\sum_{k'=1}^{r}(\boldsymbol{V}_{ik'}\boldsymbol{H}_{k'j})}\boldsymbol{V}_{ik}\right)\boldsymbol{H}_{kj};\tag{7}$$

Step4 若 iter=Max_iter,结束;否则,iter++,转 step3。

从公式(6)和(7)可以看出,与 NMF 不同的是,这里 $\boldsymbol{V}$ 基于已知的 $\boldsymbol{\Phi}$ 和当前 $\boldsymbol{H}$ 进行更新,而 $\boldsymbol{H}$ 基于 $\boldsymbol{\Phi}$、$\boldsymbol{V}$ 及 $\boldsymbol{H}$ 自身进行更新,所以 NCMF 在初始阶段只需初始化 $\boldsymbol{H}$,然后再基于此更新,且不对称的更新规则使得 $\boldsymbol{H}$ 在更新时几乎不受约束。这说明在 NCMF 算法中,不需通过显式的稀疏处理即可获得较稀疏的 $\boldsymbol{H}$。

训练集及测试集的文本通过 NCMF 特征提取方法降维后,得到一个低维语义空间中的向量。尽管各词在不同主题中出现的频度可以表明该词与各主题的相关性,但是由于没有考虑到语义信息,直接基于此权重矩阵进行分类的效果往往不尽理想,因而在本文的分类模型中进一步引入特征词与类别的模糊相关性分析,为后面分析文本及其类别间的相似度提供更多的语义信息。

1.3 特征与类的模糊隶属度

由于特征词集能够表征数据集的主要特征,而基于特征词集可以得到数据集的类别属性,因而在特征词和类别之间存在某种相关性。首先采用隶属度矩阵 $\boldsymbol{Y}(\boldsymbol{W},\boldsymbol{C})=(y_1,y_2,\cdots,y_m)^{\mathrm{T}}$ 来定义特征词与类别间的相关性,其中 $\boldsymbol{y}_i(1\leqslant i\leqslant m)$ 是一个 c 维的向量,即

$$\boldsymbol{Y}=\begin{bmatrix}y_{11}&\cdots&y_{1c}\\ \vdots&&\vdots\\ y_{m1}&\cdots&y_{mc}\end{bmatrix},\tag{8}$$

其中 $0\leqslant y_{il}\leqslant 1(1\leqslant l\leqslant c)$ 表示特征词 w_i 对类别 $c_l(c_l\in C)$ 的隶属度。基于文献[18]中的讨论,综合考虑词在类别中出现的频率及特征词在文本集中的分布情况,采用如下定义

$$d_{il}=\frac{(\sum_{w_i\in p_j\wedge c(p_j)=c_l}f_{ji})}{\sum_{l=1}^{c}f_{ji}},\tag{9}$$

$$y_{il}=\frac{d_{il}\times n_i^l}{n^l}。\tag{10}$$

在公式(9)中,d_{il} 是基于统计方法计算的 w_i 在 c_l 中出现的频率,f_{ji} 表示 w_i 在网页文本 p_j 中出现的次数,若 w_i 只在属于 c_l 的文本中出现,则 $d_{il}=1$;若在多类的不同文本中出现,则 d_{il} 值较小,d_{il} 的值可以反映 w_i 在类别中的分布。在公式(10)中,y_{il} 为 $Y(W,C)$ 中的元素,n_i^l 表示在 c_l 中包含 w_i 的文本数,n^l 表示 c_l 中包含的文本总数,其计算公式为

$$n_i^l=|\{p_j\mid c(p_j)=c_l\wedge w_i\in p_j\}|,\tag{11}$$

$$n^l=|\{p_j\mid c(p_j)=c_l\}|。\tag{12}$$

由于在相应类中分布较均匀的特征词相对于只在个别文本中出现的特征词具有更好的区别能力,通过$\frac{n_i^l}{n^l}$可以使分布均匀的词具有较高权重,从而避免因文本分布不均导致分类精度降低。

通过词-文本矩阵 $\boldsymbol{\Phi}$ 与 $\boldsymbol{Y}$ 的乘积可以获得各文本的类别隶属度矩阵 $\boldsymbol{X}$,即

$$\boldsymbol{X}=\boldsymbol{\Phi Y},\tag{13}$$

其中 $\boldsymbol{X}$ 的元素 $x_{jl}(1\leqslant j\leqslant n,1\leqslant l\leqslant c)$ 表示文本 p_j 与 c_l 隶属度。根据 NCMF 算法得到 $\boldsymbol{\Phi}$ 的近似 $\boldsymbol{VH}$,公式(13)可改写为

$$\boldsymbol{X}=\boldsymbol{VHY}。\tag{14}$$

令 $\boldsymbol{U}=\boldsymbol{HY}$,则上式变为

$$\boldsymbol{X}=\boldsymbol{V}\boldsymbol{U}, \tag{15}$$

上式中的 $\boldsymbol{U}$ 可以解释为,对由 NCMF 算法得到的文本特征矩阵 $\boldsymbol{V}$ 所代表的新特征词集 $\boldsymbol{W}'=\{w'_1,w'_2,\cdots,w'_r\}$ 和类别集合 $\boldsymbol{C}$ 的相关性描述,即 $\boldsymbol{U}(\boldsymbol{W}',\boldsymbol{C})=(u_1,u_2,\cdots,u_r)^{\mathrm{T}}$,其中 $\boldsymbol{u}_k(k=1,2,\cdots,r)$ 也是一个 c 维的向量,也即

$$\boldsymbol{U}=\begin{bmatrix} u_{11} & \cdots & u_{1c} \\ \vdots & & \vdots \\ u_{r1} & \cdots & u_{rc} \end{bmatrix}, \tag{16}$$

其中 u_{kl} 表示 w'_k 对 c_l 的隶属度,其计算公式为

$$u_{kl}=\sum_{i=1}^{m}h_{ki}\times y_{il}。 \tag{17}$$

1.4　待测文本的类别判定

将待分类文本 q 的词-文本隶属向量 $\boldsymbol{\Psi}=(\Psi_1,\Psi_2,\cdots,\Psi_m)$ 经 NCMF 降维转为一个低秩向量 $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_r)$,然后利用上节分类训练得到特征类模糊相关性矩阵 $\boldsymbol{U}$,按照公式(18)来计算 q 与 $\boldsymbol{C}$ 中 c_l 的相似度

$$s_l(q,c_l)=\frac{\sum_{k=1}^{r}((\beta_k-\overline{\alpha_k})\times(u_{kl}-\overline{u_l}))}{\sqrt{\sum_{k=1}^{r}(\beta_k-\overline{\alpha_k})^2\times\sum_{k=1}^{r}(u_{kl}-\overline{u_l})^2}}, \tag{18}$$

其中 $1\leqslant l\leqslant c$,$\overline{\alpha_k}=\sum_{i=1}^{n}\frac{v_{ik}}{n}$ 表示词 α_k 的平均词-文本隶属度,$\overline{u_l}=\sum_{k=1}^{r}\frac{u_{kl}}{r}$ 表示 c_l 中的平均词类别隶属度,v_{ik} 是根据 NCMF 分解得到的词-文本隶属度,u_{kl} 是根据公式(17)得到的词类别隶属度。

通过计算 q 到所有类别的相似度,可得到类别模糊集 $\boldsymbol{S}(q)=\left\{\frac{s_1}{c_1},\frac{s_2}{c_2},\cdots,\frac{s_c}{c_c}\right\}$。在网页文本分类中,必须从这个模糊集中选择一个作为 q 的最佳类别,这里从 $\boldsymbol{S}(q)$ 中选择隶属度值最大的类别作为 q 的类别,即

$$c(q)=\max_{1\leqslant l\leqslant c}\{\boldsymbol{S}_l(q)\}。 \tag{19}$$

1.5　结合 NCMF 的模糊分类算法

基于上述讨论,给出提出的结合 NCMF 的模糊分类算法(fuzzy classifier combined with ncmf, NFC)流程如下

Step1 提取网页中的文本数据,进行预处理;

Step2 根据公式(1)构造训练文本集的向量空间模型,计算每个特征词与文本之间的模糊关系,生成词-文本隶属度矩阵 $\boldsymbol{\Phi}=(\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_2,\cdots,\boldsymbol{\Phi}_n)^{\mathrm{T}}$;

Step3 根据 NCMF 算法对 $\boldsymbol{\Phi}$ 进行分解,根据公式(6)、(7)迭代求解得到压缩的词-文本隶属矩阵 $\boldsymbol{V}$ 和投影矩阵 $\boldsymbol{H}$;

Step4 根据公式(9)和(10)计算训练集中特征-类别的隶属度矩阵 $\boldsymbol{Y}$;

Step5 根据公式(16)和(17)计算针对降解的压缩矩阵 $\boldsymbol{V}$ 的词类别隶属矩阵 $\boldsymbol{U}$;

Step6 对于输入的待分类测试文本 q,根据公式(1)构造 m 维隶属度向量 $\boldsymbol{\Psi}$,并通过 NCMF 算法转化为 r 维向量 $\boldsymbol{\beta}$;

Step7 根据公式(18)计算测试文本模糊集与各类别模糊集的相似度 $s_l(q,c_l)$,$1\leqslant l\leqslant c$,得到文本与类别的隶属度模糊集 $S(q)$;

Step8 根据公式(19)输出 q 的类别。

2　实验结果及分析

从互联网随机抽取网页组成实验数据进行验证,在数据准备阶段,以环境、教育、经济、游戏和体育为主题(记为 $c_1\sim c_5$)随机下载共 1100 个网页作为基础实验数据,各类文本数分布如表 1 所示,从每类中随机抽取 2/3 的网页作为训练集,余下 1/3 作为测试集。

表 1　各主题文本数分布

类别	c_1	c_2	c_3	c_4	c_5
文本数	215	222	220	219	224

由于中文不像英文通过空格来分词,因此需要对中文文本进行专门的分词处理,采用中科院计算所提供的汉语词法分析系统(ICTCLAS)进行分词,并由此得到一个词表。为了提高文本表征词典的质量以提高分类器训练的效率,将词表中的"停用词"过滤掉,并采用信息增益法[20]对文档进行特征词提取,,最后选取 3 000(即 m)个特征词来表达文本语义。通过实验确定了 NCMF 算法中的迭代次数 Max_iter=100。

为评价分类算法的性能,采用最通用的性能评价方法:准确率 Precision 和召回率 Recall。对于某个特定的类别,Precision 指在分类器所标记的正样本中被正确分类的正样本所占的比例,Recall 指被分类器正确分类的正样本占所有正样本的比例。Precision 和 Recall 从不同角度反映了分类质量,但是二者必须综合考虑,不可偏废,因此通常采用 Precision 和 Recall 的加权调和平均,即 F_1 测试值来评价分类算法的优劣[7]。F_1 的值越大,则表示分类器的性能越好。通常采用微平均 Micro_F_1(记为 F_1)值评价分类器在每一类上的性能,而采用宏平均

Macro－F_1 值(记为$\overline{F_1}$)评估算法在整体数据集上的性能。

假定某个类 c_l 的偶然事件表如表 2 所示,其中“＋”表示正样本即属于 c_l 的样本,“－”表示负样本即不属于 c_l 的样本,$tp(tn)$表示通过分类器正确分类的正(负)样本,$fp(fn)$表示错误分类的正(负)样本。

表 2 类 c_l 的列联表

c_l		人工分类	
		＋	－
分类器分类	＋	tp_l	fp_l
	－	fn_l	tn_l

分类器在类 c_l 上的微平均 Precision_l 和 Recall_l 的计算公式分别如公式(20)和(21)所示

$$\text{Precision}_l = tp_l/(tp_l + fp_l); \tag{20}$$

$$\text{Recall}_l = tp_l/(tp_l + fn_l), \tag{21}$$

类 c_l 上的微平均 F_{1l} 值的计算方法如公式(22)所示[7]

$$F_{1l} = \frac{2 \times \text{Precision}_l \times \text{Recall}_l}{(\text{Precision}_l + \text{Recall}_l)}。 \tag{22}$$

在 c 个类别上的宏平均$\overline{F1}$计算公式如公式(23)所示

$$\overline{F_1} = \sum_{l=1}^{c} F_{1l}/c。 \tag{23}$$

2.1 参数 r 的确定

由于 NCMF 是 NFC 算法的重要步骤之一,而 r 是影响 NCMF 性能的主要参数,因而 r 也是 NFC 算法的重要参数。通过实验测试了不同的 r 值对算法 $\overline{F_1}$ 测量值的影响,实验结果如图 1 所示。

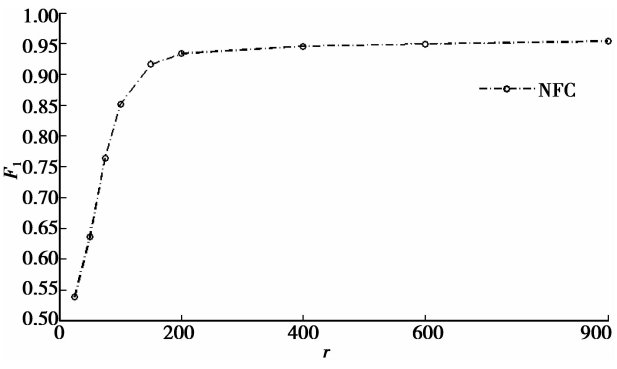

图 1 不同 r 值对 NFC 分类精度的影响

从图 1 可以看出,不同的 r 值对 NFC 算法$\overline{F_1}$值的影响,$r=25$ 时,$\overline{F_1}$只有 53.8%,$\overline{F_1}$的值随着 r 不断增大而增大,表明分类结果越来越好,但是增大 r 会相应地增加计算代价,反之 r 值越小,计算代价越小,分类的准确度也就越低。当 r 值很小时,NCMF 消减了原始词-文本矩阵中的绝大部分信息,分解后的语义矩阵不足以有效表示原始信息;当 r 逐渐增大时,降解后的低秩压缩矩阵中的语义信息越来越丰富,相对原始矩阵来说越来越能代表有效信息,当 $25 \leqslant r \leqslant 150$ 时,$\overline{F_1}$随 r 的增大而由 53.8%大幅度增到 91.7%,但当 $r>150$ 时,$\overline{F_1}$的增幅已经很小,说明在 $r=150$ 维时,降解的矩阵已经具有分类所需相对充分的语义信息,这时再增加 r 的值会使得降解矩阵中出现较多的冗余信息,这不仅不能明显提高分类效果反而会增加计算代价。因而综合考虑,在后面的分类学习中选择 $r=150$,这样既有很好的代表性又使得语义矩阵具有较少的维数,以降低分类过程中的时间和空间复杂度。

2.2 比较实验及分析

为了验证 NFC 算法的有效性,选择 KNN 分类算法[7](实验中取 $k=5$)、基于模糊关系的 FRM 算法[18]和基于 SVD 分解的 LSI 算法[19]进行比较。由于 NCMF 采用的随机初始化方法对分类结果有一定的影响,因此将算法运行 20 次,然后取 20 次运行结果的平均指标值进行比较。图 2 给出了 4 种算法在 5 大主题上的分类结果的比较。

从图 2 可以看出,NFC 只在类 c_2 上的 F_1 值劣于 FRM 算法,且在其他所有类别上的 F_1 值均优于 LSI 和 KNN 算法。

表 3 给出 4 种算法的$\overline{F_1}$测量值和运行时间(单位“s”)比较结果,其中运行时间指算法的训练时间与测试时间之和。

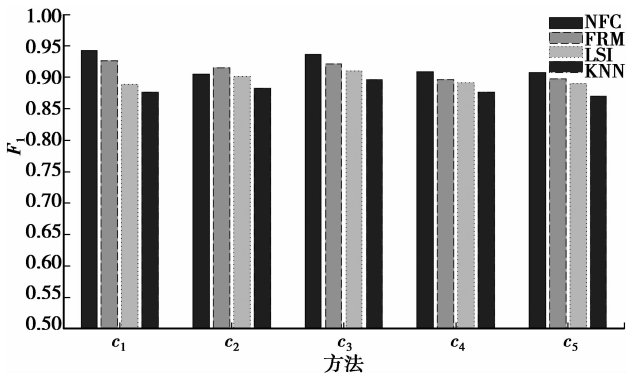

图 2 各类的 F_1 值比较

表 3　各算法的$\overline{F_1}$值与运行时间比较

算法	KNN	LSI	FRM	NFC
$\overline{F_1}$/%	87.96	89.66	90.98	91.68
运行时间/s	26.69	8.59	16.77	5.32

从表 3 可以看出，在 4 种算法中，采用模糊逻辑的 NFC 和 FRM 算法的分类准确度均优于另外 2 种方法，而 NFC 通过 NCMF 降维处理使其运行时间相对于 FRM 有了 68% 的改进，而且在$\overline{F1}$值上也有所改进。因而可以说 NFC 算法的总体性能最优。

通过进一步分析可以发现，NFC 虽然对表征样本的词-文本权重矩阵采用 NCMF 进行了降维处理，但这种数据约简的方法并没有降低分类精度，且其分类性能指标值优于未采用降维处理的 FRM 算法。这说明在初始的权重矩阵中包含大量的冗余信息，直接对该矩阵进行分类学习的效果未必理想，而通过去除一定的冗余信息，反而找出了与文本语义更紧密的特征词，所得的系数矩阵可以作为代表原始样本的特征矩阵，基于此低维特征矩阵进行模糊分类训练可以提高分类准确率。LSI 算法中虽然采用 SVD 矩阵分解方法进行降维，但其计算过程较 NCMF 复杂，运行时间较长，而且 SVD 分解所得的基向量元素中含有负值，不是真正的基于部分的分解，可解释性也较差。

3　结　论

采用一种基于 NMF 的改进算法，即 NCMF 算法，将高维的网页文本数据映射到低维语义空间。NCMF 只需要初始化投影矩阵，且不对称的更新规则可得到更稀疏的投影矩阵，实验表明 NCMF 算法简单且有效。为了进一步提高分类精度，在算法中引入对特征词与类之间的模糊相关性测度，通过模糊相关系数计算文本到各类别的隶属度，并根据最大隶属度来确定文本的类别。通过从单个类及总体的分类准确度和执行时间等方面，与其他几种相关算法的比较实验结果表明 NFC 的总体性能优于其他算法，说明 NFC 是一种有效的网页文本分类算法。如何在词-文本及词-类别的模糊相关性分析中引入更丰富的语义信息，将是下一步的研究方向之一。此外，由于采用的测试数据集的规模还不是特别大，在对 2 个数据进行连乘时并未出现溢出这种极端情况，然而在处理海量数据时可能会出现溢出。针对在处理海量数据时，2 个数据连乘可能出现的溢出问题进行研究也是一个很有意义的研究课题。

参考文献：

[1] Ozel S A. A web page classification system based on a genetic algorithm using tagged-terms as features[J]. Expert Systems with Applications, 2011, 38(4): 3407-3415.

[2] Qi X, Davison B D. Web page classification: features and algorithms [J]. ACM Computing Surveys, 2009, 41(2):10-12.

[3] Kwon O W, Lee J H. Text categorization based on k-nearest neighbor approach for web site classification[J]. Information Processing and Management, 2003, 39(1): 25-44.

[4] Lin S S. A document classification and retrieval system for R&D in semi- conductor industry-a hybrid approach[J]. Expert Systems with Applications, 2009, 36(3): 4753-4764.

[5] Yu B, Xu Z, LI C. Latent semantic analysis for text categorization using neural network[J]. Knowledge-Based Systems, 2008, 21(8): 900-904.

[6] Li Y, Hung E, Chung K. A subspace decision cluster classifier for text classification[J]. Expert Systems with Applications, 2011, 38(10):12475 -12482.

[7] Jiang S, Pang G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications, 2012, 39(1): 1503-1509.

[8] Ghiass M, Olschimke M, Moon B, et al. Automated text classification using a dynamic artificial neural network model[J]. Expert Systems with Application, 2012, 39(12):10967-10976.

[9] Chen J, Huang H, Tian S, et al. Feature selection for text classification with Naïve Bayes[J]. Expert Systems with Applications, 2009, 36(3):5432- 5435.

[10] Lo S C. Web service quality control based on text mining using support vector machine[J]. Expert Systems with Applications, 2008, 34(1): 603-610.

[11] Li C, Park S C. An efficient document classification model using an improved back propagation neural network and singular value decomposition[J]. Expert Systems with Applications, 2009, 36(2): 3208-3215.

[12] Barman P C, Lee S Y. Document classification with unsupervised nonnegative matrix factorization and supervised percetron Learning[C]//Proceedings of the International Conference on Information Acquisition. Jeju City Korea: IEEE, 2007: 2782-2785.

[13] Lee J H, Park S, Ahn C M, et al. Automatic generic document summarization based on non-negative matrix

factorization [J]. Information Processing and Management, 2009, 45(1): 20-34.

[14] Zhu Z, Guo F, Zhu X, et al. Normalized dimensionality reduction using nonnegative matrix factorization[J]. Neurocomputing, 2010, 73(10-12): 1783-1893.

[15] Yan J, Liu N, Zhang B, et al. OCFS: optimal orthogonal centroid feature selection for text categorization [C]//Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador Brazil: Association for Computing Machinery, Inc., 2005: 122-129.

[16] Salton G, Buckley C. Term weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5):513- 523.

[17] Kim J. Toward faster nonnegative matrix factorization: a new algorithm and comparisons [C]//Proceedings of 8th IEEE International Conference on Data Mining, Pisa Italy: IEEE, 2008: 353-362.

[18] 张玉芳, 娄娟, 李智星,等. 基于模糊关系的文本分类方法[J]. 计算机工程, 2011, 37(16): 149-151.
ZHANG Yufang, LOU Juan, LI Zhixing. Text classification approach based on fuzzy relationship[J]. Computer Engineering, 2011, 37(16):149-151.

[19] Lee C H, Yang H E, Ma S M. A novel multilingual text categorization system using latent semantic indexing [C]//Proceedings of the First International Conference on Innovative Computing, Information and Control, Beijing China: IEEE, 2006: 503-506.

[20] 徐燕, 李锦涛, 王斌,等. 文本分类中特征选择的约束研究[J]. 计算机研究与发展, 2008, 45(4):596-602.
XU Yan, LI Jingtao, Wang Bin, et al. A study on constraints for feature selection in text categorization[J]. Journal of Computer Research and Development, 2008, 45(4): 596-602.

(编辑 侯 湘)

(上接第 155 页)

[6] Martinek J. A Novel approach to simulate Hodgkin-Huxley-like excitation with COMSOL multiphysics [J]. Artificial Organs. 2008, 32(8): 614-619.

[7] Gray E G. The fine structure of nerve [J]. Comparative Biochemistry and Physiology, 1970: 419-422.

[8] 丁光宏, 顾全保, 主译. 电生理学基础 [M]. 上海: 复旦大学出版社, 2005, 18.

[9] Hille B. Ionic channels in nerve membranes [J]. Progress in Biophysics and Molecular Biology, 1970, 21: 1-32.

[10] Hille B. Ionic channels of excitable membranes: Current problems and biophysical approaches [J]. Biophysical Society, 1978, 22(2): 283-294.

[11] Courtney L. Computational modeling of three-dimensional electrodiffusion in biological systems: Application to the node of ranvier [J]. Biophysical Journal. 2008, 95(6): 2624-2635.

[12] Noble D, Garny A, Noble P J. How the Hodgkin-Huxley equations inspired the cardiac physiome project [J]. Journal of Physiology. 2012, 590(11): 2613-2628.

[13] Elia S. A finite element model for the axon of nervous cells [J]. COMSOL Conference 2009, October 14-16 2009, Milan, Italy

[14] Moulin C. A new 3-D finite element model based on thin-film approximation for microelectrode array recording of extracellular action potential [J]. IEEE Transactions on Biomedical Engineering, 2008, 55(2): 683-692.

[15] Mclntyre C C. Cellular effects of deep brain stimulation: Model-based analysis of activation and inhibition [J]. Journal OF Neurophysiology. 2004, 91(4): 1457-1469.

(编辑 侯 湘)