第 24 卷 第 6 期
2005 年 12 月

情 报 学 报

Vol.24,№ 6
December,2005

文本分类中的特征降维方法综述[1)]

陈 涛 谢阳群

(宁波大学商学院信息管理系,宁波 315211)

摘要 文本分类的关键是对高维的特征集进行降维。降维的主要方法是特征选择和特征提取。本文综述了已有的特征选择和特征抽取方法,评价了它们的优缺点和适用范围。

关键词 文本分类 特征降维 特征选择 特征提取

Literature Review of Feature Dimension Reduction in Text Categorization

Chen Tao and Xie Yangqun

(*Department of Information Management, Ningbo University, Ningbo* 315211)

Abstract The key to text categorization is how to reduce the high-dimension of the feature vectors. Feature reduction method involves feature selection and feature extraction. In this paper feature selection methods and feature extraction methods are colligated. Their advantage and disadvantage are evaluated.

Keywords text categorization, feature reduction, feature selection, feature extraction.

1 引 言

文本分类(Text Categorization)是指在给定的分类体系下,根据文本内容自动确定文本类别的过程[1,2,3,4]。它是文本挖掘的一个重要组成部分,在提高情报检索的速度和准确率方面有重要意义。在文本分类中,广泛使用向量空间模型(Vector Space Model)来标引文本。即文本的特征直接采用文本中的词条 T(Token)作为特征项,文本可以表示为特征的向量 $d=(t_1,t_2,\cdots,t_n)$,分量 t_i 是词条对应的权值,利用训练好的分类器将文本自动分到类别 c_i(类别集合 $C=\{c_1,c_2,\cdots,c_n\}$,n 为类别数)[5]。而这些高维的特征集对分类学习未必全是重要和有效的,同时高维特征集会加剧机器学习的负担。是否进行特征降维对文本分类的训练时间、分类准确性都有显著的影响,而且分类器的算法和实现的复杂度都随特征空间维数的增加而增加。所以,特征集的降维操作是文本分类准确率和效率的关键。特征选择(Feature Selection)和特征抽取(Feature Extraction)是特征降维中的主要方法。以下分别对特征选择和特征抽取中涉及的不同方法进行介绍。

2 特征选择(Feature Selection)

特征选择就是从特征集 $T=\{t_1,\cdots,t_s\}$ 中选择一个真子集 $T'=\{t_1,\cdots,t_{s'}\}$,满足($s'\ll s$)。其中,s 为原始特征集的大小,s' 为选择后的特征集大小。选择的准则是经特征选择后能有效提高文本准确率。选择没有改变原始特征空间的性质,只是从原

收稿日期:2005 年 3 月 14 日

作者简介:陈涛,男,1973 年生,讲师,清华大学计算机科学与技术专业硕士。研究方向:文本信息处理,数据挖掘。谢阳群,1962 年生,教授,中国科学院文献情报中心情报专业博士。研究方向:信息资源管理,数据挖掘。

1) 浙江省教育厅 2004 年度高校科研项目(项目编号 20040997)。

始特征空间中选择了一部分重要的特征,组成一个新的低维空间[6,7,8,9]。

文本分类中,用于特征选择的统计量大致有:特征频度(Term Frequency),文档频度(Document Frequency),特征熵(Term Entropy),互信息(Multi-Information),信息增益(Information Gain),X^2 统计量(Chi-square),特征权(Term Strength),期望交叉熵(Expected Cross Entropy),文本证据权(Weight of Evidence for Text),几率比(Odds Ratio)等。这些统计量从不同的角度度量特征对分类所起的作用。在下面对各统计量的解释中,概率 P 都在文本空间上进行计算。

2.1　特征频度(Term Frequency,*tf*)

特征频度指训练集中特征 t_k 出现的次数。这是最简单的特征选择方法。直观上,特征在文本集中出现次数越多,对文本分类的贡献越大。由于原始特征集中绝大部分是低频特征,因此,设定 tf 阈值对过滤低频特征非常有效,可以获得很大的降维度。就高频特征而言,特征的统计分布决定了文本分类的准确率。即当该高频特征均匀地分布在所有文本中时,对分类的作用将是有限的。因此,tf 主要用在文本标引时直接删除某些低频特征[3,4,10,11]。

2.2　文本频度(Document Frequency,*df*)

文本频度是训练集中含有词条 t_k 的文本数在总文本数中出现的概率。其理论假设为稀有词条或者对分类作用不大,或者是噪声,可以被删除。文本频度较特征频度的统计粒度更粗一些,在实际运用中有一定的效果[12]。但是如果某一稀有词条主要在某类文本中出现的情况下,可能会把该类的显著特征错误地过滤掉。文献[12]通过实验表明,用 tf 和 df 的组合进行特征选择可以得到更好的降维效果。

2.3　特征熵(Term Entropy)

在文本分类中,特征 t_k 在类型集上的熵如式(1)所示。

$$Entropy(t_k) = -\sum_{i=1}^{m} P(c_i \mid t_k)\log(c_i \mid t_k) \quad (1)$$

其中,$P(c_i \mid t_k)$ 为 c_i 中出现特征 t_k 的文本数除以训练集中出现 t_k 的文本数。特征 t_k 相当于一个事件,类型集 C 相当于一个系统。它是 t_k 发生时系统 C 的条件熵,描述了当特征在文本中出现时,确定该文本所属类型的平均不确定性。特征熵越小,对文本分类的作用越大。进行特征选择时,选择熵比较小的特征。

2.4　互信息(Multi-Information,*MI*)

在文本分类中,特征 t_k 的互信息如式(2)所示。

$$\begin{aligned} MI(t_k, c_i) &= \log\frac{P(t_k, c_i)}{P(t_k)P(c_i)} \\ &= \log P(t_k \mid c_i) - \log P(t_k) \end{aligned} \quad (2)$$

其中,$P(t_k, c_i)$ 为 c_i 中出现特征 t_k 的文本数除以训练集的文本数,$P(t_k)$ 为训练集中出现 t_k 的文本数除以训练集的文本数,$P(c_i)$ 为训练集中属于类型 c_i 的文本数除以训练集的文本数。这里将 c_i 看成是系统 C 中的一个事件。$MI(t_k, c_i)$ 表示事件 t_k 和 c_i 发生时的互信息,即特征与类型之间的相关程度。当特征的出现只依赖于某一类型时,特征与该类型的互信息很大;当特征与类型相互独立时,互信息为 0;当特征很少在该类型文本中出现时,它们之间的互信息为负数,即负相关。频度小的特征对互信息的影响大。可以看出,对于频度小的特征,$\log P(t_k)$ 的变化比 $\log P(t_k \mid c_i)$ 快,是互信息中的主要部分,这使得低频特征具有较大的互信息[33]。特征 t_k 的全局互信息如式(3)所示。

$$MI(t_k) = \sum_{i=1}^{m} P(c_i)\log\frac{P(t_k, c_i)}{P(t_k)P(c_i)} \quad (3)$$

用 MI 选择特征时,应该选择互信息大的特征。由于 MI 有利于低频特征,因此容易引起过学习(Over-fitting)[12]。

2.5　信息增益(Information Gain,*IG*)

在文本分类中,特征 t_k 的信息增益如式(4)所示。

$$\begin{aligned} IG(t_k) = &-\sum_{i=1}^{m} P(c_i)\log P(c_i) \\ &+ P(t_k)\sum_{i=1}^{m} P(c_i \mid t_k)\log P(c_i \mid t_k) \\ &+ P(\bar{t}_k)\sum_{i=1}^{m} P(c_i \mid \bar{t}_k)\log P(c_i \mid \bar{t}_k) \end{aligned} \quad (4)$$

其中,$P(\bar{t}_k)$ 为训练集中不出现特征 t_k 的文本数除以训练集的文本数,$P(c_i \mid \bar{t}_k)$ 为类型 c_i 中出现 t_k 的文本数除以训练集中出现 t_k 的文本数。特征在文本中是否出现都将为文本分类提供信息,计算不

同情况下的条件概率以确定提供的信息量的大小。信息增益是机器学习领域较为广泛的特征选择方法。利用特征取值情况划分训练样本空间,根据所获得信息量的多少选择相应特征。进行特征选择时,选择信息增益大的特征。

2.6 X^2 统计量(Chi-square,*Chi*)

在文本分类中,特征 t_k 的 *Chi* 权重如式(5)所示。

$$Chi(t_k,c_i)=\frac{n[P(t_k,c_i)\times P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\times P(\bar{t}_k,c_i)]^2}{P_d(t_k)\times P_d(c_i)\times P_d(\bar{t}_k)\times P_d(\bar{c}_i)} \tag{5}$$

其中,n 为训练集的文本数,$P(\bar{t}_k,\bar{c}_i)$为训练集中不出现特征 t_k 并且不属于类型 c_i 的文本数除以 n,$P(t_k,\bar{c}_i)$为训练集中出现特征 t_k 并且不属于类型 c_i 的文本数除以 n,$P(\bar{t}_k,c_i)$为训练集中不出现特征 t_k 并且属于类型 c_i 的文本数除以 n。它度量了特征 t_k 和类型 c_i 之间的相关程度。*Chi* 值越大,表示 t_k 与 c_i 越相关,t_k 越依赖于 c_i。X^2 统计量与 *MI* 很相似。特征 t_k 的全局 *Chi* 值如式(6)所示。

$$Chi(t_k)=\sum_{i=1}^{m}Chi(t_k,c_i) \tag{6}$$

另一个更改进的全局 *Chi* 值如式(7)所示[33]。

$$Chi(t_k)=\max_{i=1}^{m}\{Chi(t_k,c_i)\} \tag{7}$$

进行特征选择时,选择 *Chi* 值大的特征。

2.7 相关系数(Correlation Coefficient,*CC*)

对于公式(5),文献[13]认为:分子取平方使得特征与类型的正相关能力与负相关能力被同等对待,但是对于分类来说,特征的重要性主要由特征与类型的正相关能力决定。基于这一观察,提出了特征的"相关系数",如式(8)所示。

$$CC(t_k,c_i)=\frac{\sqrt{n}(P(t_k,c_i)\times P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\times P(\bar{t}_k,c_i))}{\sqrt{P_d(t_k)\times P_d(c_i)\times P_d(\bar{t}_k)\times P_d(\bar{c}_i)}} \tag{8}$$

参数与公式(5)相同。*CC* 的平方就是 *Chi*。但是在 *CC* 中,特征与类型负相关会使 *CC* 值取负。*CC* 只在小特征集上对 *Chi* 有微弱的改善。进行特征选择时,选择 *CC* 值大的特征[13]。

2.8 特征权(Term Strength,*TS*)

TS 则基于完全不同的思想来度量特征的重要性[12]。它首先定义"相关文本",然后认为同时在多个"相关文本"中出现的特征是重要的特征,具有较大的强度。特征 t_k 的权值如式(9)所示。

$$TS(t_k)=P(t_k\in d_y\mid t_k\in d_x) \tag{9}$$

其中,d_x、d_y 为任意"相关文本"对。在特征权统计量中,没有类型信息,特征权完全基于文本之间的相关性来度量。它的理论依据是基于特征 t_k 在邻近相关文本中出现的概率来测试特征的效率。进行特征选择时,选择 *TS* 值大的特征。

2.9 期望交叉熵(Expected Cross Entropy,*ECE*)

期望交叉熵不考虑词条未出现的情况。特征 t_k 的权值如式(10)所示。

$$ECE(t_k)=P(t_k)\sum_i P(c_i\mid t_k)\log\frac{P(c_i\mid t_k)}{P(c_i)} \tag{10}$$

如果类别与词条 t_k 强相关,$P(c_i\mid t_k)$值就比较大,若 $P(c_i)$在这种情况下比较小,则说明该特征词条对分类的作用大。期望交叉熵反映了文本类别的概率分布,以及在出现某种特定特征词条情况下文本类别概率分布之间的距离。进行特征选择时,选择 *ECE* 值大的特征。

2.10 文本证据权(Weight of Evidence for Text,*WET*)

为了比较类别出现概率和给定特征词条情况下类出现的概率,文本证据权值如式(11)所示。

$$WET(t_k)=P(t_k)\sum P(t_k)\left|\log\frac{P(c_i\mid t_k)(1-P(c_i))}{P(c_i)(1-P(c_i\mid t_k))}\right| \tag{11}$$

如果类别与词条 t_k 强相关,同时相应类别出现的概率比较小,文本证据权被放大的程度会比期望交叉熵更大,说明特征词条 t_k 对分类的作用被放大。进行特征选择时,选择 *WET* 值大的特征。

2.11 几率比(Odds Ratio,*OR*)

在分类过程中,对所有类别不应同等对待,重要的是关心目标类的值,也就是我们迫切想正确分类的值。几率比值如式(12)所示。

$$OR(t_k)=\log\frac{P(t_k\mid pos)(1-P(t_k\mid neg))}{P(t_k\mid neg)(1-P(t_k\mid pos))} \tag{12}$$

其中,*pos* 表示正例文本集合的情况,*neg* 表示

反例文本集合的情况。正例出现条件下,特征词条 t_k 出现概率越大;反例出现情况下,特征词条 t_k 出现概率越小。几率比越大,特征词条 t_k 对正确分类作用越大。几率比特别适合二元分类器。进行特征选择时,选择 OR 值大的特征。文献[22]的实验结果表明,几率比对文本分类的效果很好。

以上方法各有利弊。文献[13]对 df、MI、IG、Chi 及 TS 五种特征选择方法进行了比较[13]。结果显示,df、IG 和 Chi 要优于 MI 和 TS,其中以式(7)效果最好。而且,df、IG 和 Chi 之间存在很大的相关性。借助于降维统计量,特征选择方法可以取得很大的降维度而不使分类效果下降。文献[22]对文本频率、信息增益、互信息、CHI、期望交叉熵、文本证据权、几率比等特征选择方法进行了比较,结果显示,对几率比方法进行扩展的多类分类效果最好。文献[23]对 MI、CHI、期望交叉熵、文本证据权等特征选择方法进行了比较,结果显示,互信息方法效果最优。从上述研究看,没有哪种方法对分类效果有绝对优势。这是因为文本分类本身涉及到训练数据集合本身的特点,同时不同的分类器对文本分类的效果也不尽相同。

3　特征抽取(Feature Extraction)

特征抽取也叫特征重参数化(Feature Reparameterization)[14]。由于自然语言中存在大量的多义词、同义词现象,特征集无法生成一个最优的特征空间对文本内容进行描述。特征抽取是将原始特征空间进行变换,重新生成一个维数更小、各维之间更独立的特征空间。常用的特征抽取方法可以分为三类:主成分分析、潜在语义标引、非负矩阵分解。

3.1　主成份分析(Principle Component Analysis, PCA)

主成分分析方法应用线性代数中的 KL(Karhunen-Loeve, KL)变换将原始特征空间映射到一个低维的正交空间[15]。设 $d_1,\cdots,d_n$ 为训练文本的 n 个 p 维特征向量,这时得到协方差矩阵:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i,j=1}^{n} d_i * d_j^T \tag{13}$$

它的前 $q(q<p)$ 个最大的特征值及其对应的特征向量分别为 $\lambda = \{\lambda_1,\cdots,\lambda_q\}$ 和 $e = \{e_1,\cdots,e_q\}$。e 为新特征空间的基向量,维数为 q。训练集 $d_1,\cdots,d_n$ 映射到新特征空间以后,得到 q 维特征向量集 $\xi_1,\cdots,\xi_n$,实现了特征降维。$\xi_k = e^T d_k$,$(k=1,\cdots,n)$。在 Li 等人[15]的实验中,PCA 法没有取得令人满意的效果。

3.2　潜在语义索引(Latent Semantic Indexing, LSI)

文本中存在的同义词和多义词现象,导致特征向量构造的空间存在"斜交"的特点。也就是说,特征向量的各个分量存在一定的相关性。潜在语义索引通过挖掘文本与特征之间潜在的高阶语义结构,将文本特征矩阵分解为一个低维的正交矩阵,实现特征空间的降维。文本和特征被转换到低阶语义空间上进行描述,它们之间的操作转化为语义操作[16,17,18]。原始特征文本集合矩阵如式(14)所示,维数为 k。

$$D = (w_{ij})_{s\times n} \tag{14}$$

其中,w_{ij} 为特征 t_i 在文本 d_j 中的权重,$1\leqslant i\leqslant s$,$1\leqslant j\leqslant n$,s 为特征个数,n 为训练文本数。通过奇异值分解(Single Value Decomposition, SVD),矩阵 D 被分解为三个矩阵的积:

$$D = T_0 \times U_0 \times V_0 \tag{15}$$

其中,T_0 为 $s\times k$ 的单位正交矩阵,称为左奇异向量矩阵;U_0 为 $k\times k$ 的降序正定对角矩阵,称为奇异值矩阵;V_0 为 $k\times n$ 的单位正交矩阵,称为右奇异向量矩阵,k 也称为矩阵 D 的秩。选择 $k'\ll k$,保留矩阵 U_0 中的前 k' 个最大值,保留 T_0 中的前 k' 行 k' 列,保留 V_0 中的前 k' 行,分别得到矩阵 U、T、和 V。它们的积是在 k' 阶平方误差内原矩阵的最相似矩阵,如式(16)所示。它在 k' 维正交特征空间上描述了原矩阵中潜在的、最重要的语义结构,文本、特征之间的关系可以在该空间上进行运算。

$$D' = T \times U \times V \tag{16}$$

通过控制语义空间的维数,LSI 可以得到较大的降维度[44]。LSI 将原信息进行重新组合,很少丢失原特征空间中的信息。但是,LSI 计算复杂度高,在大规模数据集上进行奇异值分解非常困难,而且降维后,分类效果下降[15,19]。

3.3　非负矩阵分解(Non-negative Matrix Factorization, NMF)

由于潜在语义索引中的奇异值分解方法存在对数据变化敏感、运算速度慢以及左、右奇异矩阵的存储要求高的缺点,限制了在大规模文本分类的特征

抽取中的应用。同时,奇异值分解缺乏语义解释的直观性。非负矩阵分解方法将一个非负的矩阵分解成左右两个非负矩阵的乘积[20,21]。原矩阵中的一列可以解释为对左矩阵中所有向量的加权和,其中权重系数为右矩阵中列向量的元素,以达到获取同义词之间的关联关系。原始特征文本集合矩阵仍如式(14)所示,维数为 k。寻找一个 $m \times r$ 非负矩阵 $U = (u_{ij})_{m \times r}$ 和一个 $r \times n$ 非负矩阵 $V = (v_{ij})_{r \times n}$,满足式(17):

$$D = U \times V \tag{17}$$

其中,$(m + n)r < nm$。每个文本向量 d_j,$d_j \in D$ 可以表示成 U 中的列向量和 V 中列向量中元素的线性组合。其中 U 中的所有列向量称为基向量。对于 U 中每一列向量 u_j $(1 \leqslant j \leqslant r)$,分量值大的对应语义相似的特征项。同时,$U$ 中的列向量通常满足正交和近似正交。文献[25]将 NMF 方法也称为潜在语义索引。当然 NMF 和潜在语义索引有一定的相似性,但使用熟知的已有称谓欠妥。

文献[24]比较了文本频度、基于分类频率的文本频度、$tf \times idf$、主成分分析四种降维方法,其结果以主成分分析方法最好。本文认为将特征选择和特征抽取方法混合进行比较欠妥当。用以上特征抽取方法对相同数据集进行特征降维效果比较是今后研究工作的一部分。

4 结束语

文本分类的关键是对高维的特征集进行降维,现在越来越被关注,并广泛运用在各个领域。现有的特征选择和特征抽取方法从运行结果看对特征空间都有不同程度的降低,对分类器的效率和使用效果都有一定的提高。但是对于不同的训练数据集,特征降维的效果也不尽相同。没有一种降维方法对所有或者是大部分训练数据集合都有比较好的效果。进一步的工作包括研究某种特征降维方法适合于哪一类数据集合,如何将不同的特征降维方法进行组合并改进以提高准确率,以及选择不同的降维方法和不同分类器结合,从而取得更好的分类效果。

参 考 文 献

1 David D Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92), 1992. 37 ~ 50

2 Fuhr N, and Buchley C. A probabilistic learning approach for document indexing. ACM Transactions on Information Systems, 1991, 9(3): 223 ~ 248

3 Dumais S T, Platt J, Heckerman D, et al. Inductive learning algorithms and representations for text categorization. Technical Report, Microsoft Research, 1998

4 Joachims T. A probabilitic analysis of the rocchio algorithm with TFIDF for text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML-97), 1997

5 Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1 ~ 47

6 David W Aha, and Richard L Bankert. A comparative evaluation of sequential feature selection algorithms. In: Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics, 1995: 1 ~ 7

7 Ron Kohavi, and George H John. Wrappers for feature subset selection. Artificial Intelligence Journal. Special Issue on Relevance, 1997: 273 ~ 324

8 Tao Liu, Shengping Liu, Zheng Chen, et al. An evaluation on feature selection for text clustering. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003. 488 ~ 495

9 Lei Yu, and Huan Liu. Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003. 856 ~ 863

10 L Douglas Baker, and Andrew Kachites McCallum. Distributional clustering of words for text classification. In: Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98), 1998. 96 ~ 103

11 Apte C, Damerau F J, and Weiss S M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 1994, 12: 233 ~ 251

12 Yang Yiming, and Pedersen J O. A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML-97), 1997. 412 ~ 420

13 Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In: Proceedings of the 20th ACM International Conference on Research and Development in Information Retrieval (SIGIR-97), 1997. 67 ~ 73

14 Schutze H, Hull D A, and Pedersen J O. A comparison of classifiers and document representations for the routing problem. In: Proceedings of the 18th ACM International

Conference on Research and Development in Information Retrieval (SIGIR-95). 1995. 229 ~ 237

15 Li Y H, and Jain A K. Classification of text document. The Computer Journal, 1998, 41(8): 537 ~ 546

16 Deerwester S, Dumais S, Furnas D, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391 ~ 407

17 Thomas Hofmann. Probabilistic latent semantic indexing. In: Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99), 1999. 50 ~ 57

18 Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. Discourse Processes, 1998, 25: 259 ~ 284

19 Douglas L Baker, and Andrew Kachites McCallum. Distributional clustering of words for text classification. In: Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98), 1998. 96 ~ 103

20 Lee D, Seung H. Learning the Parts of Objects By Non-negative Matrix Factorization. Nature, 1999, 401: 788 ~ 791

21 Lee D, Seung H. Algorithms for non-negative matrix factorization. Adv. Neural Info. Proc. Syst, 2001, 13: 556 ~ 562

22 周茜,赵明生等. 中文文本分类中的特征选择研究. 中文信息学报,2004,18(3): 17 ~ 23

23 秦进,陈笑蓉等. 文本分类中的特征抽取. 计算机应用,2003,23(2): 45 ~ 46

24 陈莉. 文本挖掘与降维技术. 西北大学学报,2003,33(3): 267 ~ 271

25 黄钢石,张亚非等. 基于 NMF 的潜在语义索引模型在文本检索中的应用. 解放军理工大学学报,2004,5(2): 36 ~ 39

(责任编辑　许增棋)

文本分类中的特征降维方法综述

作者： 陈涛， 谢阳群， Chen Tao， Xie Yangqun
作者单位： 宁波大学商学院信息管理系,宁波,315211
刊名： 情报学报 ISTIC PKU CSSCI
英文刊名： JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION
年，卷(期)： 2005,24(6)
被引用次数： 60次

参考文献(25条)

1. David D Lewis An evaluation of phrasal and clustered representations on a text categorization task 1992

2. Fuhr N;Buchley C A probabilistic learning approach for document indexing 1991(03)

3. Dumais S T;Platt J;Heckerman D Inductive learning algorithms and representations for text categorization 1998

4. Joachims T A probabilitic analysis of the rocchio algorithm with TFIDF for text categorization 1997

5. Fabrizio Sebastiani Machine learning in automated text categorization 2002(01)

6. David W Aha;Richard L Bankert A comparative evaluation of sequential feature selection algorithms 1995

7. Ron Kohavi;George H John Wrappers for feature subset selection 1997(Special Issue)

8. Tao Liu;Shengping Liu;Zheng Chen An evaluation on feature selection for text clustering 2003

9. Lei Yu;Huan Liu Feature selection for high-dimensional data:a fast correlation-based filter solution 2003

10. L Douglas Baker;Andrew Kachites McCallum Distributional clustering of words for text classification 1998

11. Apte C;Damerau F J;Weiss S M Automated learning of decision rules for text categorization[外文期刊] 1994

12. Yang Yiming;Pedersen J O A comparative study on feature selection in text categorization 1997

13. Hwee Tou Ng;Wei Boon Goh;Kok Leong Low Feature selection,perceptron learning,and a usability case study for text categorization 1997

14. Schutze H;Hull D A;Pedersen J O A comparison of classifiers and document representations for the routing problem 1995

15. Li Y H;Jain A K Classification of text document 1998(08)

16. Deerwester S;Dumais S;Furnas D Indexing by latent semantic analysis 1990(06)

17. Thomas Hofmann Probabilistic latent semantic indexing 1999

18. Thomas K Landauer;Peter W Foltz;Darrell Laham An introduction to latent semantic analysis 1998

19. Douglas L Baker;Andrew Kachites McCallum Distributional clustering of words for text classification 1998

20. Lee D;Seung H Learning the Parts of Objects By Non-negative Matrix Factorization 1999

21. Lee D;Seung H Algorithms for non-negative matrix factorization 2001

22. 周茜;赵明生 中文文本分类中的特征选择研究[期刊论文]-中文信息学报 2004(03)

23. 秦进;陈笑蓉 文本分类中的特征抽取[期刊论文]-计算机应用 2003(02)

24. 陈莉 文本挖掘与降维技术 2003(03)

25. 黄钢石;张亚非 基于NMF的潜在语义索引模型在文本检索中的应用[期刊论文]-解放军理工大学学报 2004(02)

本文读者也读过(2条)

1. 王博.贾焰.杨树强.韩伟红.WANG Bo.JIA Yan.YANG Shu-qiang.HAN Wei-hong 文本多分类中的特征选择研究[期刊论文]-计算机工程与科学2010,32(8)
2. 胡洁.HU Jie 高维数据特征降维研究综述[期刊论文]-计算机应用研究2008,25(9)

引证文献(60条)

1. 张鸿彦 基于CCIPCA-LSSVM的文本自动分类算法[期刊论文]-科学技术与工程 2013(10)
2. 胡涛.刘怀亮 中文文本分类中一种基于语义的特征降维方法[期刊论文]-现代情报 2011(11)
3. 毛嘉莉 文本聚类中的特征降维方法研究[期刊论文]-西华师范大学学报（自然科学版） 2009(4)
4. 柴忠.常晓明 一种基于CFN的特征选择及权重算法[期刊论文]-微计算机信息 2009(3)
5. 宣照国.党延忠 文本分类中粗分类数据噪声修正的网络算法[期刊论文]-情报学报 2008(5)
6. 张元虹.郭剑毅.龚华明.薛征山 基于DF与LSA相结合的降维法的文本分类系统的研究[期刊论文]-山西电子技术 2008(4)
7. 杨丽玲 基于概率的覆盖算法在文本分类器中的应用[期刊论文]-漳州职业技术学院学报 2009(2)
8. 孟春艳 用于文本分类和文本聚类的特征抽取方法的研究[期刊论文]-微计算机信息 2009(9)
9. 刘海峰.王元元.张学仁.刘守生 基于散度差准则的文本特征降维研究[期刊论文]-计算机应用研究 2008(7)
10. 陈国松.黄大荣 基于信息熵的TFIDF文本分类特征选择算法研究[期刊论文]-湖北民族学院学报(自然科学版) 2008(4)
11. 严莉莉.张燕平 基于类信息的文本聚类中特征选择算法[期刊论文]-计算机工程与应用 2007(12)
12. 王倩倩.段震.张燕平 基于交叉覆盖算法的文本分类[期刊论文]-计算机技术与发展 2007(6)
13. 秦锋.赵彦军.程泽凯.陈奇明 基于词条数学期望的词条权重计算方法[期刊论文]-计算机应用与软件 2011(4)
14. 胡改蝶.马建芬 文本分类中一种特征选择方法的改进[期刊论文]-计算机与现代化 2011(5)
15. 李凯齐.刁兴春.曹建军.李峰 基于改进蚁群算法的高精度文本特征选择方法[期刊论文]-解放军理工大学学报（自然科学版） 2010(6)
16. 李家兵 中文文本分类特征选择的研究[期刊论文]-皖西学院学报 2009(2)
17. 何海斌.李新福.赵蕾蕾 基于CCIPCA和ICA降维的文本分类研究[期刊论文]-计算机工程与应用 2008(29)
18. 甄志龙.韩立新.陆佃龙 基于模糊关系的文本分类特征选择方法[期刊论文]-情报学报 2008(6)
19. 贾花萍 基于神经网络的特征选择与提取方法研究[期刊论文]-办公自动化（综合版） 2008(7)
20. 唐歆瑜.乐文忠.李志成.李军义 基于知网语义相似度计算的特征降维方法研究[期刊论文]-科学技术与工程 2006(21)
21. 陈平.廖玉霞 基于小样本条件下线性判别分析图像增强算法研究[期刊论文]-科学技术与工程 2013(6)
22. 李凯齐.刁兴春.曹建军 基于信息增益的文本特征权重改进算法[期刊论文]-计算机工程 2011(1)
23. 赵延平.谢丽聪 面向电信领域的文本分类研究[期刊论文]-计算机与现代化 2011(2)
24. 涂伟.甘丽新.周雪梅 数据挖掘在网络教学平台中的研究与应用[期刊论文]-科技广场 2011(1)
25. 陈炎龙.张志明 基于向量空间模型的英文文本难度判定[期刊论文]-电脑知识与技术 2010(12)
26. 袁晓峰 一种基于主题的Web文本聚类算法[期刊论文]-成都大学学报（自然科学版） 2010(3)
27. 陈集.樊兴华.王鹏 中文文本分类的两步特征选择法[期刊论文]-计算机辅助工程 2008(3)
28. 刘海峰.王元元.张学仁.姚泽清 文本分类中基于位置和类别信息的一种特征降维方法[期刊论文]-计算机应用研究 2008(8)

29. 李家兵 基于交叉覆盖算法的文本分类研究[期刊论文]-滁州学院学报 2008(5)

30. 林永民. 朱卫东 基尼指数在文本特征选择中的应用研究[期刊论文]-计算机应用 2007(10)

31. 刘洋 中文文本分类中特征选择方法的比较研究[期刊论文]-科技信息（科学·教研） 2007(3)

32. 刘立月. 黄兆华. 刘遵雄 高维数据分类中的特征降维研究[期刊论文]-江西师范大学学报（自然科学版） 2012(2)

33. 张小艳. 宋丽平 论文本分类中特征选择方法[期刊论文]-现代情报 2009(3)

34. 刘海峰. 王元元. 张学仁. 刘守生 一种基于聚类和LSA相结合的文本特征降维方法[期刊论文]-情报杂志 2008(2)

35. 刘洋. 张秋余 基于LSI和SVM相结合的文本分类研究[期刊论文]-计算机工程与设计 2007(23)

36. 周炎涛. 唐剑波. 王家琴 基于信息熵的改进TFIDF特征选择算法[期刊论文]-计算机工程与应用 2007(35)

37. 张伟刚. 谭建豪 基于人工免疫系统的网络文本分类研究[期刊论文]-科学技术与工程 2006(22)

38. 况夯. 罗军 基于遗传FCM算法的文本聚类[期刊论文]-计算机应用 2009(2)

39. 熊忠阳. 付玲玲. 张玉芳 文本分类中基于概念映射的二次特征降维方法[期刊论文]-计算机工程与应用 2012(1)

40. 韩永峰. 郭志刚. 陈翰. 许旭阳 基于领域特征词的突发事件层次分类方法[期刊论文]-信息工程大学学报 2012(5)

41. 耿俊成. 牛霜霞. 张才俊 使用进化神经网络进行文本自动分类[期刊论文]-计算机与现代化 2011(11)

42. 肖可. 奉国和 1999～2008年国内文本分类研究文献计量分析[期刊论文]-情报学报 2010(4)

43. 李小波 肿瘤基因表达谱分类技术研究[期刊论文]-计算机时代 2008(6)

44. 任克强. 张国萍. 赵光甫 基于相对文档频的平衡信息增益降维方法[期刊论文]-江西理工大学学报 2008(5)

45. 刘海峰. 王元元. 王倩 基于特征选择的文本分类方法评述[期刊论文]-情报科学 2007(z1)

46. 林永民. 吕震宇. 赵爽. 朱卫东 文本特征加权方法TF·IDF的分析与改进[期刊论文]-计算机工程与设计 2008(11)

47. 褚力. 张世永 基于集成合并的文本特征提取方法[期刊论文]-计算机应用与软件 2008(10)

48. 张秋余. 刘洋 使用基于SVM的局部潜在语义索引进行文本分类[期刊论文]-计算机应用 2007(6)

49. 廖开际. 叶东海. 席运江 基于知识模式的企业文本知识自动分类研究[期刊论文]-情报杂志 2010(9)

50. 余俊英. 王明文. 盛俊 文本分类中的类别信息特征选择方法[期刊论文]-山东大学学报（理学版） 2006(3)

51. 范少萍. 郑春厚. 王召兵 基于元样本稀疏表示分类器的文本资源分类[期刊论文]-图书情报工作 2011(16)

52. 廖开际. 叶东海. 闫健峻. 吴敏 基于加权语义网的专家知识发现及表示方法[期刊论文]-情报学报 2012(1)

53. 谌志群 XML文档相似度计算方法研究[期刊论文]-情报学报 2009(1)

54. 杨彦闯 基于联合特征提取的粗糙集文本分类的研究[学位论文]硕士 2006

55. 王圆 文本内容过滤的关键技术研究[学位论文]硕士 2006

56. 张雪英 基于机器学习的文本自动分类研究进展[期刊论文]-情报学报 2006(6)

57. 柯慧燕 Web文本分类研究及应用[学位论文]硕士 2006

58. 高淑琴 Web文本分类技术研究现状述评[期刊论文]-图书情报知识 2008(3)

59. 蒲筱哥 Web自动文本分类技术研究综述[期刊论文]-情报学报 2009(2)

60. 席运江 组织知识的网络表示模型及分析方法[学位论文]博士 2006

引用本文格式：陈涛. 谢阳群. Chen Tao. Xie Yangqun 文本分类中的特征降维方法综述[期刊论文]-情报学报 2005(6)