Doi: 10.3969/j.issn.1003-5060.2011.08.014

基于命名实体的 Web 新闻文本分类方法

潘正高[1,2], 侯传宇[2], 谈成访[2]
(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009; 2. 宿州学院 智能信息处理重点实验室, 安徽 宿州 234000)

摘 要: 文章对 Web 新闻领域的文本自动分类问题进行了研究, 提出一种基于新闻实体要素的分类方法; 在应用空间向量模型的基础上, 充分考虑命名实体对 Web 新闻文本分类的特殊作用, 并进行了实验。实验结果表明, 以新闻实体要素为特征的文本分类系统可得到较高的分类精度, 该方法具有一定的实用价值。

关键词: 文本分类; 向量空间模型; 特征选择; 命名实体
中图分类号: TP391.12 文献标识码: A 文章编号: 1003-5060(2011)08-1178-05

Text categorization of Web news based on named entity

PAN Zheng-gao[1,2], HOU Chuan-yu[2], TAN Cheng-fang[2]
(1. School of Information and Computer, Hefei University of Technology, Hefei 230009, China; 2. Key Laboratory of Intelligent Information Processing, Suzhou University, Suzhou 234000, China)

Abstract: In this paper, the method of automatic text categorization of Web news is researched, and a categorization method based on the named entity of news is proposed. The special effect of the named entity on the text categorization of Web news is analyzed by applying the vector space model(VSM) and an experiment is carried out. The experimental result shows that the presented method has better precision of text categorization and can perform well.

Key words: text categorization; vector space model(VSM); feature selection; named entity

随着 Internet 的飞速发展, 网络信息剧增, 其中的绝大多数是以文本方式存在的, 如何在海量的网络信息中准确查找用户真正感兴趣的内容已经成为研究热点。文本分类能够处理大量的文本, 可以一定程度地解决信息紊乱的现状, 方便用户准确地定位所需要的信息, 成为处理和组织大量文档数据的关键技术。

文本分类研究首先要解决的问题是将非结构化的文本数据表示成结构化数据, 目前的文本分类系统基本上都是采用基于词语为特征项的向量空间模型(Vector Space Model, 简称 VSM)来表示文本[1]。这种方法的特点是文本向量的维数很高, 一个文本向量通常可以达到几千甚至上万维的量级。文本向量的高维数容易导致数据稀疏、数据噪音等问题, 造成文本分类效果较差。因此, 在文本分类前, 有必要对文本的原始特征进行降维处理, 特征降维的主要方法是特征选择和特征抽取。

特征选择是从文本的原始特征中选择一个特征子集来提高分类器的训练速度, 提高分类的精度。常用的特征选择方法有文档频率、信息增益、互信息、卡方统计、期望交叉熵等。特征抽取又称特征重参数化[2], 它在考虑到文本中存在大量的多义词、同义词等现象时, 将原始特征空间进行变换, 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包括: 主成分分析[3](Principle Component Analysis, 简称 PCA)、潜在语义标引[4](Latent Semantic Indexing, 简称

收稿日期: 2010-11-08; 修回日期: 2010-12-03
基金项目: 安徽省高校优秀青年人才基金资助项目(2010SQRL193); 宿州学院自然科学研究基金资助项目(2008yzk04); 宿州学院硕士科研启动基金资助项目(2009YSS12)和宿州学院科研开放平台课题(2011ykf10)
作者简介: 潘正高(1978—), 男, 安徽六安人, 合肥工业大学硕士生, 宿州学院讲师.

LSI)和非负矩阵分解(Non-negative Matix Factorization，简称 NMF)。特征选择相对于特征抽取而言，保留了原始特征空间的性质，选择后的特征仍有实际含义，故本文采用特征选择的方法来降低文本特征维度。

1　特征选择

文本数据的预处理是文本分类的关键，直接影响文本分类的效果。文本预处理是对文本内容进行处理，包括去除标点、分词等工作。分词后的文本数据用向量的形式来表示，常用的表示方法为向量空间模型。组成一篇文本的词语成千上万，如果这些词都作为特征，显然维数过大，会影响分类算法的性能。因此，需要对文本的原始特征进行降维处理。特征选择是从文本的原始特征集 $T=\{t_1, \cdots, t_s\}$ 中选择一个真子集 $T'=\{t_1, \cdots, t_{s'}\}$，满足 $s' \leqslant s$。特征选择是在不改变原始特征空间性质的前提下，选择一部分有利于提高文本分类准确率的重要特征，组成一个新的低维空间。目前，在文本分类中常用的特征选择方法[2]包括：文档频率、信息增益、互信息等。

文档频率(Document Frequency，简称 DF)[5]是指在语料中出现该词条的文档的数目，文档频率方法是基于一个广泛承认的规则标准：低的文档频次被认为和文本分类任务不相关。将低于某个阈值的低频词从原始特征空间中去除，不但能够降低特征空间的维数，而且还有可能提高分类的精度。DF 方法形式简单，缺点也很明显：低频词可能包含更多有用的信息，而高频词可能包含较少的信息。DF 方法一般不直接用于文本分类中，而是将其作为评价其他特征选择方法的参照。

信息增益(Information gain，简称 IG)[5-6]被广泛地应用在机器学习领域，它表示了某一个特征项的存在与否对类别预测的影响，定义为考虑某一特征项在文本中出现前后的信息熵之差。设 $\{c_i\}_i^m=1$ 为目标空间中类别的集合，那么词条 t 对类别的信息增益为：

$$G(t)=-\sum_{i=1}^{m}p(c_i)\lg p(c_i)+p(t)\sum_{i=1}^{m}p(c_i\mid t)\times \lg p(c_i\mid t)-p(\bar{t})\sum_{i=1}^{m}p(c_i\mid \bar{t})\lg(p(c_i\mid \bar{t}),$$

其中，$p(\bar{t})$为文本中不出现特征 t 的概率，可以用训练集中不包含 t 的文本占整个文本集的比例计算；$p(c_i\mid \bar{t})$ 为不包含特征 t 时第 i 个类别的条件概率，一般等于不包含特征 t 且属于类别 c_i 的文本数量除以所有不包含 t 的文本数量。信息增益在文本分类中经常用来大规模地移走“无用的”单词，它能在不降低文本分类性能的前提下移走高达 98%的单词[5]。信息增益的不足在于降低了负类对分类的贡献。

互信息(Mutual Information，简称 MI)[7]是信息论中一种衡量 2 个变量间相互关系的方法。对于类别 c_i 和词条 t，它们之间的互信息定义为：

$$I(t, c_i)=\lg\frac{p(t\wedge c_i)}{p(t)p(c_i)},$$

其中，$p(t\wedge c_i)$是 t 和 c_i 的联合概率，在文本分类中该概率可用属于类 c_i 并包含特征 t 的文本占整个文本集的比例来近似计算；$p(t)$近似等于包含特征 t 的文本占整个文本集的比例；$p(c_i)$为属于类 c_i 的文本占整个文本集的比例。当特征 t 依赖于类别 c_i 时，互信息较大；当特征 t 与类别 c_i 相互独立时，互信息等于 0；互信息还可能是负值，表示两者是负相关的。互信息的不足之处在于词条的得分受其边缘概率的影响大，对于有相等条件概率的一些词，稀有词比常用词的得分还要高，因此对于频率相差很大的词，得分是不具备可比性的，这样导致互信息评估函数不选择高频的有用词，而有可能选择稀有词作为文本的最佳特征。

在文本分类中，其他常用的特征选择算法还包括 Chi-Square 统计量(简称 CHI)[5]、期望交叉熵[8]、文本证据权(Weight of Evidence for Text，简称 WET)[9]、几率比(Odds Ratio，简称 OR)[10]等。这些算法的理论基础不同，基于大量真实数据的实验证明，各个算法各有利弊，不存在任何一种算法在所有的数据集上都是最优的[11]。由此可知，特征选择算法的设计需要考虑数据集本身的特性和分类器的工作原理，不存在某一种算法适合于所有的应用。

网络新闻文档篇幅短小，语言简练，特别是人物、时间、地点等新闻实体要素对新闻文本内容有着很强的限定作用。相对于新闻报道的内容而言，新闻实体要素的规模要小得多。本文提出了一种基于新闻实体要素特征选择方法，该方法是将识别出的新闻实体要素作为特征来表示新闻文本。

2　构建新闻实体向量空间模型

2.1　命名实体识别

命名实体(Named Entity，简称 NE)[12-13]是文本中的固有名称、缩写及其他唯一标识，通常包

括 7 种类别:人物、机构、地点、日期、时间、金钱以及百分比。在具体的一篇文章中,实体是基本的信息元素,往往指示了文章的主要内容。实体识别是判断一个文本串是否命名实体,并确定其类型的过程。现有的实体识别技术包括基于规则的、基于统计的、基于规则和统计相结合的 3 类方法。

基于规则的方法是从各种命名实体的构成规则、实体与上下文的关系等方面来进行实体的识别。该方法领域性较强,需要语言专家的参与,主观意味比较重,缺乏鲁棒性和可移植性。基于统计的方法是利用标注的语料进行训练,寻找实体识别模型。该方法虽然具有一定的客观性,但是人类语言的使用不是一个单纯的随机过程,严重的数据稀疏和系统处理能力的限制也使得统计模型适用的范围很有限,使用统计的方法搜索空间往往非常大而导致过大的开销。基于规则和统计相结合的方法充分发挥前 2 种方法的优势,具有较好的可训练性和可适应性,而且保持性能所花费的代价要比基于规则的系统低得多,是目前普通采用的实体识别方法。

命名实体识别系统性能的衡量标准指标主要有 2 个:查全率和查准率。查全率是系统正确识别的结果占所有正确结果的比例;查准率是系统正确识别的结果占所有识别结果的比例。其计算公式如下:

查全率 = count(正确)/(count(正确)+count(丢失)),

查准率 = count(正确)/(count(正确)+count(虚假))。

为了综合评价系统的性能,通常还计算查全率和查准率的加权几何平均值即 F 指数,这里查全率和查准率同等看待,权值为 1,则 F 的计算公式为:F = (2× 查准率× 查全率)/(查准率 + 查全率)。

2.2 新闻实体要素的提取

新闻学的普遍规律表明[14],人物、时间、地点等新闻实体要素都会出现在报道的标题或首段中,并在报道后续部分展开介绍。从语义学的角度而言,发生在特定时间、特定地点,涉及指定人物或机构的新闻具有确定性。

新闻文本中日期、时间实体的表示方式通常比较规范,因此本文采用基于规则的方法来识别新闻文本中的日期、时间实体,类似于利用正则表达式进行字符串的匹配。下面给出一些常见的日期实体识别规则:

ind::=[1~9|一|二|三|四|五|六|七|八|九|十,

sepa::=年|月|日|/|-|.|,

emon::=jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec,

regular1::=(ind){1~4}(sepa)(ind){1~4}(sepa)(ind){1~4},

regular2::=(ind){2~4}(emon)(sepa)(ind){1~2}(emon)(sepa)(ind){1~2}(ind){2~4},

DataPattern::=regular1|regular2,

其中,{}表示集合中多个项的“或”关系。针对 Web 新闻文本中可能出现的特殊日期格式,如:“…天前”、“…天后”等日期表示形式,可以手工添加规则进行识别。

人物、机构、地点等实体识别是基于中科院计算所的汉语词法分析系统 ICTCLAS (Institute of Computing Technology Chinese Lexical Analysis System)完成的,HMM 模型训练均以北京大学计算语言所加工的 1998 年《人民日报》语料库作为训练语料。

本文通过设置权值来描述实体要素对新闻报道内容的表达及区分能力,根据新闻学及语义学的领域知识,在同类型的命名实体中,出现频率高、出现位置靠前的实体对报道内容的表示和区分能力强,其权值也应越高。若 d 是报道 s 中的一个日期实体,其权值计算方法为[14]:

$$w(d,s)=\lg\left[\left(1+\frac{N_d}{N}\right)\frac{1+(N-L_d)}{N(N-1)/2}\right] \quad (1)$$

其中,N_d 为 d 在报道 s 中出现的次数;N 为报道 s 中包含的日期实体总数;N_d/N 为 d 在报道 s 中出现的频率,将 s 中的日期实体依照出现的先后次序排列;L_d 为 d 在该序列中首次出现的位置。

对报道中其他实体的权重计算方法与上述方法相同,对新闻报道中识别出的日期、时间、地点等命名实体按权值大小进行排序,选择权值大于阈值的实体作为文本特征。基于新闻实体要素(Based on News Named Entities,简称BNNE)的向量空间模型构建算法如下:

(1) 网页预处理。对采集来的网页进行“清洗”,去除 HTML 标记,抽取新闻正文,形成“干净”的新闻文本。

(2) 分词、去除停用词。对步骤(1)生成的纯文本进行分词、去除停用词以及一些新闻常用词,如“新闻”、“报道”等。

(3) 实体识别。对分词后的文本利用统计和规则相结合的方法进行实体识别,并利用(1)式计算相应的权值。

(4) 特征选择。设置特征向量的阈值,保留

大于阈值的向量。

3　文本分类器构造

常用的分类算法包括支持向量机（Support Vector Machines，简称 SVM）、贝叶斯算法（Naive Bayes，简称 NB）、KNN 等。

SVM 由文献[15]提出，是用于解决二类划分问题的模式识别模型。它基于结构风险最小化原理，其基本思想是构造一个超平面作为决策平面，使正负模式之间的空白最大。支持向量机在解决小样本、非线性及高维模式识别问题中表现出了许多特有的优势，并在很大领域得到了成功的应用，如人脸识别、手写字体识别、文本分类等。

贝叶斯分类[16−17]方法是一种简单而又非常有效的分类方法。NB 法的一个前提假设是：在给定的文档类语境下，文档属性是相互独立的。假设 d_i 为一任意文档，它属于文档类 $C=\{C_1, C_2, \cdots, C_k\}$ 中的某一类 C_j。根据 NB 分类法有：

$$P(C_j \mid d_i)=\frac{P(C_j)P(d_i \mid C_j)}{P(d_i)} \quad (2)$$

$$P(d_i)=\sum_{j=1}^{k}P(C_j)P(d_i \mid C_j) \quad (3)$$

对文档 d_i 进行分类，就是(2)式计算所有文档类在给定 d_i 情况下的概率，概率值最大的那个类就是 d_i 所在的类，即

$$d_i \in C_j \quad \text{if} \quad P(C_j \mid d)=\max_{l=1}^{k}\{P(C_l \mid d_i)\}。$$

由(2)式、(3)式可知，对于给定分类背景和测试文档，用 NB 法分类的关键就是计算 $P(C_j)$ 和 $P(d_i \mid C_j)$，其计算过程就是建立分类模型的过程。

KNN（k-Nearest Neighbors，简称 KNN）是一种基于实例的文本分类算法，因简单、有效，在文本分类中被广泛使用。KNN 的基本思路是：在给定某测试文本后，考虑在训练文本集中与该文本距离最近（最相似）的 k 篇文本，根据这 k 篇文本所属的类别判定该文本所属的类别，KNN 算法中文本相似度通过欧几里德距离或向量间夹角来度量。

本文采用 KNN 算法来构造基于新闻实体向量空间模型的文本分类器，具体算法如下：

（1）利用新闻实体向量模型的特征项来描述训练文本向量。

（2）在测试文本到达后，根据特征词分词测试文本，确定测试文本的向量表示。

（3）在训练文本集中选出与测试文本最相似的 k 个文本，计算公式为：

$$\mathrm{Sim}(d_i, d_j)=\frac{\sum_{k=1}^{M}W_{ik}W_{jk}}{\sqrt{\left(\sum_{k=1}^{M}W_{ik}^2\right)\left(\sum_{k=1}^{M}W_{jk}^2\right)}} \quad (4)$$

（4）在测试文本的 k 个邻居中，依次计算每一个类的权重，计算公式为：

$$p(x, C_j)=\sum_{d_i \in \mathrm{KNN}}\mathrm{Sim}(\boldsymbol{x}, d_i)y(d_i, C_j) \quad (5)$$

其中，$\boldsymbol{x}$ 为测试文本的特征向量；$\mathrm{Sim}(\boldsymbol{x}, d_i)$ 为相似度计算公式，与（4）式的计算公式相同；而 $y(d_i, C_j)$ 为类别属性函数，即如果 d_i 属于类 C_j，那么函数值为 1，否则为 0。

（5）比较各个类的权重，将文本分到权重最大的类别中。

基于新闻实体向量空间模型的新闻文本分类系统工作流程，如图 1 所示。

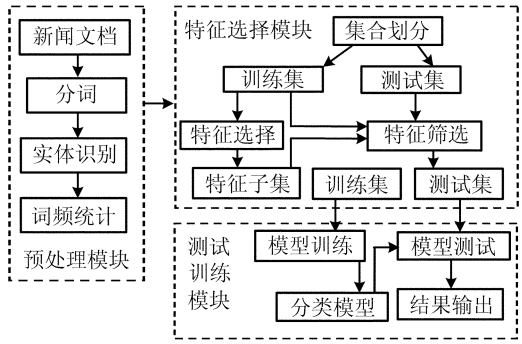

图 1　新闻文本分类系统工作流程

4　实验结果及分析

4.1　实验数据及性能评价

本文的实验数据取自搜狐、网易、新浪、新华网等国内著名新闻网站共 5 842 个网页，其中，“墨西哥湾漏油事件”主题网络新闻报道 2 753 个网页，其他主题的网络新闻报道 3 089 个网页。实验时，从中随机选取 3 895 个网页作为训练集，1 947 个网页作为测试集。

本文采取宏平均 F1（Macro-F1）值[18]来评价新闻网页分类的性能。

4.2　实验结果与分析

在 CPU 奔腾 1.73 GHz，内存 1 GB，VC++ 6.0 环境下，采用向量空间模型表示文本，特征选择方法分别采用文档频率和基于新闻实体的特征选择方法，分类算法采用 KNN 方法，其中 k 近邻值取 30，选取的特征维数为 600～2 000。文档频率（DF）和新闻实体要素（NNE）2 种特征选择方

法下,文本分类系统随着特征维数变化的宏 F1 值比较图,如图 2 所示。

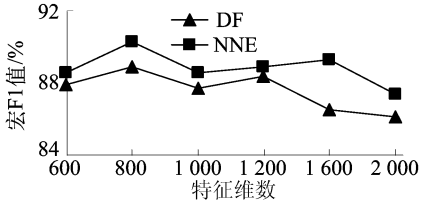

图 2 DF 和 NNE 宏 F1 值比较图

由图 2 可以看出:

(1) 当特征维数为 800 时,系统分类效果最好,随着特征维数的升高和下降,文本分类的准确率有所下降。这说明,通过特征选择在特征维数 800 时,能够比较好地反映训练文本信息。

(2) 分类系统在 NNE 方法下的分类效果明显高于 DF。其原因在于 DF 方法文本中的低频词去除,这样会将一些出现频率低而信息量大的命名实体丢弃,造成文本内容的大幅度消减,进而影响分类效果。NNE 方法在选取较少特征的情况下,也能很好地反映文本的内容信息,故而能够达到较好的分类效果。

5 结束语

文本特征的高维度极大地影响着文本分类系统的性能,特征降维处理是文本分类必要的工作。特征选择能在保留特征基本性质不变的情况下,达到降低特征维度的目的。文中在分析 Web 新闻文本内容的基础上,提出基于新闻实体要素为特征来构建 Web 新闻文本表示模型。实验结果表明,这种方法能在保留较少特征的情况下达到较好的分类效果。

[参 考 文 献]

[1] Salton G, McGill M G. Introduction to modern information retrieval and data mining[M]. California: AAAI Press, 1996: 335—338.

[2] Schutze H, Hull D A, Pedersen J O. A comparison of classifiers and document representations for the routing problem[C]// Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1995: 229—237.

[3] Yang Y. Noise reduction in a statistical approach to text categorization [C]// Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1995: 256—263.

[4] Deerwester S. Indexing by latent semantic analysis[J]. Journal of American Society for Information Science, 1990, 41(6): 391—407.

[5] Yang Yiming, Pederson J O. A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97). Nashville: Morgan Kaufmann, 1997: 412—420.

[6] Quinlan J R. Constructing decision tree, C4 5[J]. Programs for Machine Learning, 1993, 3: 17—26.

[7] Cover T M, Thomas J A. Elements of information theory [M]. New York: John Wiley and Sons, 1991: 274.

[8] Koller D, Sahami M. Hierarchically classifying documents using very few words[C]// Proceedings of the 14th International Conference on Machine Learning (ICML). San Francisco, USA: Morgan Kaufmann, 1997: 170—178.

[9] Mladenic D, Grobelnik M. Feature selection on hierarchy of web documents[J]. Decision Support System, 2003, 35: 45—87.

[10] Rijsbergen C J V. The selection of good search terms[J]. Information Processing and Management, 1981, 17: 77—91.

[11] Jain A, Zongker D. Feature selection: evaluation, application and small sample performance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(2): 153—158.

[12] Grishman R, Sundheim B. Message understanding conference-6: a brief history[C]// Proceedings of the 16th International Conference on Computational Linguistics, COLING-96, 1996: 24—29.

[13] 愈鸿魁, 张华平, 刘 群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87—94.

[14] 付 艳, 杨冬青, 唐世渭, 等. 基于实体识别的在线主题检测方法[J]. 北京大学学报: 自然科学版, 2009, 45(2): 227—232.

[15] Vapnic V. The nature of statistical learning theory[M]. New York: Springer, 1995: 127—129.

[16] Lewis D D. Naive (Bayes) at forty: the independence assumption in information retrieval[C]// The 10th European Conf on Machine Learning (ECM98). New York: Springer-Verlag, 1998: 4—15.

[17] Heckerman D. Bayesian networks for knowledge discovery [M]. Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al. Advances in Knowledge Discovery and Data Mining. Cambridge, MA: MIT Press, 1996: 273—305.

[18] 张筱丹, 胡学钢. 基于向量空间模型的自动摘要冗余处理研究[J]. 合肥工业大学学报: 自然科学版, 2010, 33(9): 1355—1358.

(责任编辑 张秋娟)