基于优势率的改进二元特征提取方法①

杜一平　刘燕君 (中国科学技术大学 计算机科学与技术学院 安徽 合肥 230027)

摘　要： 主题网络爬虫研究中一个重要问题是文本特征的提取，其好坏会直接影响主题特征的提取及网页的相关性计算。在研究了文本分类特征提取方法的基础上，分析优势率特征提取方法的优缺点，把频度、分散度作为判断要素加以考虑，提出一种改进的二元分类特征选择方法 EOR，并使用得到的 EOR 值结合词频 TF 即 TF-EOR 来计算文档特征词的权重，应用于主题网络爬虫。仿真实验证明，EOR 在中低维数下能提升文档分类准确率达 5%，而 TF-EOR 权重计算方法好于 TF-IDF 方法，实验中提高了网络爬虫的抓取准确率和查全率达 4%。

关键词： 特征提取；优势率；主题网络爬虫；频度；分散度

An Enhanced Odds Ratio Dualistic Feature Extraction Method

DU Yi-Ping, LIU Yan-Jun

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: An important issue in topical crawler research is feature extraction, which makes great impact on topic description and page relevance scoring. The existing Odds Ratio method shows high performance on high dimension vectors, whereas it does not work well on low dimension condition. An enhanced method EOR based on Odds Ratio method, with word frequency and distribution rate taken into account, is proposed. The simulation shows a 5% increase on text categorization precision on low and middle feature dimension. Furthermore, by combining EOR score and TF value, namely, TF-EOR to calculate word weight and applying it to topical crawler, 4% increases on both precision and recall are obtained.

Keywords: feature extraction; odds ratio; topical crawler; word frequency; distribution rate

面对网络的信息爆炸和个性化搜索的需求，如何在海量的网页中找到与用户主题相关的内容成为一个挑战。主题网络爬虫通过启发式搜索策略来指导抓取主题相关的网页，取得良好的效果。启发式搜索策略中的一个重要问题是如何对抓取网页进行主题相关性计算，从而预测待抓取 URL 的抓取价值。而文本特征提取方法的好坏对文本分类、主题表达及网页相关度计算至关重要。因此，文本特征提取也成为搜索引擎设计中的一个重要问题。

1　引言

目前最常用的文本特征表示模型是 G. Salton 提出的向量空间模型(Vector Space Model, VSM)[1]。它把文本或者文档表示成为一个特征向量，向量元素为文本中出现的单字，词语或者短语(在中文搜索中，我们常用词语作为向量元素)，而元素的排列顺序是无关紧要的，因此我们可以把文本看成一系列的无序的词条的集合，而文本空间则可以看做是一组正交词条向量组成的向量空间。文本向量的维数通常是非常大而且包含噪声，占据大量系统资源，因此降低向量维数非常必要。通常可以使用特征提取方法选取高价值词条过滤掉低价值词条来减小特征空间的维数。因此，特征提取方法可以看成从测量空间到特征空间的一种映射或变换[2]。

特征提取的目的是剔除测量空间中对类别区分贡献小的词条，保存贡献大的词条，在降低特征空间的维数的同时获得最能表现类别特征的词条。一般通过

① 收稿时间:2009-05-18

设计一个特征评估函数对测量空间中的特征词条进行类别相关度评估，然后根据评估结果选择高于某个阈值的 N 个词条作为表示文档的特征词条。目前特征评估函数主要有以下几种：文档频率(Document Frequency，DF)、信息增益(Information Gain，IG)、互信息(Mutual Information，MI)，词条的统计(CHI)、期望交叉熵(Expected Cross Entropy)及优势率(Odds Ratio)等。文献指出，IG 和 CHI 的效果相对较好，MI 的效果则相对较差，优势率方法则在中低维数时表现差，在高维数是性能迅速改善[3-5]。

本文通过改进优势率算法，加入词频及分散度因素，提出了一种综合的二元分类特征提取方法 EOR，并把获得的 EOR 特征权值结合词频 TF 得到 TF-EOR 来计算特征词条的权重，应用于主题网络爬虫中。

2 优势率特征选择方法

优势率(Odds Ratio, OR)特征选择方法又称几率比特征选择方法。其特点是只把训练文本分成两类，一类是正面类，我们又称为主题类或目标类，即与给定主题内容一致的文档，而其他的文档都归为负面类，因此可以认为优势率方法是一种二元特征选择方法。它通过词条对正面样本的贡献与对负面样本的贡献进行对比，得到特征评估函数如下：

$$OR(t)=\log\left(\frac{odds(t\mid pos)}{odds(t\mid neg)}\right)=\log\left(\frac{P(t\mid pos)(1-P(t\mid neg))}{P(t\mid neg)(1-P(t\mid pos))}\right) \quad (1)$$

其中 $OR(t)$ 为词条 t 的优势率分值，$P(t\mid pos)$ 表示正面样本中出现词条 t 的概率，而 $P(t\mid neg)$ 表示负面样本中出现词条 t 的概率。从公式(1)中可以看出，优势率方法注重于对目标类的评估，词条 t 在正面样本中概率越高，在负面样本中出现概率越低，则该词条 t 对目标类的分类贡献度越大。

设 A 为词条 t 和正面样本同时出现的文档数目，B 为词条 t 和负面样本同时出现的文档数目，C 为所有的正面样本的文档数目，D 为所有负面样本的文档数目。则我们可以把公式改造如下：

$$OR(t)=\log\left(\frac{A(D-B)}{B(C-A)}\right) \quad (2)$$

由于优势率方法专注于目标类，而把其他不相关的文档都标记为负面类，这种二元性质使得它特别适合用于主题网络蜘蛛的主题特征提取及文档描述。从公式(2)中我们可以看出，优势率方法给那些只出现在目标类中而几乎不出现在负面类中的词条打高分。而且由于没有考虑词条在文档中的出现的次数，即词频，而只考虑了词条的文档频率，导致了该评估函数倾向于有选择低频或者中等频率的词条作为文本的最佳特征，因为高频词条通常会出现分布在多个类中[6]。而二元分类问题中，负面类占得比例较大，因此往往高频词条在负面类中出现的次数较多，这类词条也有一定的价值，但是在优势率方法中往往被忽视，特别是在中低维数的情况下。

3 新算法EOR的提出

3.1 影响特征权值的因素

影响特征权值的因素有很多，在此我们把它分为两类：语义因素和统计因素。语义因素包括词性，标题，位置，句法结构，词语长度，词语间关联等等，本文中主要考虑统计因素，因此不再赘述；统计因素包括词频(Term Frequency，TF)、集中度、分散度等。

① 词频，我们把词频分成两种：类内词频与整体词频。前者是指词条在某一类文档中出现的频率，显然，该词频越高，说明该词条越能作为特征项代表该类文档。而后者表征词条在所有文档中的出现频率，在此不具有参考意义。

② 集中度，所谓集中度，是指在多类文档中，一个词条只在一个或者几个类别的文档中出现，而不在其他类别的文档中出现，则认为该词条集中于一个或者几个类别。显然，集中度越高，该词条对所在文档类型的表达能力就越强。

③ 分散度，是指词条在其出现的文档类中的分散的均匀程度，一个词条在所在类中分布越均匀，说明该词条对该类来说越具有代表性。

显然对于某个词条，其类内词频越高，集中度越好，分散度越高，则对于区分所处类别文档来说贡献价值越大。因此我们可以考虑通过加入词频、集中度和分散度的因素来提升特征提取方法的性能。

3.2 新算法 EOR 的提出

如前所述，优势率方法倾向于选择在低频或中频词条，而往往忽视高频词条。这种特性使得它在特征

维数较低的情况下选择大量的低频词条，因而造成分类效果欠佳，而直到维数增大的一定程度才显示出优异的性能[6]。为了去除这种对低频词条的偏向性，应该考虑增加高频词条的分值。我们通过引入类内词频因素，可以降低类内词频低的词条的打分，相对提升高频词条的价值，从而改变高频词条被忽略的状况；而分散度的引入可以进一步加强词条的可信度。

我们定义类内词频为 $tf(t,pos)$，表示词条 t 在正面类 pos 中出现的频率，其计算方法如下：

$$tf(t,pos)=\frac{1+\sum_{j=1}^{N_{pos}} tf(t,d_j)}{V+\sum_{k=1}^{V}\sum_{j=1}^{N_{pos}} tf(t,d_j)} \quad (3)$$

其中 V 为正面类 pos 中出现的词条总数，N_i 为属于类 pos 的文档总数，$tf(t,d_j)$ 为词条在文档 d_j 中出现的次数。在实际应用中，由于文档的长度越长，文档中的词条的词频就越大，因此要计算词频 $f(t,d_j)$ 时要针对文档长度做归一化处理。

定义词条分散度为 $sp(t,pos)$，计算方法如下：

$$sp(t,pos)=\frac{df(t,pos)}{N_{pos}} \quad (4)$$

其中 $df(t,pos)$ 为词条 t 在目标类 pos 中的文档频率，N_{pos} 为属于目标类的文档总数。最后结合词频因素和分散度因素，我们改进公式(1)得到 EOR(Enhanced Odds Ratio)的计算公式：

$$EOR(t,pos)=OR(t)\times tf(t,pos)\times sp(t,pos) \quad (5)$$

引入类内词频因素，EOR 算法要解决的是在特征维数较低时算法也能选择到一些高频词条作为特征词条。而引入分散度因素则可以提升选取特征的质量及代表性。

3.3 文档特征权重的计算

在搜索引擎应用中，网页往往用向量空间模型 VSM 来表示。给定一个文档 d，我们可以把它表示成为 $d=\{(t_1,w_1),(t_2,w_2),\ldots,(t_i,w_i),\ldots,(t_n,w_n)\}$，其中 t_i 为特征词条，w_i 为词条的权重。通常我们使用经典的 TF-IDF 算法来计算词条的权重。该方法是建立在这么一个假设上的：一个词条它在文档中出现的次数越多，则该词条越能代表该文档，而一个词条出现的文档频率越高，则该词条对区分文档的贡献越小。但是，在主题搜索引擎中，仅仅考虑词条在文档中的出现次数，显然还不能完全表达该词条对于区分文档类别的重要性。在应用于主题搜索引擎时，特征词权重计算必须考虑该词与给定主题的相关性。考虑到特征选择函数计算所得的特征权值很大程度上表征的是词条对类型区分的贡献度，特别是在二元分类问题中。因此可以考虑结合特征提取过程中计算所得的特征权值 EOR 和 TF 值来计算文本中特征词条的权重，我们称为 TF-EOR。

在本文中我们使用以下公式来计算特征值的权重：

$$w_i=tf(t_i,d)\times EOR(t_i) \quad (6)$$

其中 $tf(t_i,d)$ 为词条 t_i 在文档 d 中出现的频率，应用中需除以文档的长度，$EOR(t_i)$ 为词条 t_i 的优势率特征权值。同样，我们用 VSM 来表示一个给定的主题，只是主题中特征词的权重我们设为 1，表示其在主题中出现而已。

4 仿真实验及分析

4.1 实验设计

实验分为两部分，其一为对改进特征选择算法 EOR 的测试，其二为对 TF-EOR 用于特征权重计算对网络爬虫抓取效果影响的测试。实验环境：一台联想 PC 机，操作系统为 Ubuntu8.04，CPU 为 T7300，2.0G，内存大小 512M。

4.2 文本分类器选择

目前的流行的文本分类器有很多，如 KNN，朴素贝叶斯，SVM 支持向量机，神经网络等。文献[7]研究表明，KNN 与 SVM 的分类效果明显好于其他分类器。本文中对 EOR 测试的文本分类器以及网络爬虫使用的文本分类器均为 KNN，因为它容易实现，分类效果较好。

4.3 中文处理

中文处理的主要目的是对文档进行分词以获取候选特征词条以及剔除影响分类效果的停用词。试验中使用的中文分词系统是中科院的 ICTCLAS2009 共享版，辅以自行编辑的停用词集合和专用词集合，用于过滤常见无区分度词条以及识别低频专有词汇。

4.4 评价标准

目前文档分类效果及网络爬虫的爬取效果的主流评价标准有两个：准确率 P(Precision)和查全率 R(Recall)。对于文本分类，准确率 P 的定义：

$$P=\frac{\text{被正确分类的文档数}}{\text{被分类为正确的文档数}} \tag{7}$$

查全率 R 的定义：

$$R=\frac{\text{被正确分类的文档数}}{\text{所有属于该类的文档数}} \tag{8}$$

对于网络爬虫的抓取效率，准确率 P 的定义：

$$P=\frac{\text{抓取到的主题相关页面数}}{\text{所有抓取到的网页}} \tag{9}$$

查全率 R 的定义：

$$R=\frac{\text{抓取到的主题相关页面数}}{\text{测试集中所有主题相关页面数}} \tag{10}$$

4.5 实验一

实验对文中提出的 EOR 算法在训练集和测试集上测试，与改进前的 OR 算法进行对比。采用的训练文档是用人工手段从新浪，搜狐，QQ，MSN 等门户网站上下载的包括军事，体育，健康，娱乐四个类别的共 2400 篇网页，每个网页均剔除掉无效成分，只留下中文文本信息。每个类别各包含 600 篇文档，其中 400 篇作为训练文档，剩余 800 篇作为测试文档。我们分别以四个门类文档做目标类，其余文档做负面类，在不同的特征维数的情况下进行特征提取和文本分类，取四个 P 值的平均值。

实验结果如表 1 所示：

表 1 改进优势率准确率随维数变化表

维数	1000	2000	4000	6000	8000
OR(%)	65.34	69.23	75.35	84.53	85.67
EOR(%)	70.81	75.14	79.77	85.65	86.32

实验结果表明，改进后的 EOR 方法在文本分类的准确率上比原有 OR 方法有了全面提高，特别是在维数较低(低于或等于 4000)时，改善幅度比较大，达到 5%左右，而当维数比较大时，则不明显，其原因是原 OR 算法在维数高时受低频词的影响减小，高频词也被得到选择的机会。因此在牺牲一定准确率的情况下，EOR 方法能够明显降低特征维数。

4.6 实验二

实验二是比较 TF-EOR(3.3 节中公式 6 所述)作为文档词条的权重计算方法和只用 TF-IDF 作为权重计算方法对网络爬虫抓取效率的影响。实验分别使用实验一中的军事类和体育类训练文档进行特征提取，获得主题的特征词条向量 T。任意被抓取网页 D，都用训练所得的特征词条向量来表示，而网页的主题相关度通过余弦夹角来计算，公式如下：

$$sim(T,D)=\frac{\sum_{k=1}^{N} w_{tk}w_{dk}}{\sqrt{\sum_{k=1}^{N} w_{tk}^{2}\sum_{k=1}^{N} w_{dk}^{2}}} \tag{11}$$

实验中的网络爬虫是改自开源网络爬虫 larbin 的主题网络爬虫 Tlarbin。我们使用从真实网络上用广度优先算法下载的镜像网站来仿真真实的网络环境，网页数量为 20 万，以站点为单位存储在硬盘上。网络爬虫分别以军事，体育为主题向量及对应 URL 种子集为输入，取父网页相关度和锚文本作为 URL 相关度预测因素，以 Best First 为启发策略来对抓取路线进行剪枝。

实验分别使用 TF-IDF 和 TF-EOR 作为权重计算方法对测试网页进行抓取，并计算两个类别的平均 P 值和 R 值。实验结果如表 2 所示：

表 2 不同权重计算方法下 P 值和 R 值表

P&R值	P值	R值
TF-IDF	73.5%	66.7%
TF-EOR	77.3%	70.9%

表 2 显示，使用 TF-EOR 作为权重计算方法能够在同等条件下提升网络爬虫的准确率和查全率各 4%左右。

5 结语

本文针对优势率方法(OR)在中低特征维数时偏向中低频词条而造成的分类效果偏低的情况，通过加入类内词频因素以及分散度因素，得到改进的 EOR，实验证明，新方法能够在低维数情况下提高准确率达 5%，而在高维数情况下则提升不明显。通过把得到的 EOR 值和 TF 值融合作为特征词权重计算因子，得到了 TF-EOR 方法，用于主题网络爬虫文档特征权重计算。实验证明，该方法比传统的 TF-IDF 方法能分别提高抓取准确率和查全率 4%左右。

参考文献

1 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975,18(11):613－620.

2 熊中阳,张鹏招,张玉芳.基于的文本分类特征选择方

(下转第 105 页)

(上接第 109 页)

法的研究.计算机应用, 2008,28(2):513－514.

3 Yang YM, Pederson JO. A comparative study on feature selection in text categorization. Proc. of the Fourteenth International Conference on Machine Learning, 1997.412－420.

4 Mladenic D, Grobdnik M. Feature selection for unbalanced class distribution and Naive bayes. Proc. of the 16th Intl Conf on Machine Learning(ICML' 99). San Francisco: Morgan Kaufmann Publishers, 1999.258－267.

5 Schütze H, Hull DA, Pedersen JO. A comparison of classifiers and document representations for the routing problem. Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information retrieval. Seattle, Washington, United States, 1995.229－237.

6 Jiang MH, Wang L, Lu YH, Liao SS. A RBF network for Chinese text classification based on concept feature extraction. 13th International Conference on Neural Information Processing (ICONIP 2006). Hong Kong, China, 2006.3.

7 Yang YM, Liu X. A re-examination of text categorization methods. Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, California, United States.1999.42－49.