基于大规模真实文本的平衡语料分析与文本分类方法[1]

陈克利* 宗成庆* 王霞+

*中科院自动化所模式识别国家重点实验室 北京 100080

+诺基亚（中国）研究中心，北京和平里东街 11 号，诺基亚 1 号楼 100013

摘 要：本文通过对大规模真实语料的统计和分析，比较了不同领域词汇量、词类比例等特征的差异。在此基础上，对 TF*IDF 文本分类器中采用的 TF*IDF 权重算法以及由此衍生的 TF*IWF*IWF 权重算法从 TF、IWF 两个角度进行了改进，提出了一种基于大规模语料库的文本分类方法，并将它与 TF*IWF*IWF 权重算法进行了对比，从实验结果看这种方法将 F1 测度值提高了 12.28%，充分验证了其有效性。

关键词：大规模语料库 语料分析 文本分类

Analysis on Balance-Corpus and Text Categorization Based on Large-Scale Realistic Corpora

Chen Keli*, Zong Chengqing*, and Wang Xia+

*National Laboratory of Pattern Recognition, Institute of Automation, CAS. Beijing, 100080

+Nokia (China) Research Center, Nokia No.1 Building, Hepingli Dongjie 11, Beijing, 100013

Abstract: Based on the statistic and analytical results, this paper compares the differences of vocabulary and the ratios of part-of-speech in different domains. And then, this paper proposes a new approach to text categorization, which improves the TF*IWF*IWF algorithm from TF, IWF respectively. The new approach is compared with the TF*IWF*IWF algorithm. From the experimental results, we can find the F1-Measure has been improved for 12.28%. The efficiency of this approach is proved.

Keywords: Large-scale corpora, Corpora analysis, Text categorization

1 前 言

语料库建设和文本分类问题是自然语言处理领域两个热点问题。由于大规模语料库包含丰富的语言现象，能够充分反映语言使用中一些普遍性规律，所以在计算语言学领域颇受青睐。同时随着网络技术的发展和可用文本资源的飞速膨胀，有关海量信息的处理、分类也成为人们越来越关心的一个话题。在这方面许多人都做过研究，尤其是上个世纪 90 年

[1]本项研究受国家 973 项目“图像、语音、自然语言理解与知识挖掘”的资助(资助号为：G1998030504)，并得到诺基亚资助。

代以来，出现了各种各样基于机器学习的文本分类方法，如：向量空间模型（SVM），K 近邻（K-NN）算法，神经网络算法等等。尤其是前两种算法以其可操作性、准确性受到了人们的青睐，它们共同的一个关键问题就是如何选择适当的权重算法求得文本的向量表示。关于这个问题，Salton（1973）年提出了计算向量权重的 TF*IDF 算法，Thorsten Joachims 提出了概率 TF*IDF 算法[Thorsten, 1997]；Roberto Basili 提出了 TF*IWF*IWF 算法[Roberto et al, 1999]等等。这些权重算法与不同的分类算法（如上面提到的 SVM, k-NN）相结合生成了不同的文本分类器，但是在已报道的实验中，这些分类器的分类结果普遍不很理想，最好的其 F1-Measure 值也只有 85%左右，其中一个重要原因是这些分类器采用的权重算法不完备。本文结合训练语料的特点，分析了 TF*IWF*IWF 算法的优缺点，并进行了改进，然后结合线性分类方法，对改进后的权重算法与 TF*IWF*IWF 权重算法通过实验进行了比较，结果证明改进后的权重算法是非常有效的。

2 语料收集与分析

2.1 语料说明

本项研究的基础是我们与诺基亚（中国）研究中心合作为欧盟项目（LC-STAR）建设的 3087 万字的汉语语料库。语料来源于五大中文网站（新浪、人民网、中青论坛、三九健康网、科学时报网），时间上主要集中在近五年以内。目前我们将语料主要分成六大领域：体育、娱乐和游戏、财经、新闻、个人交流和消费信息。每个领域的语料都在 300 万字以上。

从这些语料中共抽取了 42923 个词（除去分词、标注错误，覆盖率达到了 99.62%），建立了总词表、各领域词表、各领域常用词表、各领域专用词表共四个词表。这里首先说明几个我们约定的术语。词频：在统计范围中某词出现的次数除以所有词的次数之和；领域覆盖率：统计词汇对某领域的覆盖率，等于统计词汇中所有词在该领域的词频之和；各领域常用词表：从各领域词表中按照频率从高到低取词建立的覆盖率达 90%的词表；各领域专用词表：由本领域内出现频率大于等于 0.0005%，在其他领域出现频率之和小于等于 0.0001%的词构成的词表。

2.2 语料统计结果

2.2.1 各领域词汇量的分布

表 1 中列出了各领域词汇量、常用词汇量、专用词汇量、专用词汇覆盖率的统计结果。

领域 项目	体育	娱乐和游戏	财经	新闻	个人交流	消费信息
词汇量	20264	29597	27283	11299	34879	30985
常用词汇量	3737	5755	4091	4278	6352	5239
专用词汇量	279	392	193	460	384	415
专用词汇覆盖率（%）	0.5694	0.5457	0.3199	0.7949	0.3654	0.6695

表 1： 各领域词汇量的分布

从表 1 各领域词汇量的统计结果可以看出：

（1） 各领域词汇量差别很大。词汇量最大的个人交流领域（共 34879）是最小的新闻领域（11299）的词汇量的三倍还要多。这和个人交流具有的口语性和综合性特点有关。

（2） 相比各领域词汇量而言，各领域常用词汇量之间的差别虽然不是很大，但也是明显的，词汇量最大的个人交流领域（6352）比最小的 SPO（3737）也多出 70%。

（3） 各领域专用词汇量的比例都很低。专用词汇量最多也只有 460 个，领域覆盖率只有 0.7949%。

2.2.2 各领域词类分布的比较

各领域词类分布结果如表 2 所示。在我们的工作中，汉语词类分为如下几类：数词（NUM）、名词（NOM）、介词（ADP）、形容词（ADJ）、副词（ADV）、代词（PRO）、连词（CON）、动词（VER）、量词（MEW）、辅助词（AUW）、其他词（OTHERS）。从词类分布可以看出某些词类在各领域中的使用频率是很不相同的，如：名词在消费信息领域使用频率最高，在个人交流领域中使用频率最低；代词在个人交流领域使用频率（6.80%）是其在财经领域使用频率（2.39%）的两倍还要多。

词性 领域	NUM	NOM	ADP	ADJ	ADV	PRO	CON	VER	MEW	AUW	OTHERS
体育	3.39%	33.38%	4.50%	5.55%	8.10%	3.97%	2.08%	20.15%	3.51%	6.71%	8.41%
娱乐和游戏	3.26%	31.74%	4.47%	6.21%	8.25%	4.91%	2.36%	19.77%	3.75%	8.38%	6.70%
财经	3.30%	34.59%	4.66%	6.13%	7.14%	2.39%	2.56%	22.07%	3.31%	7.17%	6.08%
新闻	3.02%	33.54%	5.21%	5.06%	6.39%	3.11%	2.51%	20.33%	3.92%	7.18%	9.09%
个人交流	3.32%	27.27%	4.33%	5.59%	9.71%	6.80%	2.37%	21.52%	3.38%	9.33%	6.06%
消费信息	2.64%	35.17%	4.29%	7.36%	7.56%	2.68%	2.99%	21.58%	2.89%	7.27%	5.06%

表 2：各领域词类分布

3 基于大规模语料库的文本分类

3.1 概述

关于文本分类，前人提出了很多算法，其中向量空间模型（VSM）算法，k 近邻（k-NN）算法以其可操作性受到很多人的青睐。使用这两种方法的关键之一就是如何选取适当的权重算法将文本或类用向量模型表示出来。常用的词权重算法有 TF*IDF 启发式权重算法[Salton, 1973]。其权重计算公式：

$$W(f_i,d)=TF(f_i,d)*IDF(f_i)=N(f_{id})*\log\left(N(f_i)/N\right) \qquad (3.1.1)$$

其中，$W(f_i,d)$是特征 f_i在文本 d 中的权重，$N(f_i)$是出现 f_i的训练文本数，N是总训练文本数，$N(f_{id})$是文本 d 中出现 f_i的次数。

在这个算法的基础上，Roberto Basili (1999)提出了 TF*IWF*IWF（公式 3.1.2）算法[Roberto et al, 1999]，实验证明，TF*IWF*IWF 比 TF*IDF 的分类效果有很大的提高。:

$$W(w_i,d)=TF(w_i,d)*IDF(w_i)=N(w_{id})*\left(\log\left(N(w_i)/N\right)\right)^2 \qquad (3.1.2)$$

其中，$N(w_i)$是训练语料中出现w_i的次数，N是训练语料中所有词出现次数之和，$N(w_{id})$是文本d中出现w_i的次数。

3.2 算法的改进

在TF*IWF*IWF算法的基础上，我们主要从下面几个方面进行了改进：一是从TF的角度提出了利用n次方根来调整词权重对频率的倚重；二是从IWF的角度引入了方差项。因为采用的是线性分类器，所以为了消除文本长度带来的影响，权重算法中的TF中采用的是词频而非词出现的次数。

关键词在某类的权重受三个因素影响：一是该词在当前类中的出现频率，二是该词在总语料中的出现频率，三是该词在不同类别之间出现频率的差异性。在TF*IWF*IWF算法中采用TF来表示第一个因素，但是我们知道如果在同一类别中词W1的出现频率是词W2的两倍并不能说W1对该类别的重要性是W2的两倍。因此为了消弱频率过度的影响，我们采用它的n($n\geq 1$)次方根形式，并对n=1、n=2[P.P.T.M. van Mun]、n=3、n=4等几种方根情况进行了实验。

在TF*IWF*IWF算法中采用IWF*IWF来表示第二个因素，即其出发点是总训练语料中出现次数越少的关键词权重应当越高。但TF*IWF*IWF（包括TF*IDF）忽略了第三个因素，关键词在总语料中出现次数多少并不能完全说明该词在分类中的“重要性”，频率相同的关键词在分类中的“重要性”也是不同的：在各类别之间分布越均匀，其重要性就越小；反之其重要性就越大。我们又知道方差是体现数据分布是否均匀一个很好的数学指标，但从方差公式中可以看出，方差大小又受到词频大小的影响，为了消除这种影响（因为词频因素在TF中已经表示了，这里需要得到的只是词频之间的差异性表示），我们用方差除以该词在各类中词频之和，于是得到式$\sqrt{\sum_j (p_{ij}-\bar{p}_i)^2}\Big/\sum_j p_{ij}$表示关键词在不同类之间的分布差异性。从上面的分析可以得到关键词在类和文本中的权重计算公式：

$$W(w_i,Cj)=\sqrt{\sum_j (p_{ij}-\bar{p}_i)^2}\Big/\sum_j p_{ij} \times \left(\log\left(N(w_i)/N\right)\right)^2 \times \sqrt[n]{p_{ij}} \qquad (3.3.1)$$

$$W(w_i,d)=\sqrt{\sum_j (p_{ij}-\bar{p}_i)^2}\Big/\sum_j p_{ij} \times \left(\log\left(N(w_i)/N\right)\right)^2 \times \sqrt[n]{p_{id}} \qquad (3.3.2)$$

其中，$p_{ij}=T_{ij}/L_j$，L_j是类C_j含有的所有词的次数之和，T_{ij}是词i在类C_j中出现的次数；

$p_{id}=T_{id}/L_d$，L_d文本d含有的所有词的次数之和，T_{id}是词i在文本d中出现的次数；

$\bar{p}_i = \sum_j p_{ij} / m$，其中 m 是类别数；理论上 n 可以取 1，2，3，4，…。

文本特征向量 $\vec{d}$ 和类特征向量 $\vec{C}_j$：

$$\vec{C}_j = (W(w_1, Cj), W(w_2, Cj), \cdots W(w_k, Cj)) \tag{3.3.3}$$

$$\vec{d} = (W(w_1, d), W(w_2, d), \cdots W(w_{k_d}, d)) \tag{3.3.4}$$

其中，$W(w_i, Cj)$、$W(w_1, d)$ 分别是词 i 在类别 C_j、文档 d 中的权重。k 是总关键词表中关键词数目，k_d 是文本 d 中包含的关键词的数目。

C_j 和 d 的相似度函数（这里采用的是线性分类器，严格来讲不是二者的相似度函数，这里为了表示方便只是借用这个名称）：

$$\begin{aligned} S(C_j, d) &= \vec{C}_j \bullet \vec{d} \\ &= (W(w_1, Cj), \cdots W(w_k, Cj)) \bullet (W(w_1, d), \cdots W(w_{k_d}, d)) \end{aligned} \tag{3.3.5}$$

4 实 验

4.1 语料的选择

本实验采用的语料包括两部分，一部分是上面提到的 3087 万字汉语语料库，分六个领域（体育、娱乐和游戏、财经、新闻、个人交流、消费信息），这一部分语料作为训练语料；第二部分语料是 1119 个从网上随意收集的文本（共 807158 个汉字），分属于这六个领域，各领域文本数分别为：消费信息（189）、娱乐和游戏（320）、财经（52）、新闻（100）、个人交流（101）、体育（357），这一部分语料作为开放测试的语料。

在各领域词表中按照词频由高到低选择以下数目的关键词：50、100、200、500、1000、1500、2000、2500、3000、3500、4000、4500、5000、5500、6000、6500、7000、7500、8000、8500、9000、9500、10000 按以下几种算法进行了实验，取每种算法下最好的分类效果进行比较。

(a) 在公式 3.3.1 和 3.3.2 中不包含方差项时，对 n 分别取 1，2，3，4 相应结果表示为 log*log，（log*log）*平方根，（log*log）*立方根，（log*log）*四次方根。

(b) 在公式 3.3.1 和 3.3.2 中保留方差项，n 取 1（结果表示为方差*（log*log））观察方差项对算法的影响。

(c) 从(b)中实验可以看出立方根对算法的改进最好，从(c)中可以看出方差对于算法的改进也很有帮助，所以结合二者作出在公式 3.3.1 和 3.3.2 中 n 取 3 时的分类效果。

4.2 实验结果及分析

下面各表为对应的实验结果：从表 3 可以看出 TF 采用频率立方根时效果最好，其 F1 测度值为 92.42%（正确率：93.35%；召回率：91.51%），比不采用方根（80.31%）提高 12.09%。从表 4 可以看出加上方差项以后效果明显，F1 测度值上升了 5.99%。结合上面两种情况下的实验结果，得到的最佳的权重算法，即保留方差项且 n 等于 3 得 F1 测度为 92.59%。

	log*log			(log*log)*平方根			(log*log)*立方根			(log*log)*四次方根		
	正确率	召回率	F1	正确率	召回率	F1	正确率	召回率	F1	正确率	召回率	F1
max	82.15%	78.55%	80.31%	92.83%	90.26%	91.53%	93.35%	91.51%	92.42%	94.38%	89.99%	92.13%

表 3： 方根对原模型的改进结果

	log*log			方差*(log*log)		
	正确率	召回率	F1	正确率	召回率	F1
max	82.15%	78.55%	80.31%	90.82%	82.22%	86.30%

表 4： 方差对原模型的改进结果

	log*log			(log*log)*立方根			方差*(log*log)*立方根		
	正确率	召回率	F1	正确率	召回率	F1	正确率	召回率	F1
max	82.15%	78.55%	80.31%	93.35%	91.51%	92.42%	96.06%	89.37%	92.59%

表 5： 两种方法结合后的改进效果

5 结束语

本文从大规模语料库的角度出发，统计了不同领域之间用词特点的差异性，在此基础上提出了一种基于大规模语料库的文本分类算法。根据上面的介绍，本分类方法最大特点在于方差的采用。一个词对于文本分类的重要性不只取决于该词在总语料中出现的频率，同时和它们在不同类间分布差异性关系也很密切。本文引入了方差来体现这种差异性，实验证明这种方法的采用取得了不错的效果。同时对于在权重算法中频率的表示形式，本文通过在 TF 表达式中取频率的一次方、平方、立方、四次方分别进行了实验，得到了比较好的频率项表达形式。最后结合两种改进途径提出了自己的权重计算方法。

参考文献

[1] R. Basili and A. Moschitti and M. Pazienza. 1999. A text classifier based on linguistic processing. In Proceedings of IJCAI-99, Machine Learning for Information Filtering.

[2] P. van Mun. Text Classification in Information Retrieval using Winnow. http://citeseer.nj.nec.com/133034.html.

[3] Thorsten Joachims. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of ICML'97, pages 143—151.

[4] 冯志伟. 2000. 中国语料库研究的历史与现状—语料库研究回顾与问题. In proceedings of ICCC ' 2000, Singapore, pages 1—15.

[5] Rile HU, Chengqing Zong, Juha Iso-Siila, Bo Xu. 2002. Investigation and analysis on designing Chinese balance corpus. In Proceedings of ISCSLP'2002, pages 335—338.

(其他参考文献略)