一种基于特征加权语言模型的微博分类新方法

崔为娜
（长春职业技术学院信息技术分院 长春 130033）

摘 要 微博作为社交媒体的后起之秀，已经得到快速的发展。微博快速的发展在带给人们便利的同时，也使人们置身于信息的海洋。针对微博中日益呈现出的信息过载问题，微博分类已经成为一个重要的研究课题。针对微博分类，提出一种基于特征加权语言模型的微博分类新方法。在新浪微博上抽取的真实标注数据集上进行的对比实验结果表明，所提方法是一个有效的微博分类方法。

关键词 微博，微博分类，语言模型
中图法分类号 TP391 **文献标识码** A

New Method of Microblog Classification Based on Feature Weighted Language Model

CUI Wei-na
(Information Technology Branch, Changchun Vocational Institute of Technology, Changchun 130033, China)

Abstract Microblog as a new social media has been rapid development. Microblog rapid development brings convenience to people at the same time, also makes people swimming in the ocean of information. Aiming at the increase in microblog presented the problem of information overload, microblog retrieval has become an important research topic. For microblog retrieval, this paper proposed a new microblog retrieval method based on feature weighted language model, and this method is mainly used in microblog statistical characteristics and semantic characteristics of combined to solve the retrieval problem of the microblog. Experiments were performed on the real annotation data set extracted from sina microblog, and the comparative experimental results show that the proposed method is an effective retrieval method.

Keywords Microblog, Microblog classification, Language model

1 引言

随着 Web2.0 的飞速发展，微博作为新型的社交媒体平台，为用户提供了更加迅速、便捷的网络服务。微博快速的发展给整个互联网行业带来了新的机遇与挑战。微博以其信息更新快、文本较短且内容丰富而备受广大互联网用户的喜爱。但是微博爆炸式的传播速度在给网民带来便利的同时，也会给用户提供很多不相关的信息。因此，微博用户往往要从大量微博记录中查找自己所关心的相关信息，为了能使微博用户快速地找到自己感兴趣的话题和微博内容，微博分类的研究就显得至关重要。

微博分类主要是对已经产生的微博进行有效的分类，从而可以方便微博用户便利地找到自己所关心和感兴趣的微博。微博分类不仅可以为微博用户浏览带来方便，同时微博分类的研究也是情感分析[1,2]、观点摘要[3]和观点检索系统[4]的重要基础。

微博分类的研究中除了对微博内容的分类，还有相当一部分是对微博情感进行分类。情感分类是按照文本表达的情感倾向性对文本进行分类[5]。本文的主要研究工作是对微博内容进行分类，微博情感分类并不在本文研究范围之内。

针对微博这个新型的社交媒体，如何让微博用户快速找到自己所关心和感兴趣的微博是微博需要解决的关键问题。而微博分类正是这个关键问题的基础，因此做好微博分类变得十分有意义，本文正是基于此进行的相关研究。

本文第 2 节描述本文提出的基于字符语言模型的分类方法；第 3 节给出实验结果并加以分析；最后总结全文并给出未来进一步的研究方向。

2 相关工作

微博平台每天都将产生大量微博数据，因此吸引了众多学者投身于微博的研究之中。在微博的众多研究之中，最为重要和基础的为微博分类的研究。近年来大量的学者投身到微博的分类研究中，并涌现出了一些优秀的方法，其中已涌现出一批以机器学习为主流的微博主客观分类方法[6-8]。从他们的对比实验中可以看出，相对于传统的基于规则的方法有明显的优势[9]。虽然在微博主客观分类中基于机器学习的方法要优于传统的基于规则的方法，但在机器学习方法中，特征的选择对分类结果的准确性起着至关重要的作用。因此，为了提高分类的准确性，需要选取有效的特征选择方法。张珊等人[10]通过计算 n-gram 特征项在不同情感类中出现的概率熵来进行特征选择，并在微博情感分类中取得了不错的效果；刘志明等人[11]通过对比信息增益、文档概率和卡方统计 3 种不同的特征选择方法的性能，最终选择最后的特征选择方法作为最后的实验方法。以上方法均采用单一的方法进行特征

本文受吉林省自然科学基金资助课题(M6138272)资助。
崔为娜(1980－)，女，硕士，讲师，主要研究方向为文本挖掘、自然语言处理，E-mail：wicuiweina@163.com。

选择，仅从某一个方面来衡量某一个特征的重要程度，并没有考虑更多的信息的融合和更为紧密的关系。同时，这些分类方法均未考虑特征之间的冗余性以及特征选择之间的互补性。

为了弥补上述方法的不足，本文提出了一种基于特征加权语言模型的微博分类方法，该方法采用了一种新的基于语言模型的特征词权重计算策略。事实上，微博分类相对于传统的文本分类有较大不同，首先人们在微博系统中发表的微博，通畅都是有字数限制的，因为这造成发表的微博内容通常特征词都比较稀疏，这相对于传统的文本分类中有大量的文本内容在数据稀疏行方面存在严重的稀疏。为了解决这个问题，我们关注的不是一个特征词在一个句子中的分布，而是一个特征词在整个类别中的分布。对于一条短文本的微博，拥有很少的特征词去表达一个类别，并且几乎每个特征词在微博中仅出现一次。因为我们考虑一个特征词在一个类别中出现的频率和一个特征词在整个语料集合中出现的频率，所以首先利用语言模型去计算每个特征词在特定类别中出现的概率，然后对于每个特征词最终的权重是由该特征词在具体类别中出现的概率除以该特征词在其他类别中出现的概率。

3 微博分类模型构建

3.1 分类模型系统的基本架构

首先对从新浪微博中爬取的微博进行预处理，包括去噪、分词、去停用词等；然后对处理好的微博建立微博语料库，在语料库中对基于字符串语言模型的微博分类模型进行训练，对于任意的一个微博 W 通过训练好的分类模型得到分类结果；最后对得到的最终结果用相应的评价指标进行评价。微博分类系统框架流程图如图 1 所示。

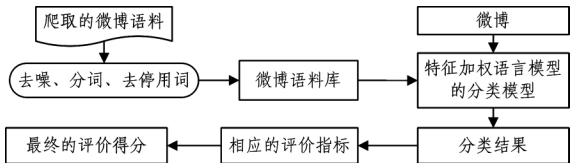

图 1 微博分类系统框架与流程

3.2 基于特征加权语言模型的微博分类

面对海量的微博数据，如何高效快速地找到自己感兴趣的微博成为目前大多数微博研究学者所面临的问题。如果能够快速准确地知道一个新发布的微博所属的分类，将其在发布时便归类在一起，可方便用户查看和寻找到自己感兴趣的微博，从而缩短用户寻找自己感兴趣微博的时间。

但在类别分类或主题分类中存在的主要挑战是如何在每个类别中有效地计算特征词的权重，解决这个问题的主要方法是利用最大似然评估特征词在类别中的分布。尤其是在像微博系统这样的应用领域，对于短文本的输入，需要更加有效地区分不同特征词之间的权重计算方法。

为了能够探索出更加有效的类别分类方法，先关注传统的文本分类领域，传统的文本分类领域并没有使用类别或主题的分布去预测和估计特征词的权重，而是使用文档的分布和语料的分布。假设一个特征词与一个具体的类别密切相关，同时与其他类别存在很少的关系，那么就认为这个特征词是在区分特定具体类别与其他类别时的重要特征。基于这样的假设，可以计算出更好质量的特征词权重，并利用这个特征词的权重去有效分类。

本文首先利用语言模型去计算每个特征词在特定类别中出现的概率，最终的特征词的权重是对于一个特征词的权重，是特征词在一个具体类别中出现的概率比上在其他类别中出现的概率。所提方法的关键点在于如何利用出现在具体类别中的诗词为类别的分类尽最大的贡献。

本文提出的模型主要是利用语言模型对微博系统中的微博进行分类，传统的语言模型主要是在信息检索中应用较为广泛，它主要是利用特征词在文档和语料集中的分布。在微博系统中，专注于一个词在一个具体类别中的分布，可得：

$$TP(w_i,c_k)=P_{ml}(w_i|c_k)=\frac{c(w_i,c_k)}{|c_k|} \tag{1}$$

其中，c_k 是当前类别，w_i 为微博中的特征词，$|c_k|$ 为当前类别的长度，即当前类别中特征词的个数，表示特征词在当前类别中出现的次数，是特征词 w_i 在当前类别中的最大似然估计。

本文在特征加权方面主要是采用以下方法，$P_{ml}(w_i|c_k)$ 表示每个特征词在当前类别中的主要作用并且可以适当地被修改。如果类别中的特征词主要在当前类别中频繁出现，而在其他类别中出现比较少或基本不出现，那么这个特征词在微博分类中具有很强的区分性。但是这样的特征词在式(1)中不能很好地被表示，因此采用的权重加权估算方法如式(2)和式(3)所示：

$$\mathrm{Re}atAvg(w_i,c_k)=\log\left(\frac{TP(w_i,c_k)}{TP(w_i,c_k')+\lambda}\right) \tag{2}$$

$$TP(w_i,c_k')=P_{ml}(w_i|c_k')=\frac{\sum_{j=1}^{n}TP(w_i,c_k)_{c_j\neq c_k}}{n-1} \tag{3}$$

其中，使用的比率是一个特征词在当前类别中出现的概率比上该特征词在其他类别中出现概率的平均值。n 表示数据集合中的类别总数目，c_k' 表示其他类别数目，$TP(w_i,c_k')$ 表示特征词 w_i 在其他类别中出现的平均似然估计。

如果一个特征词在当前类别中出现的概率高于其他类别中出现的概率，则由式(2)计算出的权重较高，相反，则由式(2)计算出的权重较低。为了避免零概率问题，在式(2)的分母中增加了一个很小的因子 λ。另外一个特征词在整个语料库中的分布作为分子添加到分布中，起到平滑的作用，整个语料库的分布通常用于语言模型。

4 实验结果与分析

4.1 实验数据

本文的实验语料来自“新浪微博”，时间为 2013 年 10 月 8 日至 2014 年 6 月 12 日，随机抓取每天上午 9 点至晚上 10 点的微博内容，共得到微博数量 10127 条。

实验的语料信息如表 1 所列。

表 1 相关语料集的统计信息情况

微博数	评论数	微博平均评论数
10127	105453	10.413

为了确保实验结果的有效性和准确性，本文实验采用交叉验证方式：首先随机将训练数据集平均分成 10 份，然后每次实验任取其中 9 份做为训练数据集，剩下的 1 份作为测试数据集，如此循环 10 次，保证每份样本既能作为训练集，又能作为测试集，最终结果取 10 次实验结果的平均值。

4.2 实验结果与分析

本文采用准确率作为分类模型性能的标准，同时为了更好地比较本文提出方法的有效性，在实验中比较了本文提出

的方法与 Naine Bayes、SVM 和基于词的语言模型方法的性能差异。为此设计了如下的实验，首先对原始的语料库进行处理，提取出每条微博的正文、发微博人、评论微博人、评论内容、评论时间、URl 等内容，并对正文和评论进行了中文分词、停用词过滤等处理。

实验结果对比如表 2－表 5 所列。

表 2　基于词的 Naive Bayes 方法的分类结果

特征个数	准确率	特征个数	准确率
100	91.3	1000	95.3
200	93.5	2000	92.6
300	94.1	5000	91.1
500	95.7	10000	90.7

表 3　基于词的 SVM 方法的分类结果

特征个数	准确率	特征个数	准确率
100	91.4	1000	96.1
200	92.5	2000	96.6
300	94.4	5000	96.3
500	95.1	10000	95.7

表 4　基于词的 n-Gram 语言模型的分类结果

特征个数	准确率	特征个数	准确率
1	94.1	5	97.4
2	97.2	6	97.3
3	97.3	7	97.3
4	97.3	8	97.2

表 5　基于特征加权语言模型的分类结果

特征个数	准确率	特征个数	准确率
1	96.4	5	99.5
2	98.6	6	99.6
3	99.5	7	99.7
4	99.7	8	99.7

(1)从上面的实验结果可知，对比表 2 和表 3 中的结果可以发现 NaivBayes 模型相对于 SVM 模型，对特征个数更加敏感。由表 2 可清晰地看到当特征数目小于 200 时，其准确率比较低，但是随着特征数目的增加，当特证数目为 300～500 时，其准确率在持续提高，然而当特征数目达到 500～1000 时，其准确率趋于平稳。随着特征数目继续增加，当特征数目超过 2000 以后，其准确率明显下降。相反在特征数目相同的情况下，从表 3 可以得到如下结论，基于 SVM 方法的鲁棒性比 NaivBayes 方法的更好，同时对特征数量的敏感度较低，从表 3 可知当特征数目较少时，采用 SVM 方法的准确度不高，但是随着特征数目的不断增加，其准确率一直趋于上升，当特征数到达一定数目时，其准确率趋于平稳，表现出了较好的鲁棒性。

(2)从表 4 可以看到，采用基于词的 n-Gram 方法，当 N 值为 4～5 时，其表现出的准确率最高，与期望基本一致。同时大量的研究表明，采用基于词的 n-Gram 方法在 N 为 2～3 时已经取得很好的效果。但是由于中文绝大多数词语都是由两个字组成的，因此对于中文，当 N 值取 4～5 时，也即包含 2 到 3 个词。

(3)最后对比表 5 和表 2－表 4 可以看出，基于特征加权语言模型比其他模型都表现出了很好的效果，主要是因为在基于特征加权语言模型中，关注的不是一个特征词在一条微博中的分布，而是一个特征词在整个类别中的分布。对于一条短文本的微博，拥有很少的特征词去表达一个类别，并且几乎每个特征词在问句中仅出现一次。因为考虑一个特征词在一个类别中出现的频率和一个特征词在整个语料集合中出现的频率，能更好地表现出一个特征在微博中的重要地位，从而表现出更好的效果。

结束语　微博自出现以来就一直备受人们关注，现在微博平台已经成为人们表达观点及信息分享的平台。随着用户数量的不断增加，产生的微博数据也日趋增加，面对海量的微博数据，如何快速有效地获取有价值的信息成为广大微博用户的迫切需求。微博分类主要是对已产生的微博进行有效地分类，从而可以方便用户根据自己的喜好和关注点选取自己喜欢的微博话题进行浏览。本文提出一种基于特征加权语言模型微博分类的新方法，多角度的对比实验结果表明，本文提出的方法能够有效解决微博系统中微博分类的问题。当然，本文描述的方法尚存在着速度等方面的问题，这需要在今后的研究中加以改进。

参 考 文 献

[1] Jiang L, Yu M, Zhou M, et al. Target-dependent Twit-ter Sentiment Classification[C]// Proceedings of the AMACL. 2011: 151-160

[2] Barbosa L, Feng J L. Robust Sentiment Detection on Twitter from Biased and Noisy[C]// Proceedings of the COLING. 2010: 36-44

[3] Hu M Q, Liu B. Opinion Extraction and Summarization on the Web[C]// Proceedings of the AAAI. 2006: 1621-1624

[4] Yu H, Hatzivassilohlou V. Towards Answering Opinion Question: Separating Facts from Opinions and Identifying the Polarity of Opinion Sebtences[C]// Proceedings of the EMNLP. 2003: 129-136

[5] Itti L, Koch C, Niebur E. A Model of Saliency Based Visual Attention for Rapid Scene Analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259

[6] Go A, Bhayani R, Huang L, et al. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, 2009, 56(3): 136-145

[7] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining [C]// Proceedings of the LREC. 2010: 1320-1326

[8] Davidov D, Tsur O. Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys[J]. Proceedings of the 23rd International Conference on Computational Linguistics: Posters(COLING '10). 2010, 58(3): 146-158

[9] 李寿山，黄居仁. 基于 Stacking 组合分类方法的中文情感分类研究[J]. 中文信息学报，2010，24(5)：56-61

[10] 张珊，于留宝，胡长军. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学，2012，39(z3)：146-148，176

[11] 刘志明，刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用，2012，48(1)：1-4

[12] Sun R, Cui H, Li K, et al. Dependency Relatio n Matching for Answer Selection[C]// Proc. of SIGIR 2005. Salvador. Brazil

[13] Feng L, Ya-qian Z, Xuan-Jing H, et al. Dependency Relation Triples Matching for Question Answering[J]. Acta Automatica Sinica, 2008, 34(11): 1410-1416

[14] Chang C, Lin C. LIBSVM: a Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems and Technology, 2001, 2(3): 389-396