

本科毕业论文（设计）

英文文献翻译

|  |  |
| --- | --- |
| **文献英文题目** | **Categorizing Blogger’s Interests Based On Short Snippets of Blog Posts** |
| **学 生 姓 名** | **秦杰** |
| **学号** | **2017081021** |
| **专业** | **软件工程** |
| **年级班级** | **2017级1班** |
| **指导教师** | **郭本俊** |
| **所在学院** | **软件工程学院** |
| **提交日期** | **2021年1月4日** |

2021 年 1 月

成都信息工程大学 软件工程学院

Categorizing Blogger’s Interests Based On Short Snippets of Blog Posts

Jiahui Liu, Larry Birnbaum, Bryan Pardo

Northwestern University

2133 Sheridan Road, Evanston, IL, 60201, USA

j-liu2@northwestern.edu, {birnbaum, pardo}@cs.northwestern.edu

摘要：博客已经成为人们表达的重要媒介在网上发表意见和分享信息。预测博主的兴趣对信息检索是有益的以及博客圈中的知识发现。在本文中，我们提出了一种两层分类模型来进行分类基于一组收集的短片段的博客作者的兴趣从他们的博客文章。实验是按照一份清单进行的，博主从博客目录和他们的片段收集自谷歌博客搜索。结果表明，该方法具有良好的实用价值并且该方法对低级错误具有较强的精确性，在博客的兴趣分类方面有令人满意的表现。

类别和主题描述符：

I.5.2[模式重组]:设计方法论-分类器

设计和评价。

一般术语算法，实验，性能。

关键词：博客，兴趣，分类

# 介绍

作为一种重要的在线出版形式，普遍适用于互联网用户们，博客已经成为一种动态的、多样化的媒体用于信息的创造、分发和积累。对某些领域感兴趣的博主会维护博客站点，发布有关他们感兴趣领域的新闻、观点和想法。识别和分类博客作者的兴趣对于从博客中检索信息和发现知识是很有价值的。博主发表的博客文章为预测他们的兴趣提供了重要线索。然而，直接对博主所写的所有文本进行分类并不能产生准确的预测。这主要有两个原因。首先，博客文章是用一种非正式的古怪风格写的。有时，博主们甚至发明新的词汇和语法来表达自己的独特之处。其次，博主们不会局限于一个话题[2]。因此，所有博客文章的混合是一个多主题和嘈杂的文本文档，很难分类。为了解决这些问题，我们提出了一个两层的分类模型来分类博主的兴趣。在第一层中，训练文本分类器来预测属于某个领域类别的博客文章的概率。虽然单个博客文章的分类并不完美，但多个博客文章的分类为预测该博客的整体兴趣提供了重要信息。在第二层，我们从博客作者所写文章的分类概率集合中获得特征。这些特性用于对博客作者的兴趣进行分类。通过整合所有类别的博客文章的成员信息，第二层分类器能够了解这些类别之间的主题相关性。我们使用从博客目录中编译的博客作者集合，以及从谷歌博客搜索中检索到的博客文章片段来实验所提出的模型。正如我们所预料的那样，第一层中的短片段的分类不是很准确。但是，从所有片段的概率集合中得出的特征对于预测博客作者的整体兴趣 是有意义和有用的。通过对所有类别进行微平均，博客兴趣分类的F1值达到了0.845。

# 拟议的技术

在这篇论文中，我们使用博客的文章片段来描述他们的兴趣。使用代码片段就不需要下载整个网页了。代码段的处理速度也比整个页面快，支持实时处理，这对web应用程序尤其关键。对于每个博客作者b，我们收集一组博客作者b最近写的博客文章记为P(b)={p1,p2,…,pn},对于每个帖子pi，我们提取一个小片段Si=Snippet(Pi)。一个片段包括标题和一篇博客文章的前几句话，大约40个单词。Sb={s1,s2,…,sn}是为博客作者 b收集的片段集。目标是将b的兴趣分类为一个或多个类，从一组类中抽取，C` = {c1,c2,...,cn}。提出的技术通过一个两层分类模型来解决这个问题。在第一层中，分类器为每个帖子片段si生成一个概率估计p(Cj | Sj)，这是片段si属于类别cj的概率。在第二层，我们从博客作者b所写的所有片段的分类概率中提取特征，并利用这些特征预测博客作者b的兴趣。

* 1. **对博客文章的片段进行分类**

为了构建片段的文本分类器，我们以片段的词干内容词为特征，去掉停止词。对于每个类别cj，我们根据信息增益[6]选择最具预测性的2000个词干。为了对文本片段进行分类，我们使用了序列最小模型(SMO)[3]，该模型已被证明是高效和有效的文本分类方法。支持向量机的输出适合sigmoid模型，以推导出合适的成员[4]的概率估计。我们为每个类别cj创建一个支持向量机。每个SVM用一个sigmoid模型对片段 si进行处理，得到一个m值向量，其中第j个特征是片段 si属于类别cj的概率(请注意，这些类别并不是相互排斥的) 。

Si = {P(C1 | Si),… P(Cj | Si),…P(Cm | Si)} (1)

* 1. **分类博客作者的兴趣**

博客作者的片段分类(公式1)提供了关于博客作者兴趣的重要线索。特征来源于代码片段的分类，以训练博客的分类器。对于每一类cj，我们取每一种可能去估算P(Cj | Si)(Si 属于Sb)。这可能的估算公式为

Ej = {P(Cj | Si),…P(Cj | Si),…P(Cj | Sn)} (2)

Ej显示了一个博主写了多少关于cj类别的文章。Ej中的概率估计被装箱并放入直方图中。对于每一类cj，我们将[0,1]范围划分为K，根据 P(Cj | Si) 计算Sb中片段的k元素分布。我们表示第j类Dkj分布中的第k个元素。Dkj是Sb中片段所占比例，P(Cj | Si)在第k个间隔内。形式上，Dkj由公式3计算。

Dkj=|Sk| / |Sb|, Sk = {s | s属于Sb},{P(Cj | s) 属于 [(k-1)/K \* (k/K)]} (3)

我们还计算了每一类cj的P(Cj | Si)的均值和方差。它们表示为Djmean。比例、均值和方差构成了一类cj的特征Dj。

Dj={Dj1,…,Djk,Djmean,Djvar} (4)

Dj描述了博客b写了多少关于cj类别的文章。然而，仅仅知道一个博主写了一些关于cj类的文章并不足以预测她对cj感兴趣。这是因为[2]博客的多主题特性。为了描述博主的兴趣，我们使用所有类别派生的所有特征。博主b被编码为Dj为了 C={c1,c2,…,cm}，如(5)所示

B = {D1,D2,…,Dj,…,Dm} (5)

为了对博主的兴趣进行分类，我们使用(5)中所示的派生特征训练第二层分类器。我们试验了许多机器学习算法，包括SMO[3]、最近邻[1]和只包含一个隐层的神经网络。我们的实验表明，神经网络达到最高的F1测度。

# 实验

在我们的实验中，我们从博客类型和雅虎目录的博客部分收集了4428个博客站点。这些网站分为八大类:博客目录中的艺术、商业、教育、卫生、法律、政治、宗教和技术。在我们的实验中，我们假设每个博客站点都由单个博主拥有。我们用博客目录中为每个博客站点分配的类别来标记每个博主。使用谷歌博客搜索收集了博客作者的博客片段。我们用每个博客站点的URL向博客搜索引擎查询，并收集每个博客最近的30个(或更少)结果。标题和搜索引擎返回的搜索结果摘要一起用作代码片段。我们总共为4428位博主收集了86598篇博文片段。在我们提出的两层分类模型的实验中，我们需要两个单独的数据集在每一层进行分类。我们将博主随机分成两组。为前一半博客作者检索的博客片段被用来训练博客片段的第一层分类器。使用代码片段分类器，我们使用10倍交叉验证来评估第二层分类器在博客的后半部分。我们使用Weka包[5]实现了上面描述的两层分类模型。为了评估每一层的分类器，我们使用传统的Precision、Recall和F1 measure。我们计算了三个度量的微观平均值，它结合了各个类别的性能，并根据类别中实例的数量进行加权。

表 1 两层分类性能

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 measure |
| 片段分类 | 0.717 | 0.416 | 0.526 |
| 博客主分类 | 0.870 | 0.822 | 0.845 |

如表1所示，短片段的第一层分类器的分类不是很准确。而在博主兴趣的第二层分类中，采用一个隐层8个节点的神经网络，分类器的微观层次F1值为0.845。实验结果表明，第二层分类器对第一层分类器的误差具有较好的精确性。换句话说，虽然第一层的精度较低，但它足以在第二层进行预测。

# 结论

在本文中，我们提出了一个基于博客文章短片段的两层分类模型来对博客作者的兴趣进行分类。在第一层中，我们预测属于每个类别的片段的概率。在第二层，我们从博客作者所写的片段的概率集合中获得特征，并使用这些特征对博客作者的兴趣进行分类。尽管有噪声的博客文章片段较短且难以分类，但两层分类模型对较低层的错误具有较强的精确性，在对博客作者兴趣进行分类方面取得了令人满意的效果。

# 参考文献

[1] Martin, B. 1995. “Instance-Based learning: Nearest Neighbor With Generalization”. Hamilton, New Zealand.

[2] Pew Internet and the American Life Project. 2006 <http://www.pewinternet.org/PPF/r/186/report_display.asp>.

[3] Platt, J. 1998. “Machines using Sequential Minimal Optimization”. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning,

[4] Platt, J. C. 1999. “Probabilities for SV machines”. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers. MIT Press.

[5] Witten, I. H. and Frank, E. 2005 "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

[6] Yang, Y., Pedersen J.P. 1997. “A Comparative Study on Feature Selection in Text Categorization”. In ICML 1997.

**附英文原文：**

