基于描述文本和实体标签的网络视频分类算法

何春辉
(湘潭大学 数学与计算科学学院，湖南 湘潭 411105)

摘 要：目前，各大社交平台和视频点播网站的网络视频数量出现了爆炸式的增长，如何快速准确地对这些网络视频进行归类和管理成为了研究的热点问题. 为了较好地解决这种分类任务，文中提出了基于描述文本和实体标签的网络视频分类算法，该算法结合了描述文本内容和知识图谱中的实体标签来构造文档-特征矩阵. 实验结果表明使用了实体标签的视频分类算法性能更好，平均精确率和平均召回率以及平均F_1值比未使用实体标签的视频分类算法要高2%以上.

关键词：特征提取；视频分类；实体标签；SVM
中图分类号：TP391 **文献标识码：**A doi:10.3969/j.issn.1672-7304.2018.03.0010
文章编号：1672–7304(2018)03–0046–03

Web Video Classification Algorithm Based on Description Text and Entity Tag

HE Chunhui
(School of Mathematics and Computational Sciences, Xiangtan University, Xiangtan, Hunan 411105, China)

Abstract: At present there has been an explosive growth in the number of web video on major social platforms and video on demand web sites. How to quickly and accurately classify and manage these web videos has become a hot spot of research. In order to solve this classification task, a web video classification algorithm based on description text and entity tag was proposed in this paper. The algorithm combines the description text and the entity tags in the knowledge graph to construct a document-feature matrix. The experimental results show that the video classification algorithm using the entity tag shows better performance, and the average precision and average recall and the average F_1 value are higher 2% than the video classification algorithm of the unused entity tag.

Key words: feature extraction; video classification; entity tag; SVM

随着互联网和大数据技术的发展，像YouTube、推特和腾讯视频这种大型社交视频网站的网络视频数量出现了爆炸式的增长. 大量的网络视频造成了数据堆积[1-2]，且目前无法及时准确对它们进行有效管理和应用. 于是，快速准确地给出网络视频的分类[3-4]，对于提升用户的体验效果和发现潜在的商业价值有重要意义. 网络视频分类是指将未分类的视频数据通过某种分类算法自动划分到事先指定类别的过程[5]. 在常见的网络视频数据中，视频所含内容的重要信息都可以通过文本形式[6-7]来进行描述. 这些文本信息经常出现在视频的简介部分、社交信息[8]以及一些实体标签数据中，通过分析挖掘与视频相关联的文本信息，可以间接对网络视频进行快速有效的分类[9]. 由上述分析，本文提出了基于描述文本和实体标签的网络视频分类算法，它同时结合描述文本和实体标签来构造分类算法的特征，可以快速准确地完成网络视频分类任务. 最后在相应的公开网络视频数据集上对算法的性能进行了测试，并给出了相应的结论.

1 特征提取

首先是计算文本中词语的权重，然后才可以根据权重对词语进行筛选得出特征词[10]. 本文采用了 TF-IDF 算法来计算文本中所有词语的权重. 其中 *TF* 是词频，用来统计文本中词出现的次数；*IDF* 为逆文档频率，它可以有效过滤一些无意义的词. *TF* 的计算如公式(1)所示.

$$TF_i = \frac{n_i}{\sum_k n_k}, \tag{1}$$

收稿日期：2018-04-27
作者简介：何春辉(1991-)，男，湖南永州人，工程师，硕士，主要从事数据挖掘及信息处理研究. E-mail: xtuhch@163.com

其中 n_i 指文档中第 i 个词语出现的次数；$\sum_k n_k$ 指文档中所有词语出现的总次数．逆文档频率 *IDF* 的计算如公式(2)所示．

$$IDF(t,D)=\log(\frac{N}{n_t}), \tag{2}$$

其中，t 指被测试词；D 指总文本集合；N 指文本总数量；n_t 指含有被测词语 t 的文本总数量．为了达到较好的效果，将 *TF* 和 *IDF* 得到的结果做积，即 W_t=*TF***IDF*，来算出第 t 个词语的权重 W_t．用 TF-IDF 算法得到所有词的权重并经过相应的过滤规则可将文本中剩下的若干特征词语加入到词袋模型里面，以便用于模型训练和预测．

2　多类 SVM 分类算法原理

在多类别的分类问题中，1 对 1 方式的支持向量机(SVM)是采用传统二分类支持向量机算法(SVM)对任意 2 类不同的数据样本间都构造一个最优的自动决策超平面[11]的方式来达到多分类的目的．对一个含有 $K(k>2)$个类别的分类问题，采用这种 1 对 1 方式的分类策略必须要先构造 k*(k-1)/2 个最优的分类超平面．该方法最根本的解决思想是采用分治策略，将多类别的分类问题分解为多个二分类问题来进行相应的求解．常用自动分类决策的超平面构造方法如下：

从样本数据中取出所有满足 y_i=s 与 y_i=t(其中 1≤s，t≤k，s≠t)的样本，通过任意 2 类 SVM 算法来构造最优的分类超平面决策函数，决策函数的形式如公式(3)所示．

$$F_{st}(x)=w_{st}\cdot\phi(x)+b_{st}=\sum_{i=st}a_i^{st}y_iK(x_i,x)+b_{st}. \tag{3}$$

再根据公式(3)对这 k 类样本数据中的每一种都构造一个最优的分类超平面自动决策函数．

根据相应的样本数据构造出分类超平面自动决策函数以后，所面对的核心难题是如何利用得到的分类超平面自动决策函数对未知样本数据给出准确的类别预测．目前常用的解决策略是采用投票选举机制：对于一个测试样本 x，为了判定它到底是属于哪种类型，该投票机制会综合考虑和分析前面得到的所有 $k(k-1)/2$ 个自动分类决策函数对 x 所属类别的判定结果，如果其中有一个自动分类决策函数将 x 判定为第 i 类，就说明第 i 类获得 1 票；以此类推，最后经过投票结果的统计，得票数量最多的那个类别就是样本 x 所对应的类别．1 对 1 的支持向量机分类方法，它的优点在于每次投入训练过程中的样本数量会相对较少，因此单个自动决策超平面的训练速度相对来说比较快，同时精度也相当高．实践中需要注意的是，由于 k 类问题需要训练 $k(k-1)/2$ 个分类超平面，当 k 的值较大时，采用 1 对 1 的支持向量机分类方式性能上会存在部分损失，且主要是体现在计算速度上．所以对于这种情况，建议使用其它策略的多分类算法来建模求解，常见的有多对多形式 SVM 分类．

3　数据集及预处理

3.1　数据集

文中使用 Google 公司所公布的一个大型公开视频数据集：YouTube-8M[1]来验证算法的分类性能．这个原始数据集共包含 8 000 000 万个 YouTube 视频链接以及视频相应的描述文本，并进行了 video-level(视频层级)的标注，总共将其标注为 4 716 种知识图谱的实体标签，平均每个实体标签对应 2 000 多个训练视频．其中，这 4 716 个实体标签总共被分为了 24 种类别．通过将实体标签对应类别映射成视频类别，就可实现通过文本内容和实体标签的挖掘分析来进行网络视频的分类任务．通过这种策略就将文本挖掘分析和视频分类完美结合起来，轻松实现跨域分析．

3.2　预处理

通过对原始数据集的分析，发现它存在以下 2 个问题：(1)某些视频的描述文本存在缺失情况；(2)某些类别对应的视频数量很少，即数据出现不平衡．为了解决以上问题，文中在数据预处理过程中对存在数据缺失情况和视频总数量小于 100 的那些类别进行相应过滤处理，最后只保留了 13 个类别总共包含 4 078 种实体标签的视频分类样本．最终将每一个视频数据都处理成只包含 ID、实体标签名称、描述文本原始内容及所属类别这 4 个属性．数据集各类别划分情况见表 1．

表 1　数据集划分

类别名称	数量	类别名称	数量
Arts&Entertainment	700	Autos&Vehicles	474
Beauty&Fitness	100	Business&Industrial	295
Computers&Electronics	339	Food&Drink	321
Games	891	Hobbies&Leisure	152
Home&Garden	134	Pets&Animals	177
Science	153	Shopping	135
Sports	207	总计	4 078

在得到上述初步预处理的视频数据集后，接下来通过使用 Python 的 jieba[12]工具包对数据集进行去除停用词、标点符号以及特殊字符等常规处理，然后利用 TF-IDF 算法来提取描述文本中对应的重要文本特征，据此得出文档-特征矩阵，为后续基于描述文本和实体标签的视频分类算法的训练和测试阶段做好相关的数据准备工作.

4 试验分析

为了充分验证文中提出的基于描述文本和实体标签的网络视频分类算法的性能，在实验验证阶段总共设计了 2 组不同的实验来进行相应的对比分析. 第 1 组是使用支持向量机 SVM 分类算法和逻辑回归 LR 算法在只使用描述文本特征作为分类特征的情况下，进行视频的分类实验；第 2 组是使用支持向量机 SVM 分类算法和逻辑回归 LR 算法在使用描述文本特征加上视频对应的实体标签作为分类特征的情况下，进行视频分类实验. 以上 2 组实验既可以反映相同模型取不同特征时的性能情况，又可反映不同模型取相同特征时的性能. 实验利用 SK-learn[13]的 SVM 和 LR 算法，并使用第 3 节所述的数据集，且所有实验中都采用 80%的样本训练模型，20%的样本验证模型. 文中使用各类别平均精确率 P、平均召回率 R 和平均 F_1 值来评价实验效果，结果见表 2.

表 2 算法实验结果对比 %

实验算法	评价指标		
	P	R	F_1
LR(不加实体标签)	79.97	79.78	79.88
LR(加实体标签)	81.12	80.51	80.79
SVM(不加实体标签)	81.41	81.37	81.39
SVM(加实体标签)	84.27	83.09	83.68

根据表 2 实验结果可以看出，SVM 算法的分类效果总体上比 LR 算法的好. 在不加实体标签情况下，SVM 算法的分类平均精确率为 81.41%，平均召回率为 81.37%，平均 F_1 值为 81.39%；在分类特征中加入实体标签之后，SVM 和 LR 的性能整体上都有较大的提升，SVM 算法的分类平均精确率达到了 84.27%，平均召回率达到了 83.09%，平均 F_1 值达到了 83.68%.

5 结论

对于社交平台和互联网上爆炸式增长的网络视频数据，如何快速准确地对它们进行归类和管理是一个有重大意义的实际问题. 对于目前的分类方法来说，大部分都是基于视频内容来进行分类，效率和准确率都难以得到保证. 为了较好地解决这种现状，文中提出了一种基于描述文本和实体标签的网络视频分类算法. 它可以较好地改善现有方法的不足，且算法的执行效率和分类准确率均得到了可靠的实验验证. 不足之处在于，只使用了 LR 算法和 1 对 1 的 SVM 算法来作为对比实验. 下一步将考虑使用其它数据集或者大规模的实体标注算法以及深度学习的优化思想来进一步提升文中提出的网络视频分类算法的性能.

参考文献：

[1]ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. YouTube-8M: A large-scale video classification benchmark[J/OL]. (2016-09-27) [2018-04-27]. https://arxiv. org/abs/1609.08675.

[2]艾丽丽, 孙明远, 孙健. 基于文本描述的视频分类系统建模[J]. 信息通信, 2013(04): 63-64.

[3]陈芬, 赖茂生. 多特征视频分类挖掘实验研究[J]. 现代图书情报技术, 2012(05): 76-80.

[4]艾丽丽. 基于文本挖掘的视频资源分类研究[D]. 成都: 电子科技大学, 2013.

[5]吴雨希. 基于文本挖掘的视频标签生成及视频分类研究[D]. 上海: 上海交通大学, 2015.

[6]齐全, 董晶. 基于描述能力的视频标题分类[J]. 华南理工大学学报: 自然科学版, 2011, 39(7): 134-139.

[7]刘璐, 贾彩燕. 基于文本扩展模型的网络视频聚类方法[J]. 智能系统学报, 2017, 12(6): 799-805.

[8]朱义明. 基于社交信息的网络视频分类[D]. 成都: 西南交通大学, 2011.

[9]桑苗杰. 基于语义的视频内容标注与检索系统研究与实现[D]. 北京: 北京工业大学, 2015.

[10]何春辉, 李云翔, 王孟然, 等. 改进的TextRank双层单文档摘要提取算法[J]. 湖南城市学院学报: 自然科学版, 2017, 26(6): 55-60.

[11]郭显娥, 武伟, 刘春贵, 等. 多类SVM分类算法的研究[J]. 山西大同大学学报: 自然科学版, 2010, 26(3): 6-8.

[12]SUN J Y. Jieba 中文分词组件[EB/OL]. (2017-08-28) [2018-04-27]. https://github.com/fxsjy/jieba.

[13]PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

(责任编校：龚伦峰)