PaperFree检测报告简明打印版

相似度：30.86%
编　号：KJQAAEDYEDBUAPQU
标　题：基于搜索关键词的用户属性分析预测
作　者：-
长　度：21730字符
时　间：2017-05-16 23:12:39
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：1.00%　篇名：《移动广告精准投放平台的设计与实现》
来源：《北京交通大学硕士论文》　年份：2013　作者：李爱飞
2. 相似度：0.87%　篇名：《支持向量机分类算法研究与应用》
来源：《湖南大学硕士论文》　年份：2007　作者：彭璐
3. 相似度：0.76%　篇名：《基于k-最近邻图的小样本KNN分类算法》
来源：《计算机工程》　年份：2011　作者：刘应东
4. 相似度：0.57%　篇名：《一类非线性系统无模型控制器的设计与分析》
来源：《东北大学硕士论文》　年份：2012　作者：王凤伟
5. 相似度：0.53%　篇名：《信息过滤系统中特征选择算法的研究》
来源：《山东师范大学硕士论文》　年份：2015　作者：王美方
6. 相似度：0.51%　篇名：《城市设计的非线性模式》
来源：《美术学报》　年份：2013　作者：钱缨
7. 相似度：0.49%　篇名：《基于rs-svm的中文文本分类研究》
来源：《天津财经大学硕士论文》　年份：2016　作者：周倩
8. 相似度：0.38%　篇名：《基于机器学习的花卉分类算法研究》
来源：《现代计算机：上下旬》　年份：2013　作者：王永波
9. 相似度：0.36%　篇名：《学习者情感挖掘：一个重要的教育技术研究领域》
来源：《软件导刊.教育技术》　年份：2014　作者：晏皓鸾
10. 相似度：0.33%　篇名：《随机森林与支持向量机分类性能比较》
来源：《软件》　年份：2012　作者：黄衍
11. 相似度：0.30%　篇名：《基于朴素贝叶斯方法的中文文本分类研究》
来源：《河北大学硕士论文》　年份：2011　作者：李丹
12. 相似度：0.27%　篇名：《基于LDA模型和SVM的文本分类研究》
来源：《网友世界》　年份：2013　作者：李小三
13. 相似度：0.27%　篇名：《自动数据挖掘算法》
来源：《计算机系统应用》　年份：2012　作者：郑盼丽
14. 相似度：0.23%　篇名：《广告精准投放的大数据玄机》
来源：《通讯世界》　年份：2013　作者：许翠苹
15. 相似度：0.23%　篇名：《基于机器学习的Web文本自动分类》
来源：《软件导刊》　年份：2015　作者：袁晓曦
16. 相似度：0.23%　篇名：《浅析科技项目查重方法的研究与现状》
来源：《中国科教创新导刊》　年份：2013　作者：史科蕾
17. 相似度：0.22%　篇名：《用于分类的支持向量机》
来源：《广西师范学院学报：自然科学版》　年份：2004　作者：黄发良
18. 相似度：0.22%　篇名：《基于小波分析的电子文献分类》
来源：《情报学报》　年份：2013　作者：张开选
19. 相似度：0.19%　篇名：《基于vsm模型和特征选择算法的中文文本自动分类研究》
来源：《江西师范大学硕士论文》　年份：2011　作者：朱坤红
20. 相似度：0.18%　篇名：《基于社会标签的文本聚类研究》
来源：《现代图书情报技术》　年份：2013　作者：何文静

21. 相似度：0.16% 篇名：《基于支持向量机的中国股指期货回归预测研究》
来源：《中国管理科学》 年份：2013 作者：赛英
22. 相似度：0.16% 篇名：《基于主题情感混合模型的无监督文本情感分析》
来源：《北京大学学报：自然科学版》 年份：2013 作者：孙艳
23. 相似度：0.16% 篇名：《文本分类中TF—IDF方法的改进研究》
来源：《现代图书情报技术》 年份：2013 作者：覃世安
24. 相似度：0.15% 篇名：《基于SVM的中小企业集合债券融资个体信用风险度量研究》
来源：《中南大学学报：社会科学版》 年份：2013 作者：曾江洪
25. 相似度：0.15% 篇名：《中文网络评论的情感分类：句子与段落的比较研究》
来源：《情报学报》 年份：2013 作者：郑丽娟
26. 相似度：0.13% 篇名：《基于机器学习的文本分类算法研究》
来源：《广西大学硕士论文》 年份：2007 作者：杨挚诚
27. 相似度：0.11% 篇名：《基于三维分类模型语义搜索的设计与实现》
来源：《电子科技大学硕士论文》 年份：2011 作者：高杰旺
28. 相似度：0.09% 篇名：《基于模糊模式的图书质量识别方法仿真研究》
来源：《计算机仿真》 年份：2011 作者：王国权
29. 相似度：0.09% 篇名：《数据挖掘与数据仓库分析》
来源：《计算机光盘软件与应用》 年份：2011 作者：李远菲
30. 相似度：0.09% 篇名：《改进的k近邻算法在网页文本分类中的应用》
来源：《安徽大学硕士论文》 年份：2010 作者：白凡
31. 相似度：0.09% 篇名：《基于SVM的高粱叶片病斑图像自动分割提取方法研究》
来源：《农学学报》 年份：2014 作者：白文斌
32. 相似度：0.08% 篇名：《改进TF—IDF算法的文本特征项权值计算方法》
来源：《图书情报工作》 年份：2013 作者：路永和
33. 相似度：0.07% 篇名：《如何进行文学类文本阅读》
来源：《中华少年：研究青少年教育》 年份：2013 作者：张乔
34. 相似度：0.07% 篇名：《基于ACO算法的SVM核函数的参数优化》
来源：《计算机工程与科学》 年份：2011 作者：赵新建
35. 相似度：0.06% 篇名：《改进朴素贝叶斯分类算法的研究与应用》
来源：《湖南大学学报：自然科学版》 年份：2012 作者：吕昊
36. 相似度：0.06% 篇名：《网络论坛教学评论的自动情感分析方法——以湖南商学院枫华论坛为例》
来源：《湖南商学院学报》 年份：2013 作者：曾伏秋
37. 相似度：0.06% 篇名：《一种基于类别分布信息的文本特征选择模型》
来源：《图书情报工作》 年份：2013 作者：刘海峰
38. 相似度：0.06% 篇名：《文本自动分类研究——基于径向基函数》
来源：《情报科学》 年份：2013 作者：黄翠玉

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)

1. 相似度：3.05% 标题：《分类算法总结 - bluenight专栏 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/chl033/article/details/5204220
2. 相似度：2.25% 标题：《文本自动分类研究进展KNN_嘟嘟和兜兜和圆圆_新浪博客》
来源：http://blog.sina.com.cn/s/blog_939ae64201011069.html
3. 相似度：1.45% 标题：《人工神经网络总结 - u013240812的专栏 - 博客频道 - ...》
来源：http://blog.csdn.net/u013240812/article/details/49361131
4. 相似度：1.33% 标题：《基于Spark的大数据精准营销中搜狗搜索引擎的用户画像挖掘 ...》
来源：http://blog.csdn.net/u011239443/article/details/53735609
5. 相似度：1.27% 标题：《分类算法总结_百度文库》
来源：http://wenku.baidu.com/link?url=x-ztH6pCPu_ASM3LlJ988H_rkDUPC3QHkwJntdG4V9dkyQ-OM9_5KA1q2TjCYFJmeemiXTXCy7QNEosTsbKl_P6x7Bkv427_YoEAULDqveK
6. 相似度：1.10% 标题：《分类算法应用场景实例二十则 - liulingyuan6的博客 - 博客频道 ...》
来源：http://blog.csdn.net/liulingyuan6/article/details/53637129
7. 相似度：0.97% 标题：《分类算法简介 - OPEN 开发经验库》
来源：http://www.open-open.com/lib/view/open1427016186870.html
8. 相似度：0.96% 标题：《tf-idf - 搜狗百科》
来源：http://baike.sogou.com/v73026138.htm?fromTitle=tf-idf

9. 相似度：0.95% 标题：《K最近邻(KNN,k-Nearest Neighbor)准确理解 - mousever的专栏 - ...》
来源：http://blog.csdn.net/mousever/article/details/51399070
10. 相似度：0.93% 标题：《YeahMobi:利用大数据提升原生广告的精准投放- 站长之家》
来源：http://www.chinaz.com/news/2015/0519/407389.shtml
11. 相似度：0.79% 标题：《KNN分类算法--python实现-布布扣-bubuko.com》
来源：http://www.bubuko.com/infodetail-526999.html
12. 相似度：0.74% 标题：《贝叶斯分类 - PhoenixZq - 博客园》
来源：http://www.cnblogs.com/phoenixzq/p/3539619.html
13. 相似度：0.70% 标题：《python中文分词:结巴分词 - C语言C++知识资源分享 - 时刻知(ShiKe...》
来源：http://www.shikezhi.com/html/2015/cbiancheng_0830/267818.html
14. 相似度：0.70% 标题：《中文分词开源项目 " 结巴 " -分析与讨论_人工智能吧_百度贴吧》
来源：http://tieba.baidu.com/p/3145509319
15. 相似度：0.64% 标题：《广告精准投放这件事 他们是怎么做的_网易科技》
来源：http://tech.163.com/15/1231/07/BC59JA7200094P40.html
16. 相似度：0.59% 标题：《机器学习之二:K-近邻(KNN)算法 - 一双拖鞋走天下 - 博客园》
来源：http://www.cnblogs.com/chensheng-zhou/p/4900123.html
17. 相似度：0.54% 标题：《2016年12月 – IT熊SEO》
来源：http://www.seozhashen.cn/?m=201612
18. 相似度：0.54% 标题：《关键词权重计算算法：TF-IDF - kalor - 博客园》
来源：http://www.cnblogs.com/likai198981/p/3344060.html
19. 相似度：0.50% 标题：《一种改进的DDAGSVM多类分类方法_熊忠阳_百度文库》
来源：http://wenku.baidu.com/link?url=dV5EPITCZq3RKxuwY5utwS-L2d5H5NbNVs2o3V9HAJQBuMhGaApdEyePg6eJzGxX5A5HxSI_OLjlN8K5pfg2jEJCJuvGo_8NteD6pZ1izl
20. 相似度：0.50% 标题：《第5章 支持向量机及其学习算法-2016_图文_百度文库》
来源：
http://wenku.baidu.com/link?url=0mzCi3LjJkcsQKErN8gxvyVTPCeRpvFDG0z5GJU_iZhkGz7BW1-QuZlowVFS3RjhAcRrx4dxvsQnmeMabufBiouI_nM7iEO5JKCnK3kxnpy
21. 相似度：0.46% 标题：《TF-IDF - 博客频道 - CSDN.NET》
来源：http://m.blog.csdn.net/article/details?id=49425201
22. 相似度：0.46% 标题：《特征值提取之 -- TF-IDF值的简单介绍 - Xs酱~ - 博客园》
来源：http://www.cnblogs.com/rausen/p/4142838.html
23. 相似度：0.45% 标题：《二维空间点索引数据结构 - New Day New Plan - 博客频道 ...》
来源：http://blog.csdn.net/dingyaguang117/article/details/7323171
24. 相似度：0.45% 标题：《KD-Tree源代码 - 下载频道 - CSDN.NET》
来源：http://download.csdn.net/detail/wangjiannuaa/2670031
25. 相似度：0.44% 标题：《关键字提取算法之TF-IDF扫盲{转载}_赵延宾_新浪博客》
来源：http://blog.sina.com.cn/s/blog_75a4f95a0101ead2.html
26. 相似度：0.44% 标题：《贝叶斯文本分类器 - 行走的逗比 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/b11040805/article/details/39085957
27. 相似度：0.44% 标题：《TF-IDF简介 - Solidfish的专栏 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/ididcan/article/details/6657977
28. 相似度：0.43% 标题：《学习笔记(1)-数据挖掘及其应用浅谈 - 锦年的博客 - 博客频道 ...》
来源：http://blog.csdn.net/qjc937044867/article/details/50273625
29. 相似度：0.39% 标题：《分类算法简介 - jediael_lu的专栏 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/jediael_lu/article/details/44152293
30. 相似度：0.38% 标题：《Libsvm和Liblinear的使用经验谈 - 止战 - 博客园》
来源：http://www.cnblogs.com/zhizhan/p/5001689.html
31. 相似度：0.38% 标题：《基于朴素贝叶斯的兴趣分类 - Totoro1745的博客 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/Totoro1745/article/details/68958044?locationNum=13&fps=1
32. 相似度：0.36% 标题：《SVM实际上是一种两分类算法,请问:如何用它来解决多类分类问题?》
来源：http://expert.sgst.cn/questionDetail.do?id=48911
33. 相似度：0.31% 标题：《KNN算法实现及其交叉验证 - 简书》
来源：http://www.jianshu.com/p/48d391dab189
34. 相似度：0.29% 标题：《机器学习经典算法详解及Python实现--K近邻(KNN)算法》
来源：http://blog.csdn.net/suipingsp/article/details/41964713

35. 相似度：0.29% 标题：《文本分类,数据挖掘和机器学习-nese-ChinaUnix博客》
来源：http://blog.chinaunix.net/uid-446337-id-94440.html
36. 相似度：0.25% 标题：《聚类算法和分类算法总结 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/a1061747415/article/details/48634395
37. 相似度：0.23% 标题：《TF-IDF - 牧马人夏峥 - 博客园》
来源：http://www.cnblogs.com/573177885qq/p/4511837.html
38. 相似度：0.23% 标题：《TF-IDF及其算法- 一座青山的专栏- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/sangyongjia/article/details/52440063
39. 相似度：0.22% 标题：《基于Spark的大数据精准营销中搜狗搜索引擎的用户画像挖掘 - ~ ...》
来源：http://www.baidu.com/link?url=KI6XlIkU2wEopjDAJyq5tZklqLvFdRq2sVsOmi5mu-E-m1FdeS8UCkymWcqJ_2veBeZumWaaiW5q71-2LfC1CCM2ZTtoQdw1PQKYqvZaMGq
40. 相似度：0.22% 标题：《Bag of words model (词袋模型)转 - 有何不可的日志 - 网易博客》
来源：http://blog.163.com/mageng11%40126/blog/static/140808374201181810936827/
41. 相似度：0.21% 标题：《python 中文分词:结巴分词-布布扣-bubuko.com》
来源：http://www.bubuko.com/infodetail-584132.html
42. 相似度：0.20% 标题：《基于树型结构的SVM多类组合分类器在文本分类中的应用_百度文库》
来源：
http://wenku.baidu.com/link?url=R5ja3vlEdjZh_IV6q4OtpbELMsJDz2qgMDD6KgzIw1g0mdEUa2_E5AC
43. 相似度：0.19% 标题：《支持向量机中的战斗机 — One class SVM-一起大数据》
来源：
http://www.17bigdata.com/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA%E4%B8
one-class-svm.html
44. 相似度：0.18% 标题：《...--朴素贝叶斯分类及其在文本分类、垃圾邮件检测中的应用 - ...》
来源：http://blog.csdn.net/suipingsp/article/details/41897901
45. 相似度：0.18% 标题：《第二学历_百度百科》
来源：http://baike.baidu.com/item/%E7%AC%AC%E4%BA%8C%E5%AD%A6%E5%8E%86
46. 相似度：0.18% 标题：《LSI模型的文本分类_百度文库》
来源：http://wenku.baidu.com/link?url=uWk3uGvUmXiR998CaFBLg21vmziTvWt-J-BxPFXBT1qRV4Q0-hb9AyvJ1F9vZWGToIR6ihsa9lNuGMXgEizerowtmwZsGdwEu_YidodGskm
47. 相似度：0.17% 标题：《文本表示+向量空间模型_SUPERTRAMP_新浪博客》
来源：http://blog.sina.com.cn/s/blog_795b865e0102v984.html
48. 相似度：0.14% 标题：《中文分词 - HackerVirus - 博客园》
来源：http://www.cnblogs.com/Leo_wl/p/4793351.html
49. 相似度：0.14% 标题：《中文分词的基本原理以及jieba分词的用法 - JohnSon - 博客 ...》
来源：http://blog.csdn.net/john_xyz/article/details/54645527
50. 相似度：0.14% 标题：《机器学习算法之朴素贝叶斯(Naive Bayes)--第二篇》
来源：http://m.blog.csdn.net/article/details?id=51264829
51. 相似度：0.13% 标题：《文本分类算法的究与改进.pdf文档全文免费阅读、在线看》
来源：http://max.book118.com/html/2015/1230/32342288.shtm
52. 相似度：0.13% 标题：《中文文本分类_图文_百度文库》
来源：http://wenku.baidu.com/view/db1e763a866fb84ae45c8d9c.html
53. 相似度：0.13% 标题：《支持向量机（SVM：support vector machine） - AI-ML-DL ...》
来源：http://www.cnblogs.com/taojake-ML/p/6129056.html
54. 相似度：0.13% 标题：《一种改进的支持向量机多类分类方法_百度文库》
来源：http://wenku.baidu.com/view/d079aca0284ac850ad0242ba.html
55. 相似度：0.12% 标题：《支持向量机SVM核函数的选择(七) - BigBzheng的博客 - 博客频道 - ...》
来源：http://blog.csdn.net/bigbzheng/article/details/51055374
56. 相似度：0.12% 标题：《主要分类方法介绍_百度文库》
来源：http://wenku.baidu.com/view/c7fe30edbceb19e8b8f6ba96.html
57. 相似度：0.12% 标题：《Kd-Tree算法原理 - lysuns - 博客园》
来源：http://www.cnblogs.com/lysuns/articles/4710712.html
58. 相似度：0.11% 标题：《朴素贝叶斯分类 - ACdreamer - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/acdreamers/article/details/44662347
59. 相似度：0.11% 标题：《stick-learn朴素贝叶斯的三个常用模型:高斯、多项式、....》
来源：http://www.cnblogs.com/Scorpio989/p/4760281.html

60. 相似度：0.10% 标题：《R语言与朴素贝叶斯算法 - R中国用户组-炼数成金-Dataguru专业数据...》
来源：http://www.dataguru.cn/thread-563060-1-1.html
61. 相似度：0.09% 标题：《KNN算法简述 - Lx85416281的专栏 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/lx85416281/article/details/40656877
62. 相似度：0.09% 标题：《Spark技术在京东智能供应链预测的应用》
来源：http://mt.sohu.com/it/d20170330/131159693_472869.shtml
63. 相似度：0.08% 标题：《分类器概述(ZT)_天下有雪_新浪博客》
来源：http://blog.sina.com.cn/s/blog_4511c21f0100grog.html
64. 相似度：0.08% 标题：《文本特征加权方法TF·IDF的分析与改进_图文_百度文库》
来源：http://wenku.baidu.com/view/4b6c7dd3240c844769eaee0f.html
65. 相似度：0.07% 标题：《核函数 - luyafei_89430的专栏 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/luyafei_89430/article/details/7629906
66. 相似度：0.07% 标题：《数据挖掘分类方法小结_fresley_新浪博客》
来源：http://blog.sina.com.cn/s/blog_a6696adf0101d0jb.html
67. 相似度：0.07% 标题：《Python中文分词工具之结巴分词用法实例总结【经典案例..._脚本之家》
来源：http://www.jb51.net/article/111244.htm
68. 相似度：0.06% 标题：《文本分类_百度百科》
来源：http://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB
69. 相似度：0.06% 标题：《TF-IDF权重-学术百科-知网空间》
来源：http://wiki.cnki.com.cn/HotWord/3088088.htm
70. 相似度：0.06% 标题：《稳定的特征选择算法分析.pdf》
来源：http://max.book118.com/html/2017/0509/105476774.shtm
71. 相似度：0.06% 标题：《中文维基百科的结构化信息抽取和词语相关度计算.pdf文档全文免费...》
来源：http://max.book118.com/html/2016/0103/32657564.shtm
72. 相似度：0.06% 标题：《分类算综述 - 道客巴巴》
来源：http://www.doc88.com/p-1837609987315.html
73. 相似度：0.05% 标题：《最近邻查找算法kd-tree - 皮皮blog - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/pipisorry/article/details/52186307
74. 相似度：0.05% 标题：《一种基于类平均相似度的文本分类算法_谭学清_图文_百度文库》
来源：http://wenku.baidu.com/view/94ce855b76eeaeaad1f330be.html
75. 相似度：0.05% 标题：《文本自动分类技术研究和实现.pdf文档全文免费阅读、在线看》
来源：http://max.book118.com/html/2015/1124/30063402.shtm
76. 相似度：0.05% 标题：《基于云模型的文本特征自动提取算法_图文_百度文库》
来源：http://wenku.baidu.com/view/1eebdd843c1ec5da51e270b1.html
77. 相似度：0.04% 标题：《分词技术》
来源：http://www.mamicode.com/info-detail-1188448.html

全文简明报告

本 科 毕 业 设 计

院 系 计算机科学与技术系

专 业 计算机科学与技术

题 目基于搜索关键词的用户属性分析预测

年 级 2013级 学 号 131220167

学生姓名 禤宝琼

指导老师 黄宜华 职 称 教授

论文提交日期

南京大学本科生毕业论文(设计、作品)中文摘要

题目: 基于搜索关键词的用户属性分析预测

计算机科学与技术 院系 计算机科学与技术 专业 2013 级本科生姓名: 禤宝琼

指导教师(姓名、职称): 黄宜华、教授

摘要:

在广告的精准投放中,根据用户的历史行为来反推用户的属性是一项基础技术。

用户在搜索引擎中查询的内容与用户的性别、年龄、学历等有着密切的关系。例如人群男性在军事、汽车主题上有更多的搜索行为,19~23岁搜索行为中较多与大学生活、社交生活有关,高学历人群更倾向于获取社会、经济方面的信息。

{ 75% : 本次研究以用户历史的查询关键词与用户的人口属性标签(性别、年龄、学历)做为训练数据集,利用搜索关键词与用户属性的关联性,通过机器学习、数据挖掘技术构建分类算法来对新增搜索用户的人口属性进行判定。 }

关键词:文本分类;朴素贝叶斯;支持向量机;用户画像

南京大学本科生毕业论文(设计、作品)英文摘要

THESIS: Analysis and Prediction of User Attributes based on User's Searching Keywords

DEPARTMENT:Computer Science and Technology

SPECIALIZATION:Computer Science and Technology

UNDERGRADUATE:BaoQiong Xuan

MENTOR:YiHua Huang

ABSTRACT:

In the accurate delivery of advertising, it is a basic technology to derive user's attributes according to the user's historical behavior.

The content that the user inquires in the search engine is closely related to the user's sex, age, educational background and so on.For example, crowds of men have more searches for military and car themes. Youth's search behavior is more related to college life and social life. Highly educated people tend to gain social and economic information.

In this study, user history query key words and user demographic attribute tags (gender, age, educational background) are used as training data sets. By using the relevance of search key and user attributes, a classification algorithm is constructed by machine learning and data mining technology to determine the population attributes of new search users.

KEY WORDS: classification; Naive Bayesian; SVM; KNN; feature selection; User Profile

目录

第1章 绪论

1.1 研究背景和意义

{81%：2000年后随着国内互联网兴起,广告营销手段也从传统时代逐步进化到了互联网时代,广告展示与搜索模式从内容与创意层面到技术层面进行了深度更迭。 }{89%：互联网发展到现在,网络广告再也不只是传统意义上的“广而告之”,而是有针对性的“有的放矢”。 }{84%：尤其在信息过载现在,狂轰滥炸式的广告使得用户体验极差,非常容易引起用户的反感,需要真正将用户对广告的反感度降到最低,那就必须走把广告投放给真正需要它的人这条路,就是所谓广告的“精准投放”。 }{89%：在大数据应用场景下,广告的精准投放对广告主、服务平台与潜在用户而言,在效率提升与商业效益方面,有着更高、更迫切的要求。 }

广告精准投放中,{ 66%：如何找到真正需要广告的人是个关键问题。 }{ 59%：每个人在现实生活当中都是一个有姓名、有性别、有年龄、有学历、有喜好、有经历、有住址的人,人们通过这些特征与他人区分开来。 }{95%：而在网络世界中其实也可以通过各种标签勾勒出与现实生活一一对应的虚拟用户,这样我们就可以在海量互联网信息中准确的找到广告投放对象。 }这就是所谓的用户画像。

用户画像是由各种用户属性组成的,用户画像挖掘技术中,根据用户的浏览、搜索等行为来反推获取用户属性是一项非常基础且非常重要的技术。本文将根据用户搜索行为与用户属性的相关性对用户属性进行分析预测。

用户在搜索引擎中查询的内容与用户的性别、年龄、学历等有着密切的关系,例如人群男性在军事、汽车主题上有更多的搜索行为,19~23岁搜索行为中较多与大学生活、社交生活有关,高学历人群更倾向于获取社会、经济方面的信息。{ 74%：本次研究以用户历史一个月的查询关键词与用户的人口属性标签(性别、年龄、学历)做为训练数据集,利用搜索关键词与用户属性的关联性,通过机器学习、数据挖掘技术构建分类算法来对新增搜索用户的人口属性进行分析预测。 }

1.2 研究现状

本文中对用户属性分析预测的根据是用户历史一个月的搜索关键词,也就是根据用户的历史搜索行为对用户进行分类,属于分类预测的范畴。{ 75%：分类预测是数据挖掘的一个重要手段, }也是对各种类型的数据进行分

析分类的一个关键工具,它广泛应用于统计学、图像处理、医疗诊断、信息检索、机器学习等多个领域。进一步来说,本数据集中用户的历史搜索行为是由用户过去一个月在搜索引擎中输入的关键词组成的文本,本文的分类预测任务本质上是一个文本分类的任务,{ 55% : 那么现在主要的文本分类算法包括贝叶斯算法,支持向量机算法,决策树, }k-近邻,基于神经网络的方法等。

{92% : 贝叶斯(Bayes)分类算法主要利用Bayes定理来预测一个未知类别的样本属于各个类别的可能性,选择其中可能性最大的一个类别作为该样本的最终类别。 }{ 70% : 支持向量机(SVM,Support Vector Machine)算法根据区域中的样本计算该区域的决策曲面,由此确定该区域中未知样本的类别。 }{88% : 决策树学习着眼于从一组无次序、无规则的实例中推理出以决策树表示的分类规则, }{100% : 构造决策树的目的是找出属性和类别间的关系,用它来预测将来未知类别的记录的类别。 }{99% : k-近邻(kNN,k-Nearest Neighbors)算法是一种基于实例的分类方法, }{86% : 该算法找出与未知样本距离最近的k个训练样本,看这k个样本中多数属于哪一类,就把该未知样本归为那一类。 }{91% : 神经网络是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型,在这种模型中,大量的节点(或称”神经元”,或”单元”)之间相互联接构成网络,即”神经网络”,以达到处理信息的目的,神经网络通常需要进行训练,训练的过程就是网络进行学习的过程,训练改变了网络节点的连接权的值使其具有分类的功能,经过训练的网络就可用于对象的识别及分类。 }

通过观察本次实验的训练数据集,我们发现一个重要的问题是类别样本分布不均匀以及存在数据缺失现象。如在年龄类别分布中,随着年龄的增加样本数据量逐渐甚至急剧减少,0-18岁有7900个样本,19-23岁有5330个样本,24-30岁有3603个样本,31-40岁有2141个样本,41-50有589个样本,51-999岁仅有82个样本。对于样本分布不均匀的情况,文献[2]中介绍了一分类支持向量机(One-class SVM)的方法,该算法构造一个高维超球面,把一类数据包起来,那么当新的数据来的时候它能判断新数据属于这类数据还是不属于这类数据,适用于有两种类型样本,但其中一类型样本数目缺失或远少于另一类型样本数目,但推而广之,也能用于多种类别的场景。

1.3 本文研究内容

{ 79% : 本次研究以用户历史一个月的查询关键词文本与用户的人口属性标签(性别、年龄、学历)做为训练数据集, }用户人口属性标签包括性别、年龄、学历,其中,性别包括有男、女2类标签,年龄包括有0-18岁、19-23岁、24-30岁、31-40岁、41-50岁、51-999岁6类标签,{ 56% : 学历包括有小学、初中、高中、大学、硕士、博士6类标签。 }

这里,对数据集惊醒了训练集和测试集的划分,也就是采用交叉验证的方法。

{ 72% : 利用向量空间模型对文本做数值化的表示, }{ 66% : 向量空间模型的基本理念是把文本化为以特征项的权重作为分量的向量表示:其中为第i个特征项的权重,以词作为特征项,本文选取词TF-IDF值表示权重。 }为此,首先对用户的搜索关键词文本做了分词、统计词频等处理。

再而对特征做降维处理,原始的特征向量高达数十万维,极大地降低了分类的效率和准确度。{ 56% : 本文分别采取了文档频率(DF)、CHI统计、信息增益三种特征选择方法, }选择保留更具代表性的特征,去除其他特征。

最后利用朴素贝叶斯、支持向量机、KNN等算法构造相应的分类模型,对测试数据进行分析预测,并做参数的调优等改进,对各分类模型的准确度、时间效率等做比较。

1.4 论文结构安排

第一章为绪论,主要介绍了在现代精准广告投放的背景下用户画像挖掘工作的背意义。同时简要地对本文所属的文本分类研究范畴所用的基本方法进行了描述,并介绍了其核心部分即分类分析的经典算法和研究现状。

第二章简单地介绍了与本文相关的工作背景和概念定义。{ 62% : 首先介绍了基于向量空间模型的文本表示方法。 }然后对降低特征向量空间维度的特征选择方法做了介绍。并分别对基于概率统计的,基于平面划分的,基于距离的、基于神经网络的分类算法进行了较为详细的介绍。

第三章将本文根据用户搜索关键词预测分析用户属性的方案分成了3部分进行详细描述,即中文分词、特征选择、基于不同分类算法构造不同的分类模型。并分析了相关部分的缺失数据处理,最后给出了该算法的整体实现流程图。

{ 56% : 第四章对实验结果进行分析,介绍了实验的相关环境和数据集情况, }统计了测试结果,对不同分类模型的预测准确度和时间成本做了对比。

第五章总体评价了本文主要的研究内容。对本文提出的解决方案的优劣进行了评价,并提出进行优化的可能方案,做出了下一步努力方向的建议。

第2章 相关工作与背景介绍

2.1 文本表示

未经处理的原始文本不能直接作为文本分类算法的输入,因此需要对文本进行预处理。{ 65% : 文本预处理阶段的目的是将文本进行数字化转化为计算机可识别的信息,即对文本进行数字化的表示。 }

2.1.1 向量空间模型

{ 58% : 本章节介绍利用向量空间模型对文本做数值化的表示。 }

{ 63% : 向量空间模型以统计学习理论和结构最小原理为基础, }将文本量化为一组线性无关的向量。向量空间模型的两个主要成分是特征项与特征权重。特征项是指构成文档的基本语言单位,主要是词或短语。特征权重是指赋予给文档中每个特征项的权重,用于体现这个特征项在文档中的重要程度,权重越大越能代表该文档[4]。

根据特征项和特征权重,{ 58% : 文档可以被表示为:其中为第个特征项的权重。 }

那么,{ 60% : 在向量空间模型中, 文本可以以多个词权重组成的向量表示,词也可以由其在多篇文本的权重组成的向量表示, } 这种对偶关系能在文本向量矩阵中得到很好的体现,在文本向量矩阵中每列代表一个词,每行代表一篇文本[4]:

在上列矩阵中,{85% : 表示第i篇文档中的第j个特征项(词)的权重, }权重的取值依据可以是词在该篇文档中的出现次数,也可以是词在该篇文档中的出现频率,本文将采用tf-idf方法对词进行加权。接下来的章节将对tf-idf加权方法进行解释介绍。

2.1.2 TF-IDF

{83% : TF-IDF是一种统计方法,用于评估一个词在一份文档中的重要程度, }这个重要程度是相对该文档所在文档集或语料库中的其他文档而言的。{99% : TF-IDF的主要思想是如果某个词或短语在一篇文章中出现的频率TF高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。 }在这种方法中,{95% : 一个词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。 }

{97% : TF-IDF实际上是:TF * IDF,TF是词频(Term Frequency),IDF是逆向文件频率(Inverse Document Frequency)。 }下面给出对于文档中某个词的TF*IDF权重的计算公式步骤:

步骤1:计算

(1)

{99% : 以上式子中 是该词在文件中的出现次数,而分母则是在文件中所有字词的出现次数之和。 }

步骤2: 计算

(2)

{ 66% : 其中 |D| 表示文件集中的文件总数, }分子表示有多少文档包含词。

步骤3:最后,将(1) (2)两式计算出的结果相乘得到

(3)

由以上公式及步骤可以看出,对于某个特定文档,若文档中的某个词在该文档中的出现频率高,{ 64% : 且该词在整个文件集中的出现频率低,则可以得到较高的TF-IDF权重。 }由此可以看出,{100% : TF-IDF倾向于过滤掉常见的词语,保留重要的词语。 }

2.2 特征选择

{95% : 文本自动分类的困难之一是特征空间的高维性和文本表示向量的稀疏性。 }{ 63% : 寻求一种有效的特征选择方法, 降低特征空间的维数, 提高分类的效率和精度, 成为文本自动分类中需要首先面对的重要问题[13]。 }

{94% : 特征选择,总是在将特征的重要程度量化之后再进行选择, 如何量化特征的重要性,就成了各种方法间最大的不同。 }{80% : 常见的特征选择方法有文档频率(DF)、信息增益 (IG)、CHI统计等[11]。 }下面对这些方法做简单的介绍。

(1)文档频率

{ 70% : 词条的文档频率(Document Frequency) 是指在训练语料中出现该词条的文档数。 }{98% : 文档频率方法提取文档频率较高的特征, 它的目的是去掉在训练集上出现次数过少的特征, 保留具有一定影响力的特征。 }{88% : 在各个特征提取方法中, DF 方法是最简单的[13]。 }

(2)互信息

互信息法用于衡量特征词为某文档类别直接带来的信息量。词对于类别的互信息值计算方式如下:

其中,{ 67% : 表示文档包含词时属于 类的条件概率,表示语料中某篇文档包含词的概率。 }

(3)信息增益

{ 61% : 信息增益度量一个特征为分类系统带来的信息量, }即这个词存在和不存在时系统的信息量差值。词t的信息增益值计算方式如下:

{ 72% : 其中,表示类文档在语料中出现的概率,表示语料中包含词的文档的概率,表示文档包含词时属于类的条件概率,表示语料中不包含词时属于的条件概率, }表示类别数。

(4)CHI 统计,即卡方校验

{90% : CHI 统计方法度量词t和文档类别c之间的相关程度。 }{85% : 词t对于c的CHI值计算方式如下: }

其中,{95% : N表示训练语料中的文档总数, c为某一特定类别, t表示特定的词条, A表示属于c类且包含t的文档频数, B表示不属于c类包含t的文档频数。 }

2.3 分类算法

本文根据用户历史搜索关键词组成的文本,对用户属性标签进行判断,也就是根据用户搜索文本对用户进行分类。因此,本文所要解决的问题,属于文本分类的范畴。下面介绍几类用于文本挖掘、文本分类的经典算法。

2.3.1 基于贝叶斯决策理论的分类算法

{ 72% : 利用概率统计知识进行分类的算法有很多种, }其中使用范围最广的是基于贝叶斯决策理论的分类算法,{ 58% : 包括朴素贝叶斯、贝叶斯网络等。 }由于在本文中主要采用朴素贝叶斯算法进行分类模型的构建,本节着重对朴素贝叶斯算法进行介绍。

朴素贝叶斯算法的概率统计理论基石是贝叶斯定理。下面直接给出贝叶斯定理:

{80% : 其中,表示在事件A已经发生的条件下事件B发生的概率, }{ 75% : 表示事件B已经发生的前提下事件A发生的概率,、分别表示事件A、B发生的概率。 }

基于上述定理,{ 78% : 朴素贝叶斯分类器的主要思想是:对于给出的待分类项,求解在此项出现的条件下各个类别出现的概率,找出其中出现概率最大的类别,就认为此待分类项属于该类别。 }下面给出朴素贝叶斯的标准分类过程:

{84% : 1. 设为一个待分类项,每个为的一个特征属性。 }

2. 有类别集合。

3. 计算, , …, 。

4. 如果,则属于类别。

其中,第3步中的 ,{ 58% : 由于对于所有类别为常数,我们只需比较分子大小, }分子大小由下列公式求得:

那么第4步最后转化为,求

2.3.2 基于超平面划分的分类算法

{ 57% : 空间的点可以被一个超平面划分为两部分, }在分类问题中,可以寻找一个超平面将两类线性可分的样本分割开来,那么,对于未知的待定项,看它与这个超平面的位置关系就可以判断出它属于哪一类。由这个数据分类问题的处理理论出发,发展出了许多基于超平面划分的分类算法,{ 78% : 主要包括感知机、支持向量机(SVM,Support Vector Machine)等。 }{ 57% : 由于本文主要利用支持向量机构造分类模型, }本节着重对支持向量机理论进行介绍。

{86% : 支持向量机方法的机理可以简单地描述为:寻找一个满足分类要求的最优分类超平面,使得超平面在保证分类精度的同时,能够使分类间隔(Margin)最大化,使得支持向量机能够实现对线性可分数据的最优分类[6]。 }{ 69% : 分类间隔是指两类中离分类超平面最近的样本且平行于分类超平面的两个超平面间的距离。 }

下图1-1中红叉与蓝圈是两类样本,从左至右是三种超平面分割方法。如图1-2所示可以直观地看出最右边的划分方式能够使超平面两侧的空白区域最大化,即分类间隔最大,容忍误差的能力最强,即支持向量机方法所要找到的最优分类平面。而感知机算法是找到随意一个能分割两类样本的超平面即可。

图1-1 三种超平面分割方式

图1-2 超平面两侧空白区域

{ 56% : 对于线性可分的分类问题,设训练集为 , } 其中, 设超平面,则训练集对超平面的集合间隔为。{ 60% : 寻找最大间隔的超平面问题可描述为如下最优化问题: }

{ 77% : 对于线性不可分的问题,支持向量机通过引入核函数,可以在一个高维的空间中来实现对于原空间为非线性的算法。 }支持向量机核函数就是某个高维空间的内积,{ 60% : 其在支持向量机中起着至关重要的作用[6]。 }

2.3.3 基于距离的分类算法

{100% : K最近邻(k-Nearest Neighbor,KNN)分类算法, }是基于距离的分类算法中的典型代表,是本文采用的主要算法之一,因此接下来着重介绍KNN算法。

并且,{ 63% : KNN是向量空间模型中最好的文本分类算法之一[8]。 }{88% : 该方法的思路是:如果一个样本在特征空间中的k个最相似的样本中的大多数属于某一个类别,则该样本也属于这个类别。 }

{ 75% : KNN通过计算新数据与训练数据特征值之间的距离,然后选取 K(K1)个距离最近(即特征空间中最邻近)的邻居进行分类判断(投票法)。 }下图2是KNN对新数据进行分类的过程示意图,如图所示,样本空间有三类样本,在计算出新数据与其他训练数据的距离后,选取出了5(K为5)个距离最近的样本,即黑色箭头指向的5个样本,统计出这5个样本中出现次数最多的类别是类,最后判定新数据属于类。

图2 KNN分类过程示意图

{ 57% : 关于K值的选择,值对文本的分类效果有很大影响 , }{ 57% : 但目前没有很好的确 定值的方法,通常是根据试验测试的结果调整值的大小[15]。 }

关于样本间距离的计算,{ 71% : 由于本文采用了向量空间模型来表示文本, }我们只要计算文本向量间的距离即可。在KNN算法中,常用的距离有两种,分别为曼哈顿距离、欧式距离。

曼哈顿距离公式为:

D

欧式距离公式为:

D

2.4 小结

分类模型的构造是本文的关键,{ 62% : 对于上面介绍的几种基本分类方法, }其各有各的优劣。{100% : 朴素贝叶斯模型发源于古典数学理论, }有稳定的分类效率,算法也比较简单,常用于文本分类,{96% : 但贝叶斯定理的成立本身需要一个很强的条件独立性假设前提,而此假设在实际情况中经常是不成立的,因而其分类准确性就会下降, }{97% : 而且需要知道先验概率且先验概率很多时候取决于假设,假设的模型可以有很多种,因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。 }k-近邻方法简单,易于理解,易于实现,无需估计参数,{99% : 该算法在分类时有个主要的不足是,当样本不平衡时,如一个类的样本容量很大,而其他类样本容量很小时,有可能导致当输入一个新样本时,该样本的K个邻居中大容量类的样本占多数, }{94% : 此外其是一种懒惰学习方法,它存放样本,直到需要分类时才进行分类,如果样本集比较复杂,可能会导致很大的计算开销,因此无法应用到实时性很强的场合。 }{95% : SVM最大特点是根据结构风险最小化准则,以最大化分类间隔构造最优分类超平面来提高学习机的泛化能力,较好地解决了非线性、高维数、局部极小点等问题, }但本质上是二分类算法,处理多分类问题效率会有所下降。

本章先介绍了向量空间模型,并介绍了用于降低特征空间的维数的一些常见的特征选择方法,接着分别介绍了基于贝叶斯理论的、基于超平面划分的和基于距离的分类算法的主要思想、代表算法及其优劣,为下文描述的根据用户搜索文本预测用户属性标签的方法提供理论依据。

第3章 具体实现方案

3.1 方案概览

本章将介绍根据用户搜索关键词预测分析用户属性的方法及其具体的实现细节。由于本文的提出的解决方法主要分成三步,所以下面对于具体方案的介绍也主要分成三部分进行描述。

首先我们对每个用户初始的搜索关键词文本数据集进行处理,{ 58% : 利用优秀的分词工具结巴分词对文本进行分词, }并针对数据集文本的特点引入了相应的网络词汇,以及做去停用词处理。接着对分词后的文本做向量化处理,这里以词作为特征项,选取词的TF-IDF表示权重,{ 57% : 并对文本特征向量做了降维处理, }剔除了出现频率极低的词汇,最终保留了十万左右的特征项。最后分别训练用于预测年龄、性别、学历的三个分类器,对于每个分类器的分类算法选择,本文分别试验了朴素贝叶斯、SVM、KNN算法,并对比分析了它们在这个分类任务上的效果。

此外,由于本文的使用的数据集中还存在缺失的用户标签数据,本文分别采用的处理方法是根据其他非缺失数据训练出一个分类模型,对这些标签缺失数据文本进行分类,再将这些数据纳入训练数据集中重新训练出新的分类模型。

通过以上的简单描述,我们知道本文关于基于用户搜索关键词预测用户属性的主线工作有:

搜索文本预处理 -^ 特征抽取和转换 -^ 训练各种分类模型 -^ 测试 -^ 实验结果分析。

3.2 中文分词

{ 68% : 中文分词是中文文本处理的一个基础性工作,本文利用优秀的开源中文分词工具结巴分词对用户搜索文本进行中文分词, }{93% : 其基本分词原理有三点:基于Trie树结构实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)、采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合、对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。 }并且,本文还针对数据集文本的特点引入了相应的网络词汇,以及做去停用词处理。

根据切分出来的词汇数量质量等的差异,分成了不同的分词模式,常见的分词模式分两种,{87% : 一种是全模式,即把句子中所有的可以成词的词语都扫描出来, }{ 57% : 另一种是精确模式,试图将句子最精确地切开使得分词结果更符合原始语义。 }下面举例说明两种分词模式的差异:对句子“我来到北京清华大学”进行分词操作,{ 76% : 全模式下的结果是“我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学”,精确模式下的结果是“我/ 来到/ 北京/ 清华大学”。 }

两种模式各有优劣,对于本文的数据集而言,全模式的分词一定程度上对样本数据做了补充增添以及不用担心有些句子中的关键短词没有被识别出来,但是也存在增添了很多垃圾无意义词汇的缺点,而精确模式下虽然有时一些重要短语没被识别出来,但却不会对训练过程造成额外的“噪音污染”。本文通过实验发现,在本文的任务中, 使用精确模式的分类器的准确率比使用全模式的高1%。

3.3 特征工程

分词过后的文本仍然不能作为分类算法的输入,需要对其做进一步的特征处理。

{ 59% : 在第2章向量空间模型中提到过, }根据特征项和特征权重,{ 60% : 文档可以被向量化为:其中为第i个特征项的权重。 }{ 62% : 一般选取词作为特征项,常用特征项权重有多种, }包括出现次数、在文档中出现频率tf以及tf-idf等。本文分别采用tf和tf-idf方法对特征项进行加权并对两者做对比,tf-idf方法在第2.1.2节中已经做过详细介绍,在此不累赘描述。

经过以上处理之后,我们发现文本特征向量的维度高达三十多万,极大影响了分类的效率和准确率,因此我们必须采取降维措施,也就是进行特征选择。

本文第2.2小节中介绍了常见的特征选择方式,{ 65% : 包括基于文档频率(DF)的、基于信息增益(IG)和CHI统计等。 }本文通过实验发现,在本文的任务中,{86% : 虽然在各个特征提取方法中, DF 方法是最简单的, }但是DF仅考虑了频率因素而没有考虑类别因素,导致DF算法非常容易引入一些高频却没有意义的词。如在本文的数据集中,“图片”、“下载”,“电影”等,DF值排名前列,然而,在各类别中都是高频词,对分类并没有多大的意义。因此本文主要使用了信息增益(IG)、互信息(MI)、CHI统计这三种特征选择方法,并且,CHI统计有最佳的表现。

3.4 分类模型

前面已经叙述了对用户原始搜索文本进行分词处理和特征处理的方案,处理过后的用户搜索文本向量和用户标签已经可以作为分类器的输入了,本章将继续描述分类模型部分的工作。

本文的任务需要对用户的年龄、性别、学历进行分析预测,也就是给用户贴上相应的标签,其中,性别包括有男、女2类标签,年龄包括有0-18岁、19-23岁、24-30岁、31-40岁、41-50岁、51-999岁6类标签,{ 56% : 学历包括有小学、初中、高中、大学、硕士、博士6类标签。 }前文已经说过,这相当于对用户的搜索文本进行分类,属于哪个年龄类、哪个性别类、哪个学历类。

那么,针对这个情况,本文分别为这三种用户属性构造三个分类器,把数据集划分成训练集和测试机进行交叉验证,并对比分析基于不同的分类算法的分类模型在此任务上的效果差异。

3.4.1 基于朴素贝叶斯的分类模型

在本文第2.3.1小节中对朴素贝叶斯算法进行过描述,其理论基础是贝叶斯定理,{94% : 但贝叶斯定理的成立本身需要一个很强的条件独立性假设前提, }即特征属性互相独立。{ 57% : 为了将朴素贝叶斯算法应用到文本分类任务中, }{ 70% : 我们采用词袋(Bag of Words)模型, }{93% : 在这种模型中,文本(段落或者文档)被看作是无序的词汇集合,忽略语法、单词的顺序。 }

{ 56% : 假设为一个待分类项,类别集合, }从本文第2.3.1小节的推导中,我们知道对x的分类任务最后转化为求。那么,{ 61% : 如何求得成了朴素贝叶斯算法的关键问题, }不同的的计算方法区分了不同的朴素贝叶斯分类模型,{ 57% : 文本分类领域中常见的朴素贝叶斯模型有以下两种: }

(1)多项式模型(Multinomial Naive Bayes)

多项式朴素贝叶斯,会为类别集合训练出一个特征分布,该分布由每个类别向量组成,,其中,{ 67% : n是特征项总数(在本文任务中n去重后的所有单词的数量), }{ 56% : 是第项特征(第个词)出现在属于类的样本的概率, }即。的计算方法在下文会介绍。

多项式朴素贝叶斯在本文任务上的应用过程如下:

假设所预测的用户属性的标签类别集合为, m为类别总数,对于年龄属性,m=2,对于年龄和学历,m=6。

构造训练数据集,其中是类别的样本集合,,每个都是一个属于类别的用户搜索文本经过分词和特征处理后的向量表示,即。

训练过程中,对于集合C中的每个类别,假设其样本空间为,先计算,;再计算,由下列公式计算,也就是:

其中,, 是一个平滑参数,取值在[0, 1]区间,参数的设置是为了防止当时的情况,一旦,也等于0,干扰了最终概率的计算,同时alpha参数的设置不会引起额外的精度损失。

若文本向量的特征权重是词在文档中的出现次数,上式就可以理解为表示第个词在类的所有文档中共出现过多少次,{ 59% : 表示在类的所有文档中每个词出现次数的总和。 }实践证明,若特征权重选取tf-idf能达到相同甚至更优的表达效果。

测试过程,对于待分类项,计算它属于每个类别的概率并取值最大的作为预测结果:

其中,和已在训练过程中根据训练集数据算出,为该待测项的分类结果。

(2)伯努利模型(Bernoulli Naive Bayes)

伯努利模型与多项式模型最大的不同是,伯努利模型输入的样本向量中的每个元素只有两个值:1/0(出现/没出现)。也就是即使我们输入的文本向量是以tf-idf为权重的,伯努利模型依然只是根据该词tf-idf的值大于0还是等于0将其转化成1或0,表示这篇文档包含或不包含这个词。

伯努利贝叶斯在本文任务上的应用过程与多项式贝叶斯基本一致,除了在第3步中对的计算有所不同,伯努利贝叶斯的的计算如下:

的计算同多项式贝叶斯中,由于在伯努利模型中,只有0/1两种值,表示该词是否出现在某篇文档中,因此,实际上是类y下包含该词的文档总数;是类y的样本集合,是类y下的文档(样本)总数。这里,同样是一个为了避免为0的情况而设置的平滑参数,取值在[0, 1]区间。

关于两类朴素贝叶斯模型中alpha的取值,本文实验从0.01开始以0.01为步长试验不同的alpha值下观察分类模型的情况。

3.4.2 基于支持向量机的分类模型

在第2.3.2节中对支持向量机(SVM)算法的介绍中可以看出超平面的划分方法最初是为了解决两类分类问题,而在我们的任务中,除了性别只有男/女两类外,年龄和学历都分别有7类,属于多类分类问题。虽然SVM分类器的本质是两类分类器,{86% : 但当前已经有许多方法将SVM推广到多类分类问题, }这些方法大致分为两大类:

{ 57% : (1)构造并组合一系列的两类分类器来实现多分类器的构造。 }

{95% : (2)将多个分类面的参数求解合并到一个最优化问题中,通过求解该最优化问题“一次性”地实现多类分类。 }

{ 68% : 第二类方法这种方法看似简单,但其在参数求解过程中的变量远远多于第一类方法, }{92% : 计算复杂度比较高,实现起来比较困难, }{ 76% : 训练速度不及第一类方法,而且在分类精度上也没有更突出的表现。 }{100% : 当训练样本数非常大时,这一问题更加突出。 }因此,第一类方法最为常用,本文采用的也是第一类方法,通过One-against-Rest策略构造组合多个两类分类器,{ 56% : One-against-Rest策略依次用一个两类SVM分类器将每一类与其它所有类别区分开来,对于n类问题,将得到n个两类分类器。 }{100% : 分类时将未知样本分类为具有最大分类函数值的那类。 }

{ 58% : 本文第2.3.2节中介绍了对于线性不可分的问题,支持向量机通过引入核函数, }{ 66% : 可将样本从原始空间映射到一个更高维的特征空间, }{ 60% : 使得样本在这个特征空间内线性可分。 }{ 59% : 特征空间的好坏对支持向量机的性能至关重要, }因此核函数的选择是个我们构造SVM模型中的关键问题,若选择了不合适的核函数,则意味着将样本映射到了一个不合适的特征空间,从而导致构造出来的SVM模型性能不佳。

{ 59% : 常用的核函数有线性核、高斯核、多项式核等, }根据前人的经验,对文本数据通常采用线性核,情况不明时可先尝试高斯核。本文的训练数据是用户的搜索文本数据,因此我们采用的核函数是线性核。

3.4.3 基于K-最邻近算法的分类模型

在前面第2.3.3节中介绍过,{92% : KNN算法的基本思想是:如果一个样本在特征空间中的k个最相似的样本中的大多数属于某一个类别,则该样本也属于这个类别。 }{ 76% : KNN最大的缺点在于它一种懒惰学习方法,它存放样本,直到需要分类时才进行分类, }{ 55% : 需要对整个无序的训练集进行对比搜索, }而在本文的分类任务中,{ 64% : 向量维数高,训练样本集数量大,会导致很大的计算开销。 }

{ 65% : 为了解决上述问题,我们使用了改进的KNN分类模型: }基于KD-Tree的KNN文本分类算法,下面对其进行介绍。

{ 62% : KD-Tree指的是k维的二叉查找树, }{ 55% : 是一种分割k维数据空间的数据结构, }基于KD-Tree可实现对给定k维数据的快速最近邻查找。KD-Tree的每个节点代表 k 维空间的一个点并且树的每一层都根据这一层的分辨器做出分枝决策。第 层的分辨器定义为:。

KD-Tree的存储规则为:对第层的任意一个节点,{83% : 若它的左子树非空,则其左子树上所有节点的第维值均小于节点的第 维值; }{88% : 若它的右子树非空,则其右子树上 所有节点的第维值均大于节点的第维值; 并且它的左右子树也分别为KD-Tree。 }

{ 55% : 有了以上对KD-Tree的定义介绍, }下面给出算法步骤:

1. 建立一个空KD-Tree,将训练集中每个用户的搜索文本向量依次插入 KD-Tree 中。

2. 在测试集中取出一个待测文本向量,在 KD-Tree中搜索这个文本向量,得到祖先节点集。

{ 55% : 3. 依次计算待测文本向量与祖先节点的相似度, }{ 60% : 相似度最大的文本类型就是待测文本的文本类型。 }

4. 对测试集中的每个用户搜索文本,重复第2、3步,直至测试集中的每个用户相关属性都计算完毕。

以上算法步骤适用于每个用户属性(性别、年龄、学历),在此就不重复叙述了。

3.5 缺失数据处理

实验过程中,发现本文用到的数据集存在缺失的标签数据,有355个样本的年龄标签缺失,424个样本的性别标签缺失,1878个样本学历标签缺失。对于样本数据缺失的问题,本文尝试了两种处理方法。

一种做法是认为它是脏数据,认为这条用户数据记录是不可靠的,把它从类别样本数据集中剔除,这样做的好处是能够完全避免这些数据的干扰,但缺点也是明显的,由于这些缺失数据的数目并不在少数,一定程度上减少了用于训练的样本量,降低了分类模型可达到的最大精度。

另一种做法是先利用其他非缺失数据训练出一个分类模型,对这些缺失用户标签的样本进行分类预测,从而对缺失的标签数据进行填补,最后将填补后的数据样本纳入训练数据集中重新训练出新的分类模型。这样做的好处是能够利用起这些缺失的数据,扩大训练集规模,坏处是分类模型的预测准确度是有限的,填补的用户标签并非完全准确,会在一定程度上对最终分类模型的训练造成噪音干扰。

实验发现,对于性别属性而言忽略缺失标签的做法比填补缺失标签的做法有1%的增益,但对于年龄和学历而言,填补缺失标签的做法比忽略缺失标签的做法提高了1%的准确率。原因应该是,性别标签只有两种,样本数量也比较平衡,所以填补缺失标签增加样本的益处小于其噪音干扰带来的坏处。而年龄和学历各有6种标签,样本数目也非常不平衡,所以填补缺失标签增加样本的益处大于其噪音干扰的坏处。

3.6 分类系统总体结构

经过上文的介绍,本文基于用户搜索关键词的用户属性分析预测系统的总体结构就如下图所示:

图3-分类系统整体结构图

第4章 实验

4.1 实验环境

本次实验中的计算环境如下所示:

硬件环境:

CPU: 1.6 GHz Intel Core i5

内存: 8 GB 1600 MHz DDR3

SSD闪存空间: 256GB

软件环境:

操作系统:Mac OS 10.11.6

开发语言:python

开源工具:Scikit-learn, Liblinear, Libsvm, jieba

其中,jieba是之前介绍过的结巴分词工具;{ 55% : Scikit-learn是基于python的机器学习模块, }提供了许多有用的数据挖掘与分析工具;{ 65% : Libsvm和Liblinear都是国立台湾大学的Chih-Jen Lin博士开发的基于SVM的分类器,Liblinear是为大规模数据所设计的线性模型,而Libsvm主要用来解决通用典型的分类问题。 }

4.2 数据集

本次实验中所使用的数据集来源于搜狗,数据集共有20000条记录,其中每条记录都是由5个字段组成的,前四个字段是用户的信息,分别是用户的ID、年龄(Age)、性别(Gender)、学历(Education),第5个字段是该用户历史一个月的搜索记录(Query List),字段的详细情况见表1-1。示例如下:

C98C47F106 1 1 5 空格符号复制 临沂是哪里 lol李青 镇魂街拉棺图片纹身 视频合成app 长颜草

9D8C36B494 2 2 4 快速选中不连续的段落并为其套用样式 微微一笑很倾城 你是我学生又怎样

表1 字段详细情况

经过观察发现,数据集中用户各个属性标签的样本分布不大平衡。在性别属性中,男性样本有11365个,女性样本有8211个,男性样本比女性样本多了三千多个样本。在年龄类别分布中,随着年龄的增加样本数据量逐渐甚至急剧减少,0-18岁有7900个样本,19-23岁有5330个样本,24-30岁有3603个样本,31-40岁有2141个样本,41-50岁有589个样本,51-999岁仅有82个样本。学历样本的分布与年龄一样严峻甚至更恶劣,初中学历的有7487个样本,高中有5579个样本,大学有3722个样本,小学有1150个样本,硕士有119个样本,博士仅有65个样本。这种样本空间的不平衡情况造成了分类器的性能不佳。

4.2 分类模型性能分析与比较

本文对分类结果的评估涉及到准确率与召回率两个指标,并以准确率为主要评价手段,下面给出它们的计算方式。

为了不复杂化结果,这里给出当训练集样本数量与测试集样本数量的比例是9:1时的分类结果比较分析。

(1)特征选择对分类模型性能的影响

这里,选取的性别属性作为预测对象,分类模型是多项式朴素贝叶斯(平滑参数取0.01),分类模型是“定量”,以特征选择算法以及特征数量作为“变量”,观察分析不同特征选择算法、不同特征数量下分类模型在预测用户性别时的平均准确率趋势。下图中,横轴是特征数量,纵轴是平均准确率,蓝色折线代表CHI2统计,绿色折线代表IG统计,红色折线代表MI。

图4 特征选择算法、特征数量对性别分类平均准确率的影响

由图可以看出,{ 61% : 分类器的性能随着特征选择的数量的增加, }先大幅增长,随后到达顶峰,之后略微下降,总体趋势是:1)在特征数量较少的情况下,{ 62% : 不断增加特征的数量,会显著提高分类器的性能, }呈现“上升”趋势;{ 56% : 2)随着特征数量的不断增加, }将会引入一些不重要的特征,甚至是噪声,{ 61% : 此时再增加特征数量已难以提高分类器性能, }因此,分类器的性能将会呈现略微“下降”并趋于平稳的趋势。

CHI统计达到的平均分类准确率峰值是80.83%,MI和IG都是80.04%。虽然这三种特征选择算法下分类平均准确率的峰值只是相差0.79%,但是CHI统计在特征数量为10000时,分类平均准确率就已升到最高值,但IG和MI要在特征数量高达33000左右时分类平均准确率才能达到最高值,而原始的特征数量是349000。这说明了,在本文数据集上,与互信息以及信息增益方法相比,{ 56% : CHI统计方法具有巨大的优势, }CHI统计方法更能提取出最重要的特征。

综合来看,上图的分类器性能走向趋势体现出了特征选择的重要性:选择出重要的特征,{ 58% : 并降低噪声,提高分类算法的泛化能力。 }

(2)伯努利朴素贝叶斯与多项式朴素贝叶斯

这里使用伯努利朴素贝叶斯与多项式朴素贝叶斯来预测用户属性,图1、图2、图3分别是年龄、性别、学历的平均分类准确率折线图,特征集是由chi2统计法选出来的10000个特征。横轴是平滑参数(称为alpha),纵轴是平均准确率,蓝色折线是多项式贝叶斯的预测结果,绿色折线是伯努利贝叶斯预测结果。

图5-1 alpha值对两种朴素贝叶斯模型的性别分类平均准确率的影响

图5-2 alpha值对两种朴素贝叶斯模型的性别分类平均准确率的影响

图5-3 alpha值对两种朴素贝叶斯模型的学历分类平均准确率的影响

综合以上三张图,可以看出:1)在本文分类任务中,伯努利贝叶斯的性能比多项式贝叶斯更优。2)alpha的大小能影响两种贝叶斯分类模型的预测精度,对多项式贝叶斯的影响尤其大。3)预测精度一开始随着alpha值的增大而波动上升,到达峰值后,随着alpha值得增大而波动下降,选取最优的alpha值对构造贝叶斯分类器意义重大。

(3)KNN分类器中K值的选取

下图是使用KNN分类器来来预测用户的年龄、性别、学历得到的分类平均准确率折线图,特征集是由chi2统计法选出来的10000个特征。横轴是K值(KNN算法选取K个最邻近邻居来做类别统计),纵轴是平均准确率,蓝线代表性别,红线代表学历、绿线代表年龄。

图6 K值对KNN分类模型的影响

由图可以直观地看出:1)KNN对性别的预测准确度远高于年龄和学历。2)对学历的预测准确度在K为80时达到最大,对年龄的预测准确度在K为110时达到最大,之后随着K值的增加,学历和年龄的预测准确度有着明显的下降趋势。而性别的预测准确度在K在1160时达到最大,之后随着K值的增加虽有略微下降,但并不明显。

首先,性别只有2类,这两类样本数量分布并不悬殊且数量较大,但性别和学历都各有6类且分布较为悬殊。性别和学历这种有限的样本数量以及样本空间的极不平衡的数据集情况是造成它们的平均准确率不佳的根本原因。

再而,上述样本空间不平衡情况导致了性别和学历的准确率达到峰值后随着K值的增加有非常明显的下降,{ 61% : 因为当K值变大时待测样本的K个邻居中更大容量类的样本更容易占多数。 }而性别的样本空间分布相对均衡,所以其预测准确率并未随着K值的增加产生明显的下降趋势。

(4)各分类模型间的对比

下表记录着,三种构建好的分类模型下,三种用户属性里每种类别的分类准确率和召回率。其中,根据上文的实验结果,朴素贝叶斯分类模型是伯努利朴素贝叶斯分类模型,其alpha值以及K-最近邻分类模型的K值也已调整到最佳值,支持向量机使用线性核。 由于各种原因,有些分类器完全没有预测出某种类别的样本,如研究生和博士,准确率为0,召回率的分母为0无法计算,这部分召回率在表中登记仍然登记为0。

表2 各类别在不同分类模型下的准确率和召回率

表2中数值为0的准确率和召回率,基本上是来自学历中的研究生和博士类别,以及年龄中的51-99岁类别。它们的共同点是样本数量极低,20000条记录中,硕士、博士仅分别占119、65条,51-99岁的记录也仅有82条。

此外,年龄中41-50岁类别的准确率和召回率也普遍比较低,虽然41-50岁类别在数据集中分别有589,比51-99岁要多出许多,但比起其他大容量类别,41-50岁的样本数目还是相当少。

性别属性中,分类模型在男性类别的表现均优于女性类别,原因还是在于样本数量不均,男性类别以及女性类别的样本数量都非常大,这是它们取得较高预测精度的原因,但同时存在男性类别的样本数比女性类别多出三千左右,这种样本不均衡使得女性类别的分类精度差于男性类别。

总体来看,在本文的分类任务中,伯努利朴素贝叶斯与支持向量机分类模型表现较为优秀,K-最近邻分类模型稍微逊色。

4.3 实验结论

通过以上的实验结果表明,本文关于根据用户搜索关键词预测分析用户属性这一课题提出的基于文本分类技术的解决方案是可行且有效的。伯努利朴素贝叶斯与支持向量机分类模型表现较为优秀,K-最近邻分类模型稍微逊色。特征数量、特征选择算法、参数的设置都对分类模型的性能有较大影响。学历及年龄的样本空间分布极其不均匀,导致了相应分类模型性能不佳。

第5章 总结与展望

5.1 本文工作总结

现代互联网广告有着迫切的“精准投放”需求,其中,{ 66% : 如何找到真正需要广告的人是个关键问题。 }用户画像挖掘技术中,根据用户的浏览、搜索等行为来反推获取用户属性是一项非常基础且非常重要的技术。本文的研究内容是基于用户搜索行为与用户属性的相关性对用户属性进行分析预测,{ 56% : 由于用户搜索记录是文本数据, }用户各属性的标签都是离散值,因此本文采用文本分类的相关技术来分析和预测用户属性。本文从中文分词、特征选择、特征加权、构建分类模型等多个方面着手,至下而上搭建了一套分类系统来对用户属性进行判别。通过对分类系统的实验结果分析,{ 55% : 证明了本文采取的分析预测方法是有效可行的, }同

时对比了采用不同的特征选择方法、不同的分类模型对分类器性能的影响。

{ 67% : 本文的主要工作内容有以下几点: }

{ 57% : 1、调查并研究了几种经典的分类算法。 }由于本文是根据用户搜索记录文本对用户进行分类,故而主要考虑了在短文本分类领域表现良好的分类算法。主要是基于概率统计知识的朴素贝叶斯算法,基于最优超平面划分的支持向量机算法以及基于向量距离比较的K最邻近算法。

{ 57% : 2、介绍了关于中文文本预处理的基本流程, }{ 58% : 以及介绍了相关特征选择和特征加权的算法。 }原始的文本不能直接作为分类模型的输入,需要做相应的中文分词、特征选择、特征加权等工作。本文采用结巴分词对原始文本做分词处理,{ 59% : 并介绍了TF-IDF的文本特征加权方法。 }特征选择方面,主要介绍了文档频率、信息增益、CHI统计三种特征选择方法,同时在实验中采用并比较了多种特征选择方法对分类模型性能的影响。

{ 70% : 3、为了提高各分类模型的性能, }针对本文研究内容情况,研究了各分类算法的改进方法,比如,基于KD-Tree的改进的KNN算法,使用线性核的支持向量机等。

4、关于构建分类模型时参数的设置问题,本文根据本文的实际情况,参考了前人的经验来设置适当的参数值,同时会依据分类结果的反馈适当调整参数的值。

5、详细介绍了文中基于用户搜索关键词的用户属性分析预测方法的实现方案,{ 62% : 给出了分类系统的总体结构图, }{ 60% : 并通过实验分析比较了各种情况下的分类效果, }包括不同特征数量、不同特征选择方法、不同分类算法、不同参数值下的分类模型性能优劣。

5.2 本文的不足与展望

本文的工作在取得一定成果的同时依然有许多方面需要得到进一步的改善:

1、中文分词步骤中,由于中文词汇系统较为复杂,没有做同义词的转换和错别字的纠正,{ 64% : 可能会导致分类性能上的一些损失。 }

2、样本空间不平衡造成的分类模型性能不佳的问题还是没能得到解决,虽然分类器在具有大数量类别上取得了较高的准确率和召回率,但小数量类别的分类准确率极差。过采样的方法虽然能增加样本数,但却会破坏原本样本空间的结构,可能会降低分类器的性能。下一步工作应该尽量采取手段,增加小数量类别的样本数。

3、分类模型参数的设置,如朴素贝叶斯中的平滑参数和KNN中的K值,都是基于经验和逐步的观察来调整的,可以采用更具理论基础的方法进行最优参数搜索与设置。

4、本文使用的数据集质量存在一定问题,因为存在多人共用一个搜狗账号的现象,如一个家庭的成员都通过同一个搜狗账号进行搜索。针对这个问题,可以提前对数据集进行清洗,人工过滤掉不合理的样本。

5

参考文献

[1] Han J, Kamber M, Pei J.数据挖掘:概念与技术(原书第3版)[M].范明,孟小峰,译.北京:机械工业出版社,2012:288-320.

[2] David M.J. Tax, Robert P.W. Duin. Support vector domain description[J]. Pattern Recognition Letters,1999,20:1191-1199.

[3] 朱华宇,孙正兴,张福炎. 一个基于向量空间模型的中文文本自动分类系统[J]. 计算机工程. 2001(02)

[4] 牛玲. 一种基于向量空间模型的改进文本分类算法[J]. 情报杂志, 2006, 25(6):63-64.

[5] Vapnik V N. An overview of statistical learning theory [J].IEEE Trans Neural Network,1999,10(5):988-999.

[6] 张冬生. 支持向量机在分类问题中的应用研究[J]. 黑龙江科技信息, 2010(35):64-64.

[7] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[J]. Lecture Notes in Computer Science, 2003, 2888:986-996.

[8] 余悦蒙, 黄小斌. 一种基于KNN的文本分类算法[J]. 电脑知识与技术, 2012, 08(3):1564-1566.

[9] 刘海峰, 刘守生, 姚泽清,等. 文本分类中基于训练样本空间分布的K近邻改进算法[J]. 情报学报, 2013, 32(1):80-85.

[10] Hastie T, Tibshirani R. Discriminant Adaptive Nearest Neighbor Classification[M]. IEEE Computer Society, 1996.

[11] Y.Yang.A Comparative Study on Feature Selection in Text Categorization. In: Proceeding of the Fourteenth International Conference on Machine Learning (ICML’97),412- 420,1997.

[12] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3):17-23.

[13] 张浩, 汪楠. 文本分类技术研究进展[J]. 科技信息:科学·教研, 2007(23):99-100.

[14] 杨凯峰, 张毅坤, 李燕. 基于文档频率的特征选择方法[J]. 计算机工程, 2010, 36(17):33-35.

[15] 刘辉, 应培培. 一种改进的KNN文本分类算法[J]. 信息安全与技术, 2011(7):25-27.

[16] 刘志刚, 李德仁, 秦前清,等. 支持向量机在多类分类问题中的推广[J]. 计算机工程与应用, 2004, 40(7):10-13.

[17] 刘忠, 刘洋, 建晓. 基于KD-Tree的KNN文本分类算法[J]. 网络安全技术与应用, 2012(5):38-40.

[18] Yang, Yiming, and J. O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization." Fourteenth International Conference on Machine Learning Morgan Kaufmann Publishers Inc. 1998:412-420.

[19] “Five balltree construction algorithms”, Omohundro, S.M., International Computer Science Institute Technical Report (1989).

[20] Li, Shoushan, et al. "A framework of feature selection methods for text categorization." Joint Conference of the, Meeting of the ACL and the, International Joint Conference on Natural Language Processing of the Afnlp: Volume Association for Computational Linguistics, 2009:692-700.

[21] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3):17-23.

[22] 张宁, 贾自艳, 史忠植. 使用KNN算法的文本分类[J]. 计算机工程, 2005, 31(8):171-172.

[23] 崔伟东, 周志华, 李星. 支持向量机研究[J]. 计算机工程与应用, 2001, 27(1):58-61.

[24] 程克非, 张聪. 基于特征加权的朴素贝叶斯分类器[J]. 计算机仿真, 2006, 23(10):92-94.

[25] 熊忠阳, 张鹏招, 张玉芳. 基于χ~2统计的文本分类特征选择方法的研究[J]. 计算机应用, 2008, 28(2):513-514.

[26] 李学明, 李海瑞, 薛亮,等. 基于信息增益与信息熵的TFIDF算法[J]. 计算机工程, 2012, 38(8):37-40.

6