



基于 DPI 的用户兴趣实时分类

陈康¹, 付华峥¹, 陈翀¹, 张宇艺², 韦誉³

(1. 中国电信股份有限公司广州研究院, 广东 广州 510630;

2. 广东工业大学计算机学院, 广东 广州 510006;

3. 广州优亿信息科技有限公司, 广东 广州 510630)

摘要:近年来,随着大数据发展热潮兴起,电信运营商能获得的 DPI 数据规模越来越大。如何对如此海量的 DPI 数据进行实时的挖掘和分类是运营商研究的热点。设计了一种基于混合 n -gram 特征的 URL 分类算法和基于 Doc2Vec 模型的文本分类算法相结合的 DPI 数据分类方法,大大提高了 DPI 数据的分类速度和分类准确率,为电信运营商进行精准营销提供了良好的基础。

关键词: DPI; URL 分类; 文本分类; n -gram; Doc2Vec

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016304

A real time approach to user interest classification using DPI

CHEN Kang¹, FU Huazheng¹, CHEN Chong¹, ZHANG Yuyi², WEI Yu³

1. Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

2. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

3. Guangzhou Usease Information Technology Co., Ltd., Guangzhou 510630, China

Abstract: In recent years, with the rapid development of internet, telecom operators can gather more and more DPI data. How to mine and classify huge amounts of DPI data in real time is the key point. A DPI data classification method was presented, which combined the URL classification model based on mixed n -gram features and the Doc2Vec based text classification algorithm. The DPI data classification speed and classification accuracy were greatly improved, and good support for precise marketing was provided.

Key words: DPI, URL classification, text classification, n -gram, Doc2Vec

1 引言

随着通信行业的深度变革,国内通信行业的市场环境和竞争格局发生了根本性的变化。在企业服务质量差异较小、价格竞争瓶颈状况明显的情况下,企业自身运营能力的提升将是企业在激烈的竞争中掌握主动权的一个关键转折点。根据用户的消费信息挖掘出不同消费群体的需

求,为用户提供更加优质的服务,提升用户对电信运营商的满意度和黏性,这将是电信企业技术研究的重要方向,也是响应“互联网+”战略的重要举措。

当前,随着大数据相关技术迅速发展,电信运营商开始建立基于数据挖掘的商业智能系统。同时,电信企业能够获取的用户数据越来越丰富,其中包括:产品/服务使用信息、用户基本信息、语音话单数据、用户上网数据和信令



数据等。通过 DPI(deep packet inspection,深度分组检测)分析技术,能够较好地识别网络上的流量类别、应用层上的应用种类等。然后基于 DPI 数据进行分析挖掘,可以形成一套完整丰富、动态更新的用户行为标签体系。基于所构建的用户标签,电信运营商可以充分地了解用户的需求,从而实现精准化营销、提升营销效果^[1]。在构建用户标签体系的过程中,一个重要的工作是对能够反映用户兴趣的 DPI 数据进行实时的分类。传统的分类方法有两种,一种是基于 DPI 数据中的 URL 来分类,另一种是基于用户 DPI 数据对应的网页内容分类。基于 URL 进行分类的速度较快,但在准确性方面却略显不足,而基于网页内容进行分类的准确性较高,但在处理海量的 DPI 数据时,通过获取 DPI 数据中的 URL 字段对应的网页内容再进行分类是不现实的。在电信运营商每天要处理的信息量规模巨大的今天,迫切需要一个高效、准确的分类方法。

鉴于此,本文提出了一种基于混合 n -gram 特征的 URL 分类算法和基于 Doc2Vec 模型的文本分类算法相结合的 DPI 分类方法,实现了对超大规模的用户上网 DPI 数据的快速分类以及分类器的自动学习和更新。相比于现有的分类模型,本文实现了以下方面的提高:在 URL 分类上,URL 的类别特征表示使用的是混合 n -gram,参考文献[2]很好地阐述了 URL 的混合 n -gram 的特征表示的优越性,减少了很多关键词的遗漏,提高了分类的准确度;在文本分类上,使用 Doc2Vec 实现了文档的语义表达,并在此基础上实现了更为精准的文本分类;对于海量的用户 DPI 数据,通过结合以上的两种分类方法,即 URL 分类和文本分类,设计了一个新的 DPI 分类架构,达到了较好的实时分类效果,大大提高了对 DPI 数据分类的速度和准确度。

2 电信运营商的 DPI 数据

DPI 技术是一种基于应用层的流量监测和控制的技术。DPI 技术除了分析 IP 分组的 4 层以下的内容,即源地址、源端口、目的地址、目的端口和协议类型外,还增加了应用层分析,用以辨别各种业务和应用,从而得到业务类型、业务状态和业务内容等信息。宽带 DPI 数据格式定义见表 1。

通过 DPI 技术,电信运营商能够展开有针对性的营销,有效地进行业务识别、业务控制和业务统计,从而能充分地挖掘出自身的价值,在激烈的竞争中处于有利的地位。

由参考文献[3]得,中国电信全网根据用户互联网访问的 DPI 数据信息的差异,分为公有信息和协议特有信息。本文主要针对 HTTP 特有的协议信息进行了深入探讨。其中,对于该协议信息中的 URL 字段是本文中对用户 DPI 数据进行 URL 分类和文本分类的基础。

3 DPI 数据分类相关技术

随着大数据的蓬勃发展,针对获得的海量 DPI 数据,各大电信运营商的大数据研发团队对此进行了不同程度的深入挖掘研究,其中涉及的关键技术有 URL 分类和文本分类。

在 URL 分类的研究方面,参考文献[3]使用爬虫技术建立的丰富的 URL 类别库结合人工标注的方式构建 URL 分类体系,作为实现对 DPI 数据结构化转换与分类的基础。该方案对于电信运营商利用用户 DPI 数据进行深入挖掘分类研究有一定参考价值。然而,在 URL 分类失败的情况下,只讨论了人工识别环节,这使得在所构建的 URL 分类体系上

表 1 宽带 DPI 数据格式定义

字段名	字段中文名	类型
account	用户账号 MD5 散列字符串(唯一)	string
srcip	HTTP 标准字段,源端网络主机地址,如:113.119.141.171	string
srcport	HTTP 标准字段,源端网络主机端口,如:6768	int
dstip	HTTP 标准字段,目标网络主机地址,如:124.192.164.48	string
dstport	HTTP 标准字段,目标网络主机端口,如:80	int
url	HTTP 标准字段,指用户访问的网站链接地址,如:qt.gtimg.cn/?q=sh600129,sz000150,sz002115,sh600831,sz300267	string
referer	HTTP 标准字段,用来判断来源页面,即用户是从哪个页面来的,如:http://qqweb.gu.qq.com/	string
useragent	HTTP 标准字段,用来检查浏览页面的访问者在用什么操作系统(包括版本号)浏览器(包括版本号)和用户个人偏好的代码。如:Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.59 QQ/8.2.17724.201 Safari/537.36	string
time	HTTP 标准字段,指发起 HTTP 请求时的时间戳(为 1970-1-1 到目前的毫秒数),如:1 466 488 328 085	int

存在顽健性较差的问题。参考文献[2]与参考文献[4]提出的两种 URL 分类方法有一个共同的特点,就是将 URL 当作文本进行分类。在传统文本分类中,需要先对文本进行特征提取,因此 URL 特征提取是 URL 分类中很重要的步骤。参考文献[2]提出了一种单一 n -gram 的特征提取方式,但这会导致很多关键词的遗漏,降低分类的准确率。

参考文献[5]很好地阐述了 URL 分类的优点,如分类速度快,但是在准确性方面,由于 URL 中包含的类别信息有限,所以分类准确性不高。对于海量的 DPI 用户数据,除了直接使用 URL 进行分类,还可以通过获取 URL 对应的文本内容进行分类。在文本分类中,传统的分类算法包括:Rocchio 线性分类算法、基于实例的 k 最近邻分类算法(k -nearest neighbor, k NN)、朴素贝叶斯分类算法(naïve Bayes, NB)、决策树分类算法(decision tree, DT)、基于支持向量机的分类算法(support vector machine, SVM)。本文中的文本分类使用的是基于 SVM 的分类算法。参考文献[6]与参考文献[7]很好地深入分析了基于 SVM 的文本分类算法,并研究指出了 SVM 是一种具有很好的泛化能力和出色的分类性能的技术。在传统的分类模型中,文本分类算法是基于关键词向量的,这损失了文本中的语义信息。因此,本文中使用的基于 Doc2Vec 模型的文本分类算法,这实现了文档的语义表达,并在此基础上实现了更为精准的文本分类。对于文本分类,由参考文献[8]可知,学术界已经做了大量的技术积累。相比于 URL 分类,文本分类准确性高。但是在处理大规模的 DPI 数据时,文本分类还是存在一些值得学术界深入研究和突破的瓶颈:首先由于下载量大,给网络的带宽带来了较大的负载;其次就是分类时延高,用户体验较差。

4 DPI 数据分类算法

对于海量的 DPI 数据分类,单独使用 URL 分类或文本分类,都存在较为明显的缺陷。鉴于此,本文选择基于混

合 n -gram 特征的 URL 分类算法和基于 Doc2Vec 模型的文本分类算法相结合的 DPI 数据分类方法,实现了对超大规模的用户上网记录实时、高效、准确地分类。图 1 展示了本文设计的 DPI 数据分类流程。若 URL 分类器对待分类的 DPI 数据中的 URL 分类成功,则直接返回分类结果;若分类失败,则提取该 DPI 数据中的 URL 对应的网页正文,然后用文本分类器对其进行分类,返回分类结果并更新 URL 分类器,以此增加 URL 分类器的分类内容储备,从而提高 URL 分类器对 DPI 数据中 URL 的分类速度和分类成功率,提升整体分类效率。

4.1 URL 分类

如参考文献[9]所述,URL 一般格式为:Protocol://[user:password]@host[:port]/path/[?query] [#fragment]。据观察,大多数情况下 URL 中隐含有类别的信息。各个网站的 URL 组织的风格不一致,但相同类别网站的 URL 具有一定的规律性,即 URL 中的一些片段隐含有类别的信息。鉴于此,提出了一个基于 n -gram 的 URL 分类算法,基本思想是使用 n -gram 实现 URL 的特征提取,并结合 TF-IDF 对特征进行筛选,然后使用朴素贝叶斯分类器进行分类。 n -gram 是一种常用的语言模型,该模型假设第 n 个词的出现只与前面 $n-1$ 个词相关,而与其他任何词都不相关^[10]。因此,在对 URL 进行特征提取时,选取一系列连续的 n 个字母的子字符串作为 URL 的特征。

以网页 http://www.allwatchers.com/Topics/Info_3922.asp 中的“allwatchers”为例,进行 n -gram 的切割后得到的词组见表 2 中的前 3 行。对“allwatchers”进行切割,得到单词“all”“watch”“watchers”这样的短词组合才是有意义的。但在单一的 n 值的情况下,“all”与“watch”不能同时存在,“allwatchers”中的信息并没有被充分利用。 n -gram 算法很难选择最佳的 n 值。如果 n 太小或者太大,URL 中的关键字会被分割或被拼凑成其他的模样,信息就会丢失,而通常情况下这些关键词能够反映出 URL 类别的特征。为了

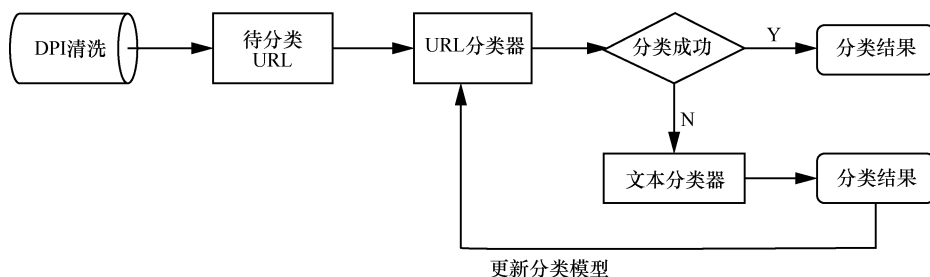


图 1 DPI 数据进行 URL 分类和正文分类流程



解决这个问题,本文不使用单一的 n 值,选用了一种 all-gram 的思想,即使用一系列的 n 来分割 URL。通过这样的方法生成的 n -gram 子串,会减少很多关键词的遗漏,如表 2 中的最后一行,选取 3、4、5 作为 n 值的时候,分割得到的词集中既有“all”,也有“watch”。

表 2 allwatchers 切割后的词集

	allwatchers
3-gram	all,llw,lwa,wat,atc,tch,che,her,ers
4-gram	allw,llwa,lwat,watc,atch,tche,cher,hers
5-gram	allwa,llwat,lwatc,watch,atche,tcher,chers
all-gram	all,llw...watch,atche,tcher,chers

经过 n -gram 分割后得到了长度不一的短词,然后通过计算得到词频,再代入由朴素贝叶斯公式变化得到的式子中即可计算出该 URL 所属分类的概率。朴素贝叶斯分类是一种较为简单的分类算法,其通过求解待分类项出现的条件下各个类别出现的概率,然后将待分类项归为概率值最大的类别:

$$P(C|w_1w_2\cdots w_n) = \frac{P(w_1w_2\cdots w_n|C)P(C)}{P(w_1w_2\cdots w_n)} \quad (1)$$

分类概率的计算式见式(1),URL 经过 all-gram 处理后得到 n 个短词,分别为 $w_1, w_2, \dots, w_i, \dots, w_n$ 。现有个类别,分别为 $C_1, C_2, \dots, C_i, \dots, C_n$ 。 $P(w_1w_2\cdots w_n)$ 代表 URL 出现的概率,它的值对所有的类别均相同。因此这个问题可以等价于求解式(2):

$$\max_{k=1,2,\dots,m} P(w_1w_2\cdots w_n | C_k)P(C_k) \quad (2)$$

本文假设所有短词之间相互独立,因此式(2)中的

$P(w_1w_2\cdots w_n|C)$ 可以转换为 $\prod_{i=1}^n P(w_i|C_k)$ 。转换后的每一项

$P(w_i|C_k)$ 的值都可以从训练数据中计算得到。由本文实验观察可知,公共且词频高的短词对分类结果影响较大。如金融、体育类别的训练集中含有大量的域名为 sina.com.cn 的 URL,则 http://sina.com.cn/finance_index1.shtml 在各分类出现的概率的大小由“sina”等词频高但类别特征模糊的短词主要贡献,而忽略了“finance”这种类别特征较强的短词。这严重影响了分类的准确性。为了解决这个问题,对于所有的短词,本文使用相应的 TF-IDF 值代替词频来计算概率。TF-IDF 值能够很好地表示一个词语在文章中的重要程度。在本文中,其值表示 URL 短词在类别中的稀有程度。稀有程度越高其分类能力越强,值越大。TF-IDF 公式在参

考文献[11]中,这里不再赘述。

4.2 基于用户浏览内容的分类

用户浏览的内容最终会被转换成文本的形式用于分类,通过文本分类来确定 URL 所属的类别。文本分类是通过训练已标注的文档集合,找出文档特征和文档类别间的关系模型,再利用训练得到的关系模型对新的文档进行分类。本文使用 Doc2Vec 模型将文章表征为向量,构建 SVM 分类器,从而实现文本分类。Doc2Vec 是在 Word2Vec 基础上提出的文章向量化方法。在文本训练阶段用 Doc2Vec 模型将大量人工分类的文章映射成一定维度的向量集合,然后使用向量集训练 SVM 分类器。

在参考文献[12]中介绍了 PV-DBOW 与 PV-DM 两种 Doc2Vec 训练模型。PV-DM、PV-DBOW 分别与 Word2Vec 中的 cbow、skip-gram 相似,区别是 Doc2Vec 中的两个模型均增加了一个代表文章的向量(paragraph vector)作为输入。两种模型在训练的时候调整文章向量,使得其能很好地表示所对应的文章。两种模型在训练的时候除了存储词的向量外也会存储代表文章的文章向量。随着训练的文章数量的增加,其存储文章向量消耗的空间呈线性增长,模型的可扩展性不好且训练阶段的文章向量对本文的文本分类用处不大。因此,基于 PV-DBOW 模型的思想设计了一个将文章表示成向量的模型。该模型把训练步骤拆分为两步,首先将文章分词、去停用词后的词集通过 Word2Vec 模型映射成向量,再使用映射得到的向量训练出文章的文章向量。本文设计的模型结构与 PV-DBOW 模型的结构相似,不同之处在于 PV-DBOW 在训练模型的时候,会同时更新词向量和文章向量,也就是词向量和文章向量是同时训练得到的;而本文只是借助了 PV-DBOW 模型的结构,利用已有的词向量去训练文章向量,也就是说词向量是需要预先训练好且不会更新。

4.3 DPI 实时分类系统

图 2 展示了 DPI 实时分类系统的架构。清洗器(cleaner)的职责是对 DPI 数据进行清洗,过滤不能反映用户行为的 DPI 数据。如用户浏览网页时,除了目标网站的 URL 请求外,浏览器还会在后台请求 css、js、gif 等无法反映用户行为的数据。分类器(classifier)的作用是对待分类的 DPI 数据中的 URL 字段进行分类,若分类成功则直接返回分类结果;若分类失败,则写入 URL 到反馈队列(queue)。反馈队列中的 URL 会被爬虫提取并抓取用户浏览的内容

转换成文本数据。得到的文本数据最终会被分类训练器(classifier trainer)进行文本分类,返回处理结果并更新分类器。

分类结果最终输出给写(writer)模块进行存储。系统架构如图 2 所示。

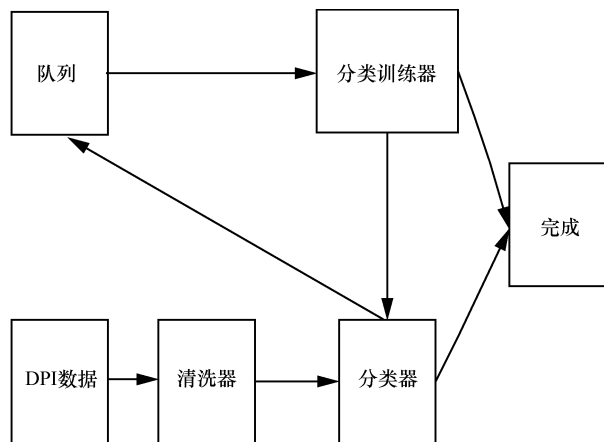


图 2 系统架构

5 实验与分析

5.1 实验数据的获取

本文所涉及的实验需要海量的 DPI 数据提供支撑,然而电信运营商所能获得的众多 DPI 数据并没有标注,给分类算法带来了较为严峻的挑战。因此,基于这样的一种思路来获取实验所需的 DPI 数据:首先,按照业务的要求把数据源划分为七大版块:教育、养老、金融、旅游、汽车、体育以及科学;然后,通过导航网站的不同板块进行搜索或通过搜索网站进行关键字搜索。例如:通过宝马、奔驰等字眼来获取汽车板块的内容;通过篮球、足球等关键字来获取体育版块的内容。由于各种类型的导航网站或者是大型的搜索网站,其提供的各个板块的内容都是广大群众所关注的内容,所以,根据本文所设计的方法获得的 DPI 数据能够较好地反映出用户的爱好和兴趣。本文对每一个板块都收集足够数量的页面,然后使用爬虫获取这些页面的 URL,这样,每一个板块都得到了足够数量的 URL,接着采用分层抽样的方法,在每一个板块中抽取实验所需的 URL,并获取该 URL 对应的页面内容。最后得到实验所需的有标注的 URL 库和相应的文本库,可以用于后续的分析。

5.2 URL 分类准确率

本实验使用单独 URL 分类以及 DPI 分类算法(即

URL 分类结合文本分类)两种方法对同一个标注好的数据集进行分类,并记录各分类的准确率然后进行比较^[13]。本实验使用 Python3 实现实验算法。本实验使用 pypspider 来实现数据的抓取,分词工具使用 jieba 提取正文使用 boilerpipe,文本表征向量的 Doc2Vec、Word2Vec 使用的是 gensim 的库。经过测试,当 Word2Vec 模型训练的迭代次数为 6 时,其发现语义相近的词语效果最好。

本实验使用第 5.1 节中获取的数据集,共 339 912 条数据。构建 URL 分类模型的时候,各个类别以 8:2 的比例随机划分训练集以及测试集,混合 n -gram 参数值为 5、6。为了使实验具有可对比性,DPI 分类算法中使用的 URL 分类模型与单独使用 URL 分类方法中的模型为同一个模型。在 DPI 分类算法中,URL 分类结果中概率最大的两个类别的预测值之差小于预设的阈值的时候,判定为分类失败,然后使用文本分类进行分类。文本分类中的 SVM 分类器使用 URL 分类模型的训练集中 URL 对应的文本数据向量化后的向量集训练得到。分类失败的时候,取 URL 对应的文本数据用 Doc2Vec 模型向量化后的向量作为 SVM 分类器的输入进行分类。分类结果以文本分类结果为准。分类结束后,统计各分类结果,实验结果如图 3 所示。

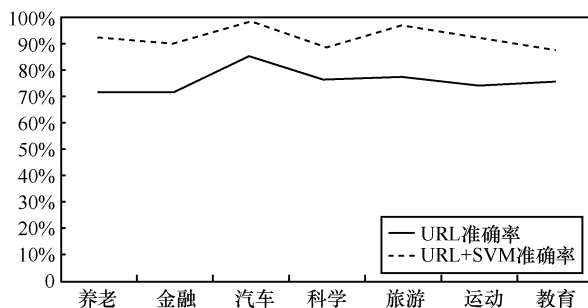


图 3 准确率记录

由图 3 可见,URL 分类结合文本分类方法相比单一的 URL 分类方法,分类的准确率有很大的提升,各分类的准确率达到 90% 以上。

5.3 基于 Storm 实现的 URL 分类吞吐量实验

实验环境由 10 台服务器构成,Storm、HDFS 部署在其中,操作系统均是 CentOS 6.4,CPU 为 32 核,内存 131 GB。所有节点使用 jdk1.8.0_45 版的 Java、apache-storm-0.10.0 版的 Storm、hadoop-2.6.0-cdh5.7.1 版的 Hadoop。10 台机器中选其中 1 台作为 Storm 的 Nimbus、HDFS 的 namenode,同时也与其他节点一样作为 Storm 的 supervisor、HDFS 的



datanode。

本实验程序仅测试 URL 分类模块功能。Spout 从 Kafka 读取 HDFS 上的数据交由 CleanerBolt 过滤不能反映用户行为的 DPI 数据,然后用 ClassifierBolt 进行分类。分类结果最后由 FeedBackBolt 输出。本实验固定 Spout、Bolt 的数目,对参数 topology.max.spout.pending (简称 pending) 的值进行修改,通过其 Ack 机制与 Metric API 计算并记录其拓扑(topology)的运行情况。本文主要关注处理时延、处理速率两个统计项。处理时延是指一个 URL 从输入到输出分类结果整个过程所花的平均时间。处理速率即每秒处理的 URL 数目。根据业务的特点与程序的实现以及实际运行过程中对 jvm 的运行情况的观察,设置每个写模块的大小为 20 GB,jvm 的垃圾回收器为 G1。Storm 吞吐量实验记录见表 3。

表 3 Storm 吞吐量实验记录

pending	处理时延/ms	处理速率/(万条·s ⁻¹)
100	15	69
300	32	91
500	51	96
800	81	98
1 000	98	99

调整 pending 的作用是设置每个 Spout 有多少 DPI 记录还没有处理完的时候停止接收新的记录,该参数可以调整整个系统中能同时被处理的记录的总数。在本实验中,当 Spout、Bolt 数目不变,随着 pending 的值增大,处理速率、处理延迟也会随之增大。当 pending 的值增大到一定程度后,处理速率会趋于平缓,处理延迟继续增大。此时的集群的占用率已经达到极限值,即使通过增大 pending 值增大输入,处理速率也不会增长,而队列中等待处理 DPI 记录数目增多,等待的总时间增大,处理延迟也会增大。等待处理的 DPI 记录过多,会引发各种问题,如 GC 时间增长,处理速率下降,worker 内存爆满重启等。因此,选取合适的 pending 值对分类系统的稳定高效运行至关重要。

6 结束语

针对如何更好地挖掘分类数据规模越来越大的 DPI 数据这一问题,提出了一种基于混合 n -gram 特征的 URL 分类方法和基于 Doc2Vec 模型的文本分类方法相结合的 DPI 数据分类实时分布式方案并在 Storm 上进行了实现。

实验证明,本文提出的 DPI 数据分类方法,在分类速度与准确率方面达到了较好的均衡。这为电信运营商进行实时精准性营销提供了一个很好的参考。

参考文献:

- [1] 罗忆祖. DPI 技术助力运营商精细化运营 [J]. 电信网技术, 2009(3): 22-24.
- [2] LUO Y Z. DPI helps operators refinement operation [J]. Telecommunication Network Technology, 2009(3): 22-24.
- [3] ABRAMSON M, AHA D W. What's in a URL? Genre classification from URLs[J]. Workshops at the Twenty, 2012.
- [3] 边凌燕, 贺仁龙, 姚晓辉. 基于 DPI 数据挖掘实现 URL 分类挂载的相关技术研究[J]. 电信科学, 2013, 29(11): 6-11.
- [4] BIAN L Y, HE R L, YAO X H. Research on URL classification with DPI data mining and related technology [J]. Telecommunications Science, 2013, 29(11): 6-11.
- [4] 朱丽娜. 中文网页分类特征提取方法研究 [D]. 北京: 中国石油大学, 2009.
- [5] ZHU L N. Research of feature selection for Chinese Web page categorization[D]. Beijing: China University of Petroleum, 2009.
- [5] KAN M Y, THI H O N. Fast webpage classification using URL features [C]//14th ACM international conference on Information and knowledge management, October 31 -November 5, 2005, New York, USA. New York: ACM Press, 2005: 325-326.
- [6] 汪光庆. 基于 SVM 的网页分类技术研究[D]. 北京: 中国石油大学, 2011.
- [6] WANG G Q. The research of Web pages classification based on SVM technique [D]. Beijing: China University of Petroleum, 2011.
- [7] 陈燃燃. 基于 SVM 算法的 Web 分类研究与实现 [D]. 北京: 北京邮电大学, 2009.
- [7] CHEN R R. Research and implementation of Web classification based on SVM algorithm [D]. Beijing: University of Posts and Telecommunications, 2009.
- [8] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1859.
- [8] SU J S, ZHANG B F, XU X. Advances in machine learning based text categorization [J]. Journal of Software, 2006, 17(9): 1848-1859.
- [9] BERNERS-LEE T, MASINTER L, MCCAHERILL M. RFC 1738: uniform resource locator[J]. Internet Engineering Task Force, 1994.
- [10] 余小军, 刘峰, 张春. 基于 n -gram 文本特征提取的改进算法[J]. 现代计算机, 2012 (23): 3-7.
- [10] YU X J, LIU F, ZHANG C. Improved text feature extraction algorithm based on n -gram [J]. Modern Computer, 2012 (23): 3-7.

- [11] 郑霖, 徐德华. 基于改进 TFIDF 算法的文本分类研究 [J]. 计算机与现代化, 2014 (9): 6-9.
ZHENG L, XU D H. Research on text categorization based on improved TFIDF algorithm [J]. Computer and Modernization, 2014 (9): 6-9.
- [12] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[J]. Computer Science, 2014(14): 1188-1196.
- [13] 庞观松, 蒋胜益. 文本自动分类技术研究综述. 情报理论与实践[J]. 信息学: 理论与应用, 2012, 35(2): 123-128.
PAND G S, JIANG S y. A survey of automatic text classification techniques[J]. Information Studies: Theory & Application, 2012, 35(2): 123-128.

[作者简介]



陈康(1972-),男,中国电信股份有限公司广州研究院大数据应用研发中心副总经理、高级工程师,主要研究方向为分布式计算、数据挖掘和机器学习。



付华峰(1989-),女,现就职于中国电信股份有限公司广州研究院,主要从事数据分析、数据挖掘以及大数据相关技术等方面的研究工作。



陈舫(1979-),女,中国电信股份有限公司广州研究院工程师,主要研究方向为电信 IT 系统研发以及数据应用等,目前专注于大数据相关技术的研究与应用。

张宇艺(1991-),男,广东工业大学计算机学院硕士生,主要研究方向为数据挖掘、智能计算、大数据应用。

韦誉(1991-),男,广州优亿信息科技有限公司大数据工程师,主要研究方向为数据挖掘、大数据应用。