



基于 DPI 数据挖掘实现 URL 分类挂载的相关技术研究

边凌燕,贺仁龙,姚晓辉

(中国电信股份有限公司上海研究院 上海 200122)

摘要:通过对 DPI 用户上网行为数据进行深入挖掘,实现与网页 URL 分类体系的归类映射,是精准锁定上网用户兴趣偏好特征的关键。在梳理 DPI 数据自动挂载 URL 分类节点流程的基础上,重点研究了过程中涉及的网页信息提取、中文分词、特征选择及文本分类等关键技术,为利用 DPI 数据提升客户洞察能力铺平了技术道路。

关键词:深度分组检测;中文分词;特征选择;文本分类

doi: 10.3969/j.issn.1000-0801.2013.11.002

Research on URL Classification with DPI Data Mining and Related Technology

Bian Lingyan, He Renlong, Yao Xiaohui

(Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China)

Abstract: In order to achieve the precise localization of internet customers' preference, the crucial point is to catch the classified mapping of URL system through digging deep into the data of DPI users' online behavior data. Based on summarizing the process of DPI data auto classification, the key technique was investigated which involved in Web information extraction, Chinese word segmentation, feature selection, text classification and so on. It may make solid technical foundations for the customer insight with DPI data.

Key words: deep packet inspection, Chinese word segmentation, feature selection, text classification

1 引言

近年来,通信产业深度变革,电信运营商在整个 ICT 产业中的主导权逐步被分化。运营商要在全新的产业格局内占据优势,必须基于自身拥有数据的规模和活性以及收集和运用数据的能力优势,挖掘得天独厚的管道数据资产价值,在保护用户隐私的前提下,为用户提供高附加值的精准目标服务,激活数据资源和客户深度洞察力的市场能量,应对越来越激烈的市场竞争。

采用 DPI(deep packet inspection,深度分组检测)技术对移动互联网的用户上网行为数据进行数据精准解析、识

别后,将相应用户访问网站归类挂载至网页 URL 分类体系,通过与分类体系内各节点特征的映射来洞察用户上网的兴趣偏好,已成为运营商顺应移动互联网发展、迁移管道优势并强化数据应用的一个热点方向。本文梳理了海量 DPI 用户上网行为数据挂载到设定的 URL 分类体系的实现流程,重点研究介绍了网页信息提取、分词及文本分类等关键的文本挖掘应用技术。

2 中国电信 DPI 用户上网行为数据说明

DPI 作为一种基于应用层的流量检测技术,除了对 IP 分组 4 层以下内容做分组检测外,还增加了应用层分析,

可以深入解析和读取 IP 分组载荷的内容,识别各种应用及其内容^[1]。中国电信全网统一部署的数据信息采集解析设备输出的用户互联网访问 DPI 数据信息,分为公有信息和协议特有信息两部分。公有信息是对所有协议都做要求的信息,必备的公有信息字段包括用户信息、终端信息、访问协议信息、用户上网行为时间属性等信息,见表 1。主要实现从业务应用、时间段等多维度挖掘洞察用户上网行为时间、频次及流量耗费等信息。

协议特有信息是针对 HTTP、WAP、RTSP、SMTP 等 14 种协议特有的协议信息,需基于公有信息的“协议类型”字段

内容进一步解析。经实际数据探测,HTTP、WAP 内容的流量占据海量用户上网行为数据总流量的 80%以上,本文的研究着重围绕该两类协议的信息数据展开,二者协议必备特有信息中的关键字段“DestinationURL”(目标网站 URL 地址)将作为后续 DPI 数据分析挖掘研究的关键输入,见表 2。

3 DPI 数据自动挂载 URL 分类体系的研究

网页 URL 分类体系作为实现 DPI 数据的结构化转化

表 1 协议公有信息必选字段

编号	字段名	字段说明	关联分析内容
1	IMSI	用户的国际移动用户识别码	手机换卡分析
2	MDN	用户手机号码	移动用户手机号码识别
3	MEID	移动台设备标识	上网用户识别,防止盗用
4	NAI	网络接入标识	识别用户的上网账号
5	DestinationIP	用户访问的目标 IPv4 或 IPv6 地址	用户使用的应用、自动连接程序分析
6	DestinationPort	用户访问的目标端口号	用户使用的应用、自动连接程序分析
7	SourceIP	用户访问外部网络时使用的 IPv4 或 IPv6 地址	用户接入地址分析
8	SourcePort	用户访问外部网络时使用的端口号	用户接入地址分析
9	BSID	基站标识	用户进行用户位置分析与定位
10	Subnet	子网号,cdma2000 1x Ev-Do 系统的子网和扇区标识	用户进行更准确的用户位置分析与定位
11	Service Option	用户所接入的 CDMA 网络类型	用户上网的方式判断,如 WAP、cdma2000 1x Ev-Do、Wi-Fi 等
12	ProtocolID	协议类型	判断用户上传内容使用的通信协议
13	Service Type	业务类型	判断用户上传的内容,系统后续细化移动互联网判断的内容
14	Start Time	业务流开始时间	判断用户上传时间
15	End Time	业务流结束时间	判断用户上传时间
16	DURATION	持续时间,单位 ms	判断用户上传时长
17	Input-Octets	发给用户的业务字节数	判断用户上传数据量
18	Output-Octets	用户发出的业务字节数	判断用户上传数据量

表 2 HTTP、WAP 特有信息必备字段

协议	编号	字段名	字段说明	关联分析内容
HTTP	19	DestinationURL	用户访问的目标网站的 URL	用户访问网址分析
	20	UserAgent	User Agent 信息	识别终端,浏览器版本信息
	21	DomainName	外部网站的域名	用户访问网址和应用识别分析
	22	HOST	主机名	主机名
WAP	23	UserAgent	User Agent 信息	识别终端,浏览器版本信息
	24	DestinationURL	用户访问的目标网站的 URL	用户访问网址分析
	25	DomainName	外部网站的域名	用户访问网址和应用识别分析
	26	HOST	主机名	主机名
	27	WAPVERSION	WAP 版本	ABP 版本识别
	28	ACT_TYPE	WAP 操作类型(5:Post, 6:Get)	ABP 操作类型识别



和分类的基础,是精准锁定客户兴趣偏好特征的关键。该体系主要以用户目的需求为主线,参照运营商本身营销需求,综合互联网门户、导航站点、自有业务门户的分类粗粒度目录,根据业务产品聚合情况进行较强的针对性设定。从 DPI 数据解析实现网页 URL 分类体系的构建,流程上按规则“DestinationURL”和无规则“DestinationURL”区分考虑,如图 1 所示。

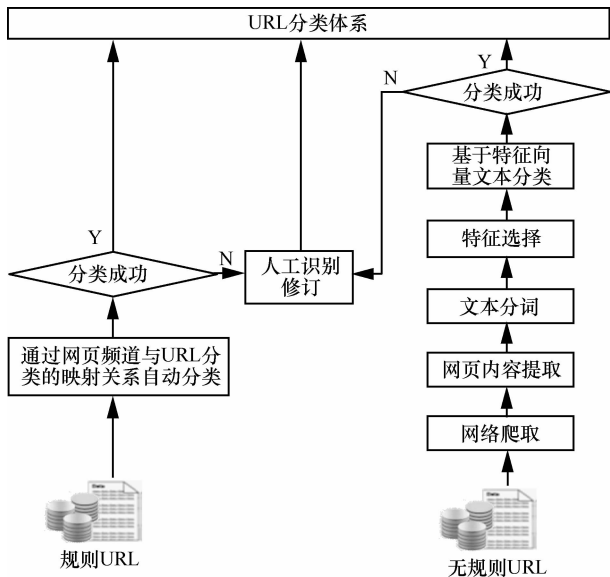


图 1 DPI 数据构建网页 URL 分类体系方案

针对规则 URL,处理方式通常借助网站本身 URL 或频道编码特征,建立与已有 URL 分类体系的映射关系,通过广度爬虫收集 URL 并进行分类自动挂载,本文不再赘述。

无规则 URL 多指没有多级域名或目录或所有栏目使用数字编码的网站地址。一般经由爬虫进行网页原始内容的解析和信息提取,通过文本分词、特征选择等步骤来实现网页文本特征向量的标定及文本分类,将未知网页映射挂载到给定的 URL 类别目录。

以上规则或无规则 URL 分类流程在系统自动实现失败的情况下,在维护流程上都加入人工识别环节,以补充自动分类器的判定不足,从而保证分类体系不断完善和及时更新,同时也可以随着用户偏好关注度及企业运营需求进行动态调整,以丰富长效稳定的 URL 分类体系。

4 无规则 URL 分类挂载的关键处理技术研究

4.1 网页内容提取

对无规则 URL 进行网页爬取后得到的页面源文件采

用超文本设计,其往往存在许多噪声,如广告、注释、导航、推荐、版权等无关信息,如果不进行过滤会直接影响后续网页分类结果。因此在做下一步挖掘分析前,网页内容提取的预处理工作显得很必要。

参考文献[2]详细介绍和对比了几种常见的网页内容提取技术。

- 基于 DOM (document object model) 树:依据专门适用于 HTML 的文档对象模型 DOM 树,解析 HTML 标签的层次关系成树状结构,通过遍历树节点的各个对象,识别网页正文信息。
- 基于文本及标签分布:参考文本及标签的分布状况编写行号与行块文本长度的分布函数,依据函数的骤升骤降,区分网页正文与非正文内容。但该方法过分依赖正文在源码中的位置分布,较易引起误提取。
- 基于视觉窗:利用导航在顶部、广告在侧边的布局常规特征,文字颜色、分隔边框、段落间距等视觉信号帮助定位网页正文信息。该方法准确率相对不高。
- 基于标记窗:先对网页标题进行分词,再取每个标签对之间的文本内容进行分词,并计算两者的相似度,设定阈值是否为提取的标准。其缺点是绝对依赖于标题的准确性。

基于 DOM 树网页内容提取技术的 HTMLParser 作为 SourceForge.net 社区的开源项目,是目前业内应用最广泛的网页解析工具。它由纯 Java 语言编写而不依赖于其他 Java 库,不仅扩展便利且可以兼容 Nutch 架构,进而超高速地实现无规则 URL 经后者爬虫抓取后的实时解析。其解析过程如图 2 所示,本文以某新闻网页为例说明,首先利用 HTMLParser 将网页文档转化为 DOM₁ 树,树的每个节点对应一个网页标签对象,进一步通过过滤树内大量的噪音信息,包括主题无关节点(通常是图片 img、对象 object、脚本 script、表单 form 等),无效节点(通常是无内容空节点),得到只含文本标签和结构标签的 DOM₂ 树。通过抓取同一站点下的两个网页,利用同一站点网页模板相关性,去除 DOM₂ 树重复内容页面信息,得到不一致文本内容包含的网页主题信息的 DOM₃ 树,即最终 DOM 树,可以基于该树的终节点实现对于 HTML 网页内容的文本转化。

如何快速有效地识别和去除网页文档中的噪音信

```

<html>
<head><title>新闻</title></head>
<body>
<table>
<tr>
<td><a href="">市委考察组考察...</a></td>
<td><a href="">潍坊2社区获评...</a></td>
</tr>
<tr>
<td><table>
<tr><td><table>
<tr><td>践行市“两会”精神提升...</td></tr>
</tr></td>
<p>刚刚顺利闭幕的...</p><p>明晰发展思路...</p>
</td></tr>
</tr></td>

</td></tr>
</tr>
<td><a href="">信泰宜蓝湾项目举...</a></td>
<td><a href="">区长姜雪松主持召开...</a></td>
</tr>
</table></td></tr>
</table>
</body>
</html>

```

(a)网页文稿

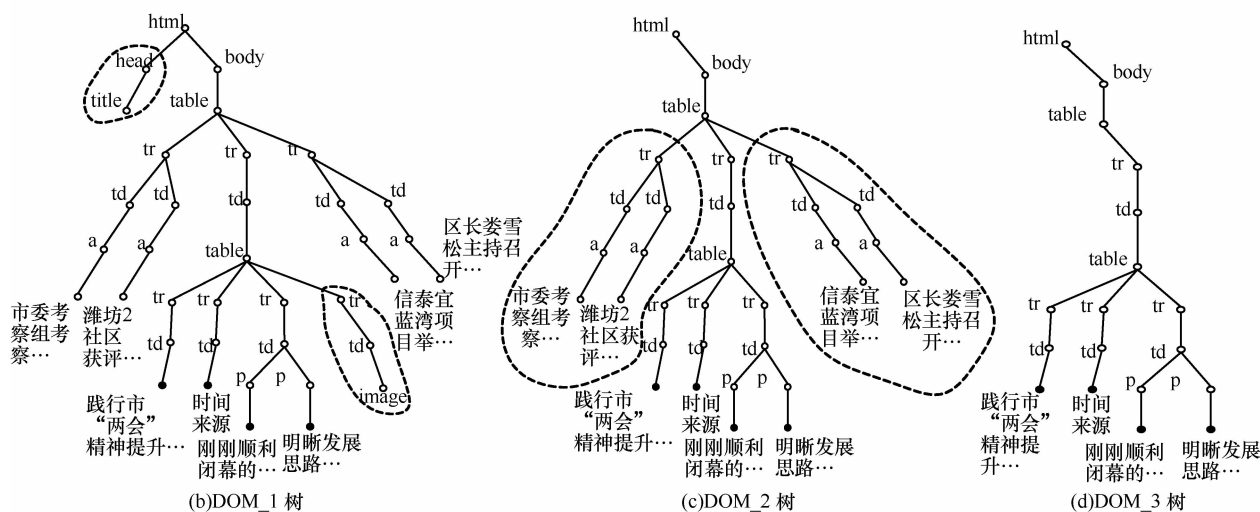


图2 HTMLParser 解析网页过程示意

息,是提高网页内容提取准确率和效率的一个关键。对于缺失标签、标签错乱等问题,HTMLParser 还能在 HTML 文档转换为 DOM 树的过程中自行修正,在此不再详述。

4.2 文本分词

文本分词作为文本挖掘的基础,主要是依据分词算法将提取后的网页内容汉字序列切分成一个一个单独的词,达到电脑自动识别语句含义的效果。

自 20 世纪 80 年代初,中文信息处理领域提出自动分词以来,很多高校科研机构都在该领域内取得了突破性的进展,研发了很多实用的分词系统,代表性的有:清华大学的 SEG 分词系统、北京大学计算语言所分词系统、复旦大学分词系统、微软研究院的多国语言处理平

台 NLPWin 及中国科学院 ICTCLAS 分词系统,具体系统评价可参考文献[3,4]。

上述分词系统实现的分词原理大多包括以下 3 种。

- 基于字符串匹配(词典)的分词:按照一定的策略将待切分的汉字序列与机器词典库中的词条进行匹配。按照扫描方向的不同,分词方法可以分为正向匹配、逆向匹配以及双向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配和最小(最短)匹配。相关统计结果表明,单纯使用正向最大匹配的错误率为 1/169,单纯使用逆向最大匹配的错误率为 1/245,建议使用逆向匹配的切分精度略高于正向匹配^[5]。



- 基于理解的分词:主要利用专家系统、神经网络等人工智能系统,分词的同时进行语义与句法的分析处理歧义现象。该类方法需要通过大量语料库的训练集学习后,在推理机中构建知识库。
- 基于统计的分词:计算词内汉字共现概率表达紧密程度,当紧密度高于阈值时,认定为词,该方法比较适用于专业文本分词。

分词技术的主要困难点在于歧义消除(不同的分词方法呈现不同的语义)和新生词汇识别问题。本文推荐利用词典分词与统计分词结合的方式,不仅具有词典匹配的快速分词效率,统计歧义词或新生词内的汉字共现频率将有助于进一步提高分词精度,如图3所示。

4.3 特征选择

分词处理后的结果,通过构建正则表达式等方法,去除常用感叹词、副词及虚词等停用词,余下的则用来表征网页文本特征向量,如式(1)所示:

$$doc_j = \{ \langle t_{j1}, w_{j1} \rangle \langle t_{j2}, w_{j2} \rangle, \dots, \langle t_{jn}, w_{jn} \rangle \} \quad (1)$$

其中 t_{ji} 是文档 j 中出现的第 i 个词, w_{ji} 是词 t_{ji} 在文档中的权值,一般可以定义为 t_{ji} 在文档中出现的频率函数。但这样得到的文档特征向量维度依然十分庞大。由于高维文本向量应用文本自动分类几乎很难实现,所以必须先经过降维处理,也就是特征选择。

经特征选择降维后的文本特征集应该包含两个特点^[6]:完全性和区分性。完全性,全面体现目标文本内容与主题;区分性,有效区分目标文本与其他文本。目前,国内外学者研究了众多的特征选择方法,其中最为常见的算法有 TFIDF (term frequency-inverse document frequency)、信息增益(IG)、互信息(MI)、统计法 CHI 等。

参考文献[7]的研究表明:信息增益主要通过特征在文本中的出现与否来度量,在遇到当类分布和特征项分布高度不平衡的情况时,由特征的不出现概率来评定该特征的信息增益,会导致该算法的特征提取表现不佳。通过学习参考文献[8,9],了解到互信息容易受到词条边缘概率密度的影响,如果两词条拥有相同的条件概率,频度低的反而有更高的相关信息量,其表现为过于倾向低频词,尤其当选择的训练文本和测试文本中有过多低频词时将直接影响后续分类效果。统计法 CHI 把特征与类别间的独立性类比为 χ^2 分布,往往偏重于考虑特征词在所有文档中出现的文档频数,对于少量文档中高频出现对分类贡献极大的特征容易被忽略^[7]。上述算法的理论基础不同,基于大量真实数据的实验证明,各个算法各有利弊,不存在任何一种算法在所有的数据集上都是最优的^[10-13]。本文考虑网页文本数据集海量的特性,推荐应用最为广泛且计算实现最为简便的特征加权技术 TFIDF 算法。

TFIDF 实际表示是 $TF \times IDF$, 其中 TF 表示词频(term frequency),即词在该文本中出现的次数;IDF 表示反文档频率(inverse document frequency),计算式如式(2)所示,表示词在整体语料库文本集中普遍重要性的度量。

$$IDF = \lg \frac{N}{n} \quad (2)$$

其中, N 为网页文本语料库全部文本数量, n 为包含词 t 的文本数量。

TFIDF 算法的主要依据是某一文本内的高频词以及该词在整个语料库文本集合中的低频率,可以产生出高权重的 TFIDF。因此,TFIDF 倾向于字词的重要性随着它

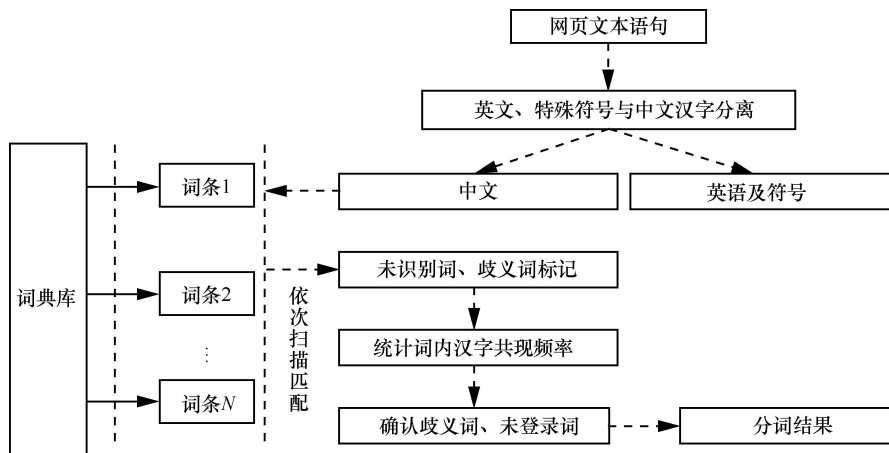


图3 词典分词与统计分词结合的分词机制

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,容易过滤掉常见的词语,保留重要的词语,适合用来作为分类的文本关键特征向量表示。

4.4 文本分类

文本分类环节主要基于网页文本的特征向量将每个网页文本归入预先定义的 URL 类别节点中。目前常见的文本分类器有以下几类:概率分类器,典型的如 Naive Bayes;决策树分类器,包括 ID3、C4.5、C5;神经网络分类器,如感知器法、logistic 回归、多层神经网络等;基于样本的分类器,即惰性学习器,典型的有 KNN;支持向量机 SVM 分类器^[14]。不同的分类算法性能具备差异:其中 Bayes、KNN 以及决策树的方法虽然效率较高,但其分类能力较弱。神经网络的最大缺点是过拟合,而防止过拟合很难实现,参见参考文献[15]。通过学习参考文献[16,17]发现,SVM 分类即便在样本分布不均衡的情况下,依然具备解决文本分类问题的出众性能,有效回避过拟合及冗余特征等问题,是进行网页文本分类的首选算法,本节进行重点介绍。

SVM 实现分类的主要途径,是通过选择非线性映射(核函数)将输入的文本向量映射到一个高维特征空间,在这个高维空间寻找最优的分类超平面,使各个样本间实现最大区分。但由于 SVM 最初就是一种典型的两类分类器,要解决的网页文本分类是个多类问题,利用 SVM 算法,以多个超平面把空间划分为多个区域,每个区域对应一个类别,一次性求解的方法计算量实在太大,大到无法实用的地步。

笔者建议采用 DAG SVM 方法,也称为有向无环图算法,实现网页文本分类。以 5 个类别的左向有向无环图算法为例:第一个分类器首先区分“1 类对 5 类”的归属判定,如果归属 5 类,分类器往左走,进入“2 类对 5 类”的分类器,如果判定还是归属 5 类,继续往左走,依次往下,直到得到最终分类结果,如图 4 所示。这样最终调用 4 个分类器(如果类别数为 K ,则只调用 $K-1$ 个),可以得到分类结果。该方法的好处是每个优化问题的规模比较小且分类效率高。

DAG SVM 算法的缺点在于如果上一个节点分类器出现错误,那么后面的分类器无法纠正错误,存在错误向下累积的现象。所以在分类器节点的布置上,笔者建议把差别大的排在前面,也就是把分类器按两类分类的

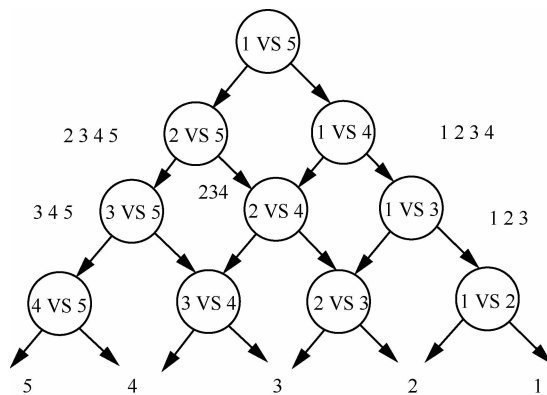


图4 有向无环图 SVM 过程说明

正确率从高到低排列,也可以考虑在每个两类分类器上都输出分类置信度,作为每个两类分类器结果准确度的参考依据。

5 结束语

移动互联网时代,电信运营商要在竞争日趋激烈的产业链上,取得更好的发展,就必须更好地适应市场和客户的需求。通过对 DPI 数据的深入文本挖掘,实现网页 URL 分类体系的自动映射挂载,从而获取用户行为特征分类,将为运营商全面洞察客户、构建客户全息视图提供依据,助力企业精准营销。

参考文献

- 1 罗忆祖. DPI 技术力助运营商精细化运营. 邮电设计技术, 2009(3)
- 2 于静. 基于页面主体提取的 Web 信息抽取技术研究. 南京邮电大学硕士学位论文, 2013
- 3 冯书晓, 徐新, 杨春梅. 国内外中文分词技术研究新进展. 情报杂志, 2002(11): 29~30
- 4 郭瞳康. 基于词典的中文分词技术研究. 哈尔滨理工大学硕士学位论文, 2010
- 5 李原. 中文文本分类中分词和特征选择方法研究. 吉林大学硕士学位论文, 2011
- 6 薛为民, 陆玉昌. 文本挖掘技术研究. 北京联合大学学报(自然科学版), 2005, 19(4): 59~63
- 7 宋江. 文本分类的特征选择方法研究. 南京航空航天大学硕士学位论文, 2010
- 8 王法波. 文本分类的特征选择和分类方法研究. 山东大学硕士学位论文, 2011
- 9 Liu H, Motoda H. Feature Extraction, Construction and Selection: A Data Mining Perspective. USA: Kluwer Academic, 1998



系数据库,借助一种动态流水线机制直接运行于 HDFS 之上。EMC 宣称,Greenplum Pivotal HD 结合了 SQL 和 Hadoop 的优势,而且速度“比 Hive 快数百倍”^[5]。

或许,在不久的将来,无需再集成不同的产品以实现分布式文件系统与 MPP 数据库的混搭,而是仅仅通过一个融合化的商业产品就能够实现两种技术的完美混搭。

参考文献

- 1 周宝曜,刘伟,范承工.大数据:战略、技术、实践.北京:电子工业出版社,2013
- 2 Apache HBase. <http://hbase.apache.org/>
- 3 Big data debate: will HBase dominate NoSQL. <http://www.informationweek.com/software/enterprise-applications/big-data-debate-will-hbase-dominate-nosql/240159475>
- 4 MapReduce and MPP: two sides of the big data coin. <http://www.zdnet.com/blog/big-datamapreduce-and-mpp-two-sides-of-the-big-data-coin/121>
- 5 Greenplum Pivotal HD 结合了 SQL 和 Hadoop 的优势. <http://www.infoq.com/cn/news/2013/03/Pivotal-HD-SQL-Hadoop>

(上接第 11 页)

- 10 Jain A, Zongker D. Feature selection: evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(2):153~158
- 11 Gorvan A. Principal Manifolds for Data Visualisation and Dimension Reduction. New York: Springer, 2007
- 12 Verleysen M, Lee J A. Rank-based quality assessment of nonlinear dimensionality reduction. Proceedings of the 16th European Symposium on Artificial Neural Networks, Bruges, Belgium, 2008:49~54
- 13 Deerwester S. Indexing by latent semantic analysis. Journal of American Society for Information Science, 1990, 41(6):391~407
- 14 陈燃燃. 基于 SVM 算法的 Web 分类研究与实现. 北京邮电大学硕士学位论文, 2009
- 15 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer, 1995
- 16 Joachims T. Text categorization with support vector machines: learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 1998:137~142
- 17 Joachims T. Transductive inference for text classification using support vector machines. Proceedings of the 16th International

[作者简介]



张雨,男,中国电信股份有限公司上海分公司高级工程师、副总工程师,主要研究方向为 IT 规划、大数据。

蔡鑫,男,中国电信股份有限公司上海研究院高级工程师、数据分析师,主要研究方向为数据规划、信息管理、数据标准。

李爱民,女,中国电信集团公司企业信息化事业部工程师、数据处处长,主要研究方向为信息管理。

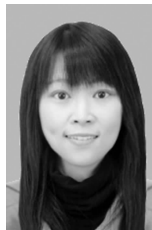
刘翼,男,中国电信集团公司企业信息化事业部工程师、业务主管,主要研究方向为数据规划、信息管理。

姚晓辉,男,中国电信股份有限公司上海研究院工程师,主要研究方向为数据规划、数据挖掘、信息管理。

(收稿日期:2013-10-25)

Conference on Machine Learning, Bled, Slovenia, 1999: 200~209

[作者简介]



边凌燕,女,中国电信股份有限公司上海研究院工程师,主要研究方向为数据分析、数据挖掘。



贺仁龙,男,中国电信股份有限公司上海研究院工程师,主要研究方向为信息管理。

姚晓辉,男,中国电信股份有限公司上海研究院工程师,主要研究方向为数据规划、数据挖掘、信息管理。

(收稿日期:2013-10-29)