电信大数据文本挖掘算法及应用

汪东升 黄传河 黄晓鹏 倪秋芬
(武汉大学计算机学院 武汉 430072)

摘 要 电信大数据中包含了大量的非结构化文本数据,无法通过常规的方法进行信息挖掘,在此情况下文本挖掘可以更好地实现对文本数据的分析挖掘。提出了基于文本的新词识别算法和命名实体识别算法,从而有效地分析用户投诉文本内容并判断其所属类别,并且从用户上网文本信息中识别出其终端型号,为电信行业提供更好的用户支撑和用户体验。最后,对模型的实际应用表明,所提方法对电信投诉文本数据的识别是高效的。

关键词 电信,大数据,文本挖掘,模型识别,用户终端机型
中图法分类号 TP392 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.042

Text Mining Algorithm and Application of Telecom Big Data

WANG Dong-sheng HUANG Chuan-he HUANG Xiao-peng NI Qiu-fen
(Computer School, Wuhan University, Wuhan 430072, China)

Abstract Major telecom data contain a large number of unstructured text data, which are difficult for conventional methods to mine information. Text mining can do better than conventional methods under this circumstance. Based on the text data, this paper proposed a new word identification algorithm and a named entity recognition algorithm. At this process, we analyzed the customers' complaint texts and judged their categories, and then identified the user's terminal types from their information, which provides better user supports and experiences for the telecom industry. Experiment results validate that the proposed algorithm achieves good performance for the identification of customers' complaint texts in the telecom.

Keywords Telecom, Big data, Text mining, Pattern recognition, User's terminal types

1 背景

随着互联网及移动互联网的快速发展,我国电信行业展现出勃勃生机,截止2015年年底,电信用户总数已超过13亿,市场饱和引起三大电信行业运营商对用户的激烈竞争。2015年5~10月间,中国移动手机用户数增加了764万,中国联通减少了304万,中国电信增加了486万[1-2]。

面对激烈的竞争环境,更好地维系现有用户并从中获得更大利益成为每个电信企业提升自身价值最重要的问题,而解决该问题的核心在于投诉处理和用户特征识别。

投诉处理主要包括客服记录用户投诉内容、形成投诉工单、分类派发至相应部门等环节。投诉处理中存在不能对大量的用户投诉文本进行自动分析和精准分类的问题。以某省运营商为例,月均电话投诉量达到10000通以上,由于客服人员能力参差不齐,对投诉的分类准确率仅60%,导致用户投诉问题无法及时得到解决。

用户特征识别依靠数据挖掘分析电信基础数据(如用户基本信息、消费情况、通话行为等),只能获得用户的一般特征且不够准确,缺乏对大量文本数据的专项分析。若某用户以自己的身份证开通家庭套餐后供一家人共同使用,则无法识别每个用户的性别、年龄等特征。还有更多用户相关的非结构化数据未能得到充分利用,如利用用户移动互联网使用行为数据可识别出每个用户使用的终端型号。

可见,以上两个问题都需要基于大数据的文本挖掘技术对大量文本进行精准分析。文本挖掘可定义为在一个知识密集型的处理过程中使用一套分析工具来处理文本集。通过识别和检索有价值的模式,从数据源中抽取有用信息[3]。本文接下来将围绕以上两个问题中涉及的文本挖掘技术的特点及应用实践进行详细的阐述。

2 文本挖掘技术

文本挖掘的主要步骤如图1所示[4]。

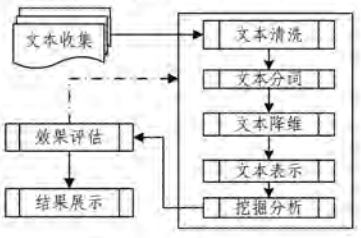

图1 文本挖掘流程

到稿日期:2016-10-10 返修日期:2017-01-12 本文受国家自然科学基金(61373040,61572370)资助。

汪东升 男,博士生,主要研究方向为数据挖掘,E-mail:wangds@tydic.com;**黄传河**(1963-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为计算机网络,E-mail:huangch@whu.edu.cn;**黄晓鹏** 男,硕士,主要研究方向为计算机网络;**倪秋芬** 女,博士生,主要研究方向为无线网络、车联网,E-mail:niqiufen@whu.edu.cn。

2.1 文本收集

目前，电信行业根据不同挖掘需求可以通过多种技术手段从多种渠道获取不同的原始文本：如从 10000 号平台、网上营业厅、工信部申诉中心等收集用户投诉文本；从大数据系统收集用户上网的浏览器信息文本、cookies 文本等；通过爬虫从互联网获取用户评论文本等。

2.2 文本清洗

收集文本数据之后，需要从以下几个方面进行清洗[5]，以减少文本规模，提升文本质量。

1)格式：如剔除长度过短的投诉文本、从复杂页面文本中提取关键信息文本等；

2)编码：如对大数据系统 DPI 数据文本进行 url 编码转换、字符集转换等；

3)业务：结合业务经验和实际情况制定特定行业的清洗规则，如电信行业中用户对同一问题进行多次投诉，只取一次进行分析。

2.3 文本分词

文本分词是基于词库使用分词算法，将非结构化的文本内容切分成单个具有独立含义的词。需要识别新词并及时补充基础词库和专业词库[6-7]，避免错误切分引起的信息损失对文本挖掘效果的影响。如电信行业的“土豪金”“e 家理财”等新词不在现有词库中，与之相关的语句容易被错误切分为“土豪/金”“e/家/理财”。

我们设计的新词识别算法 New_word 如算法 1 所示。

算法 1 New_word

```
输入:凝固度阈值、自由度阈值、高频词阈值
输出:新词集 New_word
for(i=1:N) //第几个切分词
word_left = 读入第几个切分词;
  if(count(word_left) >= cost_T) //该词是高频词
    for(idx_word_left in (1...Num_Word)) //该高频词在文本中的
    某一个位置
      if(count(word_right) >= cost_T) //其右侧是高频词
        Degree_Solid = [word_left,word_right]//凝固度;
          Degree_Free = [word_left,word_right]//自由度;
          Degree_Freq = [word_left,word_right]//频度;
If (Degree_Solid >= Const_Solid & Degree_Free >= Const_Free &
Degree_Freq >= Const_Freq)
New_Word = word_left+word_right;
```

本文提出的识别新词算法 New_word 的时间复杂度为 $O(n^2)$，比其他的识别算法有更好的新词获取率。

2.4 文本表示

分词后的文本通过计算文本中词条出现的频度来构造“文本-词条矩阵”，即向量空间模型(VSM)。以投诉文本为例，将每一个文本表示成多维向量，每一个维度即为一个特征词，维度值为该词在该文本中出现的词频(或加权处理)，如表 1 所列。

表 1　VSM 中分词的取值

文本编号	QQ	余额	IPTV	停机	彩铃	机顶盒	…
1	0	2	0	0	1	0	…
2	0	0	5	0	0	2	…
3	1	0	0	2	0	0	…
4	0		3	0	0	0	…

2.5 文本降维

由于向量空间模型中的维度数量成千上万，因此为了提升后续文本挖掘的效率，主要使用以下方法进行降维[8]：

1)基础过滤：结合词库剔除停用词、无业务价值的词或数字，合并同类字符串等。

2)特征频率：剔除低频词。实验证明仅使用部分高频词并不会降低分类器的性能[3]。

3)文档频率：如果包含某个词的文档数量过多或者过少，均需剔除该词。

4)卡方检验：使用卡方值度量特征词与类别之间的相关程度。卡方值越大，特征词对分类的贡献就越大。

2.6 挖掘分析

2.6.1 文本分类算法

通过上述步骤得到可用的文本数据集后，使用文本分类算法生成分类规则集，常用的文本分类算法有 Rocchio、KNN、决策树、贝叶斯、SVM 等[8]。文献[9]提出了两个基于文本分类的不同算法：Rocchio 和 kNN，以解决网络中的分层数据集分类问题。结果表明，Rocchio 算法的文本分类效率和精确度均比 kNN 算法要好。文献[10]提出了使用基于 Rocchio 关联反馈的 MapReduce 算法，该算法在 MapReduce 范式中提升了传统 Rocchio 算法的精确度，提升了执行速度和效率，解决了大量信息过滤问题。对于构建域指定本体论，文献[11]使用了网络上大的文本数据集，在本体构建过程完成之前将其组织成特定域，然后实现了朴素贝叶斯文本分类器并使用映射规约程序设计模型来组织大文本集。实验中，使用维基百科在线百科全书中的动物和植物域文章作为数据集。实验表明提供了一个精确度高达 98.8%和鲁棒性高的，用于在特定域本体构建时，在特定域的预处理阶段进行文件分类的方法。本文重点使用决策树和 SVM 算法。

决策树：选择信息增益最大的属性作为判定的分支节点，增益越大表示其分裂后各分支间的差异越大，有助于迅速判断样本所属的子集。信息增益的概念来自于信息熵，计算公式如下：

$$\begin{aligned} Gain(t) &= Entropy(S) - Expected\ Entropy(S_t) \\ &= \{-\sum_{i=1}^{M} P(c_i)\log P(c_i)\} - [P(t)\{-\sum_{i=1}^{M} P(c_i \mid t) \\ &\quad \log P(c_i \mid t)\} + P(\bar{t})\{-\sum_{i=1}^{M} P(c_i \mid \bar{t})\log P(c_i \mid \\ &\quad \bar{t})\}] \end{aligned} \tag{1}$$

如图 2 所示，若文本包含爱游戏则属于移动增值业务类，包含故障和送修则属于移动故障预处理类。

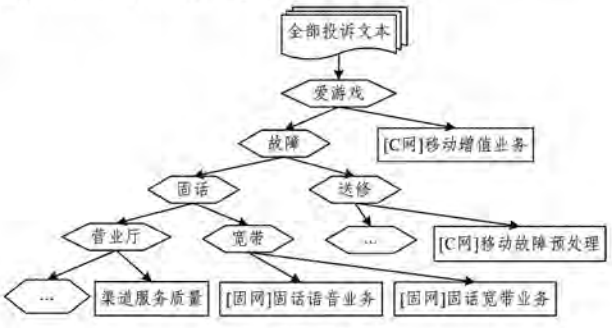

图 2　决策树分类器

SVM:在机器学习领域,支持向量机是一个有监督的学习模型,通常被用于模式识别、分类以及回归分析。该算法的关键在于核函数的选择[7]:

1)多项式核函数:$K(x,y)=[a(x\cdot y)+b]^q$

2)径向基函数:$K(x,y)=\exp(-\gamma\|x-y\|^2)$

3)双曲正切核函数:$K(x,y)=\tanh[a(x\cdot y)+b]$

由于不同的需要提供的内容不一样,因此如果把所有的投诉文本放在一起不仅不实际,而且分析效果也不理想。因此本文先分析了不同的文本,然后分别对不同的文本进行识别。如图 3 所示,实线圆表示固话宽带业务类的投诉文本,虚线圆表示渠道服务质量类的投诉文本。固化带宽业务投诉可以归结为硬件设备如光纤、上网、宽带等可以判断的因素。而渠道服务质量包含许多人为且不可量化的因素,如柜台、答应、欺骗等。高维空间中存在一个超平面,使得两个类别的样本都距离该平面尽可能远的距离。而我们需要解决的是固化宽带业务类的投诉问题。

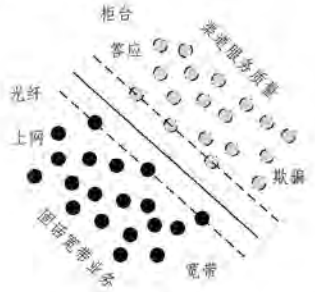

图 3 SVM 分类器

2.6.2 命名实体识别算法

从大数据中可以提取到用户的不同特征,本文仅介绍如何识别用户的终端使用特征,即用户使用的具体终端型号。

以终端型号命名实体的识别为例,设计了以下几种互为补充的算法。

(1)机型别名识别算法:电信的终端机型库中收集了多款终端的品牌、标准型号等详细信息,但在实际大数据文本中包含的机型名称并不规范,无法直接与终端机型库中的标准型号进行匹配。如终端库中的机型为 HW HUAWEI C8500,但在 DPI 中为 HW C8500 或者 HUAWEI C8500,由于它们本质上是同一款终端的不同别名,因此需要对终端库所有型号进行扩充。

机型别名识别算法如算法 2 所示。

算法 2 Term_Byname

```
输入:终端机型库 Term_Type,阈值
输出:所有机型 ALL_Term
for(i=1:n in Term_Type)
Model_res = 读入第 i 个型号;
  [seg1,seg2,seg3,seg4] =split(Model_res);
  if(seg3<>null) //可以确保 seg1 和 seg2 不为空
    if(seg1+seg2 组合未记录) then
      记录 seg1+seg2 组合,count=1;
else
      找到已记录的 seg1+seg2 组合,count++;
for(j=1:m1) //已记录的 seg1+seg2 组合数量
读入第 j 对已记录的 seg1+seg2 组合;
  if(count>=Const_Cnt)
    pair_12 = [seg1,seg2]; //记录高频的组合对
for(k1=1:n)//机型库中的型号数量
N_Model_res = 读入第 k1 个型号;
  [Nseg1,Nseg2,Nseg3,Nseg4] = split(N_Model_res);
  for(k2=1:m2) //高频 seg1+seg2 组合数量
读入第 k2 对高频组合[seg1,seg2];
    if(Nseg1=seg1)
      if(Nseg2=seg2)
        if(Nseg3<>null)
          Byname. append = Nseg1+Nseg3+Nseg4;
          Byname. append = Nseg2+Nseg3+Nseg4;
          Byname. append = Nseg3+Nseg4;
        end
      elseif(Nseg2<>seg2)
        if(Nseg2<>null)
          Byname. append = seg2+Nseg2+Nseg3;
          Byname. append = Nseg2+Nseg3;
All_Term = Remove_Repetitions(Term_Type+Byname);
```

本文提出的机型别名算法 Term_Byname 的时间复杂度为 $O(n*m^2)$。

(2)命名规则识别算法:很多品牌的手机系列有固定的命名方式,通过已有的终端机型库可以概括出这些命名规则,有助于从 DPI 数据中全面识别更多暂未收录的机型。

命名规则识别算法如算法 3 所示。

算法 3 Term_Rule

```
输入:终端机型库 Term_Type
输出:规则机型 All_Rule
for(i=1:n in Term_Type)
Model_res = 读入第 i 个型号;
  if(REGEXP_LIKE(Model_res,[\d])) //包含数字 0~9
    [seg1,seg2,seg3,seg4]=split(Model_res); //切分 4 段
    Tmp_Rule = replace([\d],'*'); //将数字替换为星号
    if(Tmp_Rule 未记录) then
    Rules. append(Temp_Rule),count=1;
  else
    Rules. find(Temp_Rule),count++;
All_Rule = Remove_Repetitions(Rules);
```

本文提出的命名规则识别算法 Term_Rule 的时间复杂度为 $O(n)$。

(3)潜在机型识别算法:上述方法得到的各种机型与大数据匹配后仍会存在识别遗漏,主要是由于有些机型不符合上述格式或两侧分隔不显著。该算法通过模糊匹配识别出疑似机型的内容供人工审核后使用。

潜在机型识别算法如算法 4 所示。

算法 4 Term_Doubt

```
输入:终端机型库 Term_Type
输出:疑似机型 Term_doubt
for(i=1:n in Term_Type)
Model_res = 读入第 i 个型号;
  for(j=1:切分成的段数)
    temp=Model_res. word(j);//取切分后的第 j 词
```

```
    if(temp未记录) then
      seg.append(temp),cnt=1;
    else
      seg.find(temp),cnt++;
for(j=1:m in seg)
  if(cnt>=Const_Cnt) //机型分词高频词
    High_Freq_Seg.append(seg(j));
UA=读取DPI中一条UA信息;
for(k1=1:length(count(UA,Splitter)))
{
  if(match(UA.word(k1),High_Freq_Seg)&&match(UA.wor-
  d(k1+1),High_Freq_Seg))//本身和右边都是高频词
  Term_Doubt.append(UA.word(k1:k1+1));
}
for(k2=1:n in Term_Type)
  if(match(UA,Term_Type(k2))) //UA中包含机型
    if(~match(UA,Term_Type(k2)+' '))
      Term_Doubt.append(机型及其后部分);
      Term_Doubt = Remove_Repetitions(Term_Doubt);
```

本文提出的潜在机型识别算法 Term_Doubt 的时间复杂度为 $O(n)$。

3　应用

3.1　投诉文本分类模型

用户拨打运营商客服号码进行投诉时，客服人员记录投诉内容，系统根据已有的方法识别出该投诉内容所属的类别，之后该投诉工单会派发至相应的岗位进行处理。

3.1.1　基于统计的分类方法

在应用投诉文本分类模型之前，客服系统主要依靠基于统计的方法对文本进行分类。该方法需要人工统计历史各类投诉文本的分词以及提取不同类别下的关键词组合，其流程如图 4 所示。

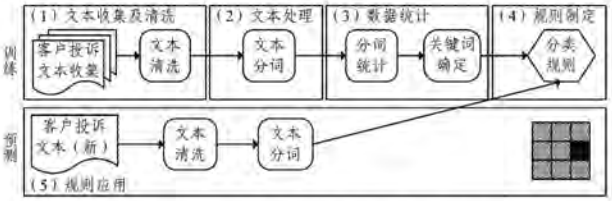

图 4　基于关键词统计的投诉文本分类流程

(1)文本收集及清洗

整合不同渠道的投诉文本并对文本进行清洗。目前主要从运营商电子渠道的客服平台收集客户历史投诉文本，删除其中内容异常的文本并人工审核纠正少量被错误分类的情况。

(2)文本处理

对清洗后的投诉文本进行文本分词，提取重要的特征信息并将其展现为向量空间格式。使用分词工具对每条投诉文本进行分词，常用的分词工具有 jieba 分词、IKAnalyzer 等。

(3)数据统计

使用分词工具对文本内容进行分词，统计每个词的频度，并确定每个类别的正反向关键词。

1)分词统计：统计每个分词在各个类别中的数量和占比（如词频、类别下的文本数等），并结合每个类别下的工单数量对占比进行修正。

2)关键词确定：选取总词频大于 100 且各类别中最大占比大于 50%的词，作为后续制定分类规则的关键词。

(4)规则制定

根据每类文本中关键词出现的情况，规定文本分类的判断规则。

1)分值计算：假设某文本 W 分词后各词为：$W1/W2/W3\cdots$，关键词集合为 K，则该文本对应类别 A 的分值为：

$$mark(A)=\sum_{i=1}^{n}WK_i\times K_{Ai},WK\in W\cap K \tag{2}$$

其中，WK_i 属于该文本分词与关键词的交集。

2)类别确定：由式(2)计算该文本所属各类别的分数，取分数最大的类别为该文本所属类别。

(5)规则应用

对新的投诉文本进行文本清洗和分词处理后，使用分类规则集判断其所属类别。

3.1.2　基于模型的分类方法

为了提升分类准确率、加快投诉处理的效率并提高客户满意度，需要识别客户投诉的文本内容，建立投诉文本分类模型。整个建模流程如图 5 所示。

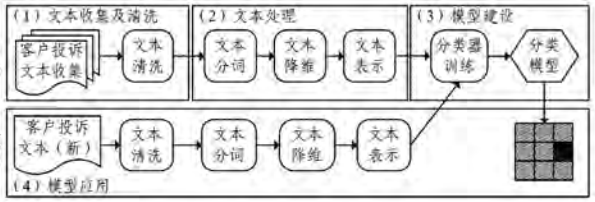

图 5　投诉文本分类挖掘过程

(1)文本收集及清洗

该过程主要包括以下 5 个子任务。

1)收集：从各渠道收集客户的投诉文本；

2)探索：分析每个月每类工单的数量；

3)初选：根据探索结果初步筛选样本；

4)清洗：与客服专家逐条阅读投诉文本，纠正少量被错误分类的情况；

5)集合：形成建模所需的训练集和测试集。

(2)文本处理

对清洗后的投诉文本进行文本分词，提取重要特征信息，并将文本表示为向量空间格式，又分为文本分词、文本降维、文本表示。

(3)模型建设

需要对训练文本集使用文本分类算法生成分类规则集，由于投诉文本的类别包含多个层级，有以下两种建模方案可供选择：

方案 1　精细模型，对每一层的每一个子类都构建一个专用模型，可以精准地预测每个工单是否属于这个子类。

方案 2　粗放模型，对每一层构建一个模型，直接预测每个工单在本层的所属类别，如图 6 所示。

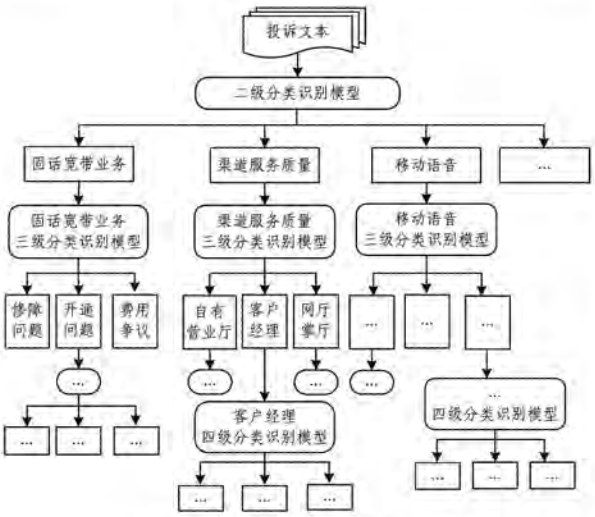

图 6 投诉文本分类的粗放模型

对各省投诉文本进行大量数据探索后，选择了方案 2；对比不同分类算法，最终确定采用支撑向量机(SVM)算法进行分类。

(4)模型应用

通过各层模型综合预测得到投诉文本所属的类别。

3.1.3 效果对比分析

取 Y 省的 47789 条已人工分类的历史投诉文本为测试集，经检验，采用基于统计的方法后分类准确的文本数为 32036 条，准确率为 67.04%；采用基于模型的方法后分类准确的文本数为 41190 条，准确率达 86.19%，效果提升显著，如图 7 所示。

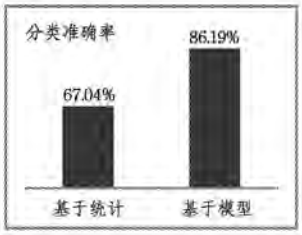

图 7 两种方法的分类准确率对比

从抽样检查的结果来看，前期基于统计的分类方法中有些文本被错误归类，但模型预测的类别正确(由客服部门审核确认)。人工分类和模型预测的对比如表 2 所列。

表 2 人工分类和模型预测的对比

投诉内容	人工错误判断的类别	模型正确预测的类别
业务号码：56＊＊＊54，56＊＊＊13 我们的电话都无法使用，你们一直都没有处理，请你们尽快安排查修。	[固网]固话宽带业务—>修障问题—>修障不及时	[固网]固话语音业务—>修障问题—>修障不及时

3.2 基于 DPI 的用户终端识别模型

运营商终端自注册平台中记录了用户从营业厅、网厅等自有渠道购买的终端的具体型号，而用户从其他渠道购买的手机则无法识别。若识别率过低则无法准确掌握用户的终端使用情况、换机轨迹，也无法挖掘用户潜在的终端使用需求。因此通过用户上网日志进行辅助识别。用户在使用智能手机上网时，浏览器及部分 APP 会记录该手机的型号，可通过文本挖掘进行识别。

现在对于实体识别的主流方法是条件随机场 CRF(Conditional Random Field)。自上而下的视觉显著性是视觉注意中的一个重要的模块，文献[12]提出了一个新颖的自上而下的显著性模块，该模块联合了一个条件随机场 CRF(Conditional Random Field)和一个视觉词典。提出的模型联合了一个从顶部到底部的分层结构：CRF，稀疏编码和图像块。将稀疏编码作为中间层，文中以监督的方式学习了随着 CRF 层结构化输出的一个字典；同时，将稀疏编码作为特征学习了 CRF 参数。对于高效的联合学习，本文提出了一个通过随机梯度凝递减算法实现的最大边界方法。

3.2.1 基于匹配的识别方法

用户使用手机等终端上网时所产生的日志中包含的 UserAgent(UA)信息记录了上网终端的操作系统及版本、浏览器及版本等内容。目前运营商普遍使用的机型识别方法将 UA 信息与已有的各种机型名称进行匹配来识别其中包含的终端型号。基于匹配的识别方法如图 8 所示。

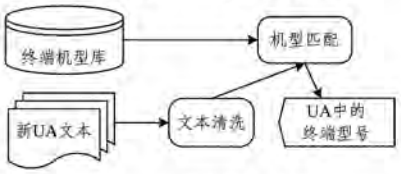

图 8 基于匹配的识别方法

1)终端机型库准备

通过网络爬取与人工录入的方式收集市场常见的各类终端信息，包括终端品牌、型号、价格、屏幕尺寸、分辨率等，当前终端机型库包含 2200 款终端。

2)新 UA 文本清洗

从待识别的大量 UA 数据中筛选出长度大于 10 个字符的记录，统一转换为大写。

3)机型匹配

将清洗后的每条 UA 与机型库中的每个型号进行循环匹配，成功则输出型号。基于匹配的识别结果如表 3 所列。

表 3 基于匹配的识别结果

UA(部分内容)	识别型号	终端库中对应型号
SCH-I679 Build/JLS36C	SCH I679	
SCH-W2013 Build/IMM76D	SCH W2013	
HUAWEI C8813 Build/Huawei		HW C8813
MI 3C Build/JLS36C	MI 3C	
YL-Coolpad_5890/4		YL 5890
HUAWEI A199 Build		HW A199
SM-N9009 Build/KOT49H		SCH N9009
vivo X3V Build/KVT	VIVO X3V	
ZTE Q701C Build/JLS36C	ZTE Q701C	
YL-Coolpad_5010/50		YL 5010
HTC D816d Build/KOT49H)	HTC D816D	

由表 3 可知，UA 中的型号与终端库中的型号不能完全匹配，因此很多 UA 中包含的型号内容无法被识别。基于匹配的识别方法严重受限于终端机型库的完整性，在 Y 省的识别率只有 65%左右，因此需要设计一种从大数据的 UA 文本内容中挖掘出用户使用的终端型号的科学识别方法。

3.2.2 基于模型的识别方法

基于模型的识别方法在文本收集及清洗、文本处理等环

节采用文本挖掘标准方法，在模型建设环节采用我们自主设计的识别算法，从而实现了用户机型智能识别、终端型号智能补充两大功能。整个建模流程如图 9 所示。

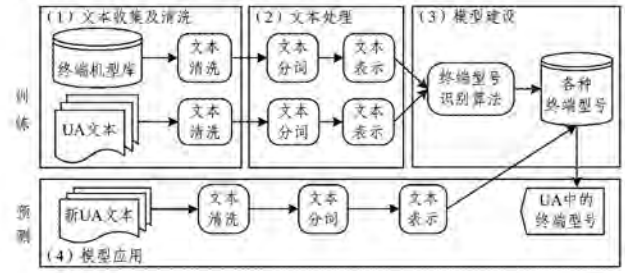

图 9　基于 DPI 的用户终端识别建模流程

(1)文本收集及清洗

准备已有的终端机型库，并提供一部分 UA 文本数据对数据进行文本清洗。终端机型库有 14000 多个型号，全部采用大写格式；UA 文本提供部分 UA 数据作为训练集进行清洗。

(2)文本处理

对机型名称进行分词表示，对 UA 文本进行处理，便于后续算法识别新机型的名称。

(3)模型建设

设计算法识别各种机型的别名和命名规则，并从 UA 文本中识别出疑似机型的内容。

(4)模型应用

算法规则固化至大数据平台之后，对新的 UA 文本的识别效果如表 4 所列。

表 4　基于模型的识别结果

UA(部分内容)	识别型号
SCH-I679 Build/JLS36C	SCH I679
SCH-W2013 Build/IMM76D	SCH W2013
HUAWEI C8813 Build/Huawei	HUAWEI C8813
MI 3C Build/JLS36C	MI 3C
YL-Coolpad_5890/4	YL COOLPAD5890
HUAWEI A199 Build	HUAWEI A199
SM-N9009 Build/KOT49H	SM N9009
vivo X3V Build/KVT	VIVO X3V
ZTE Q701C Build/JLS36C	ZTE Q701C
YL-Coolpad_5010/50	YL COOLPAD 5010
HTC D816d Build/KOT49H)	HTC D816D

3.2.3　效果对比分析

在各省份的实际应用中，基于模型的识别方法表现出了良好的识别效果。以 Y 省为例，平均每天有 300 万用户使用手机上网，产生的 DPI 数据为 3T，算法处理时长为 4h，月累计识别率高达 96%，远高于基于匹配方法 65%的识别率(见图 10)。

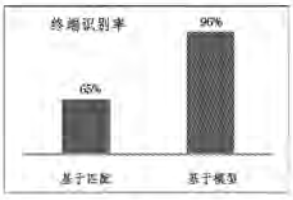

图 10　终端识别率

Y 省大数据集群正在扩容，仅采用 20 个节点进行计算，得出数据量固定在 3T 时，处理时间与节点个数呈高度拟合幂函数关系，如图 11 所示。

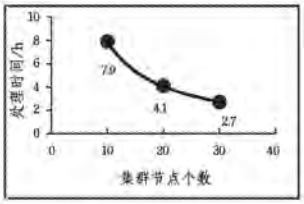

图 11　节点个数与处理时间关系图

由经验可知在节点数固定时，处理时间与数据量呈近乎线性关系。因此在用户数大、数据量亦大的省份，需大幅提升节点数来进行处理。以 S 省某运营商为例，用户上网的活跃度较高，五千万用户每天产生约 40T 的 DPI 数据，根据上述函数估算，采用 1000 个节点进行计算，约需要 1.15h。

由于近几年没有相同方法相同场景的参考文献可供比较，因此实现结果无法对比，而与其他方法或其他场景作对比也毫无意义。

结束语　用户投诉处理和用户特征识别是电信行业维系当前用户、提升用户贡献的两个重要问题。本文实现了投诉文本的精准分类、对用户使用终端型号进行精准识别这两个功能，改善了目前投诉文本人工分类效率、准确率及用户终端识别覆盖率低的现状。但现有的工作还存在以下几方面的不足：

1)没有考虑到方言的影响。比如北京话中“局器”表示仗义，陕西话中“谝”表示闲聊，四川话中“扯筋”表示吵架，有些方言词虽然频次不高但可能意义重大，容易被忽略。

2)未进行同义词合并。比如告诉与告知、安置与安放，若同义的几个词的出现次数都偏低，则可能都被剔除，从而影响分析效果。

3)未考虑词出现的位置和上下文。一般句首句尾出现的词的重要性更高[7]，且结合上下文能进一步推断该词是否为所需内容。

参 考 文 献

[1] SENBALC C, ALTUNTAS S, BOZKUS Z, et al. Big data paltform development with a domain specific language for telecom industries [C]// High Capacity Optical Networks and Emerging/Enabling Technologies. 2013: 116-120.

[2] TSENG J C, TSENG H C, LIU C W. A successful application of big storage techniques implemented to criminal investigation for telecom [C]// Network Operations and Management Symposium (APNOMS). 2013: 1-3.

[3] JONY R I, HABIB A, MOHANMMED N, et al. Big Data Use Case Domains for Telecom Operates [C]// IEEE International Conference on Smart City/SocialCom/SustainCom. 2015: 850-855.

[4] ZHONG N, LI Y F. Effective Pattern Discovery for Text Mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(1): 30-44.

[5] ELAGIB S B, HASHIM A H A, OLANREWAJU R F. CDR

analysis using Big Data technology[C]// International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE). 2015:467-471.

[6] DAM R V D. Big Data a Sure Thing for Telecommunications: Telecom's Future in Big Data [C]// Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). 2013:148-154.

[7] OUYANG Y, HU M M, HUET A, et al. Mining of leaders in mobile telecom social networks [C]// Wireless Telecommunications Symposium (WTS). 2016:1-4.

[8] HUANG W L, CHEN Z, DONG W Y, et al. Mobile Internet big data platform in China Unicom [J]. Tsinghua Science and Technology, 2014, 19(1):95-101.

[9] CHETAN S B J, SRINIVASA K G. Large Scale Multi-label Text Classification of a Hierarchical Dataset using Rocchio algorithm [C]// International Conference on Computational Systems and Information Systems for Sustainable Solutions. 2016:291-296.

[10] YANG W C, FU Y M, ZHANG D. An Improved Parallel Algorithm for Text Categorization [C]// International Symposium on Computer, Consumer and Control. 2016:451-454.

[11] SANTOSO J, YUNIARNO E M, HARIADI M. Large Scale Text Classification using Map Reduce and Naïve Bayes Algorithm for Domain Specified Ontology Building [C]// 7th International Conference on Intelligent Human-Machine Systems and Cybernetics. 2015:428-432.

[12] YANG J, YANG M H. Top-Down Visual Saliency via Joint CRF and Dictionary Learning [C]// Computer Vision and Pattern Recognition. IEEE, 2012:2296-2303.

(上接第 220 页)

可见由粗颗粒度到细颗粒度的多层次纠正，可纠正子句间、短语间的错误词对齐；与 Baseline 相比，实验结果提高了 0.99，证明了本文提出的自纠正词对齐方法能有效提高词对齐质量和机器翻译的质量。Cat_{P2} 的 BLEU 值与 Cat_{P1} 相比，依旧有所改善；与 Baseline 相比提高了 1.37，提升效果显著。

结束语 本文提出了一个针对词对齐的自纠正机制，借助于语言特征等先验知识，对词对齐进行多轮循环自纠正。在粗颗粒度到细颗粒度的纠正过程中，首先在粗颗粒度的级别上采用基于标点的纠正方法，对原始双语句对进行句子级别的切分，该方法保证了句法结构的完整，方法简单有效，准确率高达 95.6%，AER 结果改善明显；然后在子句颗粒度的级别上采用基于指示词的纠正方法，对上述切分的子句进行细颗粒度切分，发现准确率偏低，为 72.0%，而且 AER 结果并没有得到改善；另外，在短语级别的颗粒度上采用基于统计特征的纠正方法，对上一轮切分的子句进行细颗粒度切分，准确率高达 93.3%，AER 结果较基于标点的切分方法进一步得到改善。通过分析发现，在细颗粒度的词对齐纠正中，基于统计特征的切分效果明显优于基于指示词的切分，因此，本文在自动纠正词对齐过程中，先采用基于标点的方法、后采用基于统计特征的方法对双语句对进行纠正，最后将切分的子句合并成完整句子，翻译质量得到了显著提升。

参 考 文 献

[1] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003:127-133.

[2] LIU Y, LIU Q, LIN S. Tree-to-string alignment template for statistical machine translation[C]// International Conference on Computational Linguistics and, Meeting of the Association for Computational Linguistics(ACL 2006). Sydney, 2006:609-616.

[3] GALLEY M, GRAEHL J, KNIGH K, et al. Scalable inference and training of context-rich syntactic translation models[C]// International Conference on Computational Linguistics and the, Meeting of the Association for Computational Linguistics. 2012:961-968.

[4] CHIANG D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2):201-228.

[5] BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2):263-311.

[6] LIANG P, TASKAR B, KLEIN D. Alignment by agreement [C]// North American Association for Computational Linguistics (NAACL). 2006.

[7] XU J, ZENS R, NEY H. Partitioning parallel documents using binary segmentation[C]// The Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2006:78-85.

[8] BLUNSOM P, COHN T, GOLDWATER S, et al. A Note on the Implementation of Hierarchical Dirichlet Processes[C]// International Joint Conference on Natural Language Processing of the Afnlp. DBLP, 2009:337-340.

[9] GAO Q, VOGEL S. Parallel implementations of word alignment tool[C]// Association for Computational Linguistics. 2008:49-57.

[10] STOLCKE A. SRILM-an extensible language modeling toolkit [C]// Proceedings of the 7th International Conference on Spoken Language Processing. 2002:901-905.

[11] OCH F J, NEY H. A systematic comparison of various statistical alignment models [J]. Computational Linguistics, 2003, 29(1):19-51.

[12] OCH F J. Minimum error rate training in statistical machine translation [C]// Meeting on Association for Computational Liguistics. 1973:160-167.