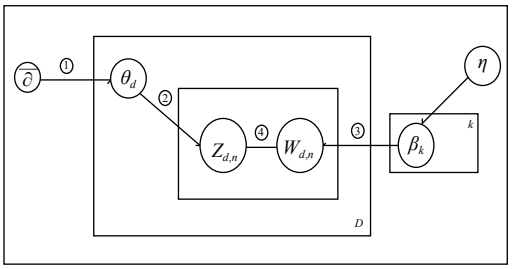

图 6 LDA 算法模型

η表示每个主题分布对应的参数, β_k表示用第K个主题来生成文字. $Z_{d,n}$表示从主题分布中产生主题, 服从多项式分布即

$$Z_{d,n} \sim \frac{n!}{n_1!\cdots n_d!} p_1^{n1}\cdots p_k^{n_k} \sum_{i=1}^{d} n_i = n \tag{7}$$

$W_{d,n}$表示从确定的主题d中产生文字, 同样服从多项式分布.

综上所述, 可以将 LDA 的算法流程整理得到:

算法 1. LDA 算法

for all topics $k\in[1,K]$ do
  sample mixture component $\beta_k \sim Dir(\eta)$ · ①
end for
for all documents $d\in[1,D]$ do
  sample mixture proportion $\theta_d \sim Dir(\vec{\partial})$····· ②
  for all words $n\in[1,N]$ do
    sample topic index $Z_{d,n} \sim Mult(\theta_d)$·····③
    sample term for word $W_{d,n} \sim Mult(\beta_{Z_{d,m}})$···· ④⑤
  end for
end for

LDA 算法属于统计模型, 使用之前需要进行预训练得到概率分布的参数. 求解模型的参数一般使用 Gibbs 采样或者 EM 算法来求解. 本文所述的 LDA 算法主要用在电网行业的文本中, 所以使用来自于北极星电力新闻网的网页组成的语料库作为训练语料进行模型训练.

③ 命名实体识别

文中在生成某些标签时, 需要关注供应商名称、机构名称或者事件发生的时间等, 这些名词在语言中被称为命名实体. 本文采用基于条件随机场的命名实体识别算法实现命名实体的识别.

条件随机场是一种在给定观察的标记序列下, 计算整个标记序列的联合概率的方法. 如$X=(X_1, X_2,\cdots,X_n)$和$Y=(Y_1,Y_2,\cdots,Y_n)$是联合随机变量, 若随机变量 Y 构成一个无向图 $G=(V,E)$表示的马尔科夫模型, 则其条件概率分布$P(Y|X)$称为条件随机场:

$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$, 其中$w\sim v$表示图$G=(V,E)$中与结点v右边连接的所有节点, $w\neq v$表示结点v 以外的所有节点. 其图结构如图 7 所示.

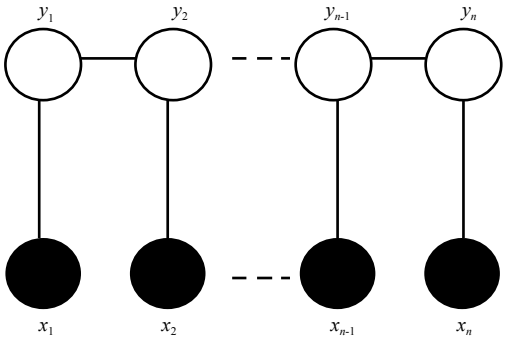

图 7 马尔科夫图模型

假设现在以企业名称识别为例. 做如下标记, 表 3:

表 3 命名实体识别结构构建

标注	含义
B	当前词为企业名称命名实体的首部
M	当前词为企业名称命名实体的内部
E	当前词为企业名称命名实体的尾部
S	当前词单独构成企业名称命名实体
O	当前词不是地理命名实体或组成部分

在训练样本中每个字的标签都在已知的标签集合中选择 (“B”, “M”, “E”, “S”, “O”),

x 是字序列, y 是字对应的标签序列. 训练条件随机场模型的过程就是将已经标注好的训练样本输入初始模型中, 迭代求解特征函数和对应特征函数权重的过程, 训练的目标函数为:

$$P(y|x)=\frac{1}{Z(x)}\exp(\sum_j\sum_i w_j f_j(y_{i-1},y_i,x,i)) \tag{8}$$

其中, $Z(x)=\sum_y \exp(\sum_j\sum_i w_j f_j(y_{i-1},y_i,x,i))$, w_j表示第j个特征函数的权重, f_j表示特征函数. 本文训练模型的数据集是开放的人民日报语料.

(2) 应用分析

与供应商企业相关的文本处理相比于传统的文本处理更加困难. 因为相关文本大多是短文本, 而传统的文本处理方法会导致文本语义特征稀疏和语义敏感等问题. 所以对企业相关的文本预处理时使用了词性标注以及拼音序列的表征.

考虑到标签数量较多且标签之间有重复使用算法

的现象, 所以选取几个典型的标签来举例. 这里上海某电器集团为来说明.

① 企业简介

本标签主要是对爬取的企业简介文本做关键词提取分析. 这部分相对于其他模型标签构建方法简单, 直接对文本进行分词处理, 分词时要对常见的企业词重点关注比如“上市”、“融资”等. 分好词的文本直接输入的训练好的 LDA 模型中然后输出相应的关键词. 原文和关键词对比见表 4.

表 4 原文与 LDA 处理结果对比

部分原文	使用 LDA 后得到的标签内容 (10 个)
xx 集团股份有限公司是中国装备制造业最大的企业集团之一, 具有设备总成套、工程总承包和提供现代装备综合服务的优势…正在成为一个主业突出、优势明显, 可持续发展的现代化、国际化大型装备集团.	装备制造业, 装备服务, …, 变压设备, 国际化

② 诉讼情况

分析企业的诉讼情况需要关注案件发生的企业双方, 缘由和最终的判决结果. 但是有关诉讼的文本比较短, 且关键性的词语和命名实体比较密集. 所以本质上需要对文本的主要的内容进行语义分析. 获取一条诉讼文本后, 首先进行句法分析得到句法分析树, 根据句法分析树和基于条件随机场的命名实体算法识别出原告和被告的关系和名称.

表 5 诉讼标签提取情况举例

部分原文	标签内容 (原告、被告、案由、结果)
原告 xxx 有限公司与被告 xx 有限公司买卖合同纠纷一案, … . 裁定如下: 准予原告 xx 有限公司撤回对被告 xx 有限公司的起诉.	xxx 有限公司, xx 有限公司 , 买卖合同纠纷, 撤回起诉

诉讼情况的得分的计算方案为:

$$\text{score}=\frac{\sum_{j=1}^{t} w_j c_j}{\sum_{i=1}^{k} w_i c_i} \tag{9}$$

其中, w_i表示i类纠纷的权重, c_i表示i类纠纷的计数, 如果裁定结果为撤诉则不参与计数. 分子的表示与实际需求最相关的t类诉讼案件, 比如当关注于供应商的产品时, 则主要选择与产品相关的诉讼案件作为分子.

2 画像效果评估

本文评估用户画像效果的方法是计算准确率、和是否有时效性机制, 这也是用户画像评估中最常用的方法.

2.1 准确率

准确率指被打上正确标签的用户比例. 准确率是用户画像最核心的指标, 计算公式是:

$$\text{准确率}=\frac{|U_{tag=true}|}{|U_{tag}|} \tag{10}$$

其中, $|U_{tag}|$表示被打上标签的用户数, $|U_{tag=true}|$表示有标签用户被打对标签的用户数.

具体的评估方法为: 随机抽取 15 家合作过的供应商企业, 行业专家首先对供应商进行标注, 并把经过两轮审核后得到的标注结果当作准确的样本. 然后再有新一批专家和自动化模型通过进行标注, 并根据准确样本计算两者标注的准确率, 为了提高评估结果的准确性, 进行 3 组相同的标注过程. 3 组的对比情况如表 6 所示.

表 6 模型准确率测试结果 (%)

组号	专家标注	模型标注
1	93.1	92.6
2	94.8	95.9
3	91.5	91.3

3 画像效果评估

假设国网现在想选择一家变压器供应商购进一批变压器, 首先给出一系列期望的变压器参数, 比如使用寿命, 价格, 安装时间等. 然后将这些参数组合成目标模板. 选择多家供应商的相关标签计算与目标模板的相似度. 根据相似度的分值, 对供应商进行排名. 排名越靠前表示推荐力度越高.

具体实验过程为: 从历史最优采购记录中选取了 20 种设备. 每种设备选取了同时期的 39 家供应商作为干扰项, 加上最优供应商一共 40 家. 然后对每家设备供应商使用 GloVe 算法提取特征, 此其中 GloVe 算法百万数量级的词典和上亿数据集上可以进行快速训练. 提取特征后进行与目标模板进行相似度计算得到一个结果. 同时使用常见的 AHP 和 Xgboot_LMT 算法进行分析得到的最终精确度比较见表 7.

表 7 模型应用准确率 (%)

算法	AHP	Xgboost_LMT	本文
准确率	87.4	86.6	91.3

4 总结与展望

本文以“辅助决策模块”为实际应用背景. 通过使用用户画像的方案对供应商的数据进行了有效的组织. 在行业专家和大数据工程师的共同参与下, 使用自然语言处理和机器学习的方法, 构建了自动更新的供应商画像标签体系, 通过评估该画像体系取得了比较高的得分. 通过使用用户画像技术简化了开发流程, 提高了系统的工作质量.

但是系统在标签构建的内容上比较繁琐, 并且在构建算法的调优上还有所不足. 后期需要逐步探索更加便捷的标签内容, 并且随着数据量的增加需要对相关算法进行重新训练提高标签内容提取的准确率.

参考文献

1 张元新, 宋婷, 何灵, 等. 基于 AHP-模糊综合评估方法的电网物资供应商评估模型构建. 管理科学与工程, 2017, 6(4): 147–153. [doi: 10.12677/MSE.2017.64018]

2 樊鹏. 基于优化的 XGboost-LMT 模型的供应商信用评价研究[硕士学位论文]. 广州: 广东工业大学, 2016.

3 席一凡, 王超, 聂兴信. 基于模糊神经网络的供应链绩效评价方法研究. 情报杂志, 2007, 26(9): 77–79. [doi: 10.3969/j.issn.1002-1965.2007.09.025]

4 贺绍鹏, 李屹, 邹兰青. 大数据环境下供应商评价设计与分析. 物流技术, 2018, 37(2): 96–100. [doi: 10.3969/j.issn.1007-1059.2018.02.013]

5 杨志和. 基于大数据技术的供应商情报管理系统. 上海电机学院学报, 2018, 21(2): 21–27. [doi: 10.3969/j.issn.2095-0020.2018.02.004]

6 刘海鸥, 孙晶晶, 陈晶, 等. 用户画像模型及其在图书馆领域中的应用. 图书馆理论与实践, 2018, (10): 92–97–23.

7 雷海燕. 基于 Hadoop 的供应商评价系统的研究与设计[硕士学位论文]. 西安: 西安科技大学, 2015.

8 徐晋, 綦振法. 基于神经网络专家系统的供应商信用等级分析. 情报科学, 2004, 22(2): 2.

9 贾安超, 周刚. 基于粗糙集和 BP 神经网络的供应商选择研究. 物流技术, 2012, 31(12): 229–232, 267. [doi: 10.3969/j.issn.1005-152X.2012.12.079]

10 刘昌法. 基于模糊神经网络的供应商选择技术研究. 航空制造技术, 2014, (1-2): 137–139.

11 Li HP, Liu J, Zhang SW. Hierarchical latent dirichlet allocation models for realistic action recognition. Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic. 2011. 1297–1300. [doi: 10.1109/ICASSP.2011.5946649]

12 Wu T, Shunk D, Blackhurst J, *et al*. AIDEA: A methodology for supplier evaluation and selection in a supplier-based manufacturing environment. International Journal of Manufacturing Technology and Management, 2007, 11(2): 174–192. [doi: 10.1504/IJMTM.2007.013190]