Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1415-1422
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116145

Hans汉斯

交易信息的数据处理及分词应用研究

涂著刚，李正军

贵阳高新数通信息有限公司，贵州 贵阳

收稿日期：2022年11月10日；录用日期：2022年12月20日；发布日期：2022年12月30日

摘　要

数据清洗技术及分词技术的应用对于挖掘海量交易信息潜在的数据价值至关重要，对不同类型交易“脏数据”按不同策略进行数据预处理，同时通过对LTP及Jieba分词技术在交易领域的应用研究，在提高交易关键信息的识别与处理效率及最终数据质量的同时，探索对信息搜索准确率及查全率的提升作用。

关键词

数据价值，数据清洗，LTP，Jieba

Research on Data Processing and Word Segmentation of Bids Information

Zhugang Tu, Zhengjun Li

Guiyang Hi-Tech Data Communication Co., Ltd., Guiyang Guizhou

Received: Nov. 10th, 2022; accepted: Dec. 20th, 2022; published: Dec. 30th, 2022

Abstract

The application of data cleaning technology and word segmentation technology is crucial to mining the potential data value of massive bids information. Data preprocessing is carried out for “dirty data” of different types of bids according to different strategies. At the same time, through the research on the application of LTP and Jieba word segmentation technology in the bids field, while improving the identification and processing efficiency of key bids information and the final data quality, the role of improving the accuracy and recall of information search is explored.

Keywords

Data Value, Data Value, Data Cleaning, LTP, Jieba Annotation

文章引用: 涂著刚, 李正军. 交易信息的数据处理及分词应用研究[J]. 软件工程与应用, 2022, 11(6): 1415-1422.
DOI: 10.12677/sea.2022.116145

1. 引言

据不完全统计，全国每天发布的公共资源及政府采购交易信息超过十五万条，信息类型多样、数据格式不一、数据结构复杂。如何有效处理海量交易信息，挖掘出潜在的数据价值，是衡量诸多招投标领域大数据应用产品竞争力的关键。中文分词是中文信息处理领域中最为关键的课题，最大熵、条件随机场等模型被用来解决序列标注任务，由于传统机器学习方法一般采用特征工程的方式，依赖专家经验，成本较高，借助于神经网络模型可以自动隐式提取特征[1]，将字向量作为输入，用一个简单的神经网络模型替代了最大熵模型，长短期记忆神经网络(Long Short-Term Memory Neural Networks, LSTM)来对句子进行建模，捕捉字与字之间的长距离依赖关系，将基于转移的思想应用于分词任务并使用大量的外部训练语料进行预训练，利用外部知识提高了分词效果[2]。理想的技术产品架构，底层是自然语言处理平台，中间一层是文本挖掘平台，最上面一层是企业智能信息处理平台，为企业提供各种智能化信息处理解决方案。其中，数据清洗技术及分词技术的应用，对于非结构化信息的识别及提取，从而实现信息的字段化、结构化、价值化尤为重要。

2. 交易数据清洗

交易数据采集完成后，对不同渠道获取的数据需要进行解析，然后对不准确、不完整、不合理、格式、字符等不规范交易数据进行过滤清洗，利用数理统计、数据挖掘和预定义清理规则等有关技术将交易“脏数据”处理掉，从数据源中检测并消除错误、不一致、不完整和重复等数据，为满足要求提供高质量的数据，清洗过的数据才能符合识别及提取需求，因此在对数据进行识别及结构化前，做好相关的数据清洗工作意义重大。

2.1. 交易数据分析

交易数据分析是交易数据清洗的前提和基础，通过人工检测或者计算机分析程序的方式对原始数据源的数据进行检测分析，从而得出原始数据源中存在的数据质量问题。交易数据的主要问题有数据重复问题(一条标讯多个来源、平台重复发布)、时间失效数据问题(发布过期数据)、测试数据问题(平台测试数据)、异常格式数据问题(pdf、附件、图片等)、无价值或低价值数据问题(类型不匹配，内容不符合)。

2.2. 定义数据清洗的策略和规则

根据分析的数据源数量和数据源中“脏”数据的程度，定义数据清理策略和规则，并选择适当的数据清理算法。如表 1 所示，对于不同类型问题数据，需要根据交易信息的特点及业务策略制定不同的清理规则。

交易实体识别：对不同数据来源中同一对象实体进行识别；

交易数据真值发现：在混有数据质量错误的冲突数据中找到属性的真实值；

交易数据不一致检测与修复：实际数据集合中通常包含了某些违反最初定义的完整性约束的数据，造成集合内或者不同数据集合间的不一致情况，本系统可利用完整性约束对数据中不一致情况进行检测和修复清洗；

交易缺失值填充：对数据集合中存在的数据缺失问题进行有效的填充修复。

Table 1. Cleaning strategies for different problem data
表 1. 不同问题数据的清洗策略

序号	数据类型	示例	清洗策略
1	测试数据	ts 全流程测试 A,勿报名招标公告 mmj + slsd (测试项目，请勿报名) 国泰测试】水利远程异地测…… 这是一条测试数据国泰测试－稠州银行(请勿报名) epointtest2 双城区五家街道办事处五家村邮储测试合作社 1.00 亩旱田使用权测试请勿报名 怒江州泸水市测试测试测试 恒瑞通-xmm 流程验证项目 test 测试－测试项目(勿投) 北京筑龙公告审核测试测试测试-BWW-新资格预审公告 招商局天翼漏洞修复－测试测试测试－不接受投标招标公告 3.0 测试-1105 招标项目流程验证测试测试测试 1 资格预审公	自动清除
2	重复数据	大连海警局中山工作站外电路改造项目竞争性磋商公告，系统采集了至少五个来源：大连市中山区人民政府、全国公共资源交易平台(辽宁省)、大连市政府采购网、辽宁大连市公共资源交易平台、中国政府采购网	多重规则去重程序识别，去重
3	低价值数据	公众号、电子卖场成交与履约信息，数据量大，价值低	识别后降低权重
4	格式异常数据	中国招标投标公共服务平台，陕西省招标与采购网等，文本为 PDF 附件	文本识别后加入
5	特定格式数据	采购意向、开标信息、候选人信息等以表格为主的信息	表格单独提取或转换成规范文本

交易数据清洗结果可视化：数据质量检测和清洗的结果以图、表形式展示给用户，让用户对交易数据集合的质量评估情况有直观的认识。

2.3. 搜寻并确定错误实例

手动检测数据集中的属性错误需要大量的时间、精力和物质资源，而且过程本身容易出错，因此需要使用有效的方法来自动检测数据集中属性错误。主要的检测方法是基于事务统计的方法、事务聚类方法和事务关联规则方法[3]。当存在多个交易数据记录源时，需要进行相关性验证。如果在数据分析过程中发现数据冲突，应调整或删除相关数据，以通过数据分析和检测使数据一致。

2.4. 纠正发现的错误

根据不同的“脏”数据存在形式的不同，执行相应的数据清洗和转换步骤解决原始数据源中存在的质量问题。需要注意的是，对原始数据源进行数据清洗时，应该将原始数据源进行备份，以防需要撤销清洗操作。

2.5. 干净数据回流

在交易数据被清理后，干净的交易数据取代了原始数据源中的“脏”数据，这可以提高信息系统的数据质量，避免未来数据提取后的重复清理。

3. 分词技术及应用

从语义相似性的角度来看，自然语言处理可以在对话或问答领域取得更好的结果。它可以迁移到事

务信息的业务字段中的数据集。由于该领域本身的知识和特点，如何选择成熟的分词技术来获得高效的带注释数据集模型，并为事务领域的机器学习奠定基础是一个迫切的问题。此外，当用户搜索交易信息时，他们需要有效的中文分词技术来提高所需数据信息的准确性和全面性。

3.1. LTP 分词

哈工大的语言技术平台提供了中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注等丰富、高效、精准的自然语言处理技术。

3.1.1. 基本原理

LTP 是基于结构化感知器[4] (Structured Perceptron, SP)，以最大熵准则建模标注序列 YY 在输入序列 XX 的情况下的 score 函数：

$$S(Y,X)=\sum_s \alpha_s \Phi_s(Y,X)$$

其中，$\Phi_s(Y,X)$为本地特征函数。中文分词问题等价于给定 XX 序列，求解 score 函数最大值对应的 YY 序列：

$\text{argmax}_Y S(Y,X)$

在 LTP 中，将分词任务建模为基于字的序列标注问题。对于输入句子的字序列，模型给句子中的每个字标注一个标识词边界的标记。

3.1.2. 应用及结果

使用代码：

```
from ltp import LTP
# 默认加载 Small 模型
ltp = LTP("LTP/small")
lines = ["南阳富达鸭河荒山光伏电站有限公司 2022 年度物资采购采购结果公告"]
output = ltp.pipeline(lines, tasks = ["cws"])
# 使用字典格式作为返回结果
print(output.cws)
```

分词结果：

```
[['南阳', '富达', '鸭河', '荒山', '光伏', '电站', '有限公司', '2022', '年度', '物资', '采购',
'采购', '结果', '公告']]
```

3.2. Jieba 分词

3.2.1. 基本原理

实现原理是隐马尔可夫模型(HMM) [5]，它结合了基于字符串匹配的算法和基于统计的算法。采用最大概率路径动态规划算法进行字符串匹配，在保持快速分割的同时保持较高的分割精度。采用 HMM 隐马尔可夫模型对新词进行分割，可以有效解决字符串匹配无法识别新词的困难。Jieba 分词依赖于汉语词库：1) 使用汉语词库来确定汉字之间的关联概率；2) 汉字之间的高概率词形成分词结果；3) 除了分词，用户还可以添加自定义字典。

Jieba 分词的三种模式：精确模式、全模式、搜索引擎模式

精确模式：把文本精确的切分开，不存在冗余单词。

全模式：把文本中所有可能的词语都扫描出来，有冗余。

搜索引擎模式：在精确模式基础上，对长词再次切分。

3.2.2. 应用及结果

分别采用三种不同模式处理部分交易信息标题，得到结果如下：

1) 精确模式：

```
import jieba
sentence = '沙井街道医用外科口罩采购招标项目的潜在投标人应在深圳高星项目管理有限公司获取招标文件'
messages = jieba.cut(sentence, cut_all = False)
print('分词结果：' + "/ ".join(messages))
```

分词结果：沙井/ 街道/ 医用/ 外科/口罩/ 采购/ 招标/ 项目/ 的/ 潜在/ 投标人/ 应/ 在/ 深圳/ 高星/ 项目管理/ 有限公司/ 获取/ 招标/文件

2) 全模式：

```
import jieba
sentence = '沙井街道医用外科口罩采购招标项目的潜在投标人应在深圳高星项目管理有限公司获取招标文件'
messages = jieba.cut(sentence, cut_all = False)
print('分词结果：' + "/ ".join(messages))
```

分词结果：沙井/ 街道/ 医用/ 外科/ 口罩/ 采购/ 招标/ 项目/ 目的/ 潜在/ 投标/ 投标人/ 应/ 在/ 深圳/ 高/ 星/ 项目/ 项目管理/ 管理/ 有限/ 有限公司/ 公司/ 获取/ 招标/ 文件

3) 搜索引擎模式：

```
import jieba
sentence = '沙井街道医用外科口罩采购招标项目的潜在投标人应在深圳高星项目管理有限公司获取招标文件'
messages = jieba.cut_for_search(sentence)
print('分词结果：' + "/ ".join(messages))
```

分词结果：沙井/ 街道/ 医用/ 外科/ 口罩/ 采购/ 招标/ 项目/ 的/ 潜在/ 投标/ 投标人/ 应/ 在/ 深圳/ 高星/ 项目/ 管理/ 项目管理/ 有限/ 公司/ 有限公司/ 获取/ 招标/ 文件

4) 调整词频

有的时候，如果按照 jieba 正常分词，会把我们不希望分开的词语给分开，这个时候就会改变句子的意思。就如以下例子，我们希望“招标文件”是一个词，不被分开：

```
import jieba
sentence = '在深圳高星项目管理有限公司获取招标文件'
messages = jieba.cut(sentence)
print('分词结果：' + "/ ".join(messages))
```

分词结果：在/ 深圳/ 高星/ 项目管理/ 有限公司/ 获取/ 招标/ 文件

5) 分词后词性标注

```
import jieba.posseg as contents
messages = contents.cut('沙井街道医用外科口罩采购招标项目的潜在投标人应在深圳高星项目管理有
```

限公司获取招标文件')

```
for message in messages:
    print(message.word, message.flag)
```

沙井 n/街道 n/医用 n/外科 n/口罩 n/采购 v
招标 n/项目 n/的 uj/潜在 t/投标人 n/应 v
在 p/深圳 ns/高星 n/项目管理 n/有限公司 n
获取 v/招标 n/文件 n/
标签和含义对照如图 1 所示：

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	t	时间
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	Loc	地名	ORG	机构名	TIME	时间

Figure 1. Comparison of part of speech meanings of labels
图 1. 标签词性含义对照

6) 添加自定义词典

```
import jieba
sentence = '沙井街道医用外科口罩采购招标项目的潜在投标人应在深圳高星项目管理有限公司获取招标文件'
jieba.load_userdict(r"C:\Users\Administrator\Desktop\user_dict.txt")
messages = jieba.cut(sentence)
print('分词结果：' + "/ ".join(messages))
```

分词结果：沙井/ 街道/ 医用/ 外科口罩/ 采购/ 招标/ 项目/ 的/ 潜在/ 投标人/ 应/ 在/ 深圳/ 高星/ 项目管理/ 有限公司/ 获取/ 招标文件

此外，Jieba 还可以方便地实现取消新词、查看文本内词语的开始和结束位置、修改字典路径等[6]。

在交易信息结构数据处理中，某些文字组合表示特定的含义，在应用中可以加入自定义词典，部分自定义交易词组如表 2 所示。

Table 2. Bids data user-defined word set (partial)
表 2. 交易数据自定义词集(部分)

序号	数据类型	部分自定义字典词集
1	招标公告	项目预算、项目预算单位、非标准金额、报名截止时间、投标截止时间(报价截止时间)
2	候选人公示	投标人(其他投标人)、投标人报价、第一候选人、第一候选人报价、第二候选人、第二候选人报价、第三候选人、第三候选人报价
3	中标结果	中标机构(供应商名称)、中标金额(采购金额)、中标金额单位(采购金额单位)、中标单位联系人(供应商联系人)、中标单位电话(供应商电话)、投标人（其他投标人）、非标准金额、第一候选人、第二候选人、第三候选人

Continued

4	合同	合同名称、合同编号、合同金额、合同金额单位、合同签订日期、供应商名称、供应商联系人、供应商电话
6	预告	项目预算、项目预算单位、非标准金额、预计招标日期
8	不分类型	项目名称、项目编号、业主、业主联系人、业主电话、代理电话、代理机构、代理联系人

4. 应用成果分析

如表 3 所示，通过制定并不断优化数据清洗策略，平台类不同类型数据的重复率及查漏率指标均得到了有效提升，另外通过两种分词工具的应用，对不同类型数据的抽取准确率得到了不同提高。可以看出，对于不同的文本格式，两种分词工具各有优势。不断尝试且优化分词方法及完善自定义词库，是提升交易数据效益的有效方式。

Table 3. Comparison of different types of bids data indicators
表 3. 不同类型交易数据指标对比

序号	数据类型	重复率 (2020 年)	查漏率 (2020 年)	抽取准确率 (2020 年)	重复率 (2022 年)	查漏率 (2022 年)	抽取准确率 (2022 年)
1	招标公告	9.32	2.11	76.12	7.21	1.74	89.65
2	候选人公示	8.67	1.96	81.79	6.71	1.42	92.51
3	中标结果	10.57	2.94	75.03	8.63	1.60	91.64
4	合同	6.57	1.34	83.23	4.63	0.98	95.16
5	预告	9.57	2.14	70.03	5.63	1.11	90.15

5. 结束语

对于绝大部分交易信息都是各种格式的复杂非结构化信息而言，对于数据的预处理及清洗是实现营销决策、数据统计及分析等各类大数据应用的前提，另外，由于汉语语言知识的笼统性、复杂性，很难将各种语言信息组织成机器可以直接阅读的形式。因此，现有的分词技术被用来使计算机模拟人们对句子的理解，以达到单词识别的效果。在分词的同时，进行句法和语义分析，并使用句法和语义信息来处理歧义。这种分词方法需要大量的语言知识和信息。除了本文中提到的 LTP 和 JIEBA，HanLP 也是一种有效的分词工具。多种分词技术的综合应用有助于对投标领域的数据进行深度挖掘和识别，构建多种应用场景，满足各类用户的数据和应用需求。

基金项目

贵州省科技计划项目(课题)黔科中引地[2021]4016；基于 BOES 开放引擎的数据分析关键技术在招投标领域的创新应用。

参考文献

[1] Matthias, S., Garham, N., Jan, N. and Alex, W. (2017) Neural Lattice-to-Sequence Models for Uncertain Inputs. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017, 1380-1389.

[2] Zhou, G.D. and Su, J. (2002) Named Entity Recognition Using an HMM-Based Chunk Tagger. *Proceedings of the*

40*th Annual Meeting on Association for Computational Linguistic*, Philadelphia, 7-12 July, 2002, 473-480. https://doi.org/10.3115/1073083.1073163

[3] 顾佼佼, 杨志宏, 姜文志, 等. 基于条件随机场的中文分词算法改进[J]. 信息与电子工程, 2012, 10(2): 184-187.

[4] 殷章志, 李欣子, 黄德根, 李玖一. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报, 2019, 33(11): 95-100, 106.

[5] Sassi, I., Anter, S. and Bekkhoucha, A. (2021) ParaDist-HMM: A Parallel Distributed Implementation of Hidden Markov Model for Big Data Analytics using Spark. *International Journal of Advanced Computer Science and Applications*, **12**, 289-303. https://doi.org/10.14569/IJACSA.2021.0120438

[6] 刘伟, 黄锴宇, 余浩, 等. 基于语境相似度的中文分词一致性检验研究[J]. 北京大学学报(自然科学版), 2022, 58(1): 99-105.