Computer Science and Application 计算机科学与应用, 2021, 11(9), 2302-2314
Published Online September 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.119235

Hans汉斯

结合标注的中文地址匹配规则链模型

李晓晰，张　伟*

北京信息科技大学，北京
Email: lixiaoxiill@163.com, *zhwei@bistu.edu.cn

收稿日期：2021年8月13日；录用日期：2021年9月10日；发布日期：2021年9月17日

摘　要

现有的中文地址匹配研究方法集中于对文本特征的研究，忽略了中文地址所包含的建筑特征、地理位置特征、统计特征和行业特征的数据，此类特征数据可以有效辅助中文地址的匹配研究。本文主要面向非规范的中文地址，以燃气行业居民用户数据为样本数据进行实验，通过分析两个数据源中用户信息的多个特征数据，提出以结合标注的中文地址匹配规则链模型。规则链的优点是链内的规则可以动态配置，通过人工和计算机结合的方式，动态管理规则，多次迭代，逐步提升匹配率。实验结果表明该模型可以一定程度提高中文地址匹配的成功率。

关键词

中文地址，中文地址匹配，标注，规则，规则链

Chinese Address Matching Rule Chain Model Combined with Annotation

Xiaoxi Li, Wei Zhang*

Beijing Information Science & Technology University, Beijing
Email: lixiaoxiill@163.com, *zhwei@bistu.edu.cn

Received: Aug. 13th, 2021; accepted: Sep. 10th, 2021; published: Sep. 17th, 2021

Abstract

Existing Chinese address matching research methods focus on text features, ignoring the data of architectural, geographic, statistical and industrial characteristics contained in Chinese addresses, which can effectively assist Chinese address matching research. This paper mainly aims at non-standard Chinese addresses, and takes gas industry resident user data as sample data to experiment. By analyzing multiple feature data of user information in two data sources, a Chinese

*通讯作者。

文章引用: 李晓晰, 张伟. 结合标注的中文地址匹配规则链模型[J]. 计算机科学与应用, 2021, 11(9): 2302-2314.
DOI: 10.12677/csa.2021.119235

address matching rule chain model is proposed to combine labeling. The advantage of rule chains is that the rules in the chain can be configured dynamically. By combining manual and computer methods, rules can be managed dynamically and iterated several times to gradually increase the matching rate. The experimental results show that the model can improve the success rate of Chinese address matching to a certain extent.

Keywords

Chinese Address, Chinese Address Matching, Label, Rule, Rule Chain

1. 引言

中文地址匹配是指分析和判断两个或多个来自不同数据源中的中文地址，是否指向现实世界中的同一个中文地址。中文地址因其存在复杂性、非规范性、非结构化等问题，给企业内业务系统之间的融合提供了很大阻碍。目前，针对非规范的中文地址匹配研究主要集中在中文地址的文本特征，例如文献[1]提出了一种基于元数据关联特征的交互式数据预处理方法。文献[2]提出了一种新的文本相似度量方法，应用自然语言处理技术对文本进行预处理。文献[3]提出了一种基于条件随机场的中文地址解析方法。文献[4]提出了一种基于动态规划的中文地址匹配方法。文献[5]提出了一种基于熵的文本相似度求解方法，在对文本间字符信息的提取基础上，建立共同子文本串度量维度，然后采用熵的方法进行相似度度量。文献[6]针对当前在电力中文地址匹配中存在部分地址歧义的问题，结合自然语言处理的基本原理，提出了一种基于贝叶斯算法的中文地址精确匹配方法。

综上，如图 1 所示现有的中文地址匹配研究方法集中于中文地址文本特征的研究，而忽略了中文地址所包含的建筑特征、地理位置特征、统计特征和行业特征的数据，此类特征数据可以有效地辅助中文地址的匹配研究。本文以燃气行业居民用户数据为例进行实验，通过分析两个数据源中用户信息的多个特征数据，提出以结合标注的中文地址匹配规则链模型，实验结果表明该模型可以一定程度提高中文地址匹配的成功率。

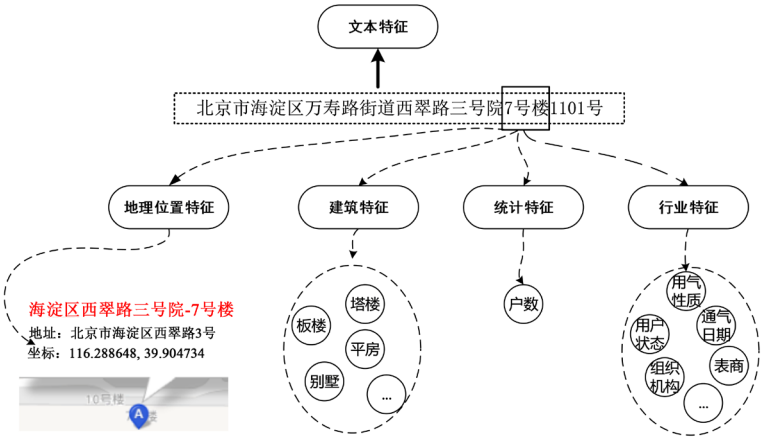

Figure 1. Feature analysis of Chinese address data
图 1. 中文地址数据特征分析

2. 结合标注的中文地址匹配规则链模型设计

本文设计一种结合标注的中文地址匹配规则链模型如图 2 所示。

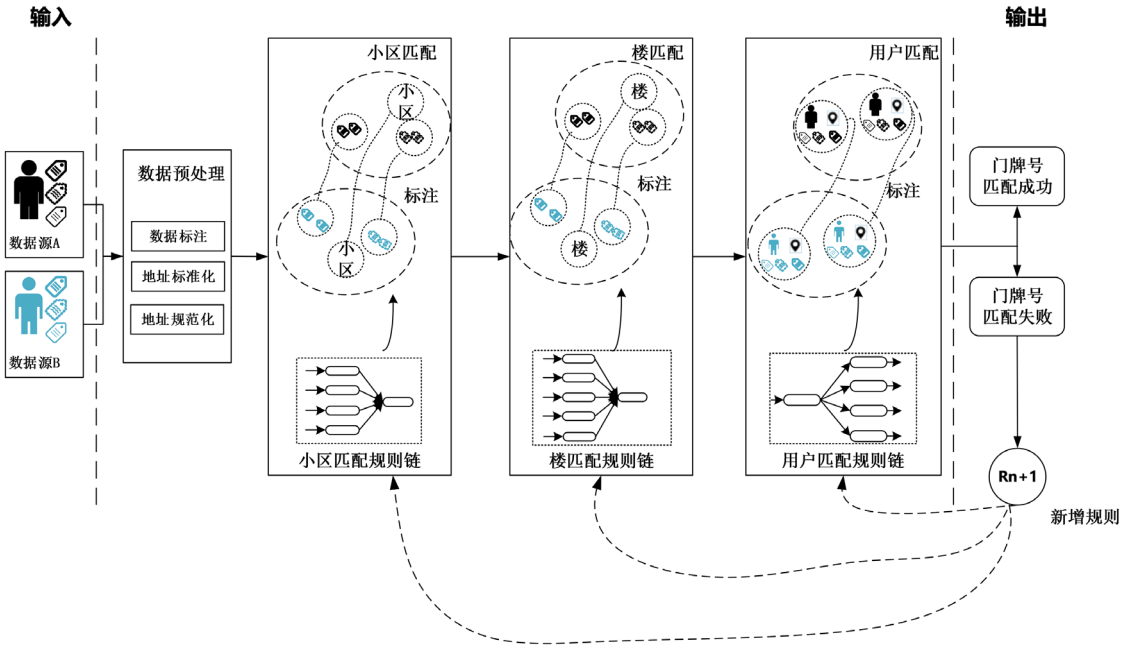

Figure 2. Chinese address matching rule chain model combined with annotation
图 2. 结合标注的中文地址匹配规则链模型

数据预处理阶段，实现两个数据源中的中文地址标注、规范化和标准化的过程。根据数据预处理的结果，首先按小区匹配，小区匹配规则链实现两个数据源中小区地址要素与北京市标准地址库的匹配。其次是按楼匹配，结合楼名称、建筑特征、地理特征、统计特征等标注信息构建规则链，实现中文地址中楼信息的匹配。最后是按用户匹配，用户匹配主要依据楼匹配的结果(二元组、三元组)进行详细地址匹配。匹配失败的可以通过增加规则，补充规则链内容，继续迭代匹配过程。规则链的优点是链内的规则可以动态配置，支持代码复用，通过人工 + 计算机的方式多次迭代，新增和优化规则，逐步提升中文地址的匹配成功率。

3. 结合标注的中文地址匹配规则链模型实现

3.1. 结合地理、行业标注的小区匹配

小区匹配是判断两个不同的数据源中的小区名称是否为同一个小区。小区匹配的难点在于小区名称存在歧义、别名、简称等多种描述方式，例如“朝阳北苑路 86 号”、“嘉铭桐城”、“桐城国际”、“朝阳嘉铭园”、“嘉铭苑”都指向同一个小区，但是因为小区的历史变革、早期的不规范录入、系统缺少统一标准等因素导致同一个小区在不同的数据源中存在多个描述的问题。本文结合了地理、行业标注数据实现小区匹配，小区匹配流程如图 3 所示。

小区匹配结合地理特征、行业特征等标注，基于小区匹配规则链实现小区信息与北京市标准地址去的匹配，如果匹配失败，通过分析匹配失败原因，增加新的规则补充至小区匹配规则链，通过多次迭代，逐步提升小区匹配的成功率。小区匹配规则链包括小区名称匹配、地址匹配和行业特征匹配三个规则集合，规则集合内的规则可以动态增加，规则链的包含的规则内容如表 1 所示。

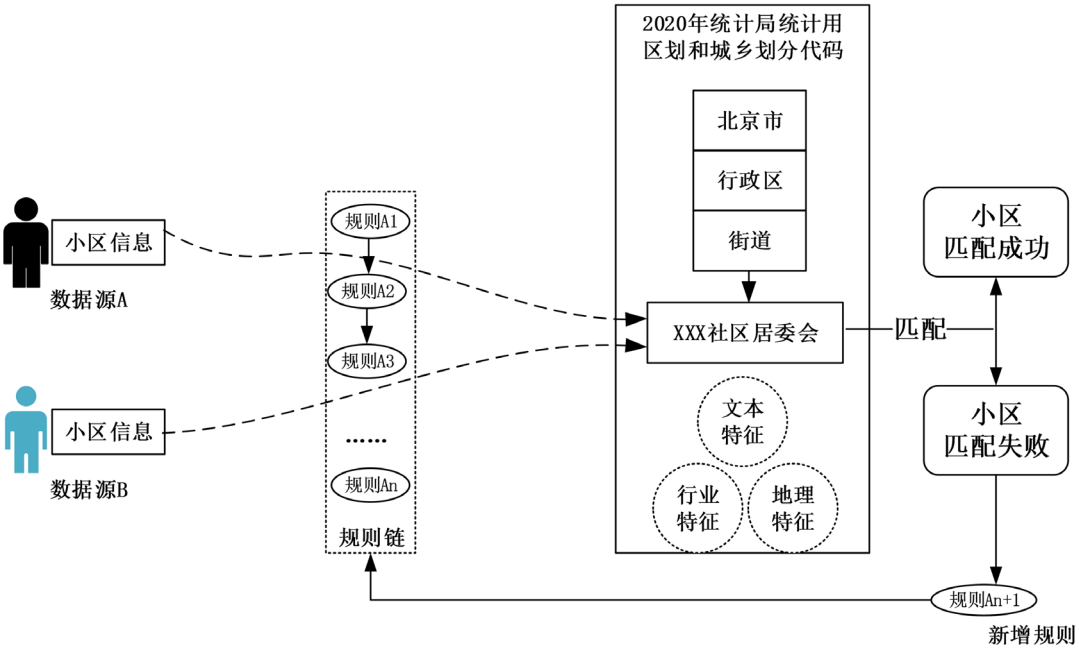

Figure 3. Community matching flow chart
图 3. 小区匹配流程图

Table 1. Community matching rule chain
表 1. 小区匹配规则链

规则链	规则编号	规则描述
A1 小区名称匹配	A1-1	小区名称相似度匹配
	A1-2	小区别名匹配
	A1-3	小区简称匹配
	.....	
	A1-n	
A2 地理特征匹配	A2-1	经纬度匹配
	.....	
	A2-n	
A3 行业特征匹配	A3-1	组织机构匹配
	.....	
	A3-n	

如表 1 所示小区匹配规则链包括以下内容：

1) 小区名称匹配

主要采用待匹配的两个数据源中的小区名称与统计局标准社区名称，如图 4 中的小区信息标准库，进行匹配，匹配成功的小区信息，可以通过关联上级的街道和行政区，同时实现小区、街道和行政区地址要素的标准化。首先通过爬虫技术获取国家统计局《2020 年度全国统计用区划代码和城乡划分代码》，逐级获取北京市→市辖区→行政区→街道→社区类别的统计用区划代码和名称。

标准库中小区信息都是以“社区居委会”结尾，需要去除“社区居委会”后，通过小区名称与两个数据源中的小区名称进行相似度匹配。

Figure 4. Community information standard library
图 4. 小区信息标准库

2) 地理特征匹配

高德开放平台提供的地理/逆地理编码服务，可以将结构化的中文地址(包含行政区、街道、小区等)解析为对应的经纬度。

根据小区名称通过接口调用方式。

a) 地理编码 API 服务地址，如表 2 所示：

Table 2. Request API
表 2. 请求参数

请求地址	https://restapi.amap.com/v3/geocode/geo?parameters
请求方式	GET

b) 请求实例：

https://restapi.amap.com/v3/geocode/geo?address=丰台区太平桥街道华源三里&output=XML&key=<用户的 key>

c) 请求参数，如表 3 所示：

Table 3. Request parameters
表 3. 请求参数

参数	值	规则描述	是否必选
address	丰台区太平桥街道华源三里	行政区 + 街道 + 小区	是
city	北京市		否

d) 返回结果

如下所示小区信息“丰台区太平桥街道华源三里”通过地理编码 API 服务地址获取到经纬度数据“116.320237, 39.879165”，分别根据两个数据源中的小区名称调用地理编码 API 服务地址可以获取小区信息的经纬度数据。

```
{
    "status" :"1",
    "info" :"OK",
    "infocode" :"10000",
    "count" :"1",
    "geocodes" :[
                "0" :{
                                "formatted_address" :"北京市丰台区华源三里",
                                "country" :"中国",
                                "province" :"北京市",
                                "citycode" :"010",
                                "city" :"北京市",
                                "district" :"丰台区",
                                "township" :[ ],
                                "neighborhood" :{ … },
                                "building" :{ … },
                                "adcode" :"110106",
                                "street" :[ ],
                                "number" :[ ],
                                "location" :"116.320237,39.879165",
                                "level" :"兴趣点"
                }
    ]
}
```


3) 行业特征匹配

组织机构特征匹配是根据数据预处理阶段标注的组织机构特征数据进行匹配，每个小区所属的组织机构是不变的，可以通过组织机构匹辅助实现中文地址匹配。

如图 5 所示，不同的两个数据源中两个组织机构“西直门网点”和“和平里网点”所包含的小区名称存在不一致的情况，数据源中 A 的“百万庄”小区和数据源 B 中的“百万庄中”和“百万庄大街”小区对应，因为都属于同一个组织机构“西直门网点”，因此通过组织机构特征的匹配，可以缩小中文地址中地址要素小区的范围，降低中文地址匹配的难度。

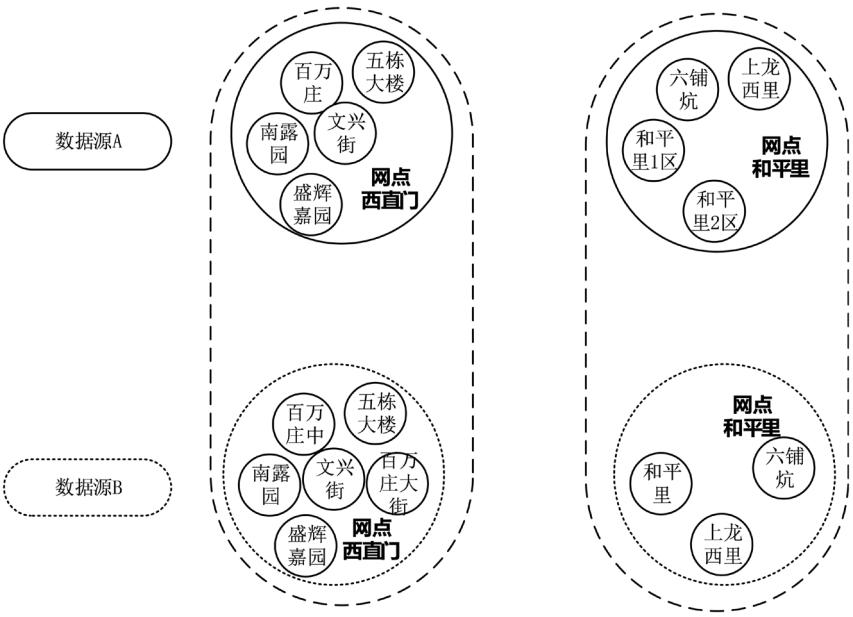

Figure 5. Organization feature matching
图 5. 组织机构特征匹配

3.2. 结合建筑、统计、行业标注的楼匹配

本节基于规则链方式实现两个或多个数据源之间的楼信息的匹配。结合了建筑、统计、行业标注数据实现楼匹配，楼匹配流程如图 6 所示。

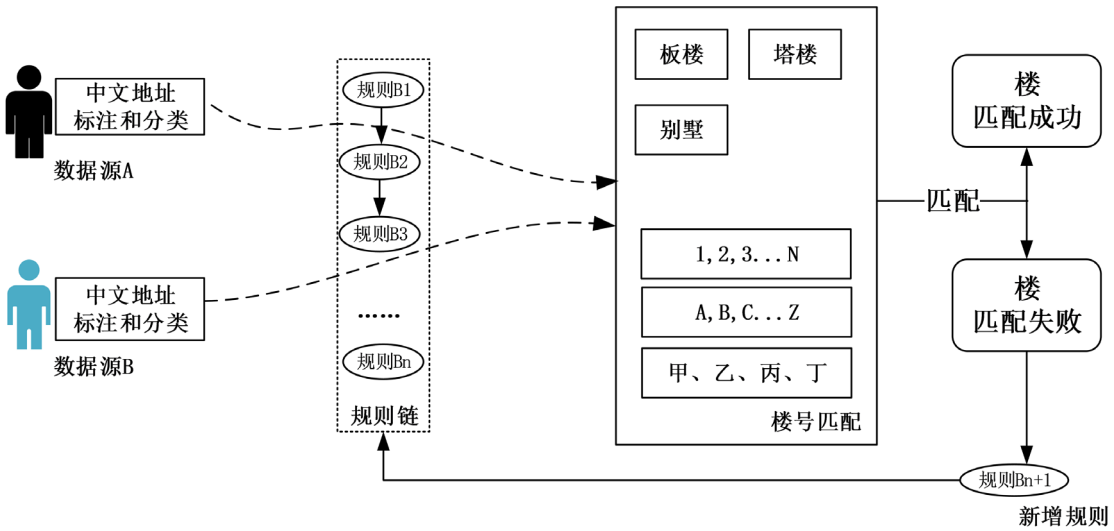

Figure 6. Building matching process
图 6. 楼匹配流程

楼匹配规则链包括名称匹配、建筑特征匹配、统计特征匹配和行业特征匹配三个规则链，规则链内的规则可以动态增加，规则链的包含的规则内容如下：

Table 4. Building matching rule chain
表 4. 楼匹配规则链

规则链	规则编号	规则描述
B1 名称匹配	B1-1	楼名称直接匹配
	B1-2	楼名称别名匹配
	……	
	B1-n	
B2 建筑特征匹配	B2-1	塔楼匹配
	B2-2	板楼匹配
	B2-3	别墅匹配
	B2-4	平房匹配
	……	
	B2-n	
B3 统计特征匹配	B3-1	包含户数相等匹配
	B3-2	包含户数差值在阀值内匹配
	……	
	B3-n	

Continued

B4 行业特征匹配	B4-1	用气性质匹配
	B4-2	通气日期匹配
	B4-3	表商匹配
	B4-4	用户状态匹配
	……	
	B4-n	

如表 4 所示楼匹配规则链包括以下内容：

1) 楼名称匹配

根据小区匹配成功的结果，验证两个数据源同一个小区内通过楼名称或别名相同的楼匹配。

如图 7 所示使用楼名称别名匹配规则 B1-2，数据源 B 数据库中的“331 号楼”与数据源 A 数据库中“1 号楼”匹配，类似“332 号楼”与“2 号楼”匹配，“333 号楼”与“3 号楼”匹配。使用楼名称直接匹配规则 B1-1，数据源 B 数据库和数据源 A 数据库的楼名称“A3 号楼”和“A4 号楼”如果相同，表示为楼匹配成功。

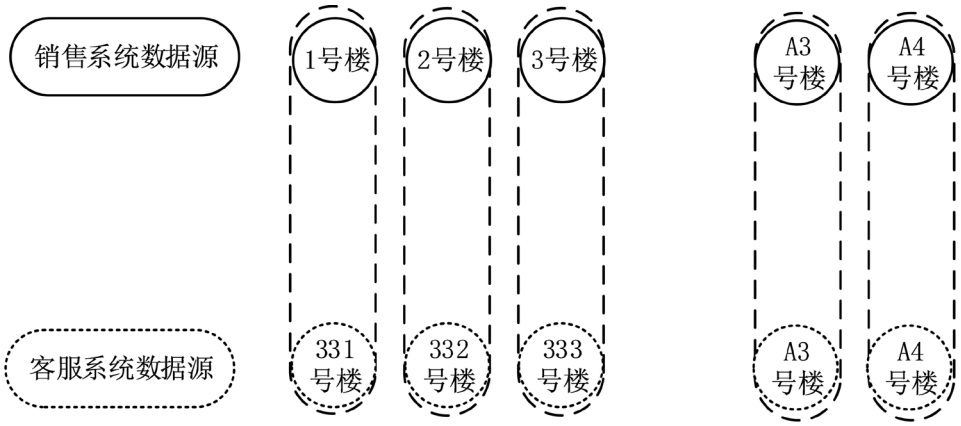

Figure 7. Building name matching
图 7. 楼名称匹配

2) 建筑特征匹配

根据数据预处理的结果和小区匹配成功的结果，验证两个数据源同一个小区内通过楼的建筑特征(塔楼、板楼、别墅、平房)是否匹配。建筑特征匹配不是唯一性匹配条件，但是可以按分类缩小楼匹配的范围。

3) 统计特征匹配

统计每个楼所包含的用户数，用户数可以作为参考值进行匹配。

如图 8 所示，在楼号相同的情况下，统计特征每个楼包括的用户数可以辅助匹配，如果楼号不相同的情况下，无法直接通过统计特征进行匹配。

4) 行业特征匹配

如图 9 所示，行业特征主要包括用气性质匹配、通气日期匹配、表商匹配、用户状态匹配四个规则。用气性质匹配按照普通用户、壁挂炉用户和大用气量用户进行分别匹配，通气日期可以直接进行匹配，表商需要分类匹配，用户状态按照正常、停用和未开户分别匹配。行业特征无法提供唯一性匹配，重要程度不高，但是通过行业特征分类匹配后，可以缩小楼匹配的范围，降低楼匹配的难度。

Cluster1 buildings

Building	Count	Assumed matched rate
1	51	0.901961
2	37	0.918919
3	52	0.884615
4	56	0.857143
5	54	0.870370
6	91	0.912088
7	50	0.940000
8	82	0.975610
9	53	0.905660
10	36	0.861111

Cluster2 buildings

Building	Count	Assumed matched rate
1	48	0.958333
2	36	0.944444
3	48	0.958333
4	48	1.000000
5	48	0.979167
6	84	0.988095
7	48	0.979167
8	84	0.952381
9	48	1.000000
10	36	0.861111

Figure 8. Building matching by count
图 8. 按楼统计用户数匹配

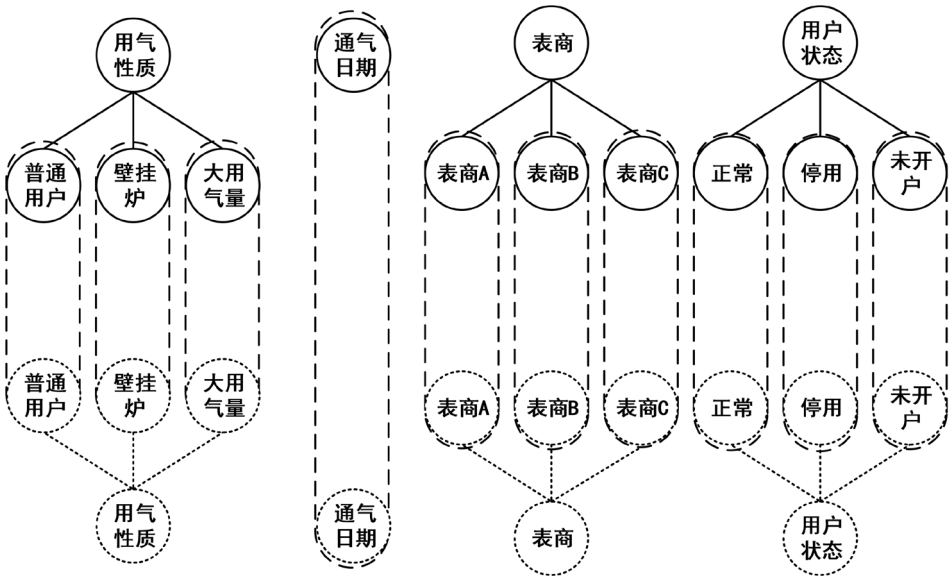

Figure 9. Matching by industry feature
图 9. 行业特征匹配

3.3. 结合建筑、行业标注的详细地址匹配

相同的中文地址信息在两个系统中的详细信息不同，如数据源 A 的中文地址描述为“北京市朝阳区建外街道永安里国际公寓 7 号楼 1 单元 1104”，而在数据源 B 中描述为“朝阳区建外街道永安里国际公寓(通用时代国际中心)7()11104”另外，在数据源 A 中的客户编码和数据源 B 中编码没有统一规范，也没有映射关系，因此无法直接通过技术手段进行关联。本文结合了建筑、行业标注数据实现详细地址匹配，匹配流程如图 10 所示。

详细地址匹配规则链包括塔楼用户匹配、板楼用户匹配、别墅用户匹配和平房用户匹配四个规则链，规则链内的规则可以动态增加，规则链的包含的规则内容如表 5 所示。

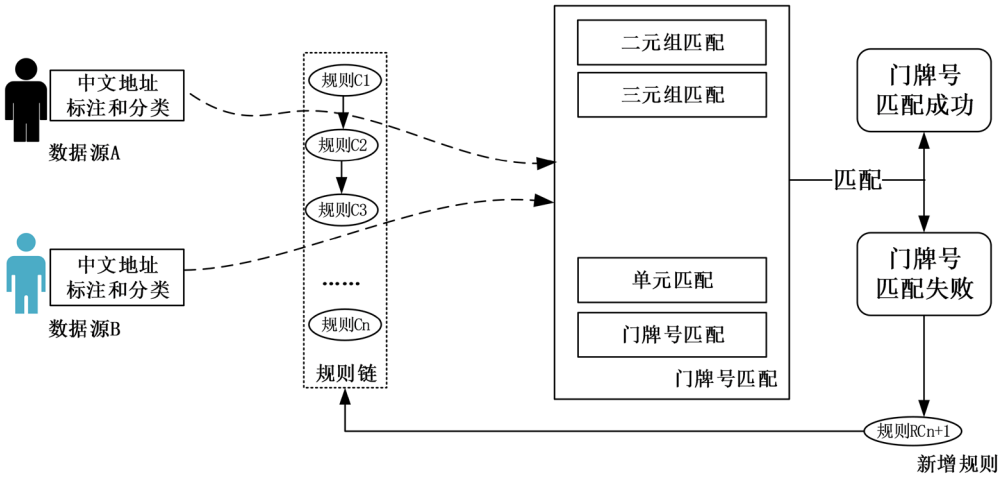

Figure 10. Detailed address matching process
图 10. 详细地址匹配流程

Table 5. Detailed address matching rule chain
表 5. 详细地址匹配规则链

规则链	规则编号	规则描述
C1 塔楼用户匹配	C1-1	二元组匹配
	C1-2	三元组匹配(包含层)
	C1-3	二元组匹配(单元为空)
	……	
	C1-n	
C2 板楼用户匹配	C2-1	三元组匹配(楼－单元－门牌号)
	C2-2	三元组匹配(楼－门－门牌号)
	C2-3	三元组匹配(分隔符)
	……	
	C2-n	
C3 别墅用户匹配	C3-1	二元组匹配
	C3-2	二元组匹配(含单元号)
	……	
	C3-n	
C4 平房用户匹配	C4-1	二元组匹配
	C4-2	二元组匹配(含楼单元)
	……	
	C4-n	

如表 5 所示详细地址匹配规则链包括以下内容：

1) 塔楼用户匹配

塔楼用户中文地址特征是“楼－门牌号”，匹配规则为“二元组匹配”。例如“车公庄北里 36 楼 1901 号”，直接通过数据库关联“楼”和“门牌号”二元组是否一致。

塔楼单元地址要素中录入了层的信息，需要通过层和门牌号结合形成唯一性的地址信息，例如：“东城区东直门街道东方银座 B 号楼 10 层 H”，塔楼，但是操作员习惯性将楼层信息录入单元字段，门牌号中不含楼层信息。这种情况匹配规则为“三元组匹配”，层的信息不能缺失。

塔楼单元地址要素中信息为空，例如：“电子城小区 20 楼单元 0807 号”，只录入中文地址要素单元，但是单元前没有具体化。这类中文地址单元无实际意义，匹配规则为“二元组匹配”。

2) 板楼用户匹配

塔楼用户中文地址特征是“楼－单元－门牌号”，匹配规则为“三元组匹配”。例如：“圣都大厦 20#楼 1 单元 305 号”。

塔楼用户中文地址特征是“楼－门－门牌号”，匹配规则为“三元组匹配”，例如：“鼓楼西大街 113 号 3 门 201”。

塔楼用户中文地址特征是“分隔符”，匹配规则为“三元组匹配”，例如：“开阳里一区-1#-5-301”。

3) 别墅用户匹配

别墅用户中文地址的特点是“小区－楼”，匹配规则为“二元组匹配”。例如“紫玉山庄 K3071 号别墅”。

别墅用户中文地址的特点是“小区－楼单元号”，匹配规则为“二元组匹配”，例如“中海瓦尔登湖别墅 247 楼单元号”，“单元号”无实际意义。

4) 平房用户匹配

平房用户中文地址特征是“小区－门牌号”，匹配规则为“二元组匹配”。例如“二外东平房 15 号”。

平房用户中文地址特征是“小区－楼单元－门牌号”，匹配规则为“二元组匹配”。例如“太平庄 10 号院平房楼单元 8 号”，“楼单元”无实际意义。

4. 实验结果及分析

为了验证模型有效性，以某燃气公司两个不同数据源中的居民用户中文地址为实验数据进行小区匹配、楼匹配和详细地址匹配验证。

4.1. 匹配率测试

分别从两个不同数据源中以组织机构为单位抽取 3 种不同容量的数据样本。

样本 A1，随机抽取数据源 A 中的 1 个服务网点的用户数据，总共包括 76 个小区，245 个楼信息，33,081 条用户数据，抽取数据源 B 中相同服务网点的用户数据，总共包括 82 个小区，231 个楼信息，32,063 条用户数据。

样本 A2，随机抽取数据源 A 中的 2 个服务网点的用户数据，总共包括 117 个小区，716 个楼信息，65,319 条用户数据，抽取数据源 B 中相同服务网点的用户数据，总共包括 126 个小区，705 个楼信息，64,972 条用户数据。

样本 A3，随机抽取数据源 A 中的 3 个服务网点的用户数据，总共包括 181 个小区，907 个楼信息，117,882 条用户数据，抽取数据源 B 中相同服务网点的用户数据，总共包括 190 个小区，924 个楼信息，120,491 条用户数据。

通过表 6 小区匹配结果分析，小区匹配因为数据量较少，虽然两个系统中小区的定义不同而存在别名情况，导致两边的小区数量不一致，但是通过规则链的方式，不断迭代增加别名替换规则，最终是可以实现小区地址要素 90%以上的匹配，小区的匹配结果为后续楼和详细地址匹配提供高质量的基础数据。

Table 6. Community address matching results
表 6. 小区匹配结果

小区匹配规则链数据集	样本数据数量(条)	小区匹配成功率
A1	76 + 82	91.46%
A2	117 + 126	91.26%
A3	181 + 190	91.05%

通过表 7 楼匹配结果分析，匹配失败的楼主要原因是原始数据缺失严重，数据源 A 的楼信息在数据源 B 无法找到，后续需要继续通过人工方式补充缺失的数据。

Table 7. Building address matching results
表 7. 楼匹配结果

楼匹配规则链数据集	总数(条)	楼匹配成功率
A1	245 + 231	93.06%
A2	716 + 705	92.59%
A3	907 + 924	92.42%

通过表 8 详细地址匹配结果分析，详细地址匹配结果作为中文地址匹配的最终结果，随着样本数量逐渐增加，匹配成功率没有显著降低，通过分析匹配失败的详细地址分析，主要原因是数据预处理阶段，无法预知所有的异常处理、缺失数据，需要后续通过不断迭代方式增加和优化规则链，提升数据预处理阶段的数据质量。

Table 8. Detailed address matching results
表 8. 详细地址匹配结果

详细地址匹配规则链数据集	总数(条)	详细地址匹配成功率
A1	33,081 + 32,063	91.89%
A2	65,319 + 64,972	91.09%
A3	117,882 + 120,491	90.46%

4.2. 匹配质量测试

对于 4.1 节中已匹配成功的中文地址进行匹配质量测试，从三组样本集中数据中，分别随机抽取 1000 条，2000 条，3000 条中文地址进行准确率统计，如表 9 所示。

Table 9. Accuracy statistics
表 9. 准确率统计表

中文地址匹配成功数据集	总数(条)	正确条数	准确率
A1	1000	982	98.2%
A2	2000	1950	97.5%
A3	3000	2934	97.8%

如表 9 中所示，三组样本集的中文地址匹配的准确率均保持在 97%以上。通过分析异常匹配数据，

主要原因是原始数据预处理阶段，中文地址标准化过程中，地址要素拆分中存在异常，需要后续优化规则链中的规则内容。

4.3. 实验小结

以某燃气公司两个不同的业务系统中的居民用户中文地址为实验数据，随机抽取三组样本数据进行中文地址匹配。通过实验结果分析中文地址匹配成功率90%左右，准确率97%左右。通过分析未匹配成功的中文地址主要有三个原因：一是中文地址存在缺失数据、重复数据、不规范数据的问题，需要不断增加和优化规则链，通过多次迭代方式逐步提升匹配成功率；二是中文地址标准化过程中地址要素拆分错误，导致后期的匹配失败；三是两个不同业务数据源的中用户数量不对等，存在数据源A中有的中文地址，数据源B中无法找到，这是因为企业缺失用户主数据的统一管理，没有形成统一的管理入口，业务人员分别从多个数据源录入，造成这种用户数据不对等的情况，因此无法实现100%的匹配。综合以上情况，实验结果表明该模型可以一定程度提高中文地址匹配的成功率，保证较高的准确性。

5. 结束语

本文面向非规范的中文地址，设计了一种结合标注的中文匹配规则链模型。该模型针对不同地址要素设计多种规则链。其中包括结合北京市地址标准库设计的小区匹配规则链，融合中文地址中的建筑特征、地理特征、统计特征构建的楼匹配规则链，以及基于二元组和三元组匹配规则设计的详细地址匹配规则链。采用分级匹配、逐级缩小、循环迭代的方式，实现了中文地址匹配。以某燃气公司两个不同的业务系统居民用户中文地址为实验数据进行匹配，实验结果验证了模型的有效性。模型的优点是针对不同特征的中文地址，可以选择不同的规则链，规则链节点内的规则可以通过迭代方式动态增减。不足之处是匹配失败的中文地址，需要人工方式更新或增加规则，继续迭代匹配。

参考文献

[1] 邓斌, 陈会平, 李凯勇. 基于元数据关联特征的交互式数据快速查询[J]. 计算机仿真, 2021, 38(7): 371-375.

[2] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864.

[3] 宋凯丽, 李云岭, 姚露露. 基于条件随机场的分词标注一体化地址解析方法[J]. 测绘地理信息, 2021, 46(S1): 185-187.

[4] 亢孟军, 杜清运, 王明军. 地址树模型的中文地址提取方法[J]. 测绘学报, 2015(1): 99-107.

[5] 李圣文, 凌微, 龚君芳, 周长征. 一种基于熵的文本相似性计算方法[J]. 计算机应用研究, 2016, 33(3): 665-668.

[6] 徐兵, 石少青, 陈超. 基于自然语言的中文地址匹配研究[J]. 电子设计工程, 2020, 28(16): 7-10, 16.