计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400074

基层社会网格治理异构数据字典融合优化方法研究

王 庆 杨万哲 张 聪
东北大学信息科学与工程学院 沈阳 110000

摘 要 数据字典(Data Dictionary,DD)是数据库系统设计内容的重要组成部分,是描述数据库中各数据属性、组成和结构的数据列表集合。一些通用性信息化系统开发过程中,设计开发人员经常遇到如何融合优化既有异构数据字典的问题,这些既有数据字典因设计时缺少行业数据标准或业务范围局限性,在数据表征定义和数据组成及结构设计上差异化明显,但其数据内涵具有高度可融合性,需要花费大量时间和资源通过人工来维护融合数据字典。文中以基层社会网格治理业务背景,针对基层社会治理推广数字化应用开发中异构数据字典融合的痛点问题,研究异构数据字典融合优化方法及相关技术;设计了考虑数据信息完备性和数据结构完整性的数据字典语义去重消歧、关键词提取、相似度计算、数据字典表结构融合方法等4个方面的数据字典融合方法和技术。基于基层社会网格治理业务相关数据字典融合优化实验验证,相较于传统的数据字典融合方法显著提升了融合效率和效果。

关键词:数据字典;数据库设计;编辑距离;相似度计算;基层社会网格治理
中图分类号 TP392

Research on Fusion Optimization Method of Heterogeneous Data Dictionary in Grass-roots Social Grid Governance

WANG Qing,YANG Wanzhe and ZHANG Cong
College of Information Science and Engineering,Northeastern University,Shenyang 110000,China

Abstract Data dictionary(DD) is an important part of the database system design content,and it is a collection of data lists that describes the attributes,composition and structure of the data in the database. In the development process of some general-purpose information systems,designers and developers often encounter the problem of how to integrate and optimize existing heterogeneous data dictionaries. Due to the lack of industry data standards or business scope limitations,these existing data dictionaries differ significantly in data representation definition,data composition and structure design,but their data content is highly convergable. It takes a lot of time and resources to manually maintain a converged data dictionary. Based on the business background of grass-roots social grid governance,this paper aims at the pain points of heterogeneous data dictionary fusion in the development of grass-roots social governance promotion digital application,and studies the optimization methods and related technologies of heterogeneous data dictionary fusion. The methods and techniques of data dictionary fusion are designed,which consider the completeness of data information and the integrity of data structure,such as semantic deduplication and disambiguation,keyword extraction,similarity calculation and table structure fusion. Based on the experimental verification of data dictionary fusion optimization of grass-roots social grid governance business,the fusion efficiency and effect are significantly improved compared with the traditional data dictionary fusion method.

Keywords Data dictionary,Database design,Edit distance,Similarity calculation,Grass-roots social grid governance

1 引言

大数据时代的来临对基层社会网格治理工作提出了新的发展要求,各项应用数据的及时同步、展示、查询都会为基层社会治理工作带来较大的便利。我国“十四五”规划明确提出,要加强公共数据开放共享、推动政务信息化共建共用、提高数字化政务服务效能。数据共享有助于推动政府治理体系和治理能力现代化,不断提高决策科学性和服务效率。

数据字典在确保数据库共享性、安全性、完整性、一致性、有效性、可恢复性和可扩充性方面发挥着关键的作用。它提供了关于数据的逻辑描述信息,为数据库的设计、实现、运行、维护和扩充提供了标准和依据。

国外从20世纪80年代开始已经把数据字典技术应用在计算机系统、信息化等方面。1984年英国的 Yvette Asscher

基金项目:国家重点研发计划(2021YFC3300300)
This work was supported by the National Key Research and Development Program of China(2021YFC3300300).
通信作者:王庆(wangqing@ise.neu.edu.cn)

首次将数据字典作为计算机的业务流程描述工具[1]；同年美国的 Julia Van Duyn 介绍了数据字典在数据安全以及节约计算机资源方面的优势[2]；1986 年美国的 Shamkant B. Navathe 等强调了数据字典在信息资源管理系统中的重要性[3]；1994 年意大利的 FioraPirri. Clara Pizzuti 介绍了通过编程设计数据字典的方法[4]；2010 年美国的 Andrew D. Arenson 等研究了如何在公共领域应用数据字典[5]；2006 年 Catherine Lai 和 Ichiro Fujinaga 设计了模拟录音的元数据数据字典[6]；2012 年 Udo doebrich 和 Roland heidel 介绍了网络物理系统中电子数据字典的概念[7]；2012 年 Alexandros bentevis 等介绍了关于智能手机的数据字典结构[8]。

数据字典融合问题指在政府部门或公司信息化系统设计开发或适配过程中，经常需要将不同来源的两个以上既有数据字典适应并与新系统集成，以确保数据字典在新业务环境中的准确性、有效性和完备性。

在传统异构数据字典融合解决方案中，面临着以下问题：

1）传统解决方案一般由原数据库设计人员或资深系统分析设计人员重新进行数据字典的适配融合调优，人员成本高、周期长。

2）数据字典主要表现形式为表格，数据字典示例如表 1 所列，字段短维度高，传统的文本融合方法准确率不高，如何实现基层社会治理数据字典实体对齐（entity matching）、如何筛选关键词、计算数据字典表间相似度、设计数据字典表间融合方法等。

基于以上问题和研究现状，以基层社会网格治理数字化应用系统需求为背景，本文提出了一种异构数据字典间融合方法流程来解决上述问题，实现基层社会网格治理异构数据字典间的融合。

表 1 数据字典示例

Table 1 Example of data dictionary

序号	数据项名称	数据项代码	数据项类型	数据项长度	选/必填	备注
1	身份证号	CARDID	Varchar	30	是	身份证编码应符合 GB 11643
2	证件类型	ID_TYPE	Varchar	10	是	编码应符合 GA/T 517
3	人口类型	PEOPLE_TYPE	Varchar	2	是	01 常住人口 02 流动人口 03 境外人口
4	姓名	NAME	Varchar	100	是	
5	性别	SEX	Varchar	2	是	编码应符合 GB/T 2261.1
6	户籍门（楼）详址	RESIDENCE_ADDR	Varchar	200	是	
7	实际居住地	INHABITED_AREA	Varchar	20	是	
8	出生日期	BIRTH_DAY	Date		是	格式为“YYYY-MM-DD”
9	出生地	BIRTH_PLACE	Varchar	200	是	
10	民族	NATIONAL	Varchar	10	是	编码应符合 GB/T 3304
11	照片	PIC	Longblob		是	

2 相关工作

2.1 文本融合相关方法

数据字典表间的相似度计算是数据字典融合模型中的一个重要环节。2011 年 Huang 等[9]提出了一种结合 NLP 技术和 TF-IDF 算法的文本相似度量方法；2012 年 Li 等[10]提出了一种基于词义的文本相似度计算方法；2013 年 Zhan 等[11]提出了一种多词条加权计算词语相似度的方法；Wang 等[12]提出了一种基于 LDA 模型的文本相似度计算方法；2015 年 Zhang 等[13]提出了一种基于义原（语言本质）语法树的文本相似度计算方法；2019 年 Ji 等[14]针对短文本扩展特征提出了一种文本分类方法。

2.2 异构数据字典融合相关技术方法

Synonyms 是一个开源的中文近义词工具包，在中文实体去重消岐上有较好效果[15]；TF-IDF（词频-逆文档频率）算法是 NLP（自然语言处理）领域常用算法之一，计算文档中关键词的权重来提取文档的关键词[16]。

2014 年黄磊加入了类内离散度（Di）来提升 TD-IDF 算法的关键词提取准确度[17]；2018 年王杰在原 TF-IDF 算法上引入了权重影响因子[18]；编辑距离，又称 Levenshtein 距离，是由俄罗斯科学家 Vladimir Levenshtein 于 1965 年提出的，具体指两个字符串之间，由一个字符串转成另一个字符串所需的最少编辑操作次数[19]；Frederickj Damau 提出了改进 Levenshtein 距离的 Damerau-Levenshtein 距离，加入了置换操作对编辑距离的影响[20-21]。

3 数据字典融合问题描述及融合优化模型

3.1 问题描述

给定异构数据字典 A 和数据字典 B，它们各自由多张数据字典表及表格间的联系组成，例如表 1 所列，这些数据字典是任何类似于基层社会网格治理应用的数字化系统设计开发及运营的基础。数据字典表中的信息包括数据项定义、数据项名称、数据项英文标识、数据项类型、数据项长度、数据项缺省值、数据项的主键和外键等数据逻辑描述信息。

本文的研究目标是如何高效地实现数据字典 A 和数据字典 B 之间的融合优化，以便集成多方异构数据，提高数字化应用系统设计开发的质量和效率。

3.2 融合优化模型结构

图 1 给出了基层社会网格治理异构数据字典间融合方法模型的具体逻辑结构，该模型从上向下共分为 5 个逻辑层，即输入文件准备、数据字典数据项预处理、数据字典语义特征提取、相似度计算、数据字典融合。

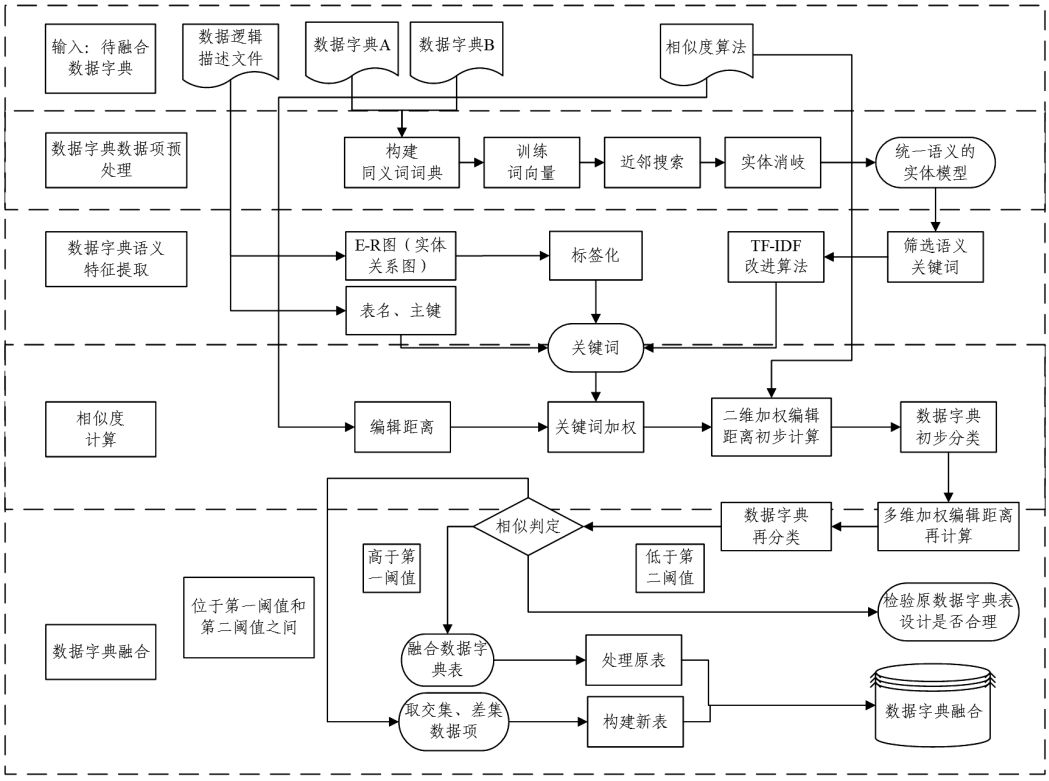

图 1　基层社会网格治理异构数据字典间融合方法模型

Fig. 1　Fusion method model of heterogeneous data dictionaries in grass-roots social grid governance

3.2.1　数据字典数据项预处理

在数据字典融合过程中，首先面临多词一义问题。举例而言，作为实体或特征“马达”和“发电机”，它们指向同一业务实体或实体特征，但在语义表达上存在明显差异。因此，在数据字典融合过程中，需要对同义词进行去重和消岐处理。本文采用建立同义词典的方法进行数据字典数据项去重消岐，利用工具包 Synonyms 加以实现。该方法首先是通过训练 word2vec、改进 gensim 函数以及最近邻相似检索等方法找到最终的相似关键词。其中训练词典容量为 435 和 729。工具包和人工比对词语相似度如表 2 所列。

表 2　工具包和人工比对词语相似度

Table 2　Toolkit and manual comparison of word similarity

词 1	词 2	Synonyms	人工评定
中午	正午	0.900	0.855
轿车	汽车	0.892	0.980
旅游	游历	0.649	0.960
工具	器械	0.881	0.737

3.2.2　数据字典语义特征提取

关键词筛选：本文从解析数据逻辑描述信息和设计关键词提取算法两个方面来筛选关键词。图 2 给出了数据字典表间关联关系 E-R 图的实例，图 3 给出了从标签化角度描述数据字典间联系的关系图谱实例。通过综合利用解析数据逻辑描述信息和关键词提取算法，可以有效地筛选出数据字典中的关键词。

在数据字典的设计文件中，数据逻辑信息提供了部分关键词的线索，例如表名、主键、实体关系图（E-R 图）中的关系和标签等特征。表名、主键是设计人员对数据字典表内容和功能的概括性描述，E-R 图和标签是对数据字典整体结构以及数据字典表内容梗概的概括性描述。

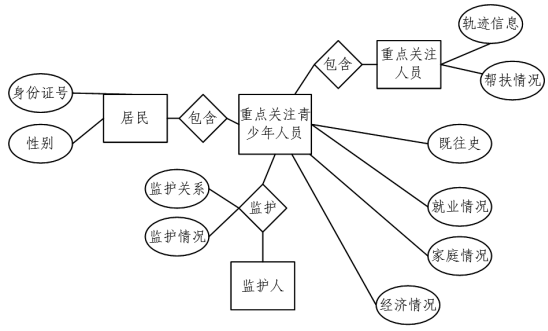

图 2　实体关系图（E-R 图）

Fig. 2　Entity relationship diagram（E-R diagram）

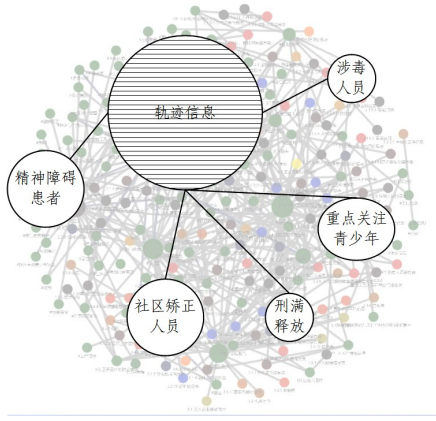

图 3　关系图谱

Fig. 3　Relationship map

关系图谱体现了数据字典表间逻辑关联关系。此外，还可以采用关键词提取算法来识别和提取关键词。

关键词提取 TF-IDF（Term Frequency-Inverse Document Frequency，词频-逆文件频率）是一种用于寻找关键词的常用加

权技术，可以用来评估某一字段在数据字典中的重要程度。改进后的加权 TF-IDF 算法在数据字典的关键词提取上表现良好。

算法 1 加权 TF-IDF 算法

TF 是词频，其表达式如下：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中，$n_{i,j}$ 是该词在文件 d_j 中出现的次数，分母则是文件 d_j 中所有词汇出现的次数总和。

λ 表示修正系数，其表达式如下：

$$\lambda=\left(\frac{1+T_i}{1+e^{pi}}\right) \tag{2}$$

其中，T_i 表示词 i 的贡献度，pi 表示词 i 所代表主题的频度，词贡献度可以用以下计算式求得。

$$\sum_{i=1}^{n} T_i C_i \tag{3}$$

IDF 是逆文档频率，其表达式如下：

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{4}$$

其中，$|D|$ 是语料库中的文件总数，$|\{j:t_i\in d_j\}|$ 表示包含词语 t_i 的文件数目。

$$TF\text{-}IDF=TF*\lambda*IDF \tag{5}$$

3.2.3 数据字典表间的相似度

相似度计算：基层社会网格治理数据字典文本构成形式多维且稀疏；在一张数据字典表中，其组成成分多元，有短语、短文本、数据类型等；语言构成形式多维：有英文、中文、阿拉伯数字、数据库语言等。因此，在计算相似度时需要考虑到不同种类关键词之间的权重占比。综上所述，本文选择了多维加权编辑距离作为数据字典间的相似度计算方法。

加权编辑距离的计算式如下：

$$lev_{a,b}(i,j)=\begin{cases} lev_{a,b}(i-1,j)+del[a(i)] \\ lev_{a,b}(i,j-1)+ins[b(j)] \\ lev_{a,b}(i-1,j-1)+sub[a(i),b(j)]_{(a_i\neq b_j)} \end{cases} \tag{6}$$

其中，$lev_{a,b}(i,j)$ 表示 a 的前 i 个字符与 b 的前 j 个字符之间的编辑距离（i 和 j 都从 1 开始）。

等号右侧第一个公式表示删除、第二个表示插入、第三个表示替换，分别是 3 种编辑条件下的编辑距离。

多维加权编辑距离的计算式如下：

$$d=\sum_{i=1}^{n} u_i l_i \tag{7}$$

其中，l_1 表示关键词 1 对应的加权编辑距离，其中关键词 1 对应的是数据字典表名；u_1 表示关键词 1 对应的权重系数；d 表示数据字典表间的相似度计算结果。

权重设计：这里的权重可以是基于统计的权重表格，也可以是基于一定规则的运算，例如数据字典里来源于国标的字段权重大于来源地方或行业标准的字段。

本文在计算数据字典之间的相似度时，采用了一种多阶段的方法。首先，基于数据逻辑描述信息调整预设权重，以得到每个目标关键词对应的第一权重。然后，根据第一权重，计算不同待融合数据字典中目标关键词之间的二维加权编辑距离。通过这一步骤，本文对目标关键词对应的数据字典表进行了初次分类，并得到了每个数据字典表对应的粗类别。在同一粗类别中，进一步判断数据字典表中的每个字段是否为目标关键词，然后确定字段的第二权重，并根据第二权重计算字段间的多维加权编辑距离。通过多维加权编辑距离，对字段对应的数据字典表进行了再次分类，得到每个数据字典表对应的细类别。通过以上的多阶段处理过程，能够更精确地计算数据字典之间的相似度，并将它们进行分类。

3.2.4 数据字典融合

融合阈值设计的步骤如下：在计算数据字典表间的相似度后，根据实际模型需求，将相似度计算结果按照区分度设计阈值，由于本文在数据字典表间预想实现融合、取交集生成新表、不融合 3 个处理模式，因此选择将两个区分度明显的数值设为第一相似度阈值和第二相似度阈值，以实现相应的数据字典表处理。

在同一细类别中，计算不同数据字典表之间的表格相似度。如果表格相似度大于设定的第一相似度阈值，将融合这些数据字典表；如果表格相似度小于第一相似度阈值但大于第二相似度阈值，提取数据字典表字段交集和差集，保留原表差集并将交集生成新的数据字典表；如果表格相似度小于第二相似度阈值，将结束融合优化过程，并生成相应融合提示信息，用于指示数据字典之间的相似性程度。同时返回到获取每个字段的设计信息的步骤，以获取新的设计信息。这些新的设计信息包括字段类型、字段长度以及缺省值等以升维的方式重新计算相似度，从而判断是否进行数据字典表融合。

通过以上的步骤，能够根据表格相似度阈值对数据字典进行融合和提取，并根据不同的相似度情况得到最终的目标数据字典。这种方法根据相似度的不同情况进行灵活处理，并提供相应提示信息，以辅助设计开发人员进行数据字典融合优化决策。

4 融合实例分析

4.1 数据集

实验选用的数据集共 104 张数据字典表（数据字典表如表 1 所列），包含 1504 个独立字段，由多个基层社会治理专业领域收集而来的不同数据字典组成：数据字典 A 和数据字典 B。实验数据集如表 3 所列。

表 3 实验数据集

Table 3 Experimental dataset

数据集	数据字典表数量	字段数量
A	87	1176
B	17	328

4.2 数据项去重消岐

如图 4 所示，首先通过去重方法去除了其中的 130 个重复项（占所有字段的 8.64%）。

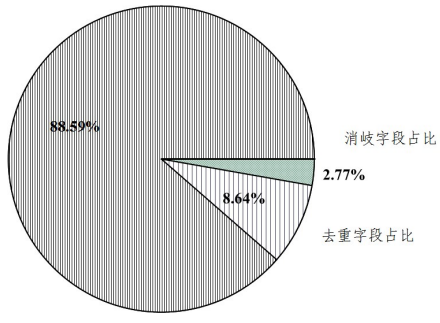

图 4 去重、消岐占所有字段比例

Fig. 4 Weight removal, disambiguation accounted for the proportion of all fields

此外，还进行了消岐处理，将类如“联系方式”和“电话号

码”以及“出生日期”和“生日”等 42 个字段（占所有字段的 2.77%）进行了消歧，确保数据字典的准确性和一致性。经过去重和消岐处理后，原数据字典表剩下 1 374 个独立的数据项。

4.3　关键词提取

如图 5 所示，从数据逻辑描述信息中提取了 157 个关键词（占所有信息的 10.41%），其中包括表名和主键，例如“重点青少年”和“身份证件”。同时基于关键词提取方法加权 TF-IDF 算法得到“年龄”等 52 个关键词（占所有信息的 3.47%）。基于 TF-IDF 算法的关键词权重实例如表 4 所列。

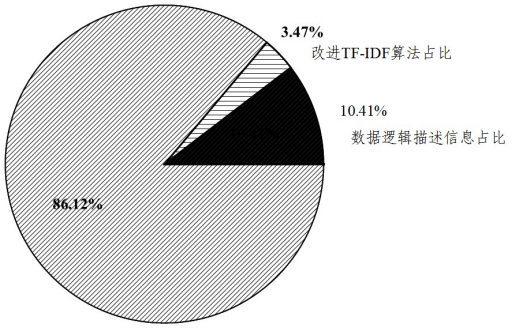

图 5　不同关键词提取方法占比

Fig. 5　Proportion of different keyword extraction methods

表 4　基于 TF-IDF 算法的关键词权重实例

Table 4　Keyword weight example based on TF-IDF algorithm

（%）

数据项	TF-IDF 值
身份证号	8.806
年龄	7.072
证件类型	2.935

4.4　数据字典粗分类

截取数据字典 A 和数据字典 B 中的各 5 个数据字典表及其对应的信息集合进行了标识为 $a-e$ 的编号，同时对来自数据字典 B 的 5 个数据字典表及其对应的信息集合进行了标识为 $f-j$ 的编号。信息集合包括表名、主键、关键词等所有数据字典表，$a-j$ 表现为数据字典表名＋主键，表名的权重设置为主键的 1.5 倍，以更准确地评估表之间的相似性。具体的相似度计算结果如表 5 和图 6 所示。以相似度 8－10 为分界线，数据字典表被分成了两类。

表 5　数据字典的粗分类

Table 5　Coarse classification of data dictionary

	A	b	c	d	e
f	12	12	13.5	**7.5**	17
g	**6**	**6**	**7.5**	14.5	17
h	**6**	**3**	**7.5**	13.5	17
i	9	9	**4.5**	16	17
j	15.5	15.5	15.5	14.5	**6**

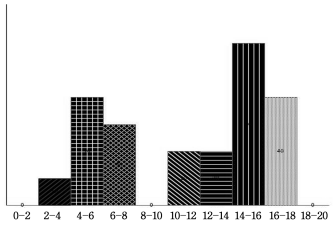

图 6　数据字典表间相似度数量分布

Fig. 6　Distributions of the number of similarity between data dictionary tables

4.5　数据字典细分类

在 4.4 节中，不同来源的数据字典表完成了粗分类，在此基础上加入字段类型并按照表名：关键词：字段类型结构关系，赋予 6∶3∶1 的权重比例，再次计算相似度，相似度计算结果如表 6 所列。其中，a，g 和 c，i 之间的相似度更为明显。

表 6　数据字典细分类

Table 6　Fine classification of data dictionary

	a	b	c
g	**4.5**	14.5	7.9
h	9.3	14.5	9.9
i	8.5	15.7	**3.1**

4.6　数据字典融合

基于步骤 4.5 节的相似度计算结果（见表 6），其中相似度值为 4.5 和 3.1 对应的数据字典表格两两判定为相似，进行数据字典表的融合：将两张数据字典表的所有内容进行对比，设置阈值，当相似度大于阈值时进行数据字典的融合。数据字典表 E-R 图融合如图 7 所示。

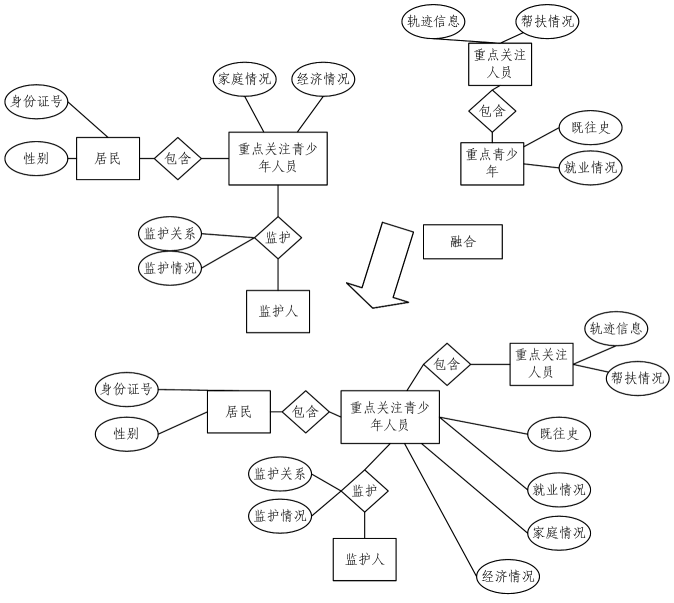

图 7　数据字典表 E-R 图融合

Fig. 7　Data dictionary table E-R diagram fusion

5 实验结果分析

5.1 阈值分析

对比不同阈值下的实验融合效率，阈值为 70%时融合优化模型的整体性能效果表现最好。不同阈值下融合效率对比如图 8 所示。

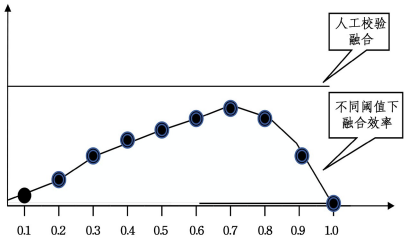

图 8 不同阈值下融合效率对比

Fig. 8 Comparison of fusion efficiency under different thresholds

5.2 融合效率对比

在 6 次数据字典融合优化流程分组实验中，将每次融合耗费的时间作为融合效率高的低评估依据。实验结果如图 9 所示，6 个融合流程中优化方法平均花费时间为 205 s，而传统融合流程平均花费时间是优化方法所用时间的 7 倍以上，融合效率提升效果明显。

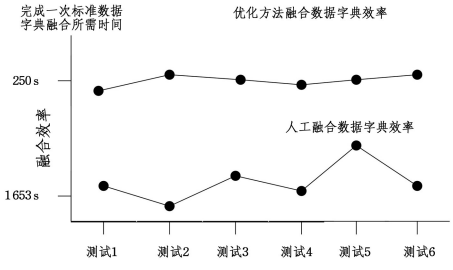

图 9 数据字典融合效率对比

Fig. 9 Data dictionary fusion efficiency comparison

5.3 融合准确率对比

设置人工融合后的数据字典检验细化调整后的准确率为 100%，如图 10 所示，传统优化方法的平均准确率为 98.3%，模型优化方法的组平均准确率为 87.5%，准确率小幅下降。

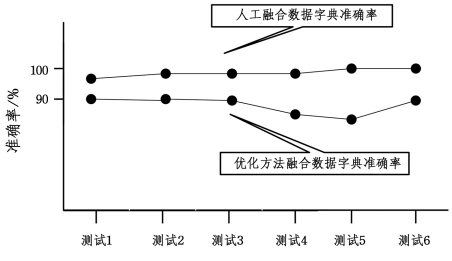

图 10 数据字典融合准确率对比

Fig. 10 Data dictionary fusion accuracy comparison

本节主要从阈值设计、效率、准确率 3 方面对数据字典融合方法的性能进行评估，并将其与传统数据字典融合方法进行对比分析。实验表明，本文所设计数据字典融合优化模型显著提升了多源异构数据字典融合效率，同时保持了较高的准确率。

结束语 本文在考虑数据信息完备性和数据结构完整性的情况下，设计了一套从数据字典语义去重消岐、关键词提取、相似度计算、数据字典表结构融合方法等 4 个方面的数据字典融合方法和技术，用于解决异构数据字典间融合问题。同时基于基层社会网格治理业务相关数据字典融合优化实验进行验证，相较于传统的数据字典融合方法显著提升了融合效率和效果。但相较于人工融合效果，准确率略有下降的原因主要在于同义词词典相对于基层社会治理领域知识表达的完备性不足，所设计的相似度匹配算法不足以充分表达复杂的业务信息映射结构关系，同时对基于深度学习、大模型的数据字典融合方法缺乏考虑，后续研究可以重点关注如何构建专业领域的同义词词典、优化相似度匹配算法、在数据字典融合流程中应用深度学习、大语言模型等方法。

参 考 文 献

[1] YVETTE A. Describing businesses with data dictionaries [J]. Data Processing, 1984, 26(6): 17-19.

[2] JULIA V D. Data dictionaries as a tool to greater productivity [J]. Data Processing, 1984, 26(6): 14-16.

[3] SHAMKANT B N, LARRY K. Role of data dictionaries in information resource management [J]. Information & Management, 1986, 10(1): 21-46.

[4] FIORA P, CLARA P. Explaining incompatibilities in data dictionary design through abduction [J]. Data & Knowledge Engineering, 1994, 13(2): 101-139.

[5] ANDREW D. ARENSON. Implementation of a shared data repository and common data dictionary for fetal alcohol spectrum disorders research [J]. Alcohol, 2010, 44(7/8): 643-647.

[6] CATHERINE L, ICHIRO F. Metadata Data Dictionary for Analog Sound Recordings [C]// Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries(JCDL'06). 2006: 344.

[7] ALEXANDEROS B, IOANNIS K, VANA K. Dictionary data structures for smartphone devices[C]// Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments. 2012: 1-4.

[8] UDO D, ROLAND H. Cyber-physical system description model [J]. Chinese instrument, 2013(10): 41-47.

[9] HUANG C H, YIN J, HOU F. A text similarity measurement method combining lexical semantic information and TF-IDF method[J]. Journal of Computer Science, 2011, 34(5): 856-864.

[10] LI M T, LUO J Y, YIN M J. A method to calculate the weight of text feature words combined with their meaning[J]. Computer Application, 2012, 32(5): 1355-1358, 1365.

[11] ZHAN Z J, LAING L N, YANG X P. Word similarity calculation based on Baidu Encyclopedia[J]. Computer Science, 2013, 40(6): 199-202.

[12] WANG Z Z, HE M, DU Y P. Text similarity calculation based on LDA topic model[J]. Computer Science, 2013, 40(12): 229-232.

[13] ZHANG H Y, LIU D B, WEN C Y. Research on word semantic similarity improvement algorithm based on Knownet [J]. Computer Engineering, 2015(2): 151-156.

[14] XIN Y F, FU Y X, MA L. Short text classification based on frequent item feature extension [J]. Computer Science, 2019, 46(z1): 478-481.

[15] WANG H L. Predicts 2023: Synonyms [EB/OL]. (2017-09-27) [2023-11-22]. https://github.com/huyingxi/Synonyms/doc.

[16] LIU G Z，ZHANG J H，WANG H D. The application of TF-IDF algorithm in e-commerce simulation training platform is improved [J]. Computer Simulation，2023，40(7)：273-277.

[17] HUANG L，WU Y P，ZHU F Q. Research and improvement of automatic keyword extraction method[J]. Computer Science，2014，41(6)：204-207.

[18] WANG J. LI X J. Improved TFIDF label extraction algorithm [J]. Software Engineering，2018，21(2)：4-6.

[19] GRAVANO L，IPEIROFIS P G，JAGADISH H V. Approximate String Joins in a Database[C]// Proceedings of the 27th International Conference on Very Large Data Bases. 2001：491-500.

[20] SONDIK E J. The optimal control of partially observableMarkov processes over the infinite horizon：discountedcosts [J]. Operations Research，1978，26(6)：282-304.

[21] SYAROFINA S，BUSTAMAM A，YANUSRA，et al. The distance function approach on the MiniBatchKMeans algorithm for the DPP-4 inhibitors on thediscovery of type 2 diabetes drugs [J]. Procedia Computer Science，2021(179)：127-134.

WANG Qing，born in 1969，Ph.D，associate professor. His main research interests include modeling and optimization，manufacturing and service planning and scheduling，logistics & supply chain resources planning，e-Commerce business optimization，intelligent optimization algorithm.