

## 引文格式：

韩铖山, 李红辉, 闫佳和, 等. 基于知识图谱的多源异构信息通道耦合技术研究 [J]. 集成技术, 2023, 12(3): 48-60.

Han CS, Li HH, Yan JH, et al. Research on coupling technology of multi-source heterogeneous information channels based on knowledge graph [J]. Journal of Integration Technology, 2023, 12(3): 48-60.

## 基于知识图谱的多源异构信息通道耦合技术研究

韩铖山<sup>1</sup> 李红辉<sup>1,2\*</sup> 闫佳和<sup>1</sup> 林映利<sup>1</sup> 屈靖淇<sup>1</sup> 贾志伟<sup>1</sup>

<sup>1</sup>(北京交通大学计算机与信息技术学院 北京 100044)

<sup>2</sup>(高速铁路网络管理教育部工程研究中心 北京 100044)

**摘要** 政务数据资源来源广泛、类型多样、数据量大且分布情况不清晰, 缺乏统一管理, 数据获取和使用效率较低, 无法持续释放数据价值。为解决上述问题, 实现不同来源和不同类型数据的关联融合, 该文采用多源异构数据融合技术, 为政务数据的交换和共享提供技术支持。该文不仅提出了信息通道和通道耦合的概念, 还提出一种基于知识图谱的多源异构信息通道耦合的方法, 可先实现多源异构数据统一化, 再进行数据融合。该方法包括构建初始通道耦合知识图谱、基于通道耦合知识图谱实现通道数据耦合和基于通道耦合知识图谱实现知识更新 3 个模块。基于知识图谱的多源异构信息通道耦合的方法将图谱构建、知识抽取、知识融合、知识加工、知识更新等技术相结合, 通过一种自顶向下的排序检索模型, 加快了知识融合和数据检索的速度和准确度。

**关键词** 多源异构信息; 数据融合; 通道耦合; 知识图谱

中图分类号 TP 274; TP 391; D 035.1 文献标志码 A doi: 10.12146/j.issn.2095-3135.20221026001

### Research on Coupling Technology of Multi-source Heterogeneous Information Channels Based on Knowledge Graph

HAN Chengshan<sup>1</sup> LI Honghui<sup>1,2\*</sup> YAN Jiahe<sup>1</sup> LIN Yingli<sup>1</sup> QU Jingqi<sup>1</sup> JIA Zhiwei<sup>1</sup>

<sup>1</sup>(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

<sup>2</sup>(Research Center for High-speed Railway Network Management, Ministry of Education, Beijing 100044, China)

\*Corresponding Author: hqli@bjtu.edu.cn

**Abstract** Government data resources are characterized by a wide range of sources, diverse types, large data volumes and unclear data distribution, which can lead to lacks of unified management, high efficiency in data access and use, and ability to continuously promote the data release value. In order to solve the above problems

收稿日期: 2022-10-26 修回日期: 2023-01-06

基金项目: 国家重点研发计划项目 (2019YFB2102500)

作者简介: 韩铖山, 硕士研究生, 研究方向为知识图谱、通道耦合; 李红辉(通讯作者), 教授, 研究方向包括大数据技术与工程、人工智能、轨道交通信息技术等, E-mail: hqli@bjtu.edu.cn; 闫佳和, 博士研究生, 研究方向为数据融合; 林映利, 硕士研究生, 研究方向为协议转换; 屈靖淇, 硕士研究生, 研究方向为区块链; 贾志伟, 硕士研究生, 研究方向为联邦学习。

and realize the correlative fusion of data from different sources and different types, this paper adopts the fusion of multi-source heterogeneous data to provide technical means for the exchange and sharing of government data. This paper proposes the concept of information channel and channel coupling, and provides a multisource heterogeneous information channel coupling method, which includes three module for constructing the Initial channel coupling knowledge graph, realizing the channel data coupling based on the channel coupling knowledge graph, and realizing the knowledge updating based on the channel coupling knowledge graph, respectively. In the proposed method, the technologies of atlas construction, knowledge extraction, knowledge fusion, knowledge processing and knowledge updating are integrated together, and a top-down sorting retrieval model is proposed to accelerate the speed and accuracy of knowledge fusion and data retrieval.

**Keywords** multi-source heterogeneous information; data fusion; channel coupling; knowledge graph

**Funding** This work is supported by National Key Research and Development Program of China (2019YFB2102500)

## 1 引言

自 20 世纪 90 年代起, 政府的数字化建设一直是公共管理的热门领域<sup>[1]</sup>, 其经历了“电子政府”“数字政府”和“智慧政府”3 个阶段。近年来, 政务数据量迅速增大, 政府数字化建设进程进一步加快, 但现阶段政务数据仍存在缺乏统一标准、获取和使用效率较低等问题。为使数字政务更好地面向公共服务, 提高政务数据协同能力, 跨部门、跨行业的不同格式的数据融合处理变得尤其重要。

目前, 我国在政务数据融合方面的研究已经取得了一定成果。国家信息中心信息与网络安全部提出的“1+1+N”模式的政务网络安全监测平台架构<sup>[2]</sup>, 支持多源数据采集、集成接口和规范化处理, 并提出基于多成分引擎的数据关联分析模型和大数据融合分析技术, 实现了协同检测和态势感知。南京市国土资源信息中心<sup>[3]</sup>搭建了一体化服务系统, 整合构建“一门户、三系统”的政务服务应用体系, 实现了多系统间服务对接和数据系统。上海市浦东新区大数据中心<sup>[4]</sup>利用

大数据技术构建了政务数据融合服务平台, 探索出体系化的地址数据标准化融合管理方式和技术方法。综合来看, 各平台已形成了具有针对性、先进性的数据融合管理方案, 但更侧重于数据交换共享, 缺乏对数据的知识层语义化提取, 融合集成程度不足。

2012 年, 谷歌公司发布了知识图谱项目<sup>[5]</sup>, 知识图谱的概念被提出。知识图谱本质上是一个具有属性的实体通过关系链接形成的网状知识库<sup>[6]</sup>, 可通过推理实现概念检索。知识图谱通过信息抽取方法从各类型的数据中抽取实体、关系和属性, 形成基于特征的知识表达; 获取知识后, 消除歧义和矛盾, 实现知识融合; 每次融合新的知识, 都需要识别知识是否合格, 并将合格的知识加入知识库, 以推理拓展现有知识。与传统方法相比, 知识图谱具有更高的效率和更好的扩展性, 为多源异构数据的融合提供了知识层面融合的可能。2019 年, 中国电子技术标准化研究院发布的《知识图谱标准化白皮书》<sup>[7]</sup>提出了一种智慧政务生态的方式, 以实体为基本单位对政务数据进行深度挖掘, 实现数据层面的融合集

成, 打破了上述平台的政策单一获取模式, 转型升级到多维立体知识检索发现服务模式, 可深度探索政策关系。

针对政务数据在融合过程中缺乏知识层面提取的问题, 受《知识图谱标准化白皮书》的启示, 本文提出一种以知识图谱为核心搜索引擎的多源异构政务数据通道耦合方法, 利用知识图谱高效率 and 可拓展的特性, 将大规模、碎片化的多源异构政务数据关联融合, 以图的形式体现客观世界中机构、政策、公文、法律法规、解读等实体之间的关系, 实现多源数据在知识层面的融合, 以解决政务领域数据融合程度不足的问题。该方法基本思路如下: 首先, 构建初始通道耦合知识图谱, 使用命名实体识别(named entity recognition, NER)模型<sup>[8]</sup>进行实体抽取, 完成图谱构建; 然后, 基于构建完成的通道耦合知识图谱, 实现多源异构数据的耦合, 通过知识抽取、知识融合、知识加工等方式深度挖掘政务数据特征, 完成知识层面的数据融合, 不停地扩建知识图谱; 最后, 基于通道耦合知识图谱实现知识更新, 采用增量更新的方式, 对每轮输入的目标词与通道耦合知识图谱作语义匹配搜索, 更新迭代通道耦合知识图谱中的数据。

## 2 多源异构数据融合现状

### 2.1 多源异构数据融合

在信息化建设过程中, 数据存储方式和管理系统的升级变化, 致使数据呈现出零散异构的缺点, 不利于处理得到有价值的信息。多源数据融合就是综合分析不同格式的数据信息, 获取数据特征, 并从中提取比单一数据更丰富的信息。20世纪80年代, 美国已经开始重视数据融合, Waltz等<sup>[9]</sup>构建了最初的理论数据融合框架, 为后续的数据融合技术发展奠定了基础。近些年, 数据融合算法得到了改良, Pandey等<sup>[10]</sup>基于

Web技术, 对异构的警报、天气、地址等信息进行融合, 提出了安全警示系统 Safety Check; 冀振燕等<sup>[11]</sup>基于社交应用场景, 提出一种可扩展的融合多源异构数据推荐模型, 可对数据源层面的评分、评论和社交等异构数据进行融合; 金鼎等<sup>[12]</sup>提出一种将图数据库技术与政务空间技术相结合的方法, 通过分析政务空间数据之间的关系实现融合。上述算法为不同领域的的数据融合提供了支撑, 但大多基于特定的场景, 数据的预处理较为简单, 因此并无拓展研究。

随着数据融合研究的不断深入, 数据融合算法的研究开始趋向于多应用场景, 愈加关注多源异构数据统一格式的问题。丁玥等<sup>[13]</sup>提出的多源关系数据融合框架 MSF, 利用属性间对齐发现机制、多样性取样策略和实体特征抽取方法, 对数据格式进行统一, 从而提高模式匹配效率。郭帅童<sup>[14]</sup>提出一种基于动态探针的多源异构数据融合技术, 对多源异构数据建立统一关联关系并进行数据融合, 此外, 还提出一种基于模糊逻辑的数据融合框架。周琥晨<sup>[15]</sup>提出一种关系感知实体匹配方法, 通过获得属性列之间的标准化结构信息, 实现面向多源异构数据的实体匹配系统。

综上所述, 已有部分多源异构数据的融合技术提出了先统一数据格式, 再进行数据融合的方法, 但数据汇聚方法较为传统, 缺乏对数据特征的描述, 融合效率较低。

### 2.2 知识图谱

近年来, 知识图谱在数据融合方面也有着突出的表现。吴运兵等<sup>[16]</sup>通过构建多源异构数据知识图谱, 融合多种数据源, 构建不同领域的领域本体库, 并结合相似性检测、冲突解决等规则, 将不同领域本体映射成全局本体库, 利用实体对齐和实体链接, 拓展了多源数据融合的知识图谱。王颖等<sup>[17]</sup>针对科技大数据应用场景, 提出以大数据平台分布式存储和高性能计算为支撑环境的科技大数据知识图谱的构建模型和技

术架构, 利用数据融合和汇聚技术, 实体抽取、实体对齐和语义丰富等方法, 构建了 3 亿实体和 11 亿关系的知识图谱, 有效支撑了科技大数据知识发现平台和“慧科研”智能随身助手的服务。韩普等<sup>[18]</sup>针对医疗领域, 从医疗实体获取、实体识别、实体链接及实体关系挖掘等技术入手, 提出了医疗多源异构数据知识图谱理论框架。雷瑛等<sup>[19]</sup>为解决军事人员面对海量战场信息的理解问题, 使军事领域人员快速准确获取军事相关知识, 构建了多源数据融合的军事知识图谱。

上述研究均利用知识图谱的特性在数据的知识层面融合发挥了重要作用, 但异构数据实体抽取采用多种抽取方式, 数据处理难度较大, 降低了数据融合的效率。综上所述, 如若将多源异构数据融合方法中先标准化格式, 再数据融合的方法, 与知识图谱的方法相结合, 不仅解决了融合过程中因格式不统一导致效率低的问题, 还解决了缺乏知识层面的数据融合问题。

### 3 基于知识图谱的多源异构信息通道耦合方法设计

#### 3.1 方法设计

信息管道指在一次数据调用过程中, 联通某个部门或系统与数据资源池的数据传输通路。信息通道指同一种数据类型信息管道的抽象集合, 即一个信息通道可以传输多个来自不同部门或系统的同一种类型的数据。通道耦合指不同通道数据之间构建关联关系, 多源异构数据的通道耦合是一种多源异构数据融合过程。

现状分析可知, 政务领域数据融合当前存在效率低、融合程度低等问题, 究其原因是政务数据规模大、类型繁多且分布不均匀, 导致数据处理难度增加。信息管道和信息通道的作用在于将多源异构的政务数据先进行划分, 使同种类型的数据汇聚在同一通道内进行处理并标准化, 然后

实现数据融合, 以提高数据处理的效率。知识图谱方法在数据融合过程中以图的形式体现实体之间的关系, 实现知识层面的融合, 具有高效率和可扩展的特性。但其在抽取多源异构数据的实体关系时, 受数据类型的影响较大, 数据预处理过程较为复杂。本文结合通道和知识图谱的特性, 提出一种基于知识图谱的多源异构信息通道耦合技术。首先, 该技术将不同类型的数据划分至不同通道, 并提取关键词, 标准化处理后得到格式统一的数据(通道 ID; 数据格式; 元数据; {关键词 1, 关键词 2, 关键词 3}; 物理存储位置); 然后, 利用知识图谱方法对标准化的数据进行实体抽取和数据融合。

多源异构通道耦合方法的基本结构如图 1 所示。具体地, 该方法根据政务信息资源目录自动或人工添加关键词, 自顶向下地建立初始通道耦合知识图谱, 实现政务知识图谱的本体构建; 根据数据类型, 将多源异构数据划分到不同通道, 通过数据预处理将所有通道内的数据标准化为通道标准数据, 再作实体抽取; 基于一种自顶向下的排序检索模型, 通过数据匹配将抽取到的通道标准数据实体融入扩展通道耦合知识图谱, 通过“词目相连”策略和“目目关联”策略, 实现知识融合和知识加工, 构建通道耦合知识图谱实体之间关联关系。通过自顶向下的排序检索模型, 实现通道耦合知识图谱的知识更新。

#### 3.2 构建初始通道耦合知识图谱

多源异构通道耦合方法架构中, 通道耦合知识图谱用于描述政务数据之间的关系, 是实现该方法的核心搜索引擎。政务领域具有领域细分、政务标准完整、数据沉淀量大的特点。通道耦合知识图谱利用政务领域的特点, 将部分现有的政府结构化知识库作为基础知识库, 其构建与业务场景无关, 因此, 通道耦合知识图谱可面对各种业务场景。在模式层, 通道耦合知识图谱与通用知识图谱的本质区别是知识边界和置信度。

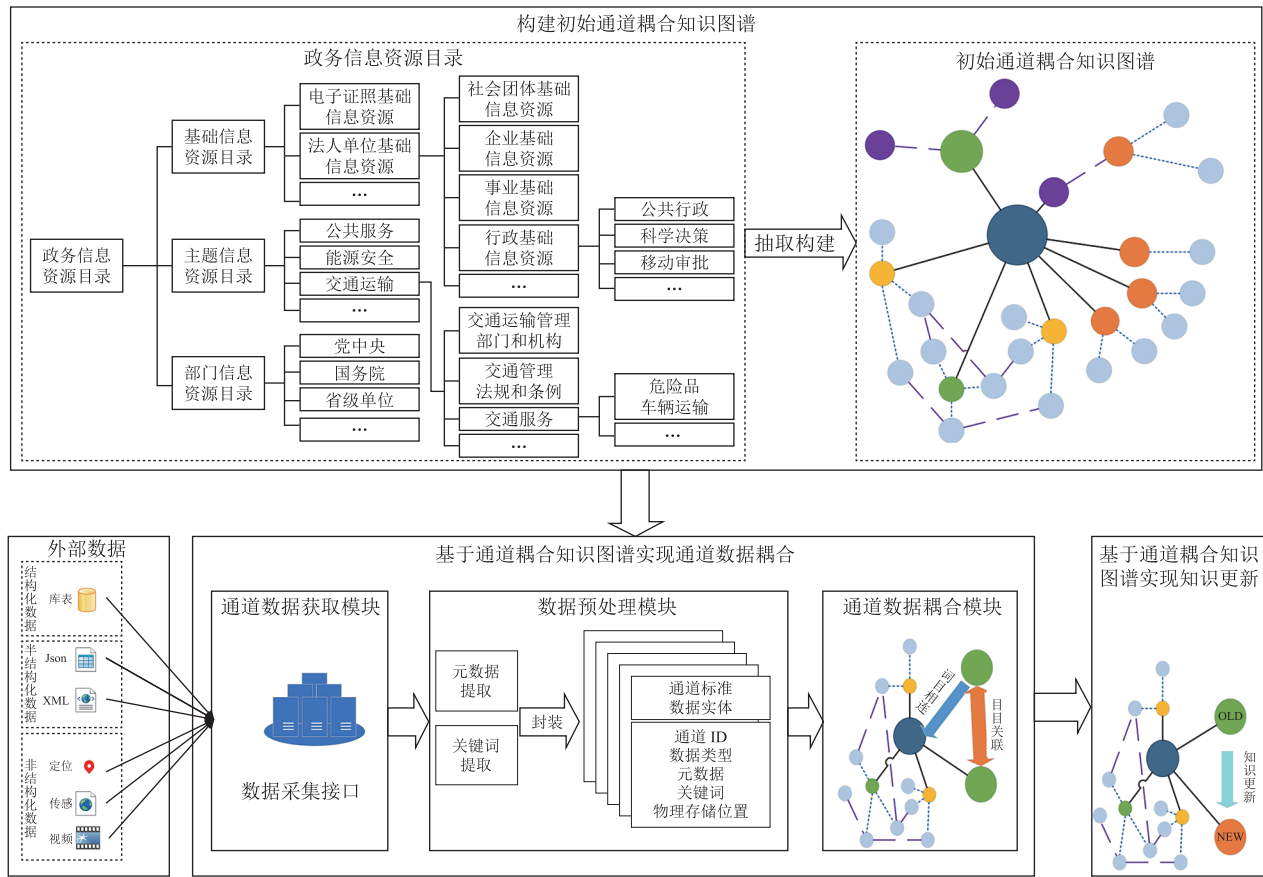


图1 多源异构通道耦合方法架构图

Fig. 1 Architecture diagram of multi-source heterogeneous channel coupling method

通用知识图谱与通道耦合知识图谱的构建方法不同，其中，通用知识图谱从公开采集数据中提取资源构建知识库，通道耦合知识图谱模式层存储的知识则依据国家发展改革委负责制定的《政务信息资源目录编制指南》<sup>[20]</sup>以及政府文件、国家标准、各政府部门规定的标准等制定而成，具有严格的知识边界。作为核心的模式层负责存储由政务信息资源目录抽取的知识，并将所抽取的公理、规则和约束条件作为本体库管理模式层，规范“类”“项”“目”和“细目”之间的包含关系。本研究使用NER模型对政务信息资源目录进行实体抽取，政务信息资源目录示例如图2所示。

为避免图谱构建过程存在歧义，在构建过程中引入冲突解决机制——通过判断不同标准的公

信力实现实体消歧。即以《政务信息资源目录编制指南》规定的政务信息资源代码的前段码为判断依据，判断抽取的文件标准所在位置，将位置不在指南中的实体替换为在指南中的实体；若两实体都不在，则由政务信息资源顺序码所构成的后端码大小顺序决定。

为保证通道耦合知识图谱的准确度，本文采用基于无监督词典规则方法的实体抽取，结合人工处理的方式添加“细目”实体，对初始通道耦合知识图谱的实体进行补充。实体抽取时，依照《政务信息资源目录编制指南》约定的“细目”为目标文本词表，对于有歧义的词汇，采用最大匹配法先进行分词，再进行NER任务。构建初始通道耦合知识图谱是本方法的先决条件，应尽量保证添加“细目”实体的种类，使“目”实体

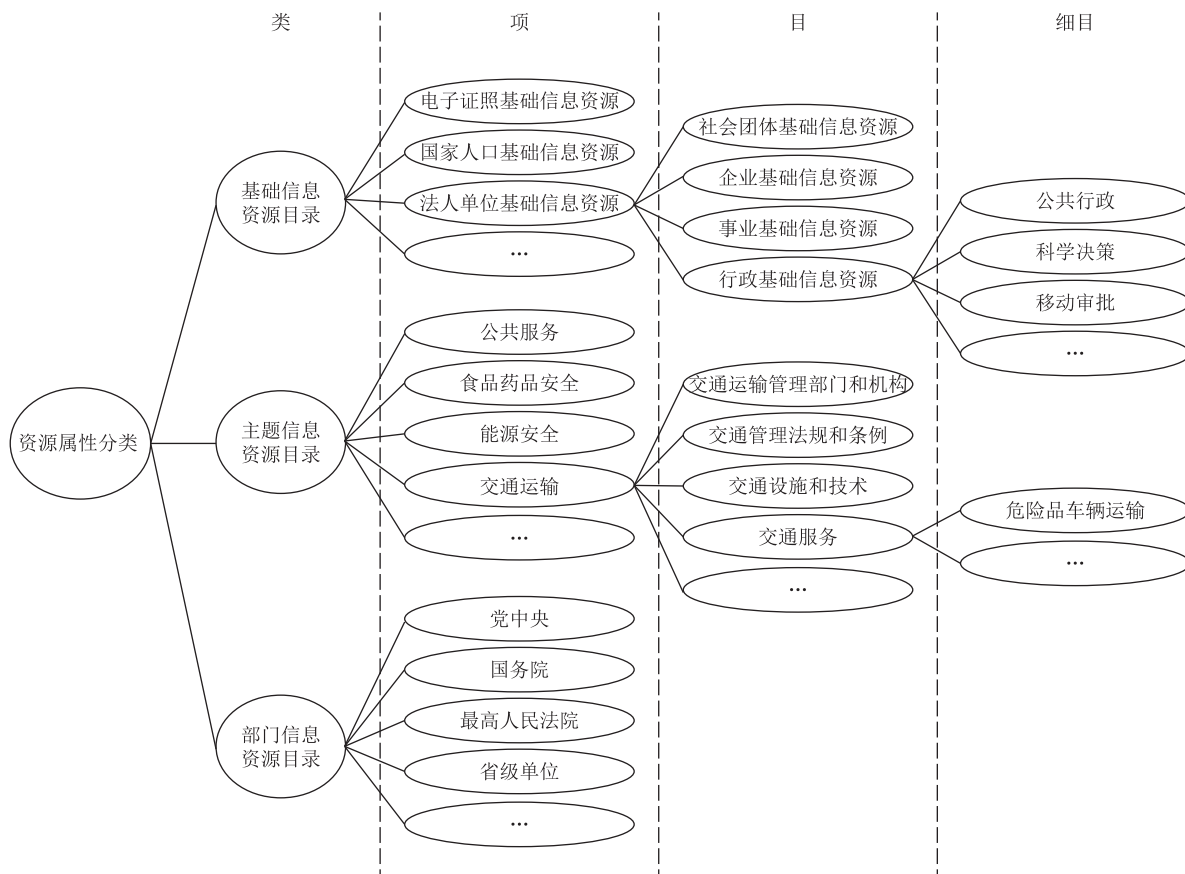


图 2 政务信息资源通道耦合知识图谱示例

Fig. 2 Example of coupled knowledge graph of government information resource channel

下有较多的“细目”实体。

### 3.3 基于通道耦合知识图谱实现通道数据耦合

通道数据耦合的过程主要分为两个方面：

(1) 通道数据标准化。政务数据来源广、格式繁多，知识图谱对其直接进行实体抽取存在困难，不利于融合过程。(2) 充分利用知识图谱的特性，拓展实体结构、层次和实体之间的因果关系，实现知识层面融合，形成动态完整的知识图谱。该过程的功能模块主要包括：通道获取数据模块、数据预处理模块、通道数据融合模块。前两个模块本质上是多源异构数据划分到不同通道，并标准化为通道标准数据的过程，通道将获取到的数据进行数据预处理，得到元数据和关键词，并将元数据传入元数据库，最后将通道 ID、数据格式、元数据、关键词和获取数据的物理存

储位置标准化为通道标准数据形式。通道数据融合模块则是将处理得到的通道标准数据融合，该模块为本方法的核心模块，通过“词目相连”和“目目关联”两个策略完成。

#### 3.3.1 通道获取数据模块

政务数据融合场景中，数据是最基本的保障<sup>[21]</sup>，这些数据包含政府开展工作产生和采集的数据、因管理服务需求而采集的外部数据、政府自有或面向政府的数据。从数据格式上看，政务数据分为结构数据(链接数据和数据库等)、半结构数据(图表和列表等)、非结构数据(文本数据、网页数据、图片数据、视频数据、传感数据和定位数据等)。通道获取数据时通过一个数据采集接口将多源异构数据汇聚，再根据数据格式通过数据划分接口将多源异构数据划分到不同的

通道进行下一步处理。

### 3.3.2 数据预处理模块

数据预处理：通道数据采集完成后，格式构成较为复杂，需要对采集数据进行预处理。数据预处理模块是为了得到每条信息通道数据的元数据，并清洗无效数据。数据预处理过程包括：

(1) 缺失数据处理：当某条通道数据被判定缺失通道 ID、数据格式、物理存储位置或数据内容为空时，应补全或删除此条数据。(2) 元数据提取：元数据即描述数据的数据。每条通道均有对应的元数据管理引擎，其中，结构化数据采用 D2R 转换方法<sup>[22]</sup>实现数据的语义模式映射，该方法将数据库表名直接映射为 RDF 中的类，字段映射到类的属性；半结构化数据通过包装器<sup>[23]</sup>对标注的数据生成 XPath 集合空间并进行归纳，形成若干个子集，以实现元数据提取；对非结构化数据进行分词和词性标注，利用联合抽取模型<sup>[24]</sup>抽取描述数据特征的关键词。该方法提出了全新的全局特征作为算法的软约束，在统一的模型中共同优化以提高元数据抽取的准确度。在政务元数据目录体系下，元数据负责对其内在内容、资源属性和数据结构等方面数据资源的内在描述，以及数据分类、数据内容属性和数据质量的外在描述。(3) 关键词抽取：通过元数据

提取，结构化数据直接将数据描述以关键词形式呈现；半结构化数据和非结构化数据则通过对其内容进行关键词提取，在知识层面描述数据内容。

数据封装：未经过标准化处理的数据进行实体抽取时效率较低，本文将数据预处理得到的元数据和关键词封装，对通道数据标准化，形成通道标准数据。每条通道数据在通道数据获取模块保留通道 ID、数据格式和物理存储位置，保留物理存储位置的目的是便于源数据索引；通道数据在数据预处理模块处理数据得到元数据和描述数据的关键词。此时，通道标准数据是拥有多个三元组的结构化数据，采用 D2R 转化的方式将其抽取为实体，每个实体均包括属性值通道 ID、数据格式、元数据、关键词和物理存储位置。该数据处理方法可保证每条政务数据均可溯源。多源异构数据转换为通道标准数据的流程图如图 3 所示。

### 3.3.3 通道数据融合模块

知识图谱的融合是将不同来源的知识进行分析、验证、总结和集成<sup>[25]</sup>，旨在将不同来源的知识组合到统一的知识图谱中。通道数据融合模块的目的是将抽取得到的通道标准数据实体融合到通道耦合知识图谱中，实现数据的抽象保存和索引。该模块通过“词目相连”和“目

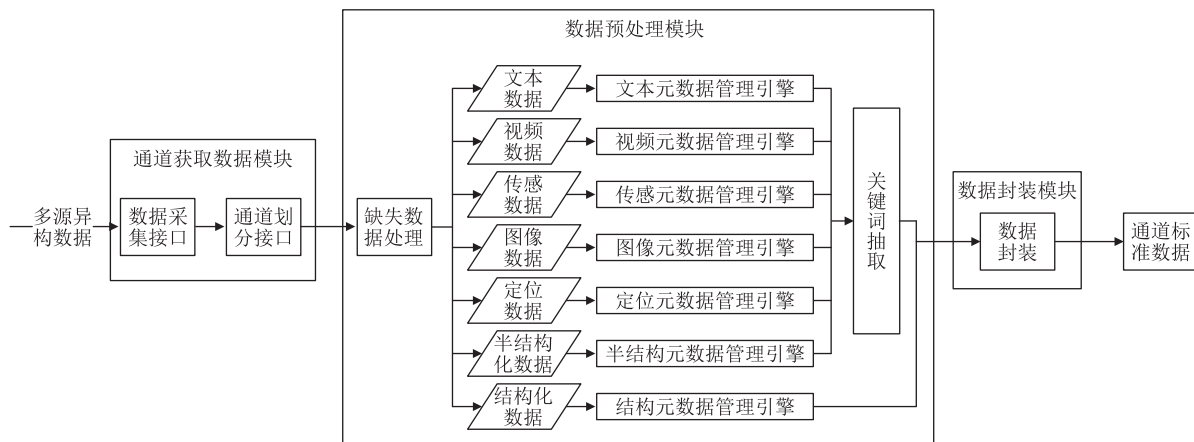


图 3 通道标准数据生成流程图

Fig. 3 Flowchart for the generation of channel standard data

关联”两个策略完成数据融合。融合模块涉及关键词匹配方法<sup>[26]</sup>, 即将两目标词映射到共同维度的语义空间过程中, 通过一个可度量的量获取两个词的向量, 并返回向量相似度, 得到匹配分数<sup>[27]</sup>。在通道耦合知识图谱构建过程中, 已经有较为完整的政务先验背景知识, 由于数据融合模块是一个持续过程, “细目”实体数量会越来越大, 若使用索引先后顺序的方法会严重影响计算速度。因此, 本文提出一种自顶向下的排序检索模型(如图 4 所示), 以实现通道标准数据实体与通道耦合知识图谱融合, 具体

流程为: (1)检索通道标准数据的关键词, 若无剩余关键词, 则退出执行; 否则依次将关键词与“类”“项”“目”“细目”匹配。(2)匹配“类”时, 设置匹配阈值, 若匹配得分高于匹配阈值, 则将所匹配到的“类”放在“类”列表中, 依次匹配, 直至匹配所有的“类”。判定“类”列表中是否有值, 若有, 则对保留的“类”列表排序, 得到关联度从高到低的“类”列表, 否则结束匹配。(3)对“项”“目”“细目”依次执行(2)的操作。(4)为优化匹配时间, 在匹配过程中, 对每一层的匹配分数进行排序,

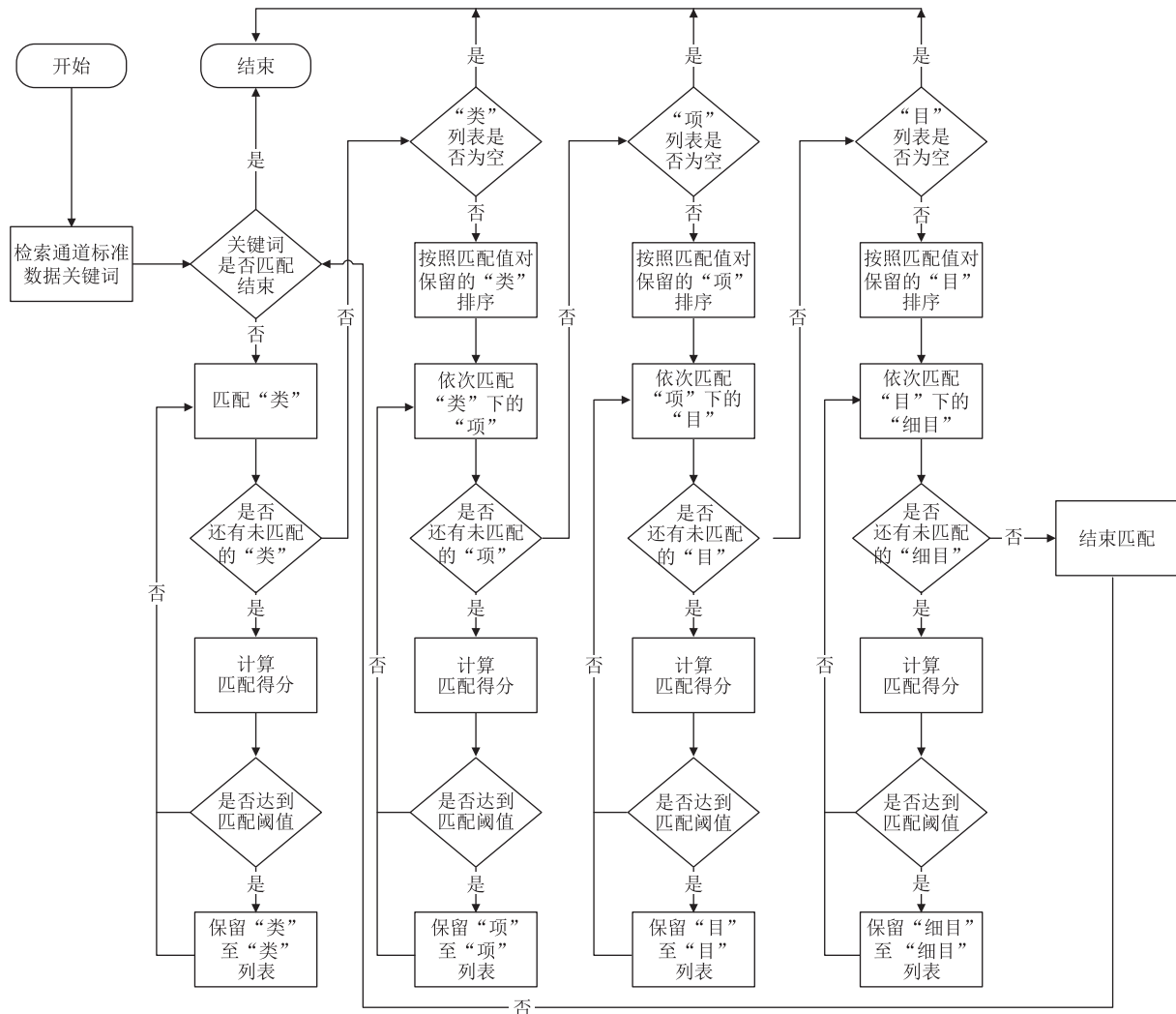


图 4 自顶向下的排序检索模型

Fig. 4 A top-down ranking retrieval model



并对得分较低的层级作剪枝处理。即在下一层级开始匹配前,对上一层级匹配分数未达到阈值的实体剪枝,直至筛选到“细目”层级,以降低时间复杂度。

词目相连:意为关键词-细目实体链接,是知识融合的一种表现方式。该过程将通道标准数据实体的关键词属性值与“细目”实体的属性值匹配,得到匹配分数最高的“细目”,并将通道标准数据实体与此“细目”实体相链接。该过程通过消除实体、关系、属性等指称项与事实对象之间的歧义实现实体链接,将通道标准数据实体链接到通道耦合知识图谱的正确实体对象<sup>[28]</sup>。基本步骤为:(1)将目标词代入自顶向下的排序检索模型,获得“细目”队列;(2)若队列中有“细目”实体,则说明当前通道耦合知识图谱中存在与通道标准数据相关联的“细目”实体,继续对队列中存在的“细目”实体排序,得到匹配值最高的“细目”实体,则此“细目”实体与目标词关联度最高,即此“细目”实体与此通道标准数据关键词属性有关联,词目相连;(3)若队列中无“细目”实体,则说明当前通道耦合知识图谱中不存在与通道标准数据相关联的“细目”实体,那么重新对“目”实体进行匹配,匹配阈值更改为创建阈值,得到创建“目”实体队列,并对此队列进行排序,得到与目标词关联度最高的“目”实体。在此“目”实体下创建属性值为(1)中目标词的“细目”实体,将该通道标准数据实体与此“细目”实体链接。

目目关联:意为细目-细目关联,是通道耦合知识图谱的推理过程,该策略自动构建“细目”实体之间的关系。“目目关联”策略本质是一种基于图结构的知识推理,本文利用路径排序算法(path ranking algorithm, PRA)<sup>[29]</sup>具有可解释性强和自动发现推理规则的优点,可快速构建实体之间的关联关系。该算法以“细目”实体之间的关系路径为特征推理,判断两“细目”实体之

间是否有关联关系。在“词目相连”策略下可优化PRA算法的特征抽取,每个确定的“细目”实体可选择有共同通道标准数据实体的路径,再计算每条路径的特征值,并为“目目关联”关系训练一个分类器,实现“细目”实体之间的关联关系。通道耦合模块流程图如图5所示。

### 3.4 基于通道耦合知识图谱实现知识更新

通道耦合知识图谱构建完成后,需要通过目标词索引实现数据检索,并及时迭代更新通道耦合知识图谱。从逻辑上看,知识库的更新包括逻辑层和数据层的更新<sup>[30]</sup>,但本文逻辑层框架准确度较高、修改幅度较小,暂不考虑。本文采用增量更新方法,基于通道耦合知识图谱,对目标词对应图谱中现有的知识进行检索,判断新旧知识的置信度,新增或更换当前知识。本文的数据检索方法使用自顶向下的排序检索模型,从通道耦合知识图谱中检索目标关键词相关的数据。

(1)将目标词“类”“项”“目”实体的属性值相匹配。匹配时,每一层设置匹配阈值,若匹配得分高于匹配阈值,则将所匹配到的实体存入对应队列中,并依照匹配得分对队列中的实体进行排序,最终得到关联度由高到低的“细目”实体队列。

(2)获得队列中每个“细目”实体及其关联的“细目”实体的匹配值后,将匹配值与新的目标词实体步长的倒数相乘,求得新的目标词与各“细目”实体之间的关联度,并设置关联阈值。队列对新的目标词与各“细目”实体之间的关联度进行判断,保留关联度高于关联阈值的“细目”实体。

(3)对当前队列“细目”实体的关联度进行排序,得到关联度由高到低的“细目”实体,再次设置输出阈值,依次判断队列中每条“细目”实体链接的每条通道标准数据实体与目标词的匹配值。若匹配值高于输出阈值,则索引得到目标词对应的数据,判断新旧数据的置信度,若新知识置信度高于旧知识,就对知识进行更新,否则

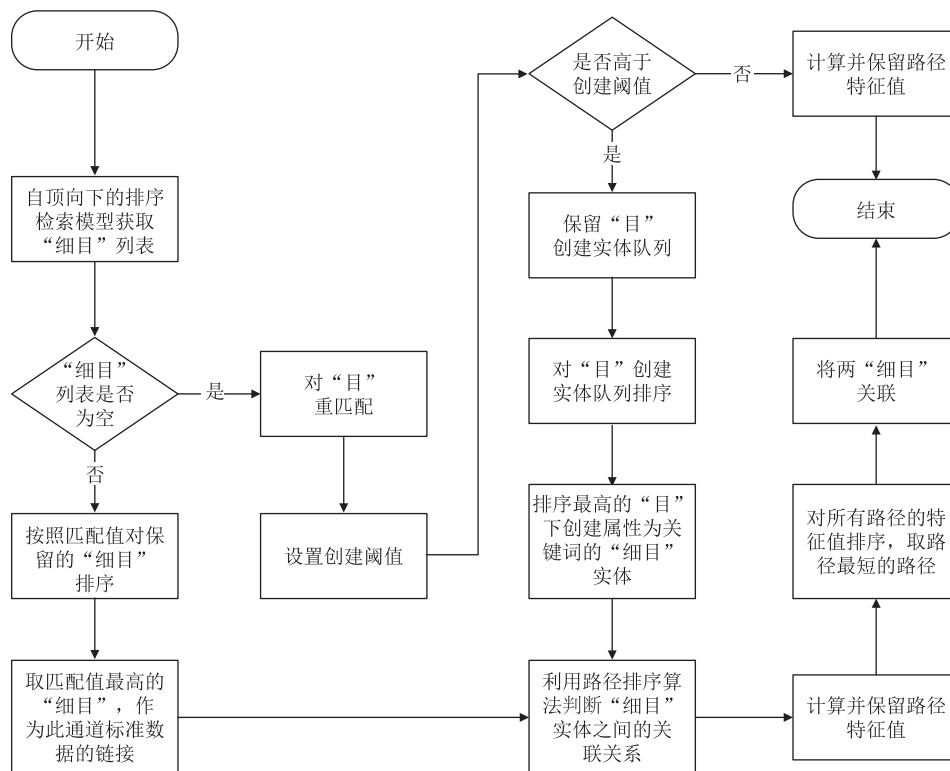


图 5 通道耦合模块流程图

Fig. 5 Flowchart of the channel coupling module

保留旧知识, 知识图谱数据完成更新。

基于通道耦合知识图谱实现知识更新流程如图 6 所示。

## 4 讨论与分析

多源异构信息通道耦合是一个较新的领域, 将通道耦合作为关键词在知网、万方等平台进行检索, 未检索到相关研究, 故本文对基于知识图谱的多源异构信息通道耦合技术进行研究。

现有的政务领域数据融合在实现信息化、数字化的职能转变过程中, 产生了大量多源异构数据。这些数据一方面存在数据格式繁多, 难以按照统一标准或方法处理的问题, 另一方面则是现有的政务数据融合方法侧重于对数据的汇聚, 缺乏对数据知识层面的融合, 耦合程度较低。

针对上述问题, 本文给出了信息通道和通道

耦合的定义, 并据此提出一种基于知识图谱的多源异构通道耦合方法。通道耦合的特点是按照数据格式, 将多源异构的碎片数据划分到不同的通道进行数据处理, 得到标准化的通道数据; 知识图谱的特点是系统化地展现与关键词相关的知识体系, 实现知识层面的融合。基于知识图谱的多源异构通道耦合方法的核心是将两者结合, 解决数据融合的难度问题。本文通过构建初始通道耦合知识图谱实现本体构建, 在通道数据融合模块, 采用通道耦合的方式实现数据预处理和标准化, 再使用“词目相连”和“目目关联”两个策略实现知识图谱的实体抽取、知识融合、知识推理过程, 完成数据融合。此外, 本文还提出了一种自顶向下的排序检索方法, 加快了数据融合过程; 在知识更新模块, 采用增量更新方式, 通过判断新旧知识置信度实现知识迭代, 该模块也可用于知识图谱, 在通道耦合知识图谱中检索目标

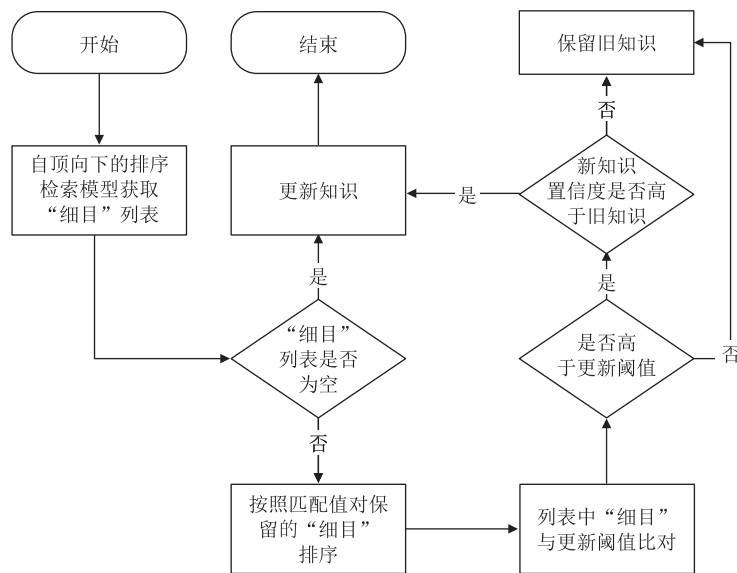


图6 基于通道耦合知识图谱实现知识更新流程图

Fig. 6 Flowchart of realizing the knowledge updating based on the channel coupling knowledge graph

词，索引并输出数据。

但是在通道耦合的过程中，多源异构政务数据仍面临着许多问题，如大量涉密政务信息资源无法参与通道耦合过程、基于通道耦合知识图谱实现数据更新较为困难等问题，这些问题还需国家层面的法律法规支持，以及政府层面的技术更新，以进一步推进知识图谱在多源异构政务数据通道耦合方向的实践与应用。

### 参考文献

- [1] 高斯芃. 整体政府视角下“互联网+政务服务”生态系统研究 [D]. 北京: 中共中央党校, 2020.  
Gao SP. The overall perspective of the government “Internet+ government services” ecosystem research [D]. Beijing: Party School of the Central Committee of CPC, 2020.
- [2] 刘蓓, 禄凯, 程浩, 等. 基于异构数据融合的政务网络安全监测平台设计与实现 [J]. 信息安全研究, 2020, 6(6): 491-498.  
Liu B, Lu K, Cheng H, et al. Design and implementation of government network security monitoring platform based on heterogeneous data fusion [J]. Journal of Information Security Research, 2020, 6(6): 491-498.
- [3] 陈波, 崔蓓, 丁鑫. 自然资源一体化政务服务系统及数据融合建设——以南京为例 [J]. 测绘通报, 2020, (12): 75-78.  
Chen B, Cui B, Ding X. Integration of natural resources government service system and data fusion construction: taking Nanjing as an example [J]. Bulletin of Surveying and Mapping, 2020, (12): 75-78.
- [4] 沈王恒. 浦东新区政务数据融合服务平台的探索 [J]. 信息技术与标准化, 2021, (6): 13-18.  
Shen WH. Exploration for Pudong New District government data integration service platform [J]. Information Technology and Standardization, 2021, (6): 13-18.
- [5] Singhal A. Introducing the knowledge graph: things, not strings [Z/OL]. (2012-05-16)[2022-10-26]. <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>.
- [6] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(3): 582-600.  
Liu Q, Li Y, Duan H, et al. Knowledge graph construction technique [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.

- [7] 中国电子技术标准化研究院. 知识图谱标准化白皮书 [EB/OL]. (2019-09-11)[2022-10-26]. <http://www.cesi.cn/201909/5589.html>.  
China Institute of Electronic Technology Standardization. Knowledge map of standardization white paper [EB/OL]. (2019-09-11)[2022-10-26]. <http://www.cesi.cn/201909/5589.html>.
- [8] 王银瑞, 彭敦陆, 陈章, 等. Trans-NER: 一种迁移学习支持下的中文命名实体识别模型 [J]. 小型微型计算机系统, 2019, 40(8): 1622-1626.  
Wang YR, Peng DL, Chen Z, et al. Trans-NER: a Chinese named entity recognition model based on transfer learning [J]. Journal of Chinese Computer Systems, 2019, 40(8): 1622-1626.
- [9] Waltz E, Linas J. Multisensor data fusion [M]. Boston: Artech House, 1990.
- [10] Pandey Y, Bansal S. Safety check: a semantic web application for emergency management [C] // Proceedings of the International Workshop on Semantic Big Data, 2017: 1-6.
- [11] 冀振燕, 吴梦丹, 杨春, 等. 可扩展的融合多源异构数据的推荐模型 [J]. 北京邮电大学学报, 2021, 44(3): 106-111.  
Ji ZY, Wu MD, Yang C, et al. Scalable recommendation models fusing multi-source heterogeneous data [J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44(3): 106-111.
- [12] 金鼎, 曹建成, 王文杰, 等. 政务空间数据关系融合技术研究和实现 [J]. 测绘地理信息, 2020, 45(5): 126-128.  
Jin D, Cao JC, Wang WJ, et al. Research and implementation of government spatial information relationship fusion technology [J]. Journal of Geomatics, 2020, 45(5): 126-128.
- [13] 丁玥, 王涓, 卢卫, 等. 面向多源关系数据的融合 [J]. 中国科学: 信息科学, 2020, 50(5): 649-661.  
Ding Y, Wang J, Lu W, et al. Multi-source relational data fusion [J]. SCIENTIA SINICA Informationis, 2020, 50(5): 649-661.
- [14] 郭帅童. 基于动态探针的多源异构数据融合与可视化技术研究及应用 [D]. 青岛: 青岛科技大学, 2022.
- Guo ST. Research and application of multi-source heterogeneous data fusion and visualization based on dynamic probe [D]. Qingdao: Qingdao University of Science and Technology, 2022.
- [15] 周琥晨. 基于文本与结构信息融合的多源异构数据实体匹配研究 [D]. 广州: 广州大学, 2022.  
Zhou HC. Based on the text and structure of information fusion of multi-source heterogeneous data entities matching research [D]. Guangzhou: Guangzhou University, 2022.
- [16] 吴运兵, 阴爱英, 林开标, 等. 基于多数据源的知识图谱构建方法研究 [J]. 福州大学学报(自然科学版), 2017, 45(3): 329-335.  
Wu YB, Yin AY, Lin KB, et al. Knowledge graph construction method based on multiple data sources [J]. Journal of Fuzhou University (Natural Science Edition), 2017, 45(3): 329-335.
- [17] 王颖, 钱力, 谢靖, 等. 科技大数据知识图谱构建模型与方法研究 [J]. 数据分析与知识发现, 2019, 3(1): 15-26.  
Wang Y, Qian L, Xie J, et al. Research on knowledge graph construction model and method of science and technology big data [J]. Data Analysis and Knowledge Discovery, 2019, 3(1): 15-26.
- [18] 韩普, 马健, 张嘉明, 等. 基于多数据源融合的医疗知识图谱框架构建研究 [J]. 现代情报, 2019, 39(6): 81-90.  
Han P, Ma J, Zhang JM, et al. The framework construction of medical knowledge graph based on multi-data source fusion [J]. Journal of Modern Information, 2019, 39(6): 81-90.
- [19] 雷瑛, 李军让, 曾熠, 等. 多源数据融合的军事知识图谱构建 [C] // 第八届中国指挥控制大会论文集, 2020: 228-233.  
Lei Y, Li JR, Zeng Y, et al. Construction of military knowledge map based on multi-source data fusion [C] // Proceedings of the Eighth China Controller in the Conference, 2020: 228-233.
- [20] 中华人民共和国国家发展和改革委员会, 中共中央网络安全和信息化委员会办公室. 政务信息资源目录编制指南(试行) [EB/OL]. (2017-07-13) [2022-10-26]. <https://www.ndrc.gov.cn/xxgk/zcfb/tz/201707/W020190905503398380313.pdf>.

- National Development and Reform Commission, Office of the Central Cyberspace Affairs Commission. E-government information resource inventory guidelines (trial) [EB/OL]. (2017-07-13) [2022-10-26]. <https://www.ndrc.gov.cn/xxgk/zcfb/tz/201707/W020190905503398380313.pdf>.
- [21] 陈静. 政务大数据融合分析的研究 [J]. 中国信息化, 2021, (10): 67-68+66.  
Chen J. Research on fusion analysis of government affairs big data [J]. China Informatization, 2021, (10): 67-68+66.
- [22] 白海燕, 梁冰. 利用 D2R 实现关系数据库与关联数据的语义模式映射 [J]. 现代图书情报技术, 2011, (7/8): 1-7.  
Bai HY, Liang B. Semantic pattern mapping between RDBMS and linked data based on open source software [J]. Modern Library and Information Technology, 2011, (7/8): 1-7.
- [23] 李效东, 顾毓清. 基于 DOM 的 Web 信息提取 [J]. 计算机学报, 2002, 25(5): 526-533.  
Li XD, Gu YQ. DOM-based information extraction for the Web sources [J]. Journal of Computer Science, 2002, 25(5): 526-533.
- [24] 付瑞, 李剑宇, 王笏辉, 等. 面向领域知识图谱的实体关系联合抽取 [J]. 华东师范大学学报(自然科学版), 2021, (5): 24-36.  
Fu R, Li JY, Wang JH, et al. Joint extraction of entity relations for domain knowledge graph [J]. Journal of East China Normal University (Natural Science Edition), 2021, (5): 24-36.
- [25] 易晓宇, 易绵竹. 知识图谱融合技术研究 [J]. 电脑编程技巧与维护, 2022, (8): 144-146+164.  
Yi XY, Yi MZ. Knowledge map fusion technology research [J]. Computer Programming Skills & Maintenance, 2022, (8): 144-146+164.
- [26] Huang PS, He XD, Gao JF, et al. Learning deep structured semantic models for web search using clickthrough data [C] // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013: 2333-2338.
- [27] Zhu HM, Tan R, Han LT, et al. DSSM: a deep neural network with spectrum separable module for multi-spectral remote sensing image segmentation [J]. Remote Sensing, 2022, 14(4): 818.
- [28] Li Y, Wang C, Han FQ, et al. Mining evidences for named entity disambiguation [C] // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013: 1070-1078.
- [29] 张美玉, 林崇, 简琤峰. 基于路径排序算法的 STEP 知识推理技术研究 [J]. 浙江工业大学学报, 2020, 48(2): 126-132+187.  
Zhang MY, Lin C, Jian ZF. Research on STEP knowledge reasoning technology based on path ranking algorithm [J]. Journal of Zhejiang University of Technology, 2020, 48(2): 126-132+187.
- [30] 张吉祥, 张祥森, 武长旭, 等. 知识图谱构建技术综述 [J]. 计算机工程, 2022, 48(3): 23-37.  
Zhang JX, Zhang XS, Wu CX, et al. Survey of knowledge graph construction techniques [J]. Computer Engineering, 2022, 48(3): 23-37.