

多源数据知识获取通用存储知识库接口

Common storage interface of knowledge acquisition for multi-source data

（征求意见稿）

2022-00-00 发布

2022-00-00 实施

广州市南沙区粤港澳标准化与质量发展促进会 发布

目 次

前言	II
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 缩略语	2
5 基本要求	2
6 多源异构大数据获取接口	2
7 知识获取引擎接口	3
8 知识库存储接口	3
9 知识推理引擎接口	3
10 接口框架	3
10.1 接口分类	3
10.2 接口结构	4
11 接口功能	4
11.1 功能定义	4
11.2 适用范围	4
11.3 功能要求	4

前 言

本文件按照GB/T 1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广东工业大学提出。

本文件由广州市南沙区粤港澳标准化与质量发展促进会归口。

本文件起草单位：广东工业大学、广东能哥知识科技有限公司、广州赛意信息科技股份有限公司、广州瑞松智能科技股份有限公司、明珞科技、广州广船国际股份有限公司、广东健新科技有限公司、佳帆科技、广东恒电信息科技股份有限公司、佛山智能装备技术研究院、佛山华数机器人、浪潮工业互联网、广州科盛隆、佛山台一精工。

本文件主要起草人：

多源数据知识获取通用存储知识库接口

1 范围

本文件规定了多源数据知识获取通用存储知识库接口的基本要求、多源异构大数据获取接口、知识获取引擎接口、知识存储接口、知识推理引擎接口、接口框架和接口功能。

本文件适用于多源数据知识获取通用存储知识库的接口管理。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

- GB/T 35274 信息安全技术 大数据服务安全能力要求
- GB/T 37973 信息安全技术 大数据安全管理指南
- GB/T 38633 信息技术 大数据 系统运维和管理功能要求
- GB/T 38672 信息技术 大数据 接口基本要求
- GB/T 38673 信息技术 大数据 大数据系统基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

接口 interface

预先定义的为应用程序提供实现一定数据访问、数据处理和数据服务等功能的程序、函数或过程。

3.2

数据爬虫 data crawler

一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

3.3

纳什均衡 Nash equilibrium

指满足下面性质的策略组合：任何一位玩家在此策略组下单方面改变自己的策略（其他玩家策略不变）都不会提高自身的收益。

3.4

皮尔逊相关系数 Pearson Correlation Coefficient

用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。

4 缩略语

下列缩略语适用于本文件。

CPU: 中央处理器 (Central Processing Unit)

GPU: 图形处理器 (Graphics Processing Unit)

JSON: 一种轻量级的数据交换格式 (Java Script Object Notation)

TF-IDF: 一种用于信息检索与数据挖掘的常用加权技术 (Term Frequency - Inverse Document Frequency)

Doc2vec: 一种无监督算法, 能从变长的文本中学习得到固定长度的特征表示

LSTM: 长短期记忆网络 (Long Short-Term Memory)

BiLSTM: 双向长短期记忆网络 (Bi-directional Long Short-Term Memory)

CRF: 条件随机场 (Conditional Random Fields)

LSTNet: 长期和短期时间序列网络 (Long and Short-term Time-series Network)

GAN: 生成式对抗网络 (Generative Adversarial Network)

HTML: 超文本标记语言 (Hyper Text Markup Language)

URL: 统一资源定位系统 (uniform resource locator)

Neo4j: 一个高性能的, NOSQL 图形数据库, 它将结构化数据存储在网上而不是表中。

BERT: 一种语言表示模型, 来自变换器的双向编码器表征量 (Bidirectional Encoder Representations from Transformers)

5 基本要求

5.1 开发性

应符合产业习惯, 兼容主流开源接口, 减小接口定制化带来的重新设计、适配成本。

5.2 易用性

应尽可能设计成抽象程度高、屏蔽底层实现、语法易理解的接口。

5.3 扩展性

同一接口可通过增加函数、操作符、语句等形式支持新的功能。

5.4 其他要求

接口其他基本要求应符合 GB/T 38672。

6 多源异构大数据获取接口

6.1 对文本、图片、视频音频等多源数据的自动获取, 包括政府管理部门报告、企业年度报告、行业领域分析报告、互联网公开数据、报刊、杂志、图书、微信、微博等数据。

6.2 对于网络数据, 定义页面抽取规则, 对指定的页面提取目标数据。具体地, 设置目标爬虫地址, 对网站的 html 语言进行解析, 按照模板通过 Beautiful soup 库完成数据爬虫。对标题、类型、日期、摘要、正文、网页以及来源等字段进行解析并存储, 构建数据库。并根据 URL 作为标识在数据库中进行查找, 以此来判断该网络数据是否会重复。

6.3 对数据进行本体分析，制定统一的资源描述框架。从数据库中抽取出关系模式，分析关系数据库中表的信息和字段信息，建立相应的概念模型。采用相似性检测规则对概念和属性进行检测，把相同或相似概念和属性进行统一。

6.4 由于多媒体数据的复杂性，需对多媒体数据进行预处理，即特征提取和压缩。然后将标量数据和预处理后的多媒体数据通过多模态深度信念网络进行特征融合，以实现初始数据的清洗，减少数据的噪音。

7 知识获取引擎接口

7.1 融合文本、图片、视频等多源数据的知识抽取。通过对抗生成网络将文本特征与对应的图像特征映射到公共语义空间，实现跨媒体数据的公共表示学习。采用 Doc2vec 模型深度挖掘语义信息，实现多源异构数据源的实体对齐，并针对已对齐的实体对进行属性融合。

7.2 采用 BiLSTM-CRF 模型进行实体识别，模型包括 BERT 编码层、BiLSTM 交互层和 CRF 推理层。将相关文本输入到 BERT 编码层进行训练，得到每个字的向量，生成向量矩阵，将得到的向量矩阵输入到 BiLSTM 交互层进行解码，最终通过 CRF 推理层进行解码，获取概率最大的标签序列，得到每个字符的类别。

7.3 采用 LSTM-Attention 对半结构化数据进行关系抽取，基于两个实体和预先定义好的关系类型，计算两个实体之间某个关系分数，归一化得到概率，把概率最高的关系定义为这两个实体的最终关系。

7.4 采用 Text-CNN 进行文本分类，通过卷积层、最大池化层处理，最后将输出外接 softmax 层来对文本进行分类。

8 知识库存储接口

8.1 对于融合后的知识采用 Neo4j 图数据库进行存储，可通过使用 Python 支持的 Py2Neo 第三方库提供的操作函数，将各类数据以节点和边等形式导入 Neo4j 中。以“实体-属性-属性值”、“实体-关系-实体”三元组进行表示存储，构建大规模知识库。

8.2 对数据和模式进行动态更新，实现多方位关联知识的迭代。采用 GAN 实现语义知识库的自优化，其优化过程是一个极小极大博弈问题，优化目标是达到纳什均衡，使生成器估测到数据样本的分布。

9 知识推理引擎接口

9.1 基于已构建的知识图谱，提供智能推送服务。对知识图谱进行实体节点介数、实体节点度数、实体节点紧密度指标的评估，最后利用熵权法对所使用的指标进行计算，实现新闻的重要度计算，以此进行推送。

9.2 挖掘时间序列数据，实现演化趋势分析。针对时间序列数据预测，采用深度学习技术，输入一段过去时间数据，输出是对未来一段时间预测数据。为了捕获长期依赖和短期依赖，采用 LSTNet 深度学习框架进行趋势预测，并通过皮尔逊相关系数分析演化的因素。

10 接口框架

10.1 接口分类

接口分类包括以下内容：

- a) 数据处理：数据爬取、数据存储；
- b) 知识抽取：实体识别、关系抽取、文本分类；
- c) 知识应用：知识推送服务、演化趋势分析。

10.2 接口结构

10.2.1 接口结构分为数据处理、知识抽取、知识应用三个部分，见表 1：

表 1 接口结构

接口	模块	输入	方法	输出
数据处理	数据爬取	URL	Beautiful soup库	<h2>、<p>
	数据存储	标签	正则表达式、Beautiful soup库	JSON(数据库)
知识抽取	实体识别	JSON	BiLSTM-CRF模型	实体类别
	关系抽取	JSON	LSTM-Attention模型	关系标签（三元组）
	文本分类	JSON	Text-CNN模型	类别标签
	演化趋势分析	JSON:时间序列数据	LSTNet、皮尔逊相关系数分析	折线图可视化
	知识推送服务	JSON:新闻信息	重要度计算	新闻列表

11 接口功能

11.1 功能定义

- 11.1.1 数据爬取：从指定的网站上收集数据信息。
- 11.1.2 数据存储：对爬取的数据进行存储。
- 11.1.3 实体识别：将预先定义好的实体类型（人名，机构，地名等）识别出来。
- 11.1.4 关系抽取：找出主体与客体之间存在的关系，并将其表示为实体关系三元组，即（主体，关系，客体）。
- 11.1.5 文本分类：对文本按照一定的分类体系进行自动分类标记。
- 11.1.6 演化趋势分析：基于构建的知识库，针对某个领域业务层进行演化趋势分析。
- 11.1.7 知识推送服务：基于构建的知识库，对海量爬取数据进行信息过滤，计算重要度，实现知识推送服务。

11.2 适用范围

政府管理部门报告数据、企业报告数据、各新闻网、报刊、杂志、图书、微信、微博等网络数据和数据平台。

11.3 功能要求

- 11.3.1 宜支持数据批量更新、删除等数据管理功能；
- 11.3.2 宜支持流式的实时数据入库，支持实时查询；
- 11.3.3 宜提供数据自动备份和手动备份的功能；
- 11.3.4 应支持对全局资源的集中管理；
- 11.3.5 应支持对任务设置优先级，并按任务优先级对资源进行调度；
- 11.3.6 应支持 CPU、内存、GPU 等异构资源调度和配置；

11.3.7 提供数据分析方法或者中间件,将数据准备模块输出的数据以及数据建模过程中产生的中间数据转变成知识或者决策;

11.3.8 系统运维和管理功能应符合 GB/T 38633;

11.3.9 数据安全应符合 GB/T 37973 和 GB/T 35274;

11.3.10 其他要求应符合 GB/T 38673。
