基于知识图谱的智能决策支持技术及应用研究

魏 瑾,李伟华,潘 炜
(西北工业大学,陕西 西安 710129)

摘 要:知识图谱是把复杂的领域知识通过数据挖掘、信息处理、知识计量和图形绘制而显示出来,解释知识领域的动态发展规律。知识图谱把所有不同种类的信息(heterogeneous information)连接在一起得到一个关系网络并从“关系”的角度去分析问题。知识图谱目前被广泛应用于智能搜索、智能问答等领域。提出了一种基于知识图谱的智能决策支持框架,用于解决传统决策支持系统存在的问题。通过大数据、知识图谱等海量知识分析和模型构建技术,结合决策支持系统,增强对问题的分解与处理、形成具有关系型网络的知识系统。最后结合电信领域中的经典决策案例,搭建基于知识图谱的欺诈电话智能决策支撑平台。和传统的决策支持系统比较,该研究方法的优点在于结合大数据处理方法提升了知识建模的算力和决策支持的效率,使实时处理大规模信息数据成为现实;基于知识图谱的关系型网络,提升了决策模型的准确性和关联相关性。

关键词:决策支持系统;知识图谱;大数据;实时计算
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)01-0001-06
doi:10.3969/j.issn.1673-629X.2020.01.001

Research on Intelligent Decision Support Technology and Application Based on Knowledge Graph

WEI Jin,LI Wei-hua,PAN Wei
(Northwestern Polytechnical University,Xi'an 710129,China)

Abstract:Knowledge graph is the display of complex domain knowledge through data mining, information processing, knowledge measurement and graphical rendering to explain the dynamic development law of knowledge field. Knowledge graph links heterogeneous Information together to form a relationship network and analyze the problem from the perspective of "relationship". Knowledge graph is widely used in intelligent search, intelligent question answering and other fields. We present an intelligent decision support framework based on knowledge graph to solve the problems existing in traditional decision support systems. Through massive knowledge analysis and model building technologies such as big data and knowledge graph, and combined with decision support system, the knowledge system with relational network can be formed by enhancing the decomposition and processing of problems. At the end, a fraud telephone intelligent decision support platform based on knowledge graph is built by combining the classical decision cases in the telecommunication field. Compared with the traditional decision support system, the advantages of the research method proposed are that it combines the large data processing method to improve the computational power of knowledge modeling and the efficiency of decision support, and makes real-time processing of large-scale information data become a reality. The relational network based on knowledge graph improves the accuracy and relevance of decision model.

Key words:DSS;knowledge graph;big data;real time computation

1 背 景

1.1 传统决策支持系统的局限性

决策支持系统(decision support system,DSS)是一个基于计算机用于支持业务或组织决策活动的信息系统。DSS 是运用计算机、数据库、多媒体、网络、类人智能提供辅助决策手段和工具,将决策这样的人类思维活动转变为决策计算和思维的结合。

传统的决策支持包括三个方面:一是需要得力的

收稿日期:2019-01-28 **修回日期**:2019-05-30 **网络出版时间**:2019-09-25
基金项目:国家国防部“十三五”预研项目(31511090403)
作者简介:魏 瑾(1985-),女,在读博士,研究方向为智能决策支持系统、机器学习;李伟华,教授,博导,通信作者,研究方向为人工智能、决策支持系统和多媒体智能决策。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1521.028.html

辅助决策机构,如顾问机构、参事机构等;其次,决策理论方法,需要正确的理论指导;三是需要使用现代化的手段和工具。随着计算机、通信、专家系统等技术的迅速发展,给决策活动带来了新的格局,决策支持使决策活动朝着程序化、科学化方向发展。

决策支持系统在国内外各领域得到了深入的研究。国内,有关大学、研究所分别研制了“分布式多媒体智能决策支持系统平台”、“基于客户/服务器的决策支持系统”、“智能决策系统开发平台 IDSDP”、“基于主体的智能协同决策支持系统”等;快速开发平台为开发实际问题的决策支持系统提供快速开发环境,产生了一批应用成果。

但是,传统的决策支持系统存在一定的局限性,主要包括:(1)获取用于决策的知识的难度较大。传统的基于规则的决策支持,例如专家系统,依赖专家的领域知识来形成规则进而进行决策,形成规则的不确定性较多,如专家对行业的理解偏差,归纳总结的能力。(2)决策支持系统的灵活性和适应性较低。如果影响决策的某种因素发生变化,已有的决策系统就无法进行准确的判断。(3)知识协同和相关性较差。正确的决策需要多种知识对于决策的关联作用,解决决策群体之间的协同问题。

1.2　知识图谱

决策支持是跨管理、人工智能、网络等多个学科的综合应用智能系统。新型的决策支持系统需要考虑引入专家系统、自然语言理解尤其是知识库系统的研究;提升决策的快速反应和自动化以提高时效性;实现知识的共享、关联及继承性[1]。

知识图谱本质上是语义网络,是一种基于图的数据结构,由节点(point)和边(edge)组成。在知识图谱里,每个节点表示现实世界中存在的“实体”,每条边为实体与实体之间的“关系”。知识图谱是关系的最有效的表示方式。通俗地讲,知识图谱就是把所有不同种类的信息(heterogeneous information)连接在一起而得到的一个关系网络。知识图谱提供了从“关系”的角度去分析问题的能力。

知识图谱示意如图 1 所示。

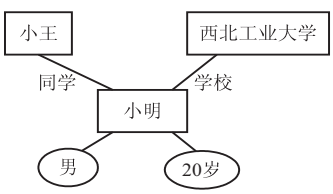

图 1　知识图谱示意

在构建知识图谱的过程中,有几个关键组成部分:

·知识的抽取:信息源主要来自两种方式:一种是业务领域本身的数据,这部分数据通常包含在数据库表并以结构化的方式存储以及业务系统中以半结构化或非结构化存在;另一种是网络上公开、抓取的数据,这些数据通常是非结构化的数据。

·知识图谱的存储:一种是基于 RDF 的存储;另一种是基于图数据库的存储。RDF 一般以三元组的方式来存储数据而且不包含属性信息。图数据库一般以属性图为基本的表示形式,所以实体和关系可以包含属性。RDF 存储的优点是数据的发布以及共享,图数据库的优点是高效的图查询和搜索。

当前,知识图谱的应用目前集中在搜索、推荐、问答、解释和辅助决策等方面[2]。知识图谱在认知领域已经得到了很多应用,例如在语义搜索方面,Google 和百度使用知识图谱应用在新一代的搜索引擎中。在人机交互和问答系统领域,有聊天机器人微软小冰和 Apple 公司的 Siri 等。在决策支持方面,IBM 在 Watson Health 中使用了知识图谱进行临床领域的决策支持[3]。此外,越来越多的企业将知识图谱作为云平台或数据中台的基础数据服务提供给上层应用消费。

2　概要说明

综合知识图谱的应用场景和传统决策支持系统的局限性,提出了一种基于知识图谱的智能决策支持技术,使用知识图谱代替了传统决策支持系统的决策模型。通过构建面向领域的知识图谱、基于知识图谱的决策模型建模,面向特定问题进行决策分析。

文中提出的基于知识图谱的智能决策支持总体架构分为四个层次结构,如图 2 所示。

·基础层:为上层的分析提供基础平台能力,包括存储能力、计算资源分配能力、实时处理能力等。其中结构化数据的存储使用关系型数据库 PostgreSql,非结构化数据(如图片、语音、视频、网页索引等)的存储使用开源的 HDFS 分布式文件系统。图文件存储使用 JanusGraph 图数据库提出图存储和索引的能力。

·数据层:数据源主要来源于四部分:(1)通用的行业知识来源于互联网的公开资料;(2)文献资料提供领域内的公开专业知识包含一些领域内的语义和语料信息;(3)行业积累的信息主要是指行业内的历史数据信息,例如存储在专业系统数据库的结构化信息,持久化存储在存储系统或光盘的行业视频、音频信息等;(4)领域专家沉淀的专家信息,通过对专家信息的收集和整理获得。

数据采集是指对各种数据源,不同数据类型采取的特定采集技术。对于互联网的数据采用网络爬虫的技术进行收集和整理。对文献资料多采用人工录入的方式。对于结构化数据,使用数据仓库分层处理模式。

对于领域专家知识使用访谈、归纳等信息抽取方式。

数据处理是通过对采集后的数据进行加工处理获取构建知识图谱所需的各种信息,包括:数据词典、语义库、语料库、通用知识库、领域知识库、规则库、关系表。

数据存储是将数据处理后的数据按各自的形式存储在存储系统中。

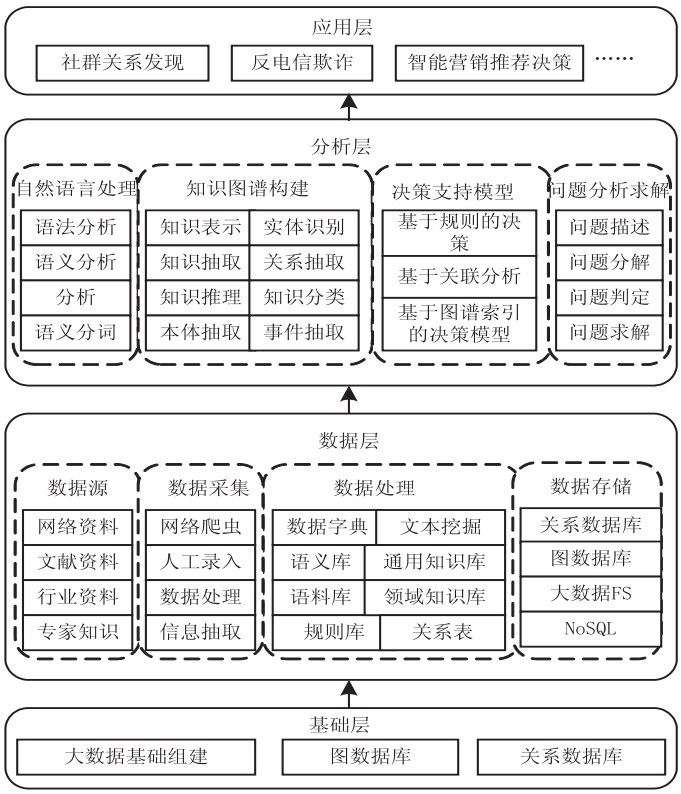

图 2　总体结构

·分析层:与传统的决策支持系统不同,文中除了基于规则的决策外,还提供了基于知识关联分析、基于图谱索引的决策。在分析层,通过自然语言处理等技术构建知识图谱,并在知识图谱之上建立决策模型。在问题分析和求解过程中,根据问题的类型使用不同的决策模型以达到最优的问题求解。

知识图谱的构建是分析层的重要部分,图谱构建的质量直接决定了上层应用的效果。知识图谱可以将多源异构的数据汇聚到一起[4]。在文中背景下的电信领域,领域知识图谱的构建比起通用知识图谱,更加依赖于结构化数据和非结构化数据的结构去构建行业领域的图谱网络。在知识图谱构建中,主要包括领域内知识表示建模、实体识别与实体链接、关系事件抽取、隐性关系发现等技术[5]。

·应用层:分析层向上通过 APIs 接口提供应用级别的决策能力开放。

3　基于知识图谱的智能决策支持技术

如图 2 所示,基于知识图谱的智能决策系统的建立主要有四个过程:数据的采集、预处理和数据加工;面向特定领域的知识图谱的构建[6];基于特定领域知识图谱的决策模型的构建;问题分析及智能决策。

3.1　数据采集加工

数据采集技术通过对数据进行提取、转换、加载,最终挖掘数据的潜在价值,给知识图谱的构建和决策建模提供数据支持。从数据源抽取出所需的数据,经过数据清洗,按照预先定义好的数据模型,将数据加载到数据仓库中,最后对数据仓库中的数据进行数据分析和处理。由于采集的数据类型各异,对于不同种类的数据进行数据分析,必须通过提取技术。将复杂格式的数据,进行数据提取,从数据原始格式中提取(extract)出需要的数据。对于数据提取后的数据,由于数据源头的采集可能存在不准确性,所以必须进行数据清洗,对于那些不正确的数据进行过滤、剔除。针对不同的应用场景,对数据进行分析的工具或者系统不同,还需要对数据进行数据转换(transform)操作,将数据转换成不同的数据格式,最终按照预先定义好的数据仓库模型,将数据加载(load)到数据平台中。数据采集加工流程如图 3 所示。

·行业业务系统数据采集:由于业务系统的数据常存放在 Oracle 或者 MySql 等传统数据库中,因此需要使用工具将数据导入到大数据平台中。文中使用

Apache Sqoop 进行数据转换,将结构化数据存储(如关系数据库、企业数据仓库和 NoSQL 系统)导入到分布式文件系统 HDFS 上,并将表填入 Hive 中。Hive 是建立在 Hadoop 之上的开源数据仓库。Hive 支持使用类似 SQL 的声明性语言(HiveQL)表示的查询。

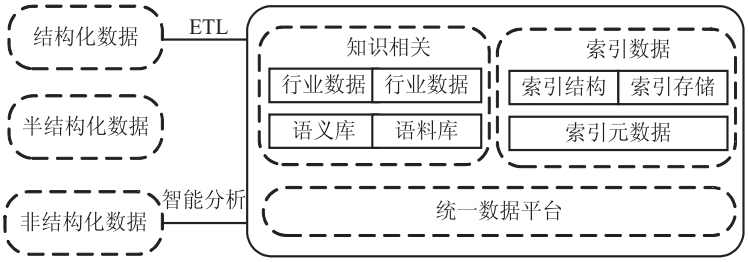

图 3　数据采集加工流程

·通用知识采集:通过网络爬虫和一些网站平台提供的公共 API 等方式从互联网上获取通用的行业数据。这样就将非结构化数据和半结构化数据的网页数据从网页中提取出来。并将其提取、清洗、转换成结构化的数据,将其存储为统一的本地文件数据。

在数据采集和加工后,将加工后的数据分为两类:一类为知识相关的数据,用于构建知识图谱,包括行业数据和通用数据、知识描述的语义库和语料库[7];另一类为间接数据,用于从知识进行索引,包括索引结构、索引元数据、索引存储和索引的数据存储。这两类数据都统一存储和管理在数据平台中。

3.2　面向领域的知识图谱构建

知识图谱的构建由知识表示、知识抽取、知识推理三个主要部分构成[7]。

·知识表示:文中用两种三元组来表示实体之间的关系以及实体的属性。用 E_i 表示第 i 个实体,E_j 表示第 j 个实体,E_i 和 E_j 之间的关系用 R_{ij} 表示,那么三元组(E_i,R_{ij},E_j)表示为实体 E_i 和 E_j 之间的联系。

用 A_x 表示实体 E_i 的第 x 个属性名称,V_x 表示该属性对应的值。那么三元组(E_i,A_x,V_x)表示为实体 E_i 的第 x 个属性描述,同理三元组(E_i,A_y,V_y)表示实体 E_i 的第 y 个属性描述。

基于这两类三元组,知识图谱就可以描述为多个实体关系网,如图 4 所示,多实体之间通过关联关系组成了一个知识图谱网络。网络的节点表示实体,实体分别与其他实体之间产生关系,并且通过属性三元组表示实体的各种属性。

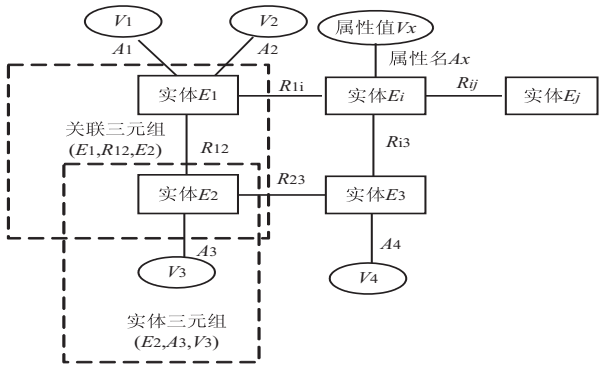

图 4　基于三元组表示知识图谱结构

·知识抽取:知识抽取[8]的目标是从预处理后的数据中识别知识图谱的实体以及实体之间的关系,从而构建本体[4],建立知识图谱。文中提出的知识抽取方法的过程为:

(1)根据数据词典标注领域数据。为了解决非结构化文本信息难于抽取实体的问题,首先通过数据字典对领域数据进行标注,获得标注数据集,以二元组($data_i$,$label_j$)表示 $data_i$ 对应的标签为 $label_j$。其中 $data_i$ 为数据词典样本中提取的数据,$label_j$ 表示样本的标签,文中表示为“实体”、“属性名”、“属性值”、“关系”等。例如:(电话号码:实体)。

(2)基于规则判断和机器学习算法识别数据中的实体及实体的属性。基于规则判断用于简单实体的抽取,如存放于数据仓库中的结构化数据。

对于大量的非结构化文本数据,采用 LSTM 深度学习的方法识别实体及属性(见图 5)。实体识别分为两部分:模型训练过程:以标注的领域数据作为样本数据,使用 LSTM 算法进行训练,获得文本分类模型;实体识别过程:使用训练好的模型对非结构化的文本数据进行识别和分类,分类的结果即为各实体三元组(E,A,V),其中 E 表示实体名,A 表示实体的属性名,V 表示该属性值。

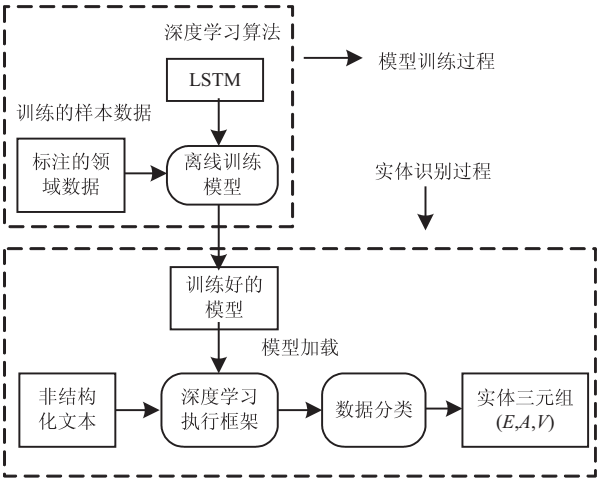

图 5　基于 LSTM 深度学习的实体识别

(3)关系抽取的过程类似于实体识别。在已获得实体列表的基础上,基于规则和启发式算法对关系和属性进行抽取[9]。最终将实体和相关实体通过关系连接组合,获得实体关系三元组(E_i , R_{ij} , E_j)。

·知识推理是指根据知识图谱中的已有知识,推断出新的、未知知识,即潜在知识的挖掘和发现[10],如图 6 所示。

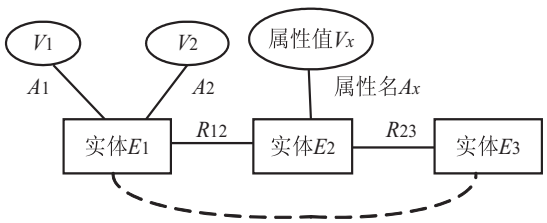

图 6　知识推理示意

实体 E_1 和 E_2 之间有关联, E_2 和 E_3 有另外一种关联,知识推理的目的就是判断 E_1 和 E_3 之间是否存在关联关系。在知识推理方面,使用基于规则技术[11](规则或模版、一阶谓词逻辑推理)。

3.3　智能决策模型

提出的智能决策模型分别为基于规则的决策支持[12]和基于关联分析的决策模型。

·基于规则的决策支持。

基于规则的决策支持有两个过程:(1)规则策略的建立。由领域专家根据领域知识库建立规则策略,并通过样本数据验证规则的准确性和有效性[13];(2)决策过程。决策者根据建立的规则对问题进行分析和判断,最终得到决策推理的结果。规则的描述根据规则引擎的不同,一般使用表达式 if(…),then。其中 if 可使用算术操作符、逻辑操作符及其他,操作符可以组合使用。

·基于关联分析的决策模型。

基于规则的决策能够针对具体的单一事件进行推理决策,但是在现实行为中,不同个体之间的协同以及群体性特征往往会影响决策的准确性。因此,引入关联分析显得尤为重要。聚类实体群决策见图 7。

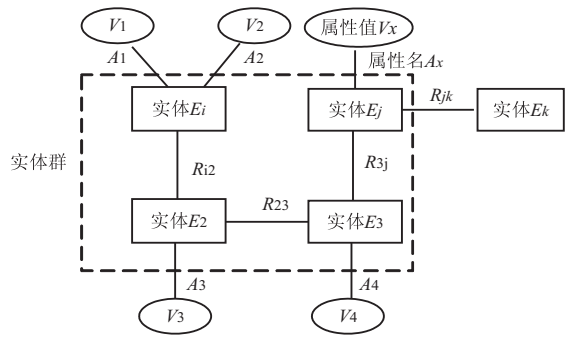

图 7　聚类实体群决策

基于关联关系的决策算法如下:

步骤一:基于 K-Means 的实体群聚类。

根据 K-Means 方法,在知识图谱中对有关联的实体进行聚类。算法如下:

Repeat

{

for i =1 to m

① $c^{(i)}$: = arg $\min_k$ | $E^{(i)}$ − μ_k | ^{2}

for k =1 to K

② $$\mu_\kappa := \frac{\sum_{i=1}^{m} I\{c^{(i)} = k\} * E^{(i)}}{\sum_{i=1}^{m} I\{c^{(i)} = k\}}$$

}

①是计算每个实体 $E^{(i)}$ 到每个中心节点 μ_k 最近的距离,将此实体划分到该中心节点对应的实体群中。

②是对每个实体群 $C^{(k)}$ 重新计算中心节点。

不断的重复①和②,最终得到实体群的聚类。

步骤二:基于实体群获得关联实体关系。

当存在(E_j , E_{jk} , E_k),且 E_i 与 E_k 同属一个实体群,获得 E_i 对 E_j 的决策。

$$R_{ik} = \frac{\sqrt{(E_i - E_j)^2}}{\arg\min \sum_{m=i,n=i+1}^{j-1} \sqrt{(E_m - E_n)^2}} * R_{jk}$$

上述公式用于表示 E_i 和 E_j 的决策关系,其中 $\sqrt{(E_i - E_j)^2}$ 表示 E_i 和 E_j 之间的欧氏距离。arg $\min \sum_{m=i,n=i+1}^{j-1} \sqrt{(E_m - E_n)^2}$ 表示 E_i 和 E_j 关联路径上所有节点的欧氏距离和最小的路径距离。也就是说,当聚类后的实体群实体 E_i 和 E_j 的关联度越高,即直接联系越近,那么基于 E_i 和 E_j 作出的决策越相近。

3.4　问题分析及求解

在对问题进行分析求解时,首先对问题进行识别和分类,通过对问题的识别,决定是采用模型求解还是

采用推理机求解。如果采用模型,则由模型自动搜索数据,产生结果;如果采用推理机,则由推理机扫描知识库(规则和事实),产生结果[14]。

3.5　面向电信领域反欺诈识别的应用

电信诈骗是一种低成本、高回报的犯罪,犯罪团伙通过拨打诈骗电话,发送短信等方式骗取受害人财产。当前电信反欺诈已成为电信领域的一个重要的研究内容。

传统的电信反欺诈往往是通过单一的规则判断进行,例如来电号码是否被标记为诈骗电话。这种识别方法在现实中的准确率并不高,容易出现漏报误报情况。文中使用基于知识图谱的智能决策分析,能够通过关系网络识别诈骗团伙,进而从群体关联的角度识别诈骗电话。

·根据实体关系构建知识图谱。

在反欺诈领域,设定实体为每个电话号码个体,实体之间的联系有多种,包括电话主叫/被叫、短信、来源地、号码同一注册人等。在运营商数据中,根据实体之间的联系搭建号码实体关系知识图谱。

·根据 K-Means 聚类诈骗团伙。

根据文中算法对号码进行聚类,假如号码 A 和号码 B 同属一个实体群,并且号码 A 对 C 的通话被规则判断为欺诈电话,那么根据算法可以计算出 B 对 C 的通话是否为欺诈电话。反欺诈决策示意如图 8 所示。

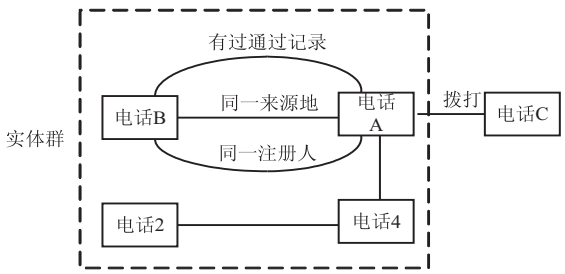

图 8　反欺诈决策示意

4　结束语

文中提出了基于知识图谱的智能决策技术,介绍了从实体识别、知识图谱构建到聚类分析、决策模型构建等技术和方法。能够结合传统决策支撑基于规则判断的方法和知识图谱对问题进行分析和求解。充分考虑了现实环境中多实体相关性对决策的影响。

知识图谱是描述现实世界的组织结构,通过引入知识图谱,能够以更易于理解并且更符合现实环境的方式进行决策建模。领域知识图谱目前在很多行业中发挥了越来越重要的作用,技术上的挑战也有不断的进展。文中提出的方法目前对领域知识图谱的知识推理还没有做深入研究,未来需要结合领域知识对知识图谱的构建、知识推理方法进行研究,以更好地支撑特定领域、行业的智能决策。

参考文献:

[1] 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.

[2] 廖祥文,刘德元,桂　林,等. 融合文本概念化与网络表示的观点检索[J]. 软件学报,2018,29(10):2897-2914.

[3] STUDER R,BENJAMINS V R,FENSEL D. Knowledge enginering:principles and methods[J]. Data & Knowledge Engineering,1998,25(1-2):161-197.

[4] JIN Xiaolong,WAH B W,CHENG Xueqi,et al. Significance and challenges of big data research[J]. Big Data Research,2015,2(2):59-64.

[5] 杨玉基,许　斌,胡家威,等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报,2018,29(10):2931-2947.

[6] 官赛萍,靳小龙,贾岩涛,等. 面向知识图谱的知识推理研究进展[J]. 软件学报,2018,29(10):2966-2994.

[7] SUCHANEK F M,KASNECI G,WEIKUM G. Yago:a core of semantic knowledge[C]//Proceedings of the 16th international conference on world wide web. Banff,Alberta,Canada:ACM,2007:697-706.

[8] 林海伦,王元卓,贾岩涛,等. 面向网络大数据的知识融合方法综述[J]. 计算机学报,2017,40(1):1-27.

[9] 杨晓慧,万　睿,张海滨,等. 基于符号语义映射的知识图谱表示学习算法[J]. 计算机研究与发展,2018,55(8):1773-1784.

[10] 刘　峤,韩明皓,江浏祎,等. 基于双层随机游走的关系推理算法[J]. 计算机学报,2017,40(6):1275-1290.

[11] 管延勇,王洪凯,徐法升. 序决策信息系统中区间决策规则的获取[J]. 控制与决策,2014,29(9):1611-1616.

[12] 陈泽华,张　裕,谢　刚. 基于粒计算的最简决策规则挖掘算法[J]. 控制与决策,2015,30(1):143-148.

[13] 张　波,向　阳. 基于语义的决策模型能力评估与选择方法[J]. 控制与决策,2010,25(9):1324-1328.

[14] 郭剑毅,李　真,余正涛,等. 领域本体概念实例、属性和属性值的抽取及关系预测[J]. 南京大学学报:自然科学版,2012,48(4):383-389.