知识图谱中的关系方向与强度研究

臧根林[1,2]，王亚强[1,2]，吴庆蓉[1,2]，占春丽[1,2]，谢新扬[1,2]
1. 拓尔思知识图谱研究院，广东 广州 510665；2. 广州拓尔思大数据有限公司，广东 广州 510665

摘要
目前普遍的知识图谱构建思路是图谱中的关系标签采用文字描述，这样很难对图谱中的关系进行计算。针对这个问题，提出了关系方向、强度因子和时态因子的概念，关系的正负、强度和时态可以通过有监督机器学习的方法形成自动模型，从而在领域知识图谱中实现关系的量化计算。这种知识图谱构建方法在计算事件舆情走向、计算企业合作与竞争情况变化、分析销售人员市场拓展情况等领域，形成了一种新的数据分析模式，对人工智能在具体行业的落地应用很有意义。

关键词
知识图谱；关系方向；关系强度；负关系；时态关系

中图分类号：TP31 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019027

Study on direction and strength of relation based on knowledge graph

ZANG Genlin[1,2], WANG Yaqiang[1,2], WU Qingrong[1,2], ZHAN Chunli[1,2], XIE Xinyang[1,2]
1. TRS Knowledge Graph Research Institute, Guangzhou 510665, China
2. Guangzhou TRS Big Data Co., Ltd., Guangzhou 510665, China

Abstract
In current popular ideas for knowledge graph construction, the relations in graphs were described by words, it is difficult to calculate the relations in graphs. To this issue, concepts of the direction, intensive factors, temporal factors of relations were proposed. Automatic models of positive, negative, intensive and temporal relations can be formed through supervised machine learning, so that the quantitative calculation of the relations can be implemented in the domain knowledge graph. This method forms a new idea in many areas such as calculating the trend of incidents, calculating the change of cooperation and competition between enterprises, and analyzing the market expansion of sales people. It is meaningful for artificial intelligence to be applied in specific industries.

Key words
knowledge graph, direction of relation, strength of relation, negative relation, temporal relation

1 引言

知识图谱本质上是一个用图数据结构等技术作为载体，描述客观事物及其关系的大型知识库。在具体实现上，知识图谱用语义网 (semantic web) 中的资源描述框架 (resource description framework, RDF) 对知识体系和实例数据两个层面的内容进行统一表示，共同构成一个完整的知识系统[1]。

世界本身不存在结构化问题或非结构化问题，在信息化时代之前也没有人区分结构化或非结构化，但人类进入信息化时代后，什么都想通过计算得到结果，这就产生了信息是否要结构化的问题，因此这是人类为了将就计算机而产生的问题。知识图谱把这个问题进一步加深了，因为大家不仅想让计算机计算信息，还想让计算机学会跟人类一样思考和判断，也就是现在很热门的人工智能。为了教会计算机看懂人类的知识，人们将知识整理成了知识图谱。

虽然知识图谱的概念由谷歌公司在2012年5月首次提出，但知识图谱这样的理念和表达思路在中国已经存在了超过2000年，五行理论和相生相克图就是一个典型的知识图谱结构。知识图谱描述的3个关键要素是实体、属性和实体之间的关系，五行相生相克图（如**图1**所示）完整表达了这3个要素，金、木、水、火、土每个字不但表达一种实体类型，而且字面本身已经蕴含了属性解释，如“水”字是指世界上一切液体性质的实体，每个实体都和另外4个实体发生联系，4条连线代表了实体之间相生或相克的关系。

由于技术原因，五行相生相克图并不能表达更多的信息，如各个实体的属性特征（如火对应红色、对应南方，水对应黑色、对应北方等）只能用很多外延说明进行表达。现在的技术可以很容易解决这样的知识表达，不仅可以用知识图谱方式表达出关系，还能把图片、声音、视频、文档等很多信息直接关联上去。

由于知识图谱技术研究刚刚兴起，目前学术界还没有找到统一的知识图谱定义，在知识表达上还存在诸多缺陷。比如在五行相生相克图中，每个实体都有4条边相连，也就是存在4个关系，但是每一条边的含义都不同。如果把作为计算中心点的实体称为主体，其他与之关联的实体称为关联体，那么每一个主体都存在2条关系指向主体，另外2条关系指向关联体。而在指向主体的2个关系中，一个是相生关系，另一个完全相反，是相克关系。从主体指向关联体的2条关系亦然，分别是相生关系和相克关系。

谷歌公司提出的知识图谱思路和存储架构在智能检索、机器问答等方面有一定优势，因此绝大多数知识图谱研究机构采用了和谷歌公司类似的知识图谱构建思路。但事实上，在很多领域人们对知识图

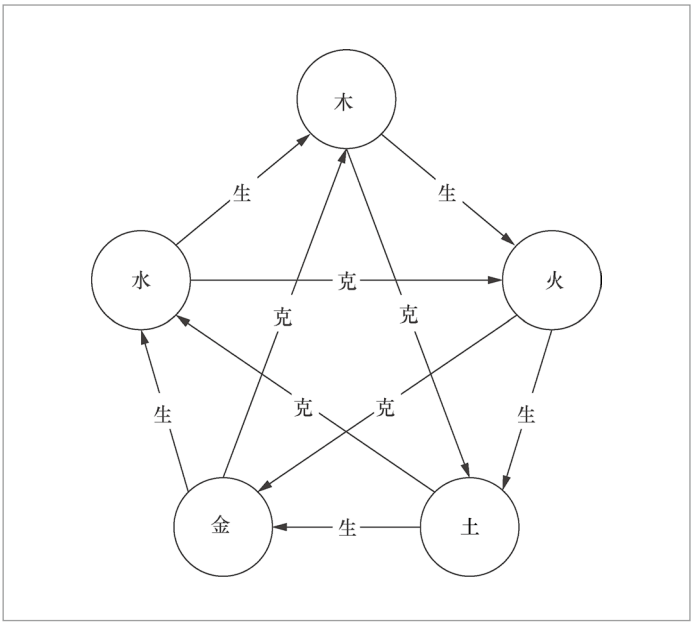

图1 五行相生相克图

谱的需求不仅仅是检索、问答等，还有计算。比如，用知识图谱结构画出一个企业的合作与竞争态势，其中有10家合作关系企业和5家竞争关系企业，经过一年的营销努力，该企业的合作企业数量和竞争企业数量虽然没有改变，但营业额发生了很大变化，这意味着该企业与合作企业的合作力度加强了，从竞争企业手上抢到了更多的市场份额。如果用目前一般的知识图谱构建方法描述这样的一种变化，合作关系企业和竞争关系企业数量没有变化，则只能在关系描述上用文字进行说明，无法描述清楚合作和竞争的变化趋势，更无法量化该趋势。同样，在销售人员与企业联系的例子中，仅使用文字描述关系，只能说明销售人员与企业进行了联系，无法表达出联系的数量和效果等情况，如图2所示。

笔者认为，在知识图谱中的关系类型上增加“负关系”，在关系属性上增加“强度因子”和“时态因子”，可以在某些具体领域知识图谱上实现量化计算，从而让知识图谱更加准确地描述事实。目前虽然有一些研究提到了“逆关系”等涉及关系方向的问题[1]，也提到了关系强度问题[2]，但并没有系统性地提出关系方向、关系强度的问题，特别是关系中存在完全相反的关系。在与时间相关的关系中，目前都是在关系上直接用文字标签进行描述，无法对关系进行计算。关系的正负、强度和时态可以通过有监督机器学习的方法形成自动模型，从而在领域知识图谱中实现关系的量化计算，在计算事件舆情走向、计算企业合作与竞争情况变化、分析销售人员市场拓展情况等很多领域形成一种新的研究思路，对人工智能在具体行业中的落地应用很有意义。

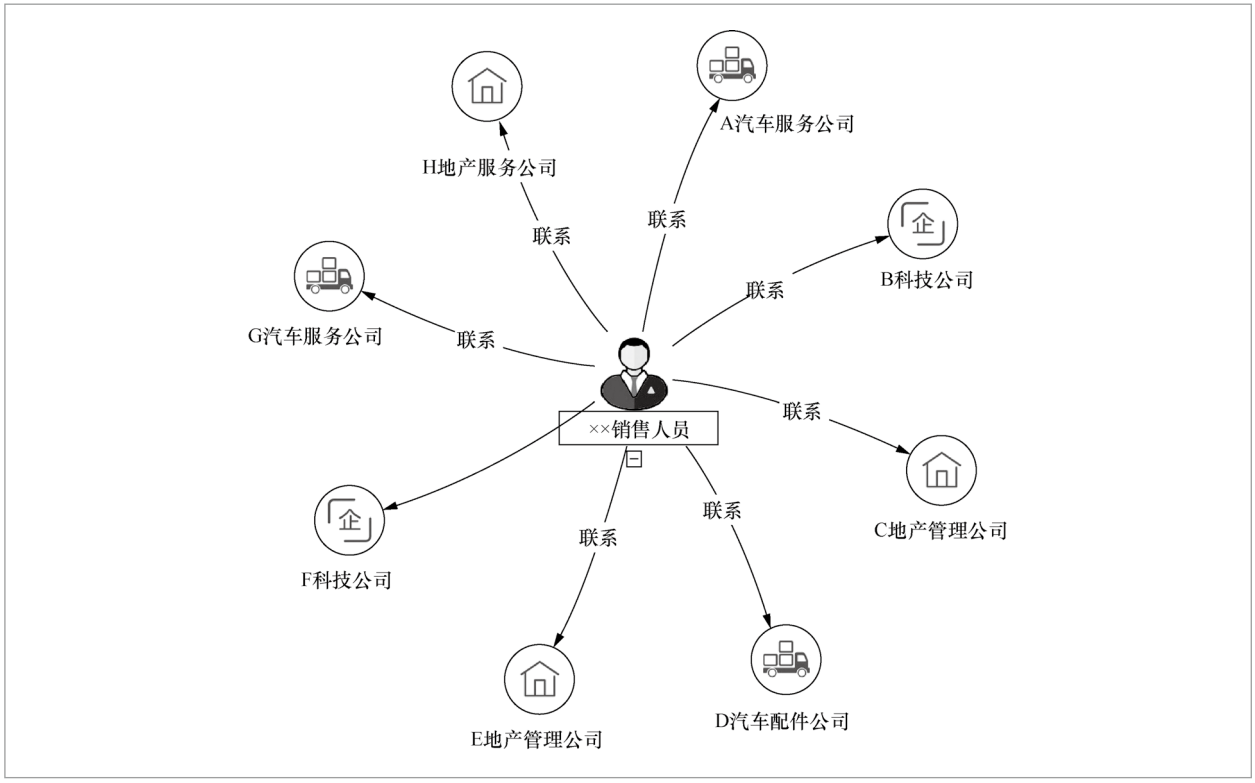

图 2 仅用文字描述销售人员与企业的关系

2 知识图谱中的关系方向

知识图谱中核心的3个元素是实体、属性和关系。实体实际上包含了客观世界的事物、概念、信息等，属性是对实体的描述，这两个元素是相对固定的，如五行中“金”的含义相对固定，而对“金”的属性描述也相对固定。关系则要复杂得多，有相对固定的关系，也有与时间强相关的、变化的关系，有对主体有利的关系，也有对主体有害的关系。

知识图谱中的关系是有方向性的。有一些关系是双向的，还有一些关系是单向的；有一些关系方向是可逆的，还有一些关系方向是不可逆的。在**图1**中，所有的关系是固定不变的，而且不可逆，如“水克火”这条边，方向不可能倒过来，因为水能灭火，是先民们总结出来的共识，已经形成了固定知识。用大火可以把水烧干，是不是就把方向逆转了？这个就是下文要研究的关系强度问题，五行相生相克图是在强度均衡情况下的共识，如果强度严重不一致，就会出现共识的破坏，会产生不正常的后果，这个正是“五行缺X”的含义，表示不正常了。再比如知识图谱中的父子关系也是不可逆的，这样的例子很多。

很多时候关系是固定的，也是可逆的。**图3**是公安部门数据存储的常用分类。数据被分为人（人口信息）、地址（门牌楼等）、案件、物品（刀、枪、毒品等）及组织（某黑社会、邪教组织等）。例如，某犯罪嫌疑人和一个手机号码，可以从机主关联到手机号码，也可以从手机号码关联到机主，这就是可逆的关系，可逆的关系不用画箭头，因为画箭头也是双向箭头，没有意义。

有一些关系方向会发生变化。比如一个犯罪分子和一个犯罪组织的关系，可能在某个时间点之前，该犯罪分子是犯罪组织的团伙成员，经过某些思想工作突然弃暗投明，成为线人，在捣毁这个犯罪组织的过程中起到了关键作用，这实际上就是关系方向发生了变化，这样的关系被称为时态关系。

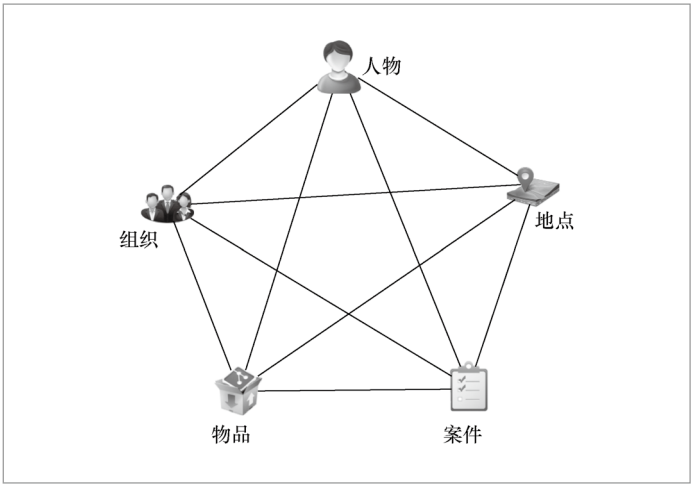

图3 公共安全领域五要素

图1中每一个实体都有4个关系：2个箭头向内，2个箭头向外。同样向内的箭头实际上含义正相反，一个是相生，另一个是相克，笔者把对主体有利的关系称为正关系，而对主体有害的关系称为负关系。虽然关系的方向相同，但可能产生的含义是相反的。

3 负关系与时态关系

在五行相生相克图中可以看到，并不是转换一下箭头方向就可以转换关系的正负的。实际上对于主体而言，每一个关系要承载相生、相克、被相生、被相克4个信息。箭头只能表达主动与被动的关系，或者表达关系的发起方和接受方，却无法表达有利还是有害。在纸面上，可以用“金克木”“木生火”等标注表达关系，但在知识图谱的存储里，

这样的非结构化表达却让计算机更难理解，笔者认为直接用“负关系”表达可以为知识图谱的知识智能计算创造条件。

有一些关系是稳定的，但有一些关系却与时间强相关。比如销售人员与某个企业曾经是友好合作关系，后来因为项目工作没有做好，企业对该销售人员表达了强烈的不满，销售人员与企业的关系变差。经过销售人员的努力挽回，企业又认可了销售人员的工作效果，双方又变成了友好关系。在知识图谱里如何表达这样的知识？笔者认为，关系中应加入时态信息，与时态强相关的关系被称为时态关系。

时态关系分为无头无尾、有头有尾、有头无尾、无头有尾4类。无头无尾的时态关系就是一般的关系类型，没有时间起点和终点，如五行中的相生相克关系；有头有尾的时态关系是指在某一个时间段内存在，但之前和之后都不存在的关系；有头无尾的时态关系是指一旦发生就永远保存的关系，比如一个人的犯罪记录，一旦产生就会一直保存下去；无头有尾的时态关系是指曾经一直存在，但随着时间推移或技术进步，突然失去了存在价值的关系。

4 关系强度

五行相生相克实际上是先民们总结出来的一套平衡理论，把世界上的客观存在划分成金、木、水、火、土五大要素，每个要素相互有生有克，从而达成大自然的平衡。这套理论也映射到了人的身体里面，五行对应人体五脏，若要身体健康就必须保持人体内五行的平衡，这实际上说的就是关系强度的问题。相生相克的前提是强度对等，比如水可以克火，但如果关系强度不对等，火就把水烧干了。

在知识图谱里也存在着关系强度问题。在不同领域、不同体系里人们对强度的定义不一样，笔者认为在大范围的知识图谱里，知识图谱计算或者知识智能的推进是没有意义的，很难形成一个有效的计算模型。但在特定的领域、特定的范围里，人们可以通过机器学习或其他方法，逐步提炼出一些关系特征，建立关系强度定义规则，并对关系进行计算。比如，如果一个企业要画出方方面面的与自身相关的关系是非常复杂的，但如果只画出该企业的客户关系、盟友关系、竞争对手关系，就可以将合同金额、合作签约金额、被竞争对手抢走的合同金额等数字作为强度计算依据，计算出一个量化的数字，若第二年还用相同的规则进行计算，根据得出的同比数字就可以看出企业生存环境的变化情况。当然这只是一个非常简单的例子，实际中的数学模型建立会复杂很多。

负关系在关系强度计算上起到负面的作用，而时态关系则需要增加一个时间系数。知识图谱计算不是简单地正关系累加后减去负关系，而是一个考虑多种约束条件和权重的复杂模型。

5 应用案例

某销售人员客户关系知识图谱的关系方向和强度实现了对销售人员和企业在客户关系中的量化分析。通过对客户联系数量、联系效果等进行情感分析，再结合时态因子和加权计算，可以实现对销售人员客户关系的量化评分，用同样的方法可以实现企业的友好度情况分析，从而表达出事件和企业、销售之间的基于时态的关系度量。

在对销售人员与企业的联系情况进行分析的过程中，在关系属性上增加强度因子（联系数量、联系效果）和时态因子（联系

月份) 可以更细化地描述实体之间的关系，进而更准确地描述事实，如**图4**所示。关系的强度因子和时态因子都可用于量化计算，其计算结果可用于对实体之间宏观变化情况的分析。

从**图4**可以看出，引入了强度因子和时态因子并进行量化计算后，销售人员与企业之间的关系有了新的分析角度，根据计算结果可以得到销售人员的企业关系变化趋势，如**图5**所示。

6 结束语

本文研究了知识图谱中的关系方向问题和关系强度问题，系统性地提出了关系中存在的负关系、时态关系等概念，并对关系强度及可能的关系计算提出了初步探索，笔者认为在某些特定领域内，知识图谱是可以建立有效的模型且可计算的。本文提出的关系类型比大部分已有研究更细化，更能精确地表达实体与实体之间的关系，但也给知识图谱的知识构建、关系抽取、图谱存储等方面增加了复杂度。目前业界大部分知识图谱平台可能还无法支持这样的关系结构，需要进一步进行研究。在知识图谱计算方面，算法模型的设计必将是一个难点，在较大范围内实现计算还不现实，在一些特别行业或领域内建模计算也需要大量的机器学习和模型训练过程，因此需要根据领域特点进行专门研究。

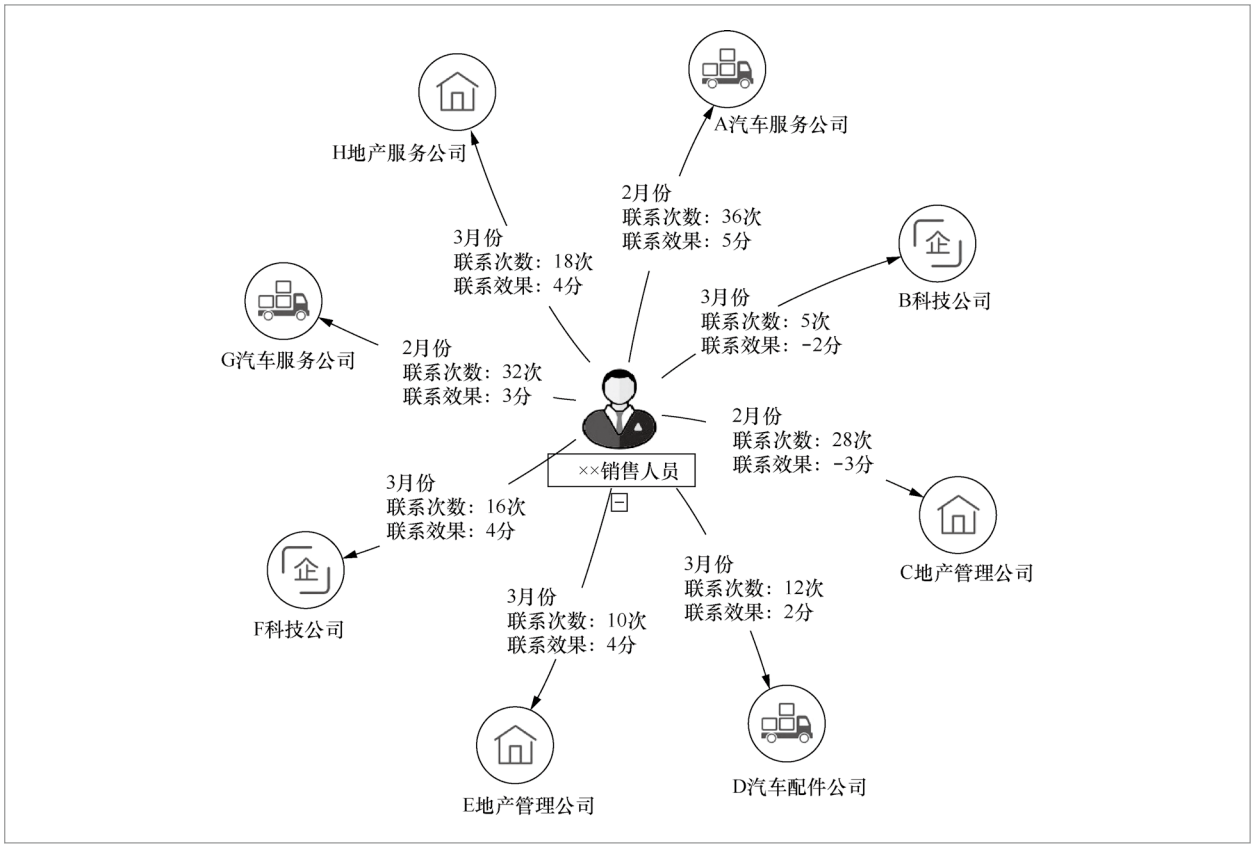

图 4 引入强度因子和时态因子后，销售人员与企业的关系

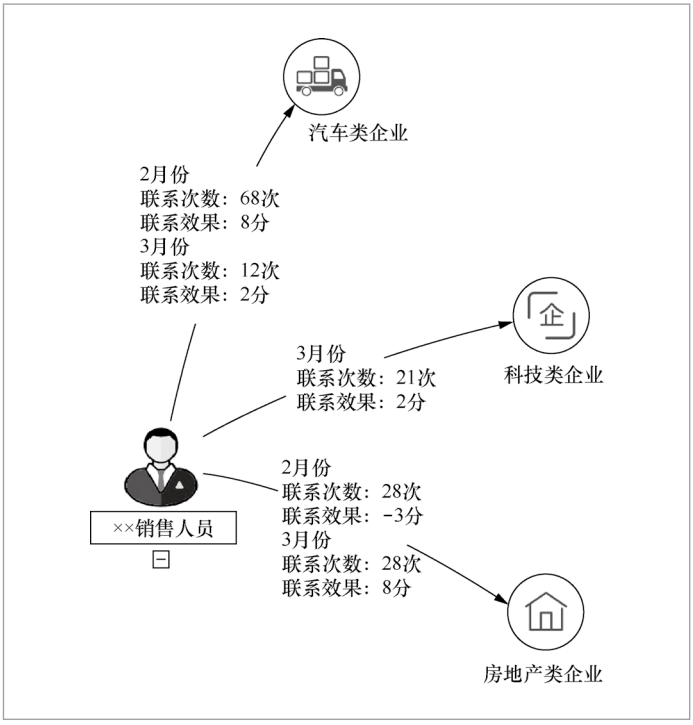

图 5 引入强度因子和时态因子后的关系运算结果

参考文献：

[1] 赵军, 刘康, 何世柱, 等. 知识图谱[M]. 北京: 高等教育出版社, 2018.
ZHAO J, LIU K, HE S Z, et al. Knowledge graph[M]. Beijing: Higher Education Press, 2018.

[2] 陈秀娟, 冷德荣. 面向语言信息处理的语义研究——合著者的社会网络分析[J]. 情报科学, 2013(7): 126-129.
CHEN X J, LENG D R. Semantic study in language and information processing based on social network analysis[J]. Information Science, 2013(7): 126-129.

[3] 秦长江, 侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报, 2009(1): 30-37, 96.
QIN C J, HOU H Q. Knowledge graph: a new area of information management and knowledge management[J]. Journal of Academic Libraries, 2009(1): 30-37, 96.

[4] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016(4): 589-606.
XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016(4): 589-606.

[5] 李涓子, 侯磊. 知识图谱研究综述[J]. 山西大学学报 (自然科学版) , 2017, 40(3): 454-459.
LI J Z, HOU L. Reviews on knowledge graph research[J]. Journal of Shanxi University(Natural Science Edition), 2017, 40(3): 454-459.

[6] 朱木易洁, 鲍秉坤, 徐常胜. 知识图谱发展与构建的研究进展[J]. 南京信息工程大学学报 (自然科学版) , 2017, 9(6): 575-582.
ZHU M Y J, BAO B K, XU C S. Research progress on development and construction of knowledge graph[J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2017, 9(6): 575-582.

[7] SOWA J F. Principles of semantic networks: exploration in the representation of knowledge[J]. Frame Problem in Artificial Intelligence, 1991(2-3): 135－157.

[8] STAAB S, STUDER R, SCHNURR H. Knowledge processes and ontologies [J]. IEEE Intelligent Systems, Special Issue on Knowledge Management, 2001, 16(1): 26-34.

[9] HENDRICKX I, KIM S N, KOZAREVA Z, et al. SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals[C]// The Workshop on Semantic Evaluations: Recent Achievements and Future Directions, June 4, 2009, Boulder, USA. [S.l.]: Association for Computational Linguistics, 2009: 94-99.

[10] HASHIMOTO K, STENETORP P, MIWA M, et al. Task-oriented learning of word embeddings for semantic relation classification[J]. Computer Science, 2015: 268-278.

[11] ZHOU Z Q, QI G L, GLIMM B. Exploring parallel tractability of ontology materialization[C]// European Conference on Artificial Intelligence, Hague, Netherlands, August 29-September 2, 2016. [S.l:s.n.], 2016: 73-81.

作者简介

臧根林（1963- ），男，博士，拓尔思知识图谱研究院院长，广州拓尔思大数据有限公司首席营销官，主要研究方向为知识图谱、知识工程、大数据应用、企业管理、企业文化。

王亚强（1971- ），男，拓尔思知识图谱研究院首席科学家，广州拓尔思大数据有限公司常务副总经理，主要研究方向为领域知识图谱、知识工程。

吴庆蓉（1972- ），女，拓尔思知识图谱研究院研究员，广州拓尔思大数据有限公司第三事业部副总经理，主要研究方向为领域知识图谱、知识工程。

占春丽（1975- ），女，拓尔思知识图谱研究院研究员，广州拓尔思大数据有限公司研发中心总经理，主要研究方向为领域知识图谱、知识工程。

谢新扬（1976- ），男，拓尔思知识图谱研究院研究员，广州拓尔思大数据有限公司副总经理，主要研究方向为领域知识图谱、知识工程。

收稿日期：2019-01-30

“2019年度TOP10大数据应用最佳实践案例”征集活动

一、组织结构

主办单位: CCF大数据专家委员会

协办单位: 中关村大数据产业联盟

二、时间安排

2019年5月25日, 正式发布并启动; 2019年5月—11月, 案例征集阶段; 2019年11月, 初审复评阶段; 2019年12月, 终审及颁奖。

三、征集范围

包含但不限于区块链、人工智能、智能制造、军民融合、电信、金融、能源、教育、交通、医疗、政企、电商、零售、物流等行业和领域。

四、案例内容结构

(一) 项目背景及挑战。包括: 案例名称、所属行业、应用类型等; 企业简介、项目背景、应用场景和技术挑战。

(二) 技术方案。包括: 解决方案及技术选型、所研发的产品及提供的服务类型; 新技术方案的技术架构、特点分析、具体部署和执行思路, 需提供拓扑图; 新技术方案应用前后的分析对比、成果展示。

(三) 经验总结。包括: 项目实施后, 企业业务能力的提升效果 (请从内增效、外增收两个方面提供量化指标) ; 技术方案执行过程中的经验和教训; 该项目的最大创新点; 项目负责人对新技术方案的点评 (需提供该负责人的姓名、照片、职务、个人简介、联系方式) 。

(四) 专家推荐语。请专家结合TOP10案例评选指标阐述推荐此案例的原因, 并分析该案例的亮点。

五、案例评选指标

- 创新性: 是否有技术突破或应用创新。
- 先进性: 是否合理利用了先进的大数据及相关技术。
- 引领性: 是否能够代表所在行业或领域的应用需求和特点。
- 效益性: 是否具有较高的推广价值。

六、特别福利

- 50多个媒体平台的联合宣传, 可极大地提升企业和团队的品牌及技术影响力, 凸显行业内的领先地位和标杆效应, 同时有助于企业推广业务、吸纳人才。
- 所有参选案例均入选《2019大数据应用最佳案例特别专题》。
- TOP30案例分享者将优先受邀在2019中国大数据技术大会 (BDTC) 以演讲嘉宾身份出席大会并解读案例, 享有全程会议通道, 出席闭门交流午宴和晚宴。
- TOP10优秀案例分享者将出席年末颁奖典礼, 获得奖牌并享受独家宣传报道。
- 案例征集及评选过程不涉及任何费用。

七、参与方式

联 系 人: 郭玉婷 (CCF大数据专家委员会秘书处)

联系方式: bigdata@ccf.org.cn, 18510777603