

知识融合研究方法

林 硕,赵 震*

(渤海大学 信息科学与技术学院,辽宁 锦州 121013)

摘 要:知识融合是知识管理的一个重要部分,它可以融合、转换分布式的信息源,是一种将多个信息源的不同知识进行合并,从而产生新知识的过程。知识融合能对不同的知识库进行深度挖掘,实现不同知识系统资源之间的共享,从而获取有价值或可用的新知识。该文主要对知识融合的相关研究方法进行了综述。首先,对国内外研究现状进行了归纳整理并对语义规则、贝叶斯网络等知识融合算法进行了总结,对所用算法的目的和未来研究方向进行了描述。其次,重点从知识融合定义、模式、框架三个方面进行综述,意在解释当前知识融合由于多学科领域交叉,概念边界模糊,没有形成统一的框架的问题,为研究知识融合的通用框架提供了新思路。再次,详细讨论了机器学习方法、深度学习方法等知识融合的前沿方法,并结合具体的应用实例对知识融合方法的特点进行了分析。最后对知识融合领域应用状况及现阶段存在的问题和发展方向进行了总结和展望。

关键词:知识融合;发展现状;信息融合;机器学习方法;综述

中图分类号:TP182

文献标识码:A

文章编号:1673-629X(2022)08-0007-08

doi:10.3969/j.issn.1673-629X.2022.08.002

Review of Knowledge Fusion Research Methods

LIN Shuo,ZHAO Zhen*

(School of Information Science and Technology,Bohai University,Jinzhou 121013,China)

Abstract:Knowledge fusion is an important part of knowledge management. It can integrate and transform distributed information sources, which is a process of merging different knowledge of multiple information sources to produce new knowledge. Knowledge fusion can conduct deep mining of different knowledge bases and realize the sharing between different knowledge system resources, so as to acquire valuable or available new knowledge. In this study, we mainly review the relevant research methods of knowledge fusion. Firstly, the research situation at home and broad is summarized, knowledge fusion algorithms such as semantic rules and Bayesian networks are summed up, and the purpose and future research direction is described. Secondly, the focus is on the three aspects of the definition, model and framework of knowledge integration, aiming to explain the current knowledge integration because of interdisciplinary intersection, vague conceptual boundaries and no unified framework, which provides new ideas for the general framework of knowledge integration. Thirdly, the cutting-edge knowledge fusion methods of machine learning methods and deep learning methods are discussed in detail, and the characteristics of knowledge fusion methods combined with specific application examples are analyzed. Finally, the application status and existing problems and development direction of knowledge integration are summarized and predicted.

Key words:knowledge fusion;development status;information fusion;machine learning methods;review

0 引言

传统的知识融合一般是静态的,在固定的应用场景下,以人为应用主体的知识融合的过程。然而21世纪是一个信息爆炸的时代,每天都有无数条信息流入网络中,传播较为迅速。这些传递的信息是多元化的,而且缺乏统一的描述规则,给不同领域的信息获取和

管理带来了很多挑战。通过信息抽取,实现了从非结构化和半结构化数据中获取实体关系及属性,然而,这些结果中可能存在大量的冗余和错误信息,因此,需要对其进行清理和融合。如何快速、准确地获取信息,让融合后的知识可以更好地满足不同需求的用户,且形成特定问题的领域知识库已成为现阶段研究的重点。

收稿日期:2021-06-21

修回日期:2021-10-26

基金项目:国家自然科学基金项目(61976027);国家社会科学基金(19BTQ028);辽宁省自然科学基金项目(2019-ZD-0496);辽宁省教育厅基本科研项目(LJKZ1028);渤海大学2019年校级教学改革研究项目(BHU-2019-55);渤海大学2021年国家网络安全研究院项目(XK202134-31)

作者简介:林 硕(1997-),女,硕士研究生.CCF会员(H4463G),研究方向为知识图谱构建等;通讯作者:赵 震(1977-),男(满族),博士,CCF会员(D0145M),研究方向为人工智能与知识图谱。

知识融合是在信息融合的基础上发展起来的一个新的概念,它可以看成是信息融合的高级领域。该文的主要贡献如下:

(1)对国内外研究现状进行了归纳整理并对语义规则、贝叶斯网络等知识融合算法进行了总结,对所用算法的目的和未来研究方向进行了描述。

(2)对知识融合模式及框架进行了综述,并详细讨论了机器学习方法、深度学习等方法知识融合的前沿方法。

(3)深入分析知识融合应用状况及现阶段面临的挑战,提出未来研究方向,为知识融合相关研究提供参考。

总体框架如图 1 所示。

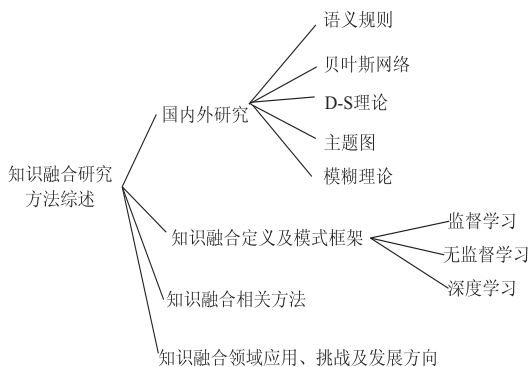


图 1 总体框架

1 知识融合发展现状

如图 2 所示,在中国知网数据库中,对关键词“知识融合”进行模糊检索,共检索 1 721 篇文献。自从 2002 年中国首次发表知识融合的论文以来,关于知识融合的相关研究不多,说明很长一段时间学者们对此的研究不够。2015 年至今,知识融合的相关研究已经引起了学者们的关注,文献数量逐渐增加,但还没有到达顶峰,说明现阶段知识融合已成为热门的研究方向。国外知识融合的研究最早出现在 20 世纪 80 年代后期。语义规则、贝叶斯网络、D-S 理论方面的融合算法是国外研究的重点。国内研究主题图、模糊理论等方面,除此之外还对国外研究的各个方面进行了深入分析。

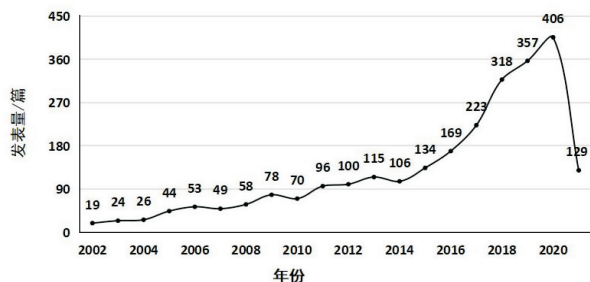


图 2 知识融合研究趋势

1.1 基于语义规则

G Jin 等^[1]在语义规则方面提出一种基于 GA 和语义规则的知识融合算法,提出调整参数和优化融合的反馈机制,融合的结果被结构化地存储在一个知识空间中。侯锦等^[2]利用语义规则将知识对象分类处理,将其转换为对应的本体描述和元知识集。整个框架总体采用分布式结构,具有很好的可扩充性、很强的安全性和实用性以及比较低的误警率。结果表明,提高了知识对象的可重用性和融合的正确率。

1.2 基于贝叶斯网络

贝叶斯网络是研究不确定性知识表达和推理的有效方法,已成为人工智能领域研究的热点之一。基于贝叶斯模型的方法在知识为真时的先验概率和从数据源观察到的条件概率都已知的情况下,求出知识为真的后验概率。后验概率最大时对应的知识就是要找的正确知识^[3]。Santosl 等人^[4]在贝叶斯网络方面,将多个贝叶斯融合成单个贝叶斯,更容易聚合和分解多个源的信息,解决了专家对关系权重意见不一致的问题。张玉洁^[5]提出了一种不需要原始数据的贝叶斯网络融合方法,在已有的贝叶斯网络融合方法的基础上,利用评分机制,得到最终的融合结果。张振海等^[6]使用 K2 算法来学习贝叶斯网络结构。根据贝叶斯定理,如公式(1):

$$p(S^h | C) = \frac{p(S^h, C)}{p(C)} = \frac{p(C | S^h)}{p(C)} \quad (1)$$

其中, $p(S^h)$ 表示网络结构的先验概率, $p(C)$ 表示与结构无关的常数, $p(S^h | C)$ 表示边界似然。通过收集不同专家的意见,使用证据理论排除无意义的因果关系,减小搜索空间,提高算法的学习效率。结果表明,基于专家知识融合的贝叶斯网络构造方法利用专家知识限制学习算法的搜索条件,有效地缩小了搜索空间。

1.3 基于 D-S 理论

D-S 证据理论的方法是融合不同观测结果的信任函数,得到基础概率分配后,再选择最大支持度的假设作为最优判断,从而选择认为正确的知识。D. Andrade 等^[7]在 D-S 理论方面研究了 3 个组合规则,包括原始的 D-S 规则、墨菲规则和基于非精确狄利克模型的规则。结果表明,前两者有汇聚的能力,而后者具有数据挖掘的能力。Sun 等^[8]利用知识融合方法 D-S 理论,对野生鸟类禽流感 H5N1 病毒全球空间的风险估计进行整合。韩立岩等^[9]提出一种新的融合方法: D2S (Dempster2Shafer) 证据理论。利用模糊的概念,选择一个函数,根据估计方法将计算出的数值与阈值的差值,转换为 $[0, 1]$ 之间的数字,此数字代表企业失败的概率。结果表明,提高了企业失败估计的准确性。

1.4 基于主题图

简单来说,就是根据图上的一组现有的边,预测其他边存在的可能性。王海栋等人^[10]提出了一种置信度理论知识融合模型,使用自动校正机制,更好地表现置信度的客观性,加强了在不确定性方面处理的缺点,解决了实体融合过程中信息歧义的问题。鲁慧民等^[11]提出一种面向多源知识的融合算法。从语法、语义和语用三个方面计算相似度,并考虑了概念结构和语境的相似性。结果表明算法在查准率(Precision)、查全率(Recall)和F值(F-measure)均有所提升。评价标准如公式(2)所示:

$$P = \frac{RN}{AN} \quad R = \frac{RN}{PN} \quad F = 2 \times \frac{P \times R}{P + R} \quad (2)$$

其中,PN是通过人工比对认为应该融合的元素对数,AN是算法判定应该融合的元素对数,RN是各元素中正确的元素对。模糊集理论的方法是在D-S证据理

论的基础上工作的。

1.5 基于模糊理论

模糊集理论的方法在D-S证据理论的基础上,进一步放宽了贝叶斯模型的限制条件^[12-13]。目前应用较为广泛的方法是基于模糊积分的方法^[14]。模糊积分是一个非线性函数,可以完成质量评估,找到置信度最高的知识作为正确的知识。Yin等人^[15]设计一种基于粗糙集算法的知识融合模型,可以自动实现复杂表面零件制造过程的质量预测。模型降低了数据的不确定性,从而提高了产品的质量。周芳等^[16]在知识融合中借鉴了信息融合的想法,用基于模糊集理论的方法对多源知识进行融合。结果表明,该工作提高了企业失败预警判别的确定性。

国内外研究现状对比如表1所示。虽然学者们对知识融合有着不同的出发点,但是本质都是为了使知识能够最大化的被利用。

表1 国内外研究汇总

| 类型 | 目的 | 未来研究方向 |
|---------|---|---|
| 基于语义规则 | (1)优化知识融合过程和空间 ^[1] (2)强化语义之间的关系,简化判断语义逻辑关系的过程 ^[2] | (1)将融合规则与其他算法进行结合 |
| 基于贝叶斯网络 | (1)解决在不确定环境下知识融合的问题 ^[4] (2)解决贝叶斯网络的结构问题,优化融合方法,提高融合效率 ^[5] (3)提高算法的学习效率 ^[6] | (1)对融合贝叶斯网络的验证 (2)可能使算法陷入局部最优解;对原贝叶斯网络存在依赖 |
| 基于D-S理论 | (1)提供有关野生鸟类准确、详细的关于病毒H5N1爆发的信息 ^[8] (2)提高企业失败估计的准确性 ^[9] | (1)只涉及一种学科知识,理解方面的发展有限 (2)没有考虑多种证据之间的相关性 |
| 基于主题图 | (1)解决多源知识不一致、不完整的不确定性等问题 ^[10] (2)提高相似性计算的准确性 ^[11] | (1)对反馈机制做形式化定义和收敛证明;对反馈参数的界限和幅度的取值进行优化 (2)对待比较元素进行归类处理,减少比较元素对数目 |
| 基于模糊理论 | (1)降低数据不确定性;证明粗糙度预测的准确性 ^[15] (2)降低企业失败预警判别的不准确性 ^[16] | (1)加强在管理领域的应用研究 (2)提高大规模数据挖掘的效率;丰富相关领域知识的评估 |

2 知识融合定义及模式框架

2.1 知识融合

知识融合是实时地融合和处理多源的信息来创造新的知识的过程,包括实体链接和知识合并两部分。一个典型的知识融合系统应该提供以下三种基本服务^[10]:

(1)知识定位服务:供用户或其他组件在网络上定位相关知识。

(2)知识转换服务:将异构知识资源转换为统一的语言或本体表示。

(3)知识融合服务:对知识资源进行组合和处理,

合并、简化知识,找出满足某种条件限制的解决方案。

其中具有代表性的框架是Preece AD的KRAFT(Knowledge Reuse and Fusion/Transform)^[17],如图3所示。将知识融合定义为从多个异构的资源中对相关的知识进行定位和提取,将其转换为统一的知识模式,使融合的知识能够解决实际问题。

(1)UA:用户为消费者。

(2)W:为系统和KRAFT代理接口提供桥梁。例如:关系数据库的传统接口是SQL/ODBC,KRAFT中的W会接受来自KRAFT中其他代理的请求信息,将其转换为SQL语句并在数据库上运行,最后返回结果。

(3)M:每个 M 从其他代理获取知识,是知识融合的核心。

(4)F:建立服务请求,每个 KRAFT 网络中至少有一个 F。

(5)R:服务资源,包括数据库和知识库。

F 根据 W 提供的信息去寻找匹配的 M 进行连接。当连接是从 W 到 M 时,M 进行知识转换;当一条路径上有多个 M,或同一 M 在多条路径上时,进行知识融合。当连接是从 UA 对应的从 W 到 M 时,M 会用统一的知识模式提供给 UA^[18]。

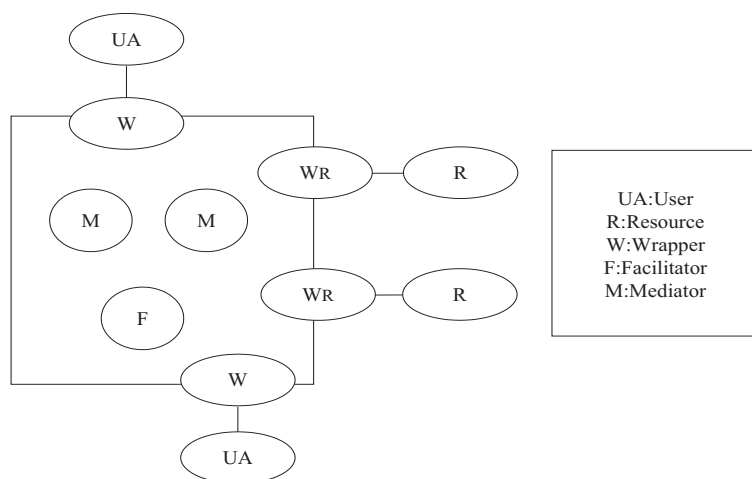


图 3 KRAFT 结构

2.1.1 实体链接

实体链接是指通过对本体库中名字是否具有相同含义来判断实体是否相对应,或者名字不同的实体是否能够表达同一个含义。实体链接又包含实体消歧和共指消解两部分^[19]。

(1) 实体消歧。

例如“张蓝心”这个单词(指称项)可以对应于作为演员身份的张蓝心这个实体,也可以对应于模特身份的张蓝心这个实体,还可以作为国家跆拳道选手身份的张蓝心这个实体。通过具体语境,对实体进行消歧。重点在于计算描述的词汇与实体之间的相似度。

(2) 共指消解。

共指消解是解决多个词汇(指称项)对应同一个实体的问题。例如某文中提到“唐纳德·特朗普”,“川普”,“特朗普”指向的是同一个实体,其中如“他”、“他的”,都有可能指向这个实体。将这些指称项通过共指消解,合并到正确的实体对象中。

2.1.2 知识合并

知识合并包括外部知识库和关系数据库。

(1) 外部知识库:包括数据层面和模式层面。

(2) 关系数据库:将关系数据库的数据转换成三元组。

2.2 知识融合模式

为了解决知识共享问题,将知识融合分为多个层次,可以更好地解决实体的属性、关系以及概念的重复等问题。周利琴^[20]从知识表示的角度,将网络知识模式分为实例、关系、域集、属性和概念融合。其中实例融合是对实体对象进行去重与合并,从而产生新的实

例。域集融合是在实例融合的基础上产生的。关系融合是对多源知识的关系进行对比分析,与属性融合是相互作用的。概念融合则是根据每一次产生新的知识概念来实现的。

2.3 知识融合框架

知识融合框架是进行知识融合的开端,为各个模块提供方向。因为知识融合的复杂性,需要对特定问题制定专门的框架,现在国内还没有统一的知识融合框架。徐赐军等^[21]设计了基于本体的知识融合框架,实现对元知识集进行构建、知识的测量标准、包含融合算法的设计以及融合后处理等功能。可以减少融合的规模,提高准确性。陈思华等^[22]提出一种文化算法框架,采用两阶段遗传算法,包括编码阶段和融合阶段。从两个层面对知识进行优化的知识融合策略,用启发式规则进行表示。谢能付^[23]提出的框架包括知识聚类模块、评估模块和融合模块。

JointDirectors of Laboratories (JDL) 由美国国防部在 1986 年首次提出^[24],主要用于军事领域。JDL 的融合框架如图 4 所示。

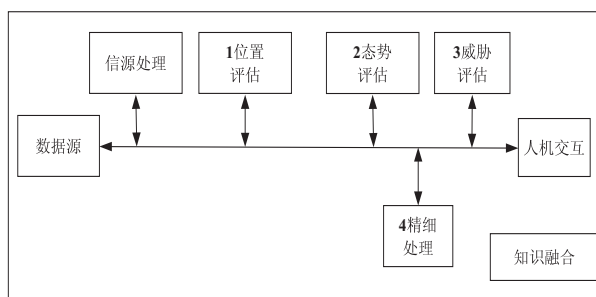


图 4 JDL 的融合框架

3 知识融合相关方法

3.1 机器学习

机器学习方法是一个比较热门的研究主题,是计算机科学和人工智能的一个分支学科。在知识融合领域也可以应用各种机器学习方法。可以根据训练样本是否有输出值,将机器学习方法分为监督学习、无监督学习、半监督学习。监督学习是机器学习中的一种训练方式,监督学习(Supervised Learning)中的常用方法为 SVM、决策树、集成学习等。

3.1.1 监督学习

3.1.1.1 SVM

SVM(Support Vector Machine)是一种二分类模型,通过找到间隔最大的超平面来对数据进行分类,可以转换为一个凸二次规划问题进行求解。Park 等人^[25]提出了一种基于分数级融合的虹膜识别方法。使用两个 Gabor 波滤波器用于局部和全局虹膜处理,用 SVM 融合了由 Gabor 波滤波器计算出的 HD(Hamming Distance)。SVM 表示成公式(3):

$$f = \text{sgn} \left[\sum_{i=1}^k \partial_i y_i K(x, x_i) + b \right] \quad (3)$$

其中, k 表示数据的数量, $y_i \in \{-1, 1\}$ 表示训练样本 x_i 的类标, ∂_i 表示求解二次规划问题的线性约束条件, b 表示偏置。利用核函数将 SVM 扩展到非线性决策面。结果表明,降低了由此产生的认证误差。

一般来说,虹膜识别精度取决于 Gabor 波滤波器的大小、频率和振幅的选择。文中为了减少时间和复杂的操作,使用了传统的 1D Gabor,如公式(4)、公式(5):

$$G(x) = A \cdot e^{-\pi \left[\frac{x-x_0}{\sigma} \right]^2} (\cos 2\pi [\mu_0(x-x_0)]) \quad (4)$$

$$\sum_{x=-N}^N G(x) = 0 (\because DC = 0) \quad (5)$$

其中, A 表示 Gabor 过滤器($G(x)$)的振幅, σ 和 u_0 分别表示 Gabor 波滤波器的大小和频率, x_0 表示移动量, $DC = 0$ 表示将波滤系数归一化为 0。

3.1.1.2 决策树

决策树(Decision Tree)有分类树和回归树。根据损失函数最小化原则建立决策树模型。Elfeky 等^[26]在 TAILOR 工具包中实现了一种 ID3 决策树算法, TALOR 是一个记录链接工具箱。用户可以调整系统参数和插入工具来构建自己的实体对齐模型。结果表明,算法匹配效果高于传统的概率模型方法。张晓丹等^[27]利用 ID3 算法分析和处理测试空间中的多源数据,建立准确的评估模型。最后的结果表明,该方法在解决多源数据问题,并且在处理大量无序和不确定数据方面非常有效。

袁雅萍^[28]用决策树模型作为土壤与环境关系中

知识融合和抽取的方法。使用一致性分析等统计方法,实现多源知识的互补和融合。利用混淆矩阵运算出生产精度(PA)和用户精度(UA),用于表示每个分类的精度指标。总精度(OA),用于表示总体分类的精度指标,通过这些指标共同检验预测土壤图的精度。三种指标的计算公式如下:

$$PA = \frac{P}{N_1} \quad (6)$$

$$UA = \frac{P}{N_2} \quad (7)$$

$$OA = \frac{n}{N} \quad (8)$$

其中, n 表示土壤图所有准确分类的样本数量, N_1 表示野外该类土壤的样本总数, N_2 表示土壤图中划分到该类土壤的样本总数, N 表示样本总数。

3.1.1.3 集成学习

集成学习通过使用一些规则将各个学习器学到的结果整合,从而得到比较好的效果。Chen 等^[29]提出了统一的决策模型,使用 Context-Extended 和 Context-Weight 方法,将两个部分用两种组合方法进行融合。实验结果表明,集成学习框架在不同领域的应用上,实现了更高的匹配质量,也证明了所提出的方法相对于其他方法的优势,提高了实体的消歧质量。

3.1.2 无监督学习

当训练样本数量不足时,可以通过无监督学习完成知识融合中的实体对齐。常用方法是聚类。聚类根据相似度或距离来判断,将相似的样本聚集在相同的类,不相似的样本分散在不同的类。Zhang 等^[30]采用基于实例的无监督学习方法,该方法提供了一个 MBL 框架。结果表明,能够对实体之间的多种关系进行精确的识别,获得了很好的结果。Bhattacharya 等^[31]提出一种关于实体的属性和关系信息的聚类算法。研究了不同关系相似性影响对实体质量的解决办法。结果表明,当数据中存在模糊引用时,关系聚类算法的效果优于属性相似度。Verykios 等人^[32]使用聚类方法,通过少量标记样本推断聚类中其他样本的情况,使用属性和关系的信息来确定实体。结果表明,通过在相似性搜索,在知识获取方面有很大提高。

3.1.3 半监督学习

半监督学习是监督学习与无监督学习相结合的一种学习方法。使用大量的未标记数据,同时使用标记样本,来进行模式识别工作。常用的方法是留一验证法和交叉验证法。Carlson 等人^[33]从网页中提取类别和关系,使用半监督学习方法和 CPL(Coupled Pattern Learner)和 CSEAL(Coupled SEAL)耦合的方式,证明了这种方法可以提高多种类型的提取器的准确性。

3.2 深度学习方法

严格来讲,深度学习属于机器学习范畴。但深度学习可以更好地处理大规模数据,所以将深度学习方法单独列出来。

神经网络(Neural Network)也叫做人工神经网络(Artificial Neural Network),由大量的节点(或神经元)直接相互关联而构成,是一种模仿动物神经网络行为特征,进行分布式并行信息处理的算法数学模型。神经网络包括监督学习和无监督学习。

Gabriel 等^[34]将不同分类方法和神经网络集成在一起,形成代理虚拟组织,用于从 E-nose 检索的参数中进行信息融合,该系统模拟人脑如何分类。利用 PCA 作为一种降维方法,对初始数据进行预处理,然后利用反向传播神经网络 BPNN 对 E-nose 进行分类,结果表明组合分类器的结果和精度均大于单个分类器。

Wang L 等^[35]通过对反向传播(BP)神经网络使用遗传算法(GA)来优化,评估创新生态系统中知识融合的风险。使用预处理后的数据作为神经网络的输入值,确定种群大小和最大迭代次数,选择交叉概率,设置权重和阈值的上下限。结果表明,GA-BP 神经网络具有更快的收敛速度和更高的稳定性,可以更快地实现目标。

Zeng 等^[36]利用分段卷积神经网络和多实例学习进行远程监督关系提取。其使用分段最大池化来自动学习特征,结合多实例学习来解决错误的标签问题。Santos 等^[37]提出一种排名分类模型 CR-CNN,使用单词嵌入作为输入要素,利用卷积神经网络来处理关系分类任务。使用新的成对排名损失函数,可以有效减少人工分类的影响。

3.3 其他方法

Hka B^[38]提出一种基于机器学习和知识图谱的 AM(Additive Manufacturing)框架,对来自国家标准和技术研究所的测量数据,采用分类和回归树的机器学习方法来解决 AM 相关的问题。Wang^[39]提出一种新的知识融合方法 HCKCF(Human-Computer Cooperative Genetic Algorithm),利用进化计算(Evolutionary Computation)融合了人类知识、先验知识和计算知识。

George^[40]提出 CKF(Collaborative Knowledge Fusion)方法,想要了解和控制信息的传播,如何促进真实信息的传播。Balemans^[41]提出了传感器融合方法,为了提高不同环境的感知精度,传感器提供关于相同特性的互补信息,通过结合两个传感器的信息来提高检测精度。

各类知识融合方法特点的总结如表 2 所示。

表 2 知识融合方法总结

| 类型 | 方法 | 特点 |
|-------|---------|--|
| 监督学习 | SVM | 用 SVM 融合两个 Gabor 滤波器计算出的 HD ^[25] 根据 TALOR 模型提出三种机器学习记录链接模型:归纳、聚类和混合模型 ^[26] |
| | 决策树 ID3 | 在处理无序、不确定性的数据效果很好 ^[27] 利用两种模型的耦合;运用忽略和夸大不确定性提高判断推理图的准确程度 ^[28] |
| | 集成学习 | 提出一种 ER 集成架构,能够适应上下文处理的数据集 使用加权投票方法和聚类算法;可自动调参 ^[29] |
| | 聚类 | 使用 K-means;离原点最近的是匹配记录对的集群,距离最远的是不匹配的; 反之是表示可能匹配的记录对的集群 ^[30] 可以处理多个实体和关系类型 ^[31] |
| 无监督学习 | 聚类 | |
| 半监督学习 | 半监督 | 通过对学习任务用联结(Couple)的方法,可以获取更高的精度 ^[33] |
| 深度学习 | 神经网络 | 将不同的分类方法和神经网络结合在一起 ^[34] 利用遗传算法,得到了最优权值和阈值,提高了精度且得到全局最优解 ^[35] 其使用分段最大池化来自动学习特征 ^[36] 利用卷积神经网络来处理关系分类任务 ^[37] |
| | | 可以基于先验知识和 AM(Additive Manufacturing)数据来构建知识 ^[38] |
| | | 在各种条件下收敛到全局真实知识状态的条件和权衡 ^[39] |
| | | 使用现成的训练网络,LRP(layer-wise relevance propagation)剪枝进行优化 ^[41] |
| 其他 | | |

4 知识融合领域应用、挑战及发展方向

知识融合应用领域十分广泛,覆盖自动问答、银行、企业发展等领域。其中由清华大学、清华同方发起的中国知识基础设施工程(CNKI),集成了各个学科的公共知识和各学科专家的个人知识,建立了一个庞大的共享知识库,旨在为科研、教学和知识服务提供基础。知识融合现在有了一定的发展,但仍不能满足人们的需求。现阶段知识融合依旧是一项具有挑战的工作,仍有很多问题需要解决。

(1)知识的不一致性。如何在异构知识情况下,对特定知识进行融合,为用户提供需求是一个艰巨的任务。

(2)知识的复杂关系。存在大规模语义表达相似的知识,导致关系的难理解。这需要更健壮的技术,并能够消除噪声。

(3)实体链接实现的准确性。目前,如何在上下文信息受到限制的情况下,准确地将实体与知识库中的实体链接成为现在普遍关注的问题。

未来知识融合领域也有更多的发展方向:

(1)实时融合大规模知识,进行多种语言的融合。

(2)建立一个统一、专门的知识融合体系结构。

(3)将深度学习应用到知识融合中。可以获得更高的性能和预测精度。深度学习能够从大数据中获取实体之间复杂、模糊的关系,是很有效率的方法。

未来的研究应该更加投入到图书情报中,将知识融合充分运用到其中。知识图谱成为智能搜索的关键技术,具有很深远的价值。知识融合是知识图谱中的一个重要环节,期待更多的研究人员可以对此进行深入研究,促进知识融合领域的发展。

参考文献:

- [1] GOU J, WU Y, LUO W. Knowledge fusion: a new method to share and integrate distributed knowledge sources [C]//Innovative approaches for learning and knowledge sharing. Crete: Springer, 2006: 609-614.
- [2] 缙 锦, 吴扬扬, 罗 伟, 等. 一种基于语义规则的知识融合方法 [C]//全国理论计算机科学学术年会. 长春:《计算机科学》杂志社, 2006: 305-306.
- [3] 杭婷婷, 冯 钧, 陆佳民. 知识图谱构建技术: 分类、调查和未来方向 [J]. 计算机科学, 2021, 48(2): 175-189.
- [4] SANTOS E, WILKINSON J T, SANTOS B E. Fusing multiple Bayesian knowledge sources [J]. International Journal of Approximate Reasoning, 2001, 52(7): 935-947.
- [5] 张玉洁. 基于评分的贝叶斯网融合方法 [D]. 昆明: 云南大学, 2011.
- [6] 张振海, 王晓明, 党建武, 等. 基于专家知识融合的贝叶斯网络结构学习方法 [J]. 计算机工程与应用, 2014, 50(2): 1-4.
- [7] ANDRADE D, HOREIS T, SICK B. Knowledge fusion using Dempster-Shafer theory and the imprecise Dirichlet model [C]//IEEE conference on soft computing in industrial applications. Muroran: IEEE, 2008: 32-39.
- [8] SUN Liqian, WARD M P, LI Rui, et al. Global spatial risk pattern of highly pathogenic avian influenza H5N1 virus in wild birds: a knowledge-fusion based approach [J]. Preventive Veterinary Medicine, 2018, 152(1): 32-39.
- [9] 韩立岩, 周 芳. 基于 D-S 证据理论的知识融合及其应用 [J]. 北京航空航天大学学报, 2006, 32(1): 65-68.
- [10] 王海栋, 郑晓庆, 张红俊. 基于置信度理论的网络知识融合系统和应用 [J]. 计算机系统应用, 2011, 20(1): 1-6.
- [11] 鲁慧民, 冯博琴, 李 旭. 面向多源知识融合的扩展主题图相似性算法 [J]. 西安交通大学学报, 2010, 44(2): 20-24.
- [12] ZADEH L A. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338-353.
- [13] ABDULGHAFOR M, CHANDRA T, ABIDI A. Data fusion through fuzzy logic applied to feature extraction from multi-sensory images [C]//IEEE international conference on robotics & automation. Atlanta: IEEE, 1993: 2-6.
- [14] GRABISCH M, MUROFUSHI T, SUGENO M. Fuzzy measures and integrals: theory and applications [M]. [s. l.]: Springer, 2000: 477.
- [15] YIN Y, ZHANG L, LIAO W, et al. A knowledge resources fusion method based on rough set theory for quality prediction [J]. Computers in Industry, 2019, 108: 104-114.
- [16] 周 芳, 刘玉战, 韩立岩. 基于模糊集理论的知识融合方法研究 [J]. 北京理工大学学报: 社会科学版, 2013, 15(3): 67-73.
- [17] PREECE A D, HUI K Y, GRAY W A, et al. KRAFT: an agent architecture for knowledge fusion and transformation [J]. Knowledge-Based Systems, 2000, 13(2-3): 113-120.
- [18] 缙 锦. 知识融合中若干关键技术研究 [D]. 杭州: 浙江大学, 2005.
- [19] 孙婉莹. 面向知识图谱的共指消解方法研究 [D]. 长春: 吉林大学, 2018.
- [20] 周利琴, 范 昊, 潘建鹏. 网络大数据中的知识融合框架研究 [J]. 情报杂志, 2018, 37(1): 145-150.
- [21] 徐赐军, 李爱平, 刘雪梅. 基于本体的知识融合框架 [J]. 计算机辅助设计与图形学学报, 2010, 22(7): 1230-1236.
- [22] 陈思华, 陶长琪. 一种新的文化算法框架下知识融合策略 [J]. 小型微型计算机系统, 2009, 30(10): 2030-2033.
- [23] XIE Nengfu, CAO Cungen, GUO Hongyu. A knowledge fusion model for web information [C]//2005 IEEE/WIC/ACM international conference on web intelligence. Compiegne: IEEE, 2005: 67-72.
- [24] PREECE A, HUI K, GRAY W A, et al. KRAFT: an agent architecture for knowledge fusion [J]. International Journal of Cooperative Information Systems, 2001, 10(1-2): 171-195.
- [25] PARK H A, KANG R P. Iris recognition based on score level fusion by using SVM [J]. Pattern Recognition Letters, 2007,

- 28(15);2019–2028.
- [26] ELFEKY M G, VERYKIOS V S, ELMAGARMID A K. TAILOR: a record linkage toolbox [C]//International conference on data engineering. [s. l.]: IEEE, 2002: 17–28.
- [27] 张晓丹, 赵海, 王刚, 等. 不确定信息的模糊决策融合算法[J]. 东北大学学报: 自然科学版, 2004, 25(7): 657–660.
- [28] 袁雅萍. 利用不确定性模型实现土壤推理制图中知识的获取与融合[D]. 武汉: 华中农业大学, 2016.
- [29] CHEN Z, KALASHNIKOV D V, MEHROTRA S. Exploiting context analysis for combining multiple entity resolution systems [C]//ACM SIGMOD international conference on management of data. Providence RI: ACM, 2009: 207–218.
- [30] ZHANG Y, ZHOU J. A trainable method for extracting Chinese entity names and their relations [C]//Proceedings of the 38th annual meeting of the association for computational linguistics. [s. l.]: Association for Computational Linguistics, 2000: 66–72.
- [31] BHATTACHARYA I, GETOOR L. Collective entity resolution in relational data [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 5–es.
- [32] VERYKIOS V S, ELMAGARMID A K, HOUSTIS E N. Automating the approximate record–matching process [J]. Information Sciences, 2000, 126(1–4): 83–98.
- [33] CARLSON A, BETTERIDGE J, WANG R C, et al. Coupled semi–supervised learning for information extraction [C]//Proceedings of the third international conference on web search and web data mining. [s. l.]: FDBLP, 2010: 101–110.
- [34] VILLARRUBIA G, DE PAZA J F, PELKI D, et al. Virtual organization with fusion knowledge in odor. classification [J]. Neurocomputing, 2017, 231: 3–10.
- [35] WANG L, BI X. Risk assessment of knowledge fusion in an innovation ecosystem based on a GA–BP neural network [J]. Cognitive Systems Research, 2020, 66(46): 201–210.
- [36] ZENG D, KANG L, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]//Conference on empirical methods in natural language processing. Lisbon: Association for Computational Linguistics, 2015: 1753–1762.
- [37] SANTOS C, BING X, HOU B. Classifying relations by ranking with convolutional neural networks [J]. arXiv: 1504.06580v1, 2015.
- [38] HKA B, PW A, YAN L A, et al. Machine learning and knowledge graph based design rule construction for additive manufacturing [J]. Additive Manufacturing, 2020, 37: 101620.
- [39] WANG Y, TENG H. Knowledge fusion design method: satellite module layout [J]. Chinese Journal of Aeronautics, 2009, 22(1): 32–42.
- [40] KAMPIS G, LUKOWICZ P. Collaborative knowledge fusion by Ad–Hoc information distribution in crowds [J]. Procedia Computer Science, 2015, 51: 542–551.
- [41] BALEMANS D, CASTEELS W, VANNESTE S, et al. Resource efficient sensor fusion by knowledge–based network pruning [J]. Internet of Things, 2020, 11: 100231.