面向网络大数据的知识融合方法综述

林海伦[1)] 王元卓[2)] 贾岩涛[2)] 张 鹏[1)] 王伟平[1)]

[1)](中国科学院信息工程研究所 北京 100093)

[2)](中国科学院计算技术研究所 北京 100190)

摘 要 网络大数据是指“人、机、物”三元世界在网络空间中交互、融合所产生并在互联网上可获得的大数据.网络大数据中蕴含丰富的知识资源,包括描述特定事物的实体、刻画实体逻辑联系的关系、用于语义标注实体的分类等.知识自身呈现出异质性、多元性和碎片化等特点.如何在网络大数据环境下海量碎片化的数据中提取出能够用于解决问题的知识,并对知识进行有效的融合计算,将从网络大数据中获得的知识有效组织起来是知识库构建亟待解决的技术难点和当前研究的热点.该文从知识融合的定义出发,介绍近年来的可用于知识融合的技术和算法的最新进展,通过分类和总结现有技术,为进一步的研究工作提供可选方案.文中首先介绍了在知识融合中用于判断知识真伪的知识评估的若干研究和评估方法;然后基于知识评估的结果,从实体扩充、关系扩充和分类扩充 3 个方面详细总结了知识融合中各种可用的知识扩充方法和研究进展;探讨了应用于网络大数据的知识融合的总体框架;基于这些讨论,总结面向网络大数据的知识融合面临的主要挑战和可能解决方案,并展望了该技术未来的发展方向与前景.

关键词 网络大数据;知识库;知识融合;知识评估;知识扩充

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2017.00001

Network Big Data Oriented Knowledge Fusion Methods: A Survey

LIN Hai-Lun[1)] WANG Yuan-Zhuo[2)] JIA Yan-Tao[2)] ZHANG Peng[1)] WANG Wei-Ping[1)]

[1)](*Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Network big data refers to the massive data generated via interaction and fusion of the ternary human-machine-thing universe in the cyberspace and available on the Internet. There is a large amount of knowledge elements in big data, such as entities representing specific objects, relations depicting logic connections between entities, classes annotating entities semantics, etc. The very fast development of knowledge in big data environment has presented the characteristics of heterogeneity, variety and fragmentation. How to extract and fuse knowledge from large and fragmented data, to effectively organize the knowledge elements obtained from the big data, have become a technical difficulty to solve and also a hot research topic in knowledge base construction. This paper presents a survey on the techniques and algorithms of knowledge fusion in decades, and expects to provide alternative options for further research by analyzing the existing methods. Firstly, the most commonly knowledge evaluation methods used to judge the authenticity of knowledge in knowledge fusion are introduced. Secondly, the research progress of knowledge population is reviewed in detail from entity population, relation population and taxonomy population

收稿日期:2016-01-26;在线出版日期:2016-08-29.本课题得到国家科技支撑计划(2012BAH46B03)、核高基项目(2013ZX01039-002-001-001)、国家重点研发计划(2016YFB1000902)、国家自然科学基金(61602467,61303056,61402442,61402464,61572469,61572473,61502478)、北京市自然科学基金项目(4154086)资助.**林海伦**,女,1987 年生,博士,助理研究员,主要研究方向为开放知识网络、信息抽取.E-mail:linhailun@iie.ac.cn.**王元卓**,男,1978 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为网络行为分析、开放知识网络等.**贾岩涛**,男,1983 年生,博士,助理研究员,主要研究方向为开放知识网络、社会计算等.**张 鹏**,男,1984 年生,博士,副研究员,主要研究方向为云计算、流数据处理.**王伟平**,男,1975 年生,博士,研究员,主要研究领域为大数据存储与处理.

aspects. Thirdly, the overall framework of knowledge fusion is discussed. Finally, this paper summarizes the key challenges and possible solutions, and further gives a future outlook on the research of knowledge fusion for network big data.

Keywords network big data; knowledge base; knowledge fusion; knowledge evaluation; knowledge population

1 引 言

随着互联网、物联网、云计算等技术的迅猛发展,网络空间(Cyberspace)中各类应用层出不穷,这些应用在改变人们生活方式的同时,也产生了巨大的数据资源,形成了网络空间的大数据(简称网络大数据)[1].根据 2014 年 EMC 公司公布的第七份数字宇宙(Digital Universe)报告[2],通过国际数据公司(International Data Corporation,IDC)的研究和分析,2013 年全球大数据总量为 4.4 ZB,预计到 2020 年全球的数据将增长 10 倍,总量达到 44 ZB,这些数据包含大量非结构化和半结构化数据,以及结构化数据,而且常常以数据流的形式动态、快速地产生,具有很强的时效性.

这些网络数据中蕴含丰富的以实体为中心的知识资源,包括描述特定事物的实体,刻画实体逻辑联系的关系以及用于语义标注实体的分类等,因此,在知识工程领域,知识通常描述为实体、关系、分类等要素及其组合的形式.这些知识对于人们理解和获取有用的信息具有重要的作用.然而,从单一数据源获取的知识并不全面、知识间缺乏深入的关联,给知识的理解和应用带来巨大的困难.因此,应当将网络数据中蕴含的知识进行有效的关联,将其转变成为一种基础知识资源来协同提供服务,从而有效利用网络大数据的价值.

正如 Google 的首席经济学家 Hal Varian 所说,在当前网络大数据时代,数据是广泛可用的,所缺乏的是从中获取知识的能力[3];有效利用网络大数据价值的主要任务不是获取越来越多的数据,而是从数据中挖掘知识,对知识进行有效的组织关联,并将其应用到实际问题解决中.知识库作为知识组织管理的一种特殊的数据库,是知识存储和计算的重要组织形式[4],得到了国内外广泛的研究.

知识库是用来描述现实世界中实体间的关系,通常以网络的形式进行组织,网络中的每个节点代表实体,每条连边代表实体间的关系.知识库是推动人工智能发展和支撑智能服务应用的重要基础技术.在过去几十年,人们曾尝试采用专家知识、利用群体智慧、自动或半自动知识抽取三类方法来构建知识库.专家知识是指根据专家的经验,获得的启发式知识,通常由领域专家参与完成.在这种情况下,受时间和经济成本的约束,很难实现大规模知识库的构建.利用群体智慧是指采用众包机制,通过互联网任何人都可以参与到知识的编辑.自动或半自动知识抽取方法是当前流行的构建知识库的主要方法,其基本思想是通过自动或半自动的算法,从网络数据中提取知识.

目前,国内外多个研究机构建立了很多知识库,并在此基础上构建了多种应用系统[5].其中,有代表性的知识库包括 WordNet[6]、KnowItAll[7]、Freebase[8]、DBpedia[9]、WikiTaxonomy[10]、YAGO[11-12]、ReadTheWeb[13]和基于网络大数据构建的概率化的知识库 Probase[14]、Knowledge Vault[15]等.此外,还有一些著名的知识搜索和计算平台,例如 Google 公司的知识图谱[16]、Wolfram 公司的知识在线自动问答系统 Wolfram Alpha① 和美国的官方政府搜索平台 Data.gov② 等.在国内,代表性的工作有搜狗知立方③、百度知心④、陆汝钤院士等人[17]提出的知件、上海交通大学 zhishi.me[18]和复旦大学 GDM 中文知识图谱展示平台⑤等.

这些知识库和知识平台试图从网络大数据中,依靠信息抽取和自然语言理解等技术,自动或半自动地构建知识库,为用户提供"所搜即所得"的智能服务,同时通过知识的分类,有效识别用户的搜索意图,提供更加精准的检索结果,优化搜索结果展示,实现智能推理,提高信息推荐质量等,通过这些应用充分体现了知识库应对面向网络大数据的上层应用时所表现的价值,展现了知识库的实用性.然而,在

① http://ww1.wolframalpha.co/
② http://www.data.gov/
③ http://baike.sogou.com/v66616234.htm
④ http://www.newhua.com/2013/0208/197829.shtml
⑤ http://gdm.fudan.edu.cn/GDMWiki/Wiki.jsp?page=CKGBDR

网络大数据背景下，对知识库的实用性提出了更高的要求，主要表现在以下几个方面：

(1) 覆盖性. 网络大数据规模巨大(Volume)，不仅体现在数据源包含的数据规模大，而且数据源的规模也很大，即使是一个领域的数据源也成千上万. 因此，这就要求知识库必须具备很强的覆盖能力，原因在于：单一来源或片面的知识缺乏深入的关联，无法全面了解知识，给知识的理解和应用带来巨大的困难.

(2) 时新性. 网络大数据变化高速(Velocity)，不仅互联网持续产生新的可用数据，而且数据源本身也是动态更新的. 因此，这就要求知识库必须具备快速扩展新知识的能力和响应知识变化的能力，保证知识库中知识的时新性，从而满足用户对知识的时新性要求.

(3) 包容性. 网络大数据多样(Variety)，不仅包含大量异构的数据，而且包含大量同义共指的数据和多义表达的歧义数据. 因此，这就要求知识库必须具备求同存异的能力，包容知识的不同表达形式，从而保证应用和用户的个性化需求得到满足.

(4) 价值性. 网络大数据中数据源各不相同(Value)，可能包含大量的老旧数据、错误数据，导致数据的价值密度各不相同. 因此，这就要求知识库必须确保知识的价值性，以防提供错误知识，给用户带来损失.

通过上述分析可以看出，来源于大数据的知识面临以下问题：(1) 知识分散在网络大数据中，要从网络大数据中公开的海量碎片化数据中获取知识无异于“大海捞针”；(2) 知识的真值随时间动态演化，知识之间可能存在新值与旧值的冲突，同时，知识的真值可能会湮没在错误值之间，导致知识真值发现难；(3) 由于自然语言表达的多样性，存在大量同义和多义表达的知识，导致知识的语义理解难；(4) 依托不同数据源的知识的质量与数据源的质量密切相关，导致知识的价值判断难. 针对这些困难和挑战，国内外工业界和学术界通过研究知识融合方法，将网络大数据获得的知识有层次、有结构、有次序的关联组织起来，构建相应的知识库来支撑上层应用，挖掘网络大数据的价值.

知识融合[19]是将从网络大数据公开的碎片化数据中获取的多源异构、语义多样、动态演化的知识，通过冲突检测和一致性检查，对知识进行正确性判断，去粗取精，将验证正确的知识通过对齐关联、合并计算有机地组织成知识库，提供全面的知识共享的重要方法. 通过知识融合的定义可以看出，知识融合建立在知识获取的基础上，知识获取为知识融合提供知识来源. 在知识融合中，如何刻画开放网络知识的质量，消除知识理解的不确定性，发现知识的真值，将正确的知识更新扩充到知识库中是研究者们关注的重点. 知识融合不同于数据融合、信息融合[19]. 数据融合处理的是最原始的、未被加工解释的记录，表现为对文本、数字、事实或图像等数据的关联、估计与合并. 信息融合处理的对象则是被加工过的建立关联关系的数据，被解释具有某些意义的数字、事实、图像以及能够解答某一问题的文本等形式的信息. 而知识融合处理的对象是知识，值得重点关注的是知识不是数据的简单累积，而是有序的可用于指导实践的信息[19].

面向网络大数据的知识融合方法的研究具有非常重要的意义. 从理论角度看，知识融合是自然语言处理、人工智能领域所面临的重要研究课题之一，知识融合研究所取得的每一个进步都有助于计算机加深对人类的智能、语言、思维等问题的理解；从知识工程角度看，知识融合为构造适应网络大数据环境下的知识库提供有效的扩展方法，保障知识库的开放性，通过知识融合方法的研究解决知识库中知识覆盖面窄，知识库难以动态扩展的难题；从应用上看，知识融合具有巨大的社会价值和经济效益，它是将来自网络大数据碎片化数据中的知识关联起来进行重构，基于网络大数据背后隐藏的知识之间的关系，建立知识库的重要手段，而知识库在很多应用中起着至关重要的作用，例如在检索方面，借助知识库进行检索具有多方面优势，如它可以使检索结果更加精准，不仅如此，借助知识库进行检索还可以智能分析用户的意图，并且进行推理与计算，直接给出用户想要的结果. 除此之外，借助知识库可以提供更加全面的检索结果，通过知识库构建的完整的知识体系，用户可以更加全面的掌握知识点；在当前快速发展地电子商务领域中，借助构建的商品知识库，融合用户的兴趣特点和行为，可以准确的向用户推荐用户感兴趣的信息和商品；在知识问答、知识推理、情报分析等方面也有重要应用. 由此可见，面向网络大数据的知识融合方法研究不但具有深远的理论价值，而且有着广泛的应用前景，可以创造巨大的社会和经济效益.

目前，国内外工业界和学术界对知识融合的关键技术展开了广泛的研究. 然而，现有工作大部分是针对知识融合中的局部问题或特定技术，如知识评

估、实体链接、分类对齐、分类合并等，还没有形成一套系统的理论方法以及完整的计算模式和框架.以知识获取为基础，本文将介绍面向网络大数据的知识融合的最新进展.首先，介绍开放网络知识评估方法的研究进展；然后，基于知识评估的结果，介绍验证为正确的开放网络知识扩充方法的研究进展；总结适用于面向网络大数据的知识融合方法的总体框架；最后，展望知识融合的未来发展方向和前景.

2 开放网络知识评估方法

开放网络知识评估建立在知识获取之上，其主要目标是解决从网络大数据不同数据源中获取的知识之间的冲突和不一致性，并从中找到反应真实世界的事实，即知识的真值.知识评估是知识融合的首要步骤，对验证为正确的知识继续进行融合计算才有意义.目前，知识评估的研究工作主要分为以下几类：包括传统的基于贝叶斯模型的方法、D-S 证据理论的方法和模糊集理论的方法，以及近几年提出的基于图模型的方法，下面详细介绍这些方法.

2.1 基于贝叶斯估计的知识评估方法

考虑到不同数据来源的知识质量可能不一定相同，基于贝叶斯模型的知识评估方法的基本思想如下[20-21]：设 $k=\{k_i \mid 1\leq i\leq n\}$ 是一组待评估的知识，其对应的先验概率为 $P(k_i)$，$S=\{S_j \mid 1\leq j\leq m\}$ 是由 m 个数据源获取的对 k 的观察，根据贝叶斯理论，在观察条件下，k_i 为真(true)的后验概率为

$$P(k_i \mid S)=\frac{P(S \mid k_i)\cdot P(k_i)}{P(S)} \tag{1}$$

其中，$P(S)=P(S_1,\cdots,S_m)$ 为观察值的联合概率分布.假设观察值 $S=\{S_j \mid 1\leq j\leq m\}$ 之间相互独立，则有

$$P(S \mid k_i)=\prod_{j=1,m} P(S_j \mid k_i) \tag{2}$$

将式(2)代入式(1)则得到

$$P(k_i \mid S)=\frac{\prod\limits_{j=1,m} P(S_j \mid k_i)\cdot P(k_i)}{\sum\limits_{i=1,n}\prod\limits_{j=1,m} P(S_j \mid k_i)\cdot P(k_i)} \tag{3}$$

通过式(3)可以看出，当知识 k_i 为真的先验概率 $P(k_i)$ 已知，并从数据源观察中获得条件概率 $P(S_j \mid k_i)$，就可以求得 k_i 为真的后验概率 $P(k_i \mid S)$(也表示为 $P(k_i \mid S_1,\cdots,S_m)$).根据最大后验概率准则(Maximum A Posteriori，MAP)，后验概率最大时对应的 k_i 即为要找的正确的知识.

基于贝叶斯模型的方法提供了一种计算假设概率的方法，基于假设的先验概率、给定假设下观察到的不同知识的概率以及观察到的知识本身而得出，计算简单、直接.然而，贝叶斯方法需要满足如下条件：不同来源的知识之间的观测是相互独立的，而且这些知识的先验概率是可预知的，这在网络大数据环境中很难得到满足，从而无法保证贝叶斯方法在具体应用中的实用性.

2.2 基于 D-S 证据理论的知识评估方法

该方法[22-23]是对贝叶斯概率论方法的进一步扩展，它具有直接表示“不知道”和“不确定”的能力，支持满足比贝叶斯概率论更弱的条件，能够处理不确定信息.基于 D-S 证据理论的知识评估方法的基本思想如下[24]：

设 Ω 表示识别框架，其由互不相容的基本命题组成，用于表示对某一问题的所有可能答案，但只有一个是正确的.Ω 的子集称为命题，所有命题组成的集合记为 2^{Ω}.m 函数表示分配给各命题的信任程度，也称为基本概率分配函数[24].对于 $\forall A\in 2^{\Omega}$，记 $m(A)$ 为基本可信函数，表示对命题 A 的信度的大小.记信任函数 $Bel(A)$ 表示对 A 的信任程度.记 $P_l(A)$ 为似然函数，表示对 A 非假的信任程度[24].在实际中，$[0,Bel(A)]$ 为 A 的支持证据区间；$[Bel(A),P_l(A)]$ 为 A 的不确定区间；$[0,P_l(A)]$ 为 A 的拟信区间；$[P_l(A),1]$ 表示 A 的拒绝证据区间.设 m_1 和 m_2 分别表示由两个相互独立的证据源导出的函数，则基于 Dempster 组合规则可计算出由这两个证据共同作用产生的表示融合信息的新的 m 函数.

基于 D-S 证据理论的方法主要根据数据源提供的知识和先验信息，处理流程如下：首先，利用数据挖掘等手段，提取不同观测结果的信任函数；其次，基于 Dempster 证据组合规则，对观测结果的信任函数进行融合；然后，得到基础概率分配，选择具有最大支持度的假设作为最优的判断，从而选择认为正确的知识.

基于 D-S 证据理论的方法能够很好地建模不确定性的知识，解决知识冲突的问题，但该方法与贝叶斯方法类似，也要求参与评估的知识源之间是相互独立的，其判别决策含有更多的主观性，而且当知识源间高度冲突时，往往产生相悖的结论，同时，该方法的时间复杂度随知识源数目的增加呈指数级增长.综上，基于 D-S 证据理论的方法难以有效处理网络大数据中大规模知识的评估问题.

2.3 基于模糊集理论的知识评估方法

该方法采用分类的局部理论，在 D-S 证据理论

方法的基础上，进一步放宽了贝叶斯概率论方法的限制条件[25-26]．较为流行的基于模糊集理论的评估方法是采用基于模糊积分的方式[27]，具体如下：

设集合 X，其 Borel 域为$\varnothing$，则定义在$\varnothing$上的测量函数 g 满足以下条件：

$$\begin{cases} g(\varnothing)=0 \\ g(X)=1 \end{cases},$$

$$g(A)\leq g(B),\ A\subseteq B \text{ 且 } A,B\subseteq X,$$

$$\begin{cases} \lim g(A_i)=g(\lim A_i) \\ A_i\subseteq X \text{ 且}\{A_i\}\text{为单调升集合序列} \end{cases}.$$

为了解决互不相容的知识子集的合集的度量问题，文献[27]提出了 g_λ 模糊测量方法，该方法在上述测量函数 g 的基础上，还需要满足附加的 λ 规则[28]：

$$\begin{cases} g(A\cup B)=g(A)+g(B)+\lambda g(A)g(B) \\ A\cap B=\varnothing \text{且 } A,B\subseteq X \end{cases},$$

其中，$\lambda>-1$，它通过求解式(4)得到

$$\lambda+1=\prod_{i=1,N}(1+\lambda g_i) \tag{4}$$

其中，设 $X=\{x_1,\cdots,x_N\}$，$g_i(i=1,\cdots,N)$表示模糊密度，则有 $g_i=g_\lambda(\{x_i\})$．

模糊积分是一个非线性函数，它定义在模糊测量集合上．设 $h:X\to[0,1]$，则在 $A\subseteq X$ 上的模糊积分定义为

$$\int_A h(x)\circ g(\cdot)=\sup_{0\leq\alpha\leq 1}[\min(\alpha,g(A\cap a_\alpha))],$$

$$a_\alpha=\{x:h(x)\geq\alpha\} \tag{5}$$

其中：h 为知识的隶属函数，X 代表知识集合，则式(5)表示的模糊积分的计算就完成了知识质量评估的过程，从而找到置信度最高的知识作为正确知识．

基于模糊集理论的方法能够同时处理不精确和不确定的信息，有效实现开放网络知识的评估．然而，基于模糊集理论的知识评估方法需要凭经验设置知识的模糊规则和隶属函数，缺乏系统性，难以保证不同知识源类型的知识评估结果的稳定性和鲁棒性．因此，基于模糊集理论的方法难以有效处理网络大数据中多源异构的知识的评估问题．

2.4　基于图模型的知识评估方法

除上述传统的知识评估方法以外，近几年比较流行的知识评估方法是基于图模型的方法，这种方法使用从其他类型的数据中获得的先验知识，如使用已有知识库中的知识来拟合先验模型，从而为知识分配一个概率，可被看作是图上的链路预测问题，也就是说，我们观察一组现有的边(连接不同实体)，预测其他边存在的可能性，从而根据预测的边指导数据源中获取的知识的质量的评价．

Lao 等人[29]提出了一种基于路径排序算法(Path Ranking Algorithm，PRA)的知识先验计算方法，该方法通过利用已有的知识去预测这些知识之间的隐含信息．以图 1 为例，该方法将实体之间的关系抽象成一种路径模型，首先，枚举实体间所有的关系路径；然后，将每条路径作为训练专家，在关系路径图上执行随机游走，计算每条路径终点的概率值；最后，利用逻辑回归对所有训练专家排序．PRA 通过利用已有的知识之间的关系预测它们之间可能产生的隐含的知识，从而与数据源中抽取的知识进行比对，识别不同来源知识中可能的真值．

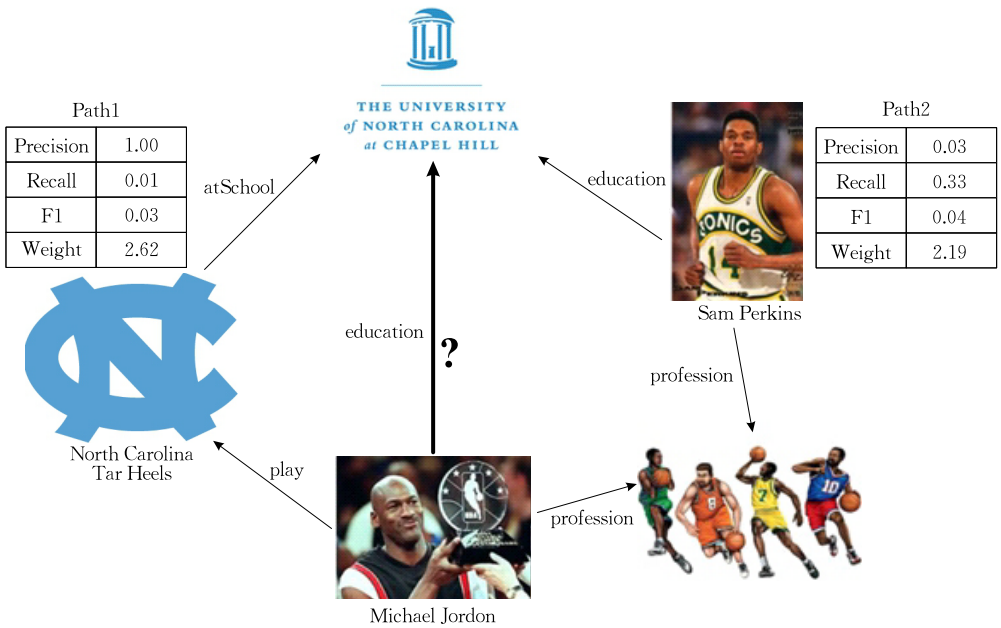


图 1　基于路径排序的知识先验计算示意图[29]

除此之外，Dong 等人[15]提出了一种基于神经网络模型的方法，该方法将上述链路预测问题转化为矩阵填充问题(matrix completion)进行求解. 具体地，将原始的知识库看作是一个稀疏表示的三维矩阵 $\boldsymbol{G}:E\times P\times E$，其中，$E$ 表示实体的个数，P 是谓词的个数，若(s,p,o)在知识库中存在，则 $\boldsymbol{G}(s,p,o)=1$，否则 $\boldsymbol{G}(s,p,o)=0$. 通过为每个实体和谓词关联一个低维潜在向量对这个矩阵进行低秩分解，然后计算元素的内积，计算方式如下：

$$\Pr(\boldsymbol{G}(s,p,o)=1)=\sigma\Big(\sum_{k=1}^{K}u_{sk}w_{pk}v_{ok}\Big),$$

其中：$\sigma(x)=\dfrac{1}{(1+e^{-x})}$是 sigmoid 函数或 logistic 函数；K 是向量的维度. 这里将元组(s,p,o)中的 s,p,o 映射到一个低维的语义空间，分别用一个 K-维的数值向量 $\boldsymbol{u}_s,\boldsymbol{w}_p,\boldsymbol{v}_0$ 表示.

基于上述表示形式，然后使用标准的多层感知机(Multi Layer Perceptron，MLP)捕获交互项，模型的表示形式如下：

$$\Pr(\boldsymbol{G}(s,p,o)=1)=\sigma(\boldsymbol{\beta}^{\mathrm{T}}f[\boldsymbol{A}[\boldsymbol{u}_s,\boldsymbol{w}_p,\boldsymbol{v}_o]]),$$

其中：$\boldsymbol{A}$ 是一个 $L\times 3K$ 维的矩阵，用来表示感知机第 1 层的权重，其中 $3K$ 的项是由 $\boldsymbol{u}_s,\boldsymbol{w}_p,\boldsymbol{v}_0$ 这 3 个 K-维向量产生的；$\boldsymbol{\beta}$ 是一个 $L\times 1$ 的向量，表示感知机第 2 层的权重. 通过该模型可以将语义相近或相关的知识归类到一起. 然后，利用这些先验知识从而完成知识质量的评估.

考虑到网络大数据中数据源的质量不尽相同并且不同知识抽取器抽取知识的正确性存在差异，Dong 等人[30]提出了一种区分数据源错误和知识抽取错误的方法，该方法使用多层概率模型进行联合推断，模型如图 2 所示.

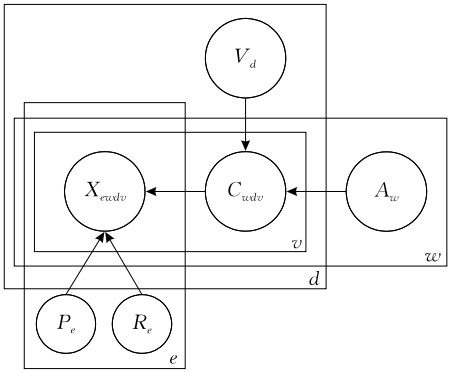


图 2 多层概率模型示意图[30]

在图 2 中，X_{ewdv} 是观察变量，表示知识抽取器 e 是否从数据源 w 中抽取到数据项 d 的值 v，即键值对(d,v)；C_{wdv} 和 V_d 是隐变量，其中 C_{wdv} 表示数据源 w 是否真正包含(d,v)，V_d 表示数据项 d 的真正取值；A_w 表示数据源 w 的准确性；P_e 表示抽取器 e 的准确率；R_e 表示抽取器 e 的召回率. 该方法通过使用上述多层概率模型实现数据源准确性和抽取到的数据值质量的联合推断.

Zhao 等人[31]提出了一种针对数据型数据真值发现的贝叶斯概率模型—高斯真值模型(Gaussian Truth Model，GTM)，该模型具有有原则地使用数值数据的特点，而且不需要任何监督信息就可以推断真值和数据源的质量. GTM 是一个生成式模型，该模型将数据源(S)质量、实体(E)的真值及其每个断言(C)的观察值综合起来建模它们之间的依赖关系，模型如图 3 所示.

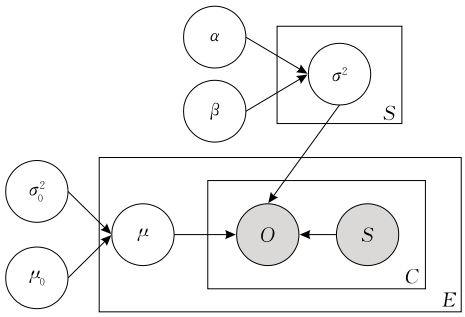


图 3 GTM 概率图模型[31]

在图 3 中，GTM 首先对每个数据源 s 采用基于超参数(α,β)的先验逆伽马分布生成它的质量 σ_s^2，其中 α 为形状参数，β 为尺度参数；对每个实体 e 采用均值为 μ_0、方差为 σ_0^2 的高斯分布生成它的真值 μ_e；对每个实体的断言 c，它是观察到的一个虚拟索引变量用于选择相应的源质量. 记 c 的来源为 s_c，GTM 利用均值为 μ_e、方差为 $\sigma_{s_c}^2$ 的高斯分布生成 c 的规范化值 o_c. 因此，在给定超参数时，观察值和未知参数的似然可写作如下形式：

$$p(o,\mu,\sigma^2\mid\mu_0,\sigma_0^2,\alpha,\beta)=$$

$$\prod_{s\in S}p(\sigma_s^2\mid\alpha,\beta)\times\prod_{e\in\varepsilon}\Big(p(\mu_e\mid\mu_0,\sigma_0^2)\prod_{c\in C_e}p(o_c\mid\mu_e,\sigma_e^2)\Big).$$

Zhao 等人将真值发现转换为计算最优的真值估计使得联合概率最大，即计算 μ 的最大后验估计(Maximum A Posteriori，MAP)：

$$\hat{\mu}_{\mathrm{MAP}}=\arg\max_{\mu}\int p(o,\mu,\sigma^2\mid\mu_0,\sigma_0^2,\alpha,\beta).$$

采用 EM 算法不断迭代求解实体 e 的最大后验估计 $\hat{\mu}_e$，根据最大后验估计预测其真值，从而实现其对应的真值的计算.

通过上述分析发现，基于图模型的方法借助外

部辅助信息保证知识评估的高准确率.然而,该方法在外部信息提供的闭环知识集合上难以扩展到网络大数据中抽取的所有知识质量的预测评估上,无法保证方法在面向网络大数据的知识评估的扩展性和适应性.

不论是传统的基于贝叶斯的方法、D-S 理论和模糊集理论的方法,还是近几年兴起的基于图模型的方法都是考虑了知识获取的不确定性,通过对获取的知识进行综合评估,计算知识的真值,在一定程度上降低了知识的不确定性,减少了错误的知识,提高了知识的可靠性和置信度,对提高知识库的实用性起到至关重要的作用.然而,来源于网络大数据的知识随着网络大数据的发展,具有动态演化特性,上一时刻正确的知识,下一时刻未必为真,而上一时刻未发生的知识,下一时刻可能就变成了现实,而现有的知识评估的方法缺乏对知识时间维度的考虑,主要是针对静态知识的评估,无法直接适用于随时间动态演化的知识的评估,缺乏针对动态知识的处理方法.而且,现有的方法在进行知识评估时,缺乏对数据源之间关系的分析,缺乏对知识获取渠道和获取方式的建模,因此难以从不可靠的知识获取方式中区分不可靠的数据源,这导致这些方法在处理网络大数据中大规模、多源异构、动态演化的知识的评估时面临准确率不高、鲁棒性较差等问题.

3　开放网络知识扩充方法

开放网络知识扩充建立在知识评估的基础上,其主要目标是将从网络大数据中获取的知识经过知识评估,验证为正确的知识更新到知识库中,与知识库中已有的知识进行关联计算与合并计算,从而实现知识的融合.以知识评估为前提和基础,本节着重探讨可用于将验证正确的知识扩充到知识库的方法.具体来讲,针对知识的组成要素,从实体扩充、关系扩充、分类扩充 3 个方面对开放网络知识扩充的相关工作进行介绍.图 4 总结了现有的知识扩充方法的分类.

3.1　实体扩充方法

实体扩充的主要目标是将从网络大数据的文本中获取的实体动态扩展到知识库中.从文本中获取的实体与知识库中的实体存在两种可能的关系:一种是知识库中存在与文本实体映射的实体(即相同实体或等价实体),对此类实体只需要找到文本实体在知识库中的映射实体,即实体链接(entity linking)[32];另

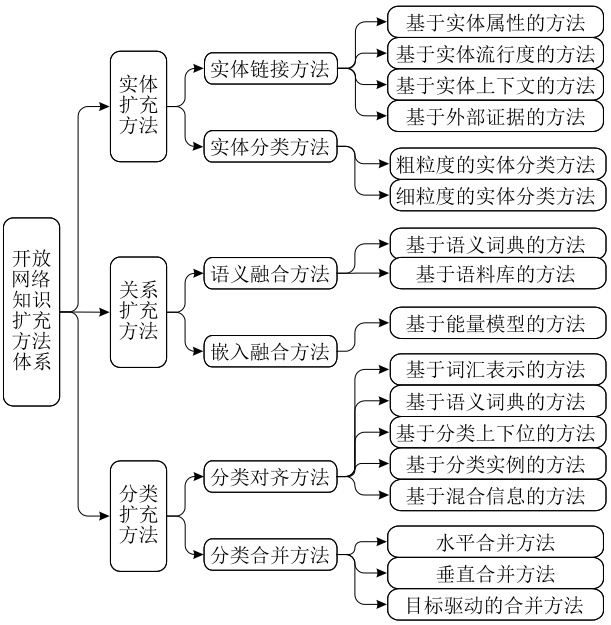


图 4　开放网络知识扩充方法

一种是知识库中不存在与文本实体映射的实体,在这种情况下,首先基于知识库中的分类为文本实体标注类别,即实体分类(entity classification)[33],然后根据分类将文本实体扩展到知识库对应的分类下,从而完成文本实体与知识库的关联合并.接下来,分别讨论实体链接和实体分类的研究现状.

3.1.1　实体链接方法

实体链接的主要作用是利用知识库中的实体对从网络大数据的文本文中获取的实体指代进行消歧,识别每个实体指代在知识库中与其对应的映射实体.这里实体指代是指实体的一种文本表示形式,一个实体可能有多种不同的表达,如全名、别名、缩写等,而一个实体指代也可能表示不同的实体.按照实体链接采用的信息不同,现有工作主要分为基于实体属性的实体链接方法、基于实体流行度的实体链接方法、基于上下文的实体链接方法和基于外部证据的实体链接方法.

(1)基于实体属性的实体链接方法

早期的实体链接方法是通过计算描述实体的属性的相似度判断实体是否相同.最直接的方式是基于实体的名字属性的字符串相似度的方法[34-35],除此之外,还有结合领域知识的计算字符串相似度的方法[36-37].这些工作主要通过编辑距离、Jaccard 系数等方式计算表示实体的名字的词语和描述实体其他属性的词语的相似度.但是基于字符串相似度的方法无法处理实体语义异构的情况,如“凤梨”和“菠萝”,它们表示同一个实体,但基于字符串相似度的

方法则将其判别为两个不同的实体.

因此,为了解决实体语义异构的问题,一些研究工作除了使用字符串相似度之外还引入了实体的语义特性,通过语义相似度来度量实体之间的相似度,典型的语义相似度计算方法是借助语义词典,通过计算两个词语在词典中的语义距离来计算它们之间的相似度,针对英文,主要借助 WordNet 实现[38-39],针对中文,则使用《知网》、《同义词词林》计算词语的语义相似度[40-41]. 典型的引入语义相似度的方法有 Chen 等人[42]提出的基于 WordNet 和模糊形式概念分析的方法. 这些方法通过词干还原,查找该词干在语义词典中的同义词集合表示和描述,通过词语在语义词典中的概念层次结构中的最短路径、同义词集合和描述计算语义相似度,将实体的字符串相似度和语义相似度进行加权平均作为实体相似度的度量.

基于实体属性的实体链接方法在属性信息丰富、没有噪音的情况下是有效的,但是从网络大数据中获取的属性难以保证完全没有噪声,而且描述实体的所有属性在度量实体相似性时所起的作用并不是完全相同的,可能存在某些属性比其他属性更典型的情况[43],因此单纯基于实体属性的方法难以满足网络大数据中实体链接对准确率的要求.

(2) 基于实体流行度的实体链接方法

基于实体流行度的方法本质上是一种基于概率统计的方法[44],它基于这样的假设:对于一个给定的实体指代,与其对应的映射实体最有可能是现实世界中最著名(most prominent)的实体. 例如,给定一个实体指代"李娜",人们可能最先想到的是著名的中国网球运动员"李娜". 通常实体的流行度是通过其出现在在线百科(如维基百科)锚文本中的频率度量的,计算方式如下:

$$P(e)=\frac{\#(e\text{ 出现的锚文本})}{\#(\text{在线百科中的锚文本})},$$

其中:e 表示给定的实体,$P(e)$表示实体 e 的流行度;$\#(\cdot)$表示数量.

Ratinov 等人[45]认为实体流行度这种简单的启发式规则是正确地进行实体链接的一个非常可靠的指标. 但显然,依靠这一特征进行实体链接的方法将会导致:不论实体出现的上下文是什么,它总会将所有具有相同实体指代的实体链接到知识库中一个固定的实体. 这种方式的缺陷在于没有考虑实体的歧义问题,鲁棒性较差.

(3) 基于实体上下文的实体链接方法

这种方法通过计算与实体相关的上下文的相似性来判断两个实体是否为同一实体. 它基于这样一种假设:如果两个实体的上下文相似,那么这两个实体就可能是同一个实体. 这种方法追溯到 Bunescu 和 Pasca[46]提出的利用在线百科对命名实体进行消歧的工作,他们定义了一种利用实体指代的上下文与所指候选实体的维基百科页面的内容之间的余弦相似度来度量实体相似度的方式. 这种方式选择上下文相似度得分最高的候选实体作为实体指代对应的映射实体.

除此之外,Ananthakrishna 等人[47]根据上下文中实体的共现关系,寻找与实体关联的实体集合,计算与实体关联的实体集合的相似度,通过这种启发式规则判断这两个实体是否是同一实体. 在此基础上,Bhattacharya 等人[48]提出了一种联合消解方法,不仅要求与实体相关的上下文相似,而且还要求这些与实体关联的实体必须是相同的实体,通过这种方式强化上述启发式规则.

基于实体上下文的方法可以弥补基于实体流行度方法的缺陷,但是上下文相似度方法要求两个被比较的文本之间存在词重叠,由于自然语言使用的灵活性,这会成为一个严格的约束条件. 不仅如此,在网络大数据中,实体的上下文可能出现稀疏或存在噪音,例如:描述实体的上下文可能包含不相关的信息,这些信息混淆实体链接任务的实现. 因此,在文本足够长并且文本相对清洁时,基于实体上下文的实体链接方法能够取得较好的准确率,但是文本稀疏或存在噪音的情况下则无法保障.

(4) 基于外部证据的实体链接方法

一种典型的基于外部证据的实体链接方法是采用"话题连贯性". Cucerzan[49]首先认识到使用话题连贯性可以有效提升实体链接的准确率,他采用基于候选实体和同一上下文中其他实体在维基百科中的分类和链接的重叠率计算实体之间的话题连贯性. Milne 等人[50]则在此基础上,通过使用规范化的谷歌距离[51]改进其对话题连贯性的定义,并且仅利用上下文中非歧义的实体计算话题连贯性. 除此之外,还存在很多其他的话题连贯性的度量方式:Bhattacharya 等人[52]利用一个实体和一个文档的潜在主题之间的联系建模话题连贯性. Sen[53]则利用实体之间的共现关系建模话题连贯性. Han 等人[54]根据同一文本的"话题连贯性"提出了一种联合推断的方法,该方法认为同一文本中的实体并不是独立的,它们之间存在语义相关性,而这种相关

性有助于提升实体链接的准确率.考虑到知识库包含的信息的有限性,Li 等人[44]基于话题提出了一个生成模型 MENED 和自动从文本中挖掘有用证据的增量算法,通过建模文本中包含的背景话题(background topic)和未知实体(unknown entities)从语料中收集额外的证据提升实体消歧准确率.

另一种典型的基于外部证据的方法是借助在线百科的结构信息,包括实体页面、重定向页面、消歧页面、分类页面和百科页面中的超链接等.Shen 等人[33]利用维基百科提取实体的不同表示形式,包括名称、别名、缩写、昵称等建立词典,然后通过查找词典的方式对实体链接完成决策.Cucerzan[55]则利用从维基百科提取的实体页面所属的分类信息作为主题、页面中包含的超链接锚文本作为概念,将描述实体的文档扩展表示为维基百科中包含的概念和主题空间的向量模型,然后利用新的文档表示模型与维基百科中的实体进行链接.

除了基于在线百科寻找实体链接的证据之外,Gottipati 等人[56]提出了一种基于查询扩展方式的实体链接方法,该方法主要是借助信息检索中的统计语言模型实现的:通过使用 KL 散度检索模型[57]和扩展查询语言模型建模实体上下文和 Web 中检索到的信息.Zhang 等人[58]提出采用缩写扩展的方式降低缩写形式的实体指代的歧义性的方法.Hoffart 等人[59]提出了利用 YAGO 知识库中实体的目录、类型、语义关系进行实体链接的 AIDA 方法.Thater 等人[60]则利用从大规模的句法分析语料库中获取的共现信息计算实体与候选实体之间的相似度,从而建立实体与候选实体间的加权关系图,然后通过寻找图中的稠密子图,获得实体链接结果.

Lee 等人[61]针对实体上下文缺乏的情况,引入了 CnD(Clean data has no Duplicates)原则和 BoF(Birds of a Feather)原则,从多种数据源中挖掘实体链接反对证据和支持证据.其中,反对证据从 Probase 的分类、Freebase 中实体类别和 Wikipedia 列表数据获得,主要依据是如果一个列表组织良好,那么它不可能包含任何重复的记录.支持证据则从实体的文本相似度、以及与该实体相关的 Wikipedia 页面的内部链接、重定向链接和非歧义页面获取,每一条支持证据都赋值一个[0,1]之间的权重,用于表示该支持证据的强度.

除此之外,针对网页中特定结构的实体链接也展开了一些研究,Limaye 等人[62]针对 Web 表格中的实体链接提出了一种新的概率图模型,该模型利用 YAGO 中的实体、关系和类型同时为表格中的每个单元格选择实体、为列选择类型、为列对选择关系.针对 Web 列表中的实体,Shen 等人[63]提出了 LIEGE 框架,该框架结合实体的先验概率和列表中实体类型的一致性将 Web 列表中的实体与知识库中的实体进行链接.

上述方法都是针对普通长文本中实体的链接问题,下面介绍在短文本方面的一些研究,如针对社交网络文本.为了克服社交网络文本短、上下文稀疏的问题,Guo 等人[64]提出利用额外的相似的微博文本丰富实体的上下文.Shen 等人[65]则通过建模用户兴趣将 Twitter 发布的 Tweets 包含的实体与知识库进行链接.Guo 等人[66]提出了一种组合结构化学习和一阶、二阶、上下文敏感的多种特征的实体链接方法.这些方法都是通过挖掘额外的特征丰富实体背景知识,这也导致特征挖掘的有效性直接关系实体链接的有效性.

表 1 汇总了实体扩充中的实体链接方法.

表 1　实体扩充中的实体链接方法分类汇总表

方法	主要特点	优点	不足
基于实体属性的方法	通过描述实体的属性的文本相似度和语义相似度度量实体是否为同一实体	在实体属性信息丰富时,实体链接的准确率高	没有考虑实体属性的区分度,并且当实体属性稀疏或存在噪音时方法无法有效工作
基于实体流行度的方法	通过利用在线百科锚文本统计的实体的流行度度量实体是否为同一实体	计算简单,时间复杂度低	将不同文本中获取的相同的实体指代判断为同一实体,没有考虑实体的歧义性,鲁棒性较差
基于实体上下文的方法	通过与实体相关的上下文的相似度度量实体是否为同一实体	引入实体的语义环境,能够弥补实体流行度方法的缺陷	上下文稀疏或存在噪音时,方法无法有效工作
基于外部证据的方法	通过挖掘与实体相关的证据(如话题、相关实体、概念等)并结合实体的上下文度量实体是否为同一实体	引入丰富的特征信息,方法的扩展能力得到有效提升	方法的有效性直接依赖于挖掘的外部证据的质量

3.1.2　实体分类方法

实体分类的主要目标是对从网络大数据的文本中获取的实体进行类别标注[67].按照分类标注粒度的不同,实体分类方法主要分为两类[68]:一类是粗粒度方法(coarse-grained);一类是细粒度方法(fine-grained).粗粒度的实体分类主要是将实体分

为人名、地名、机构名等类别，而细粒度的实体分类则根据本体或知识库包含的成千上万分类信息对实体进行更细致的类别标注.

(1)粗粒度的实体分类方法

在粗粒度的实体分类中，占主导地位的是有监督的方法，主要包括基于隐马尔可夫模型的方法[68]、基于决策树的方法[69]、基于最大熵模型的方法[70]、基于支持向量机的方法[71]和基于条件随机场的方法[72-73]等.其中，最著名的是斯坦福大学开发的 NER 工具[73]，它采用条件随机场模型，训练 CoNLL 规定的四分类模型和 MUC 会议规定的七分类模型对实体进行分类标注.国内比较著名的是中国科学院计算技术研究所开发的 ICTCLAS 工具[74]，ICTCLAS 采用层叠隐马尔可夫模型将汉语词法分析的所有环节都统一到一个完整的理论框架.然而，有监督的方法存在如下的缺陷：需要人工标注大量的语料以及人工定义实体分类的规则，费时费力.为了克服有监督方法需要人工标注语料的问题，近年来，也出现了一些基于半监督和无监督学习的实体分类方法.

半监督方法的主要思想是利用种子训练数据，通过自我学习不断标注新的样本数据，迭代改进分类方法的准确率.Collins 等人[75-76]提出了一种基于投票感知机的方法，首先通过解析一个完整的语料来搜索候选实体的模式；其次利用一个初始的拼写规则种子集，检查候选实体模式，根据满足的规则形式对候选模式进行分类并积累其出现的上下文；然后将出现频率最高的上下文变为上下文规则，通过上述方式积累的上下文规则发现更多的拼写规则[75]，通过这种方式同时学习不同类型的实体类别有助于发现实体分类的反对证据，防止过度迭代[76].Riloff 等人[77]提出基于 mutual bootstrapping 的方法，该方法以一些给定的某一类型的实体作为种子，然后在一个大规模的语料库中积累这些实体出现的所有模式，然后对这些模式进行排序以此发现新的模式.Cucchiarelli 等人[78]则在 Riloff 工作的基础上，利用句法关系在实体上下文中发现更准确的证据.Pasca 等人[79]也提出了一种基于 mutual bootstrapping 的实体分类方法.由于半监督学习的方法需要较少的人工介入，而精确率又较高，因此无论在理论上还是实践上都很有意义.

在无监督的方法中，实体分类在没有任何标注数据的条件下进行[80].Alfonseca 等人[81]利用从 WordNet 中获取的实体类别对实体进行标注，它首先利用一个语料集中频繁共现的词为 WordNet 中的每一个同义词集合分配一个主题签名，然后计算给定的实体的上下文与主题签名之间的相似度，利用相似度最高的签名对实体进行分类标注.Evans[82]则利用 Hearst Patterns[83]对开放文本中的实体进行类别标注.Etzion[84]则引入点互信息(Pointwise Mutual Information，PMI)特征判断一个命名实体是否可以为其分配一个给定的类型.虽然使用无监督的方法进行实体分类的研究取得了一定的进展，但无监督方法的准确率还无法满足人们的要求，是一个正在研究中的问题.

(2)细粒度的实体分类方法

在细粒度的实体分类中，一种传统的细粒度实体分类方法是有监督的基于分类模型的方法，通过提取一些语言特征，如词、词性和实体上下文等训练分类器，然后利用分类器判断实体的分类.典型的工作参见文献[85-87]，这些工作考虑包含上百个类别的分类体系的实体分类问题.由于一个知识库包含成千上万的分类，实体的类别不仅仅是人名、地名、机构名等.因此这种方法已经无法直接适用于细粒度的实体分类，原因在于该方法要求的训练数据规模很大，构造训练数据需要大量的时间和人力.

因此，半监督的方法更适合于细粒度的实体分类.Cimiano 等人[88]提出了基于 Harris 分布假设和向量空间模型的方法，它利用本体中与实体上下文相似度最高的分类标注实体.Tanev 等人[89]则提出利用实体上下文的词法-句法信息为本体中的每一个分类自动学习一个特征向量的方法，该方法假设测试数据中的实体是非歧义的，然而，这种假设在实际的数据集中并不总是为真.Ganti 等人[90]提出了一种基于多上下文的分类方法，首先从语料中抽取实体出现的所有上下文，其次从这些聚合的上下文中提取 n-gram 特征和 list-membership 特征，然后基于这些特征为实体进行分类标注.然而，该方法假设实体所在语料中都已表示成统一规范的形式，在实际的应用中，通常文本语料无法满足该假设.

Giuliano 等人[91]提出了一种基于实例的细粒度实体分类方法，该方法采用词汇替代技术，首先利用训练语料中出现的句子替换要进行分类的实体所在的句子，然后利用 Web 数据来估计生成的新的句子的合理性.与此类似，Giuliano[92]还提出了一种基于核函数的细粒度实体分类方法，该方法利用实体出现的所有上下文建模实体，将实体映射到从维基百科获得的隐含语义空间中.然而，为了对每个实体收集足够的上下文信息，该方法借助搜索引擎的查

询结果扩展实体的上下文，并没有考虑实体的歧义性，所以在对歧义的实体进行分类时，该方法的效果相对较差.

Nakashole 等人[93]提出了一种基于语义规则库的分类方法 PEARL，该方法首先基于 PATTY[94]定义的语义类型规则来匹配实体出现的文本，并利用语义规则定义的类别来标注实体类别，然后基于概率模型将实体分类问题转化为整数线性规划问题求解. 其中，PATTY 是基于频繁项挖掘技术，利用词法和句法分析从 Wikipedia 中生成的层次化的语义规则库. 实验结果表明该方法优于当前的许多工作，但该方法受限于 PATTY 制定的规则，当规则失效时，该方法则无法对实体进行分类. Yosef 等人[95]提出了一种面向领域的细粒度的实体分类方法，利用实体周围邻近的词、bi-grams、词性以及从知识库中获取的地名短语等解决分层结构的多标签分类问题[96]，为一个实体分配多个分类标签.

与上述多标签分类方法不同，Shen 等人[33]提出了一种基于图模型的细粒度实体分类方法 APOLLO，该方法通过实体的上下文建立文本实体与知识库实体之间的关系图，在创建的关系图上利用基于随机游走的标签传播算法获得文本实体的分类. APOLLO 主要基于文本实体与知识库实体上下文中包含的维基概念，创建文本实体与知识库实体之间的关系图，通过利用知识库中实体的分类信息采用吸收算法(Adsorption algorithm)[97]实现文本实体分类的预测. 图 5 展示了 APOLLO 利用维基概念创建的关系图实例.

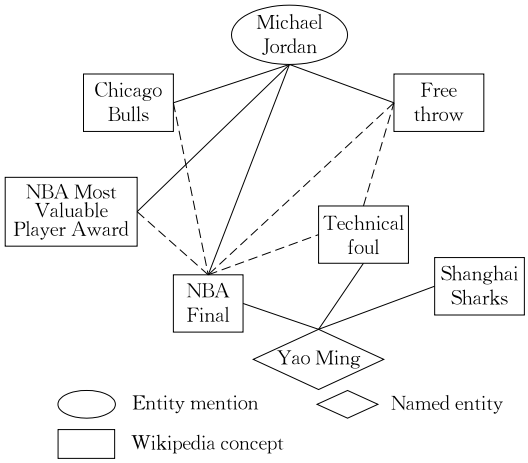


图 5　基于维基概念的实体关系图[33]

APOLLO 利用知识库中包含的信息作为标注数据，不需要额外的人工标注的训练数据即可自动完成实体的分类标注. 然而，该方法需要借助实体上下文中包含的维基概念建立与知识库实体之间的关系，但是在实际的应用中，一个描述实体的文本(尤其是短文本)中可能并不包含维基概念，在这种情况下，该方法将面临失效的问题.

表 2 汇总了实体扩充中的实体分类方法.

表 2　实体扩充中的实体分类方法汇总表

方法	主要特点	学习方式	代表工作	优点	不足
粗粒度实体分类方法	实体类别标签粒度相对较粗，主要将实体分为人名、地名、机构名等类别	有监督	SVM[71]、NER[73]、ICTCLAS[74]等	可充分利用分类分布的先验知识，控制训练样本的选择，获得较高的准确率	人为主观因素较强，训练样本的选取和评估需要花费较多的人力、时间
		半监督	基于投票感知机的方法[75-76]、基于 mutal bootstrapping 方法[77]等	能够自动在无标注的样例的帮助下训练有类别标签的样本，弥补训练样本不足的缺陷	采用的样本数据都是无噪声干扰的，但在实际中难以得到纯样本数据，导致分类准确率降低
		无监督	基于词典的方法[81]、基于 Hearst Pattern 的方法[82]、基于 PMI 的方法[84]等	不需要人工标注训练数据，人为误差的机会减少，需输入的初始参数较少	对结果需要进行大量的分析和后处理，分类的准确率无法保证满足用户需求
细粒度实体分类方法	实体类别标签粒度相对细致，将实体可能分为成千上万个类别	有监督	基于分类模型(SVM、CRF 等)的方法[85-87]	控制训练样本的选择，并可通过反复检验训练样本，提高分类的准确率	标注具有成千上万类别的训练样本需要花费大量的人力、时间
		半监督	基于向量空间模型的方法[88]、基于词法-句法的方法[89]、基于多上下文的方法[90]、基于核函数的方法[92]等	自动对未标记数据加以利用、学习整个数据分布上具有较强泛化能力的模型	要求参与分类的实体没有歧义性或实体在语料中都表示为统一规范的形式，这在实际应用中难以得到满足
		无监督	基于规则库的方法[93]、基于领域划分的方法[95]、基于图模型的方法[33]等	无需对实体类别划分有较多了解，输入的初始参数较少，所分的类别比监督的方法更均质	对实体的分类结果仍需大量的分析和后处理；与有监督学习相比，实体分类的时间复杂度较高

从以上对于实体链接和实体分类方法的分析可知，实体链接和实体分类都是实体扩充的重要组成部分，两者缺一不可. 然而，通过对相关工作的分析发现，当前缺乏有效地面向知识库的统一的实体扩充框架，现有工作大多数独立解决实体扩充两个子问题中的一个；除此之外，现有的工作主要利用基于统计的特征(实体流行度)和词汇的特征(上下文相似度、话题连贯性)，这些特征对流行的实体相对丰

富但对长尾的实体却表现稀疏，导致对流行实体和长尾实体扩充的准确率存在明显的偏差.

总体上讲，现有的实体扩充方法仍面临以下两个限制：一是这些方法不适合文本实体的背景知识(如上下文)稀疏的情况；二是这些方法不适用于文本中包含的实体之间独立性假设不成立的情况. 不仅如此，现有的实体扩充方法采用的统计特征和词汇特征的方式对长尾实体表现稀疏，导致长尾实体扩充的准确率不高，并且还面临鲁棒性和可扩展性的挑战.

3.2 关系扩充方法

在上一节，我们介绍了实体扩充的工作，而关系作为实体之间的一种逻辑联系，在网络大数据中，每时每刻随着活动、时间、场景等发生变化，会产生很多刻画实体的关系，因此，如何将这些动态产生的关系扩展到已有的知识库中，对提高知识库的时新性、覆盖能力至关重要. 由于自然语言表达的随意性，关系存在大量同义和多义的表达，这给关系的扩充带来巨大的挑战. 在本节，将着重介绍近年来在关系扩充方面取得的研究成果.

关系扩充的主要目标是将从网络大数据的文本中获取的实体关系动态扩展到知识库中. 从文本中获取的实体关系与知识库中的实体关系存在两种可能的情况：一种是知识库中存在与文本实体关系映射的实体关系，即相同或等价的实体关系，对此只需要找到文本实体关系在知识库中与之对应的实体关系；另一种是知识库中不存在与文本实体关系映射的实体关系，在这种情况下，则需要将实体关系扩展合并到知识库中，从而完成文本实体关系与知识库实体关系的关联合并.

通过对从网络大数据的文本中获取的实体关系与知识库中的实体关系的映射分析可以看出，关系扩充建立在第 3.1 节介绍的实体扩充的基础上，首先需要基于实体扩充中的实体链接建立文本实体关系关联的实体与知识库中实体之间的映射关系，若知识库中存在与文本实体关系关联的实体对应的映射实体，则需要判断映射实体之间是否存在与文本实体关系相同或等价的关系，若存在则建立它们之间的映射关系，若不存在则将该文本实体关系扩充合并到知识库中；若知识库中不存在与文本实体关系关联的实体对应的映射实体，则需要基于实体扩充中的实体分类对文本实体关系关联的实体进行分类，然后根据分类将文本实体关系关联的实体扩展到知识库对应的分类下，并建立它们之间的关系，完成文本实体关系与知识库的关联合并.

因此，通过上述分析可以看出，基于实体扩充，关系扩充的关键在于判定两个描述实体的关系是否表达同一种关系，是否是包含关系等. 针对这一问题，现有的工作主要分为两种：一种是传统的基于语义的方法，对描述实体的关系进行语义理解；另一种是近几年流行的基于嵌入学习的方法，将实体关系进行结构映射.

3.2.1 基于语义的关系扩充方法

基于语义的方法是一种通过比对描述关系的词汇之间的语义相似度来验证是否是相同关系和包含关系. 为了计算词汇之间的语义相似度，现有的方法主要分为两种：一种是基于语义词典的方法，利用词汇在词典中的距离度量语义相似度；另一种是基于语料库的方法，利用词汇在语料库中的词或是 n-grams 的分布度量语义相似度.

(1) 基于语义词典的关系扩充方法

典型的基于语义词典的方法是基于 WordNet 的方法：利用 WordNet 中定义的 isA 关系. 最直接的基于 WordNet 计算语义相似度的方式是在 WordNet 分类体系图中寻找连接两个词汇的最短路径[98]，这种方式虽然简单，但是准确率较低，原因在于它认为分类体系中的所有边都是等距离无差别的，而且这种方法没有考虑隐藏在分类节点背后的信息量(information content). 更进一步的方法则是利用两个词汇与分类体系结构相关的信息内容度量语义相似度. 最早的工作是 Resnik[99] 提出的基于两个词汇在分类树中最小公共祖先节点的信息度量相似度的方法，在这种方法中为了计算每个词汇的信息量，需要在一个大规模的文本语料中获取词的共现信息，该方法的局限性在于不管词汇的信息量是什么，在分类体系中相同概念下的所有孩子的相似度是一样的.

为了克服上述局限性，Banerjee 等人[100] 提出了一种基于 Lesk 算法[101] 的语义相似度计算方法. Patwardhan 等人[102] 提出了一种融合 WordNet 结构和内容以及词汇所在原始文本的共现信息的语义相似度计算方法. Sánchez 等人[103] 提出了利用词汇在 WordNet 中所有后代包含的叶子节点的数量和 WordNet 根节点的所有后代包含的叶子节点的数量的比值计算词汇的信息量. Li 等人[104] 提出了一种基于 Probase[14] 中概率化的 isA 分类体系的方法，该方法通过将词汇映射到 Probase 的概念空间，在概念空间中度量相似度，这种方式相比 WordNet

覆盖更多的上位-下位关系. 刘群等人[40]提出了一种基于《知网》的语义相似度计算方法，利用《知网》定义的独立义原度量词语的语义相似度. 王斌[41]则提出了基于《同义词词林》的语义相似度计算方法.

除此之外，一些研究也尝试利用图学习的算法计算语义相似度. Alvarez 等人[105]首先利用从 WordNet 中获得的词汇的上位词、其他关系和描述性注解等信息构造带权图模型 Gsim，然后基于随机游走算法选择与两个词汇相关的距离最近的两个上位词计算相似度得分. 随后 Agirre 等人[106-107]提出了一种基于 WordNet 的个性化 PageRank 算法的语义相似度计算方法，该方法首先计算每个词汇对应的 PageRank，然后将其累积到每个 synset 的概率分布中，最后利用两个概率分布之间的余弦相似度度量两个词汇的语义相似度.

基于语义词典的方法简单、直接. 然而，比较流行的词典如 WordNet、《同义词词林》等无法提供足够的词语覆盖度，原因在于这些词典大多数基于人工方式构建，它来不及收录网络大数据中每天产生的新词和新义，从而导致词典的覆盖面有限. 因此，当词典中出现词语缺失时则无法有效工作.

(2) 基于语料库的关系扩充方法

基于语料库的方法通过从大规模语料库中抽取词汇的上下文，然后归纳上下文中词或 n-gram 的分布性质. 其中，语料库可以是 Web 页面、Web 搜索片段和其他文本库. Chen 等人[108]提出了一种 double-checking 模型，利用 Web 搜索引擎返回的文本片段计算词之间的语义相似度，该方法利用词汇在其搜索文本片段中的出现次数评估语义相似度. Bollegala 等人[109]提出了一种基于搜索引擎检索到的页面计数和文本片段的新的相似度度量方法. Radinsky 等人[110]提出了基于时间的语义分析模型，该模型获取语料库的时间信息，使用更精确的表示：每个概念不再是标量，而是表示为在时间上有序的文档语料库. 这种方法可以提高皮尔逊相关系数，但是它需要大量的历史数据，从而导致在文本处理更费时间. Mikolov 等人[111-112]提出了基于连续词袋模型和 Skip-gram 模型的 word2vec 方法，将词表征为实数值向量，其利用语料库进行训练，把词转换为 k-维向量空间中的向量运算，通过计算向量空间上的相似度来表示词汇语义上的相似度.

通过实验表明，基于语料库的方法可以有效改进语义相似度度量的准确性. 然而，该方法也面临一些局限性：首先，这种度量方式存在偏差，这是由搜索引擎使用的索引和排序机制导致的. 其次，有些搜索结果导向性的相似度方法需要与搜索引擎进行交互，这导致通信开销和索引成本大大增加从而无法适用于在线应用. 再次，基于上下文的词或 n-grams 分布统计的方法忽略了如下事实：(1) 语义单元不仅可以是词或 n-grams，也可以是通常意义上的多词短语形式；(2) 许多词或短语的含义是模糊的，可以有多种解释.

表 3 汇总了关系扩充中的基于语义的关系扩充方法.

表 3　关系扩充中的基于语义的方法分类汇总表

方法	主要特点	优点	缺点
基于语义词典的方法	利用表达实体关系的词汇在语义词典中的距离度量实体关系之间的相似度	关系之间的相似度计算简单、直接；准确率较高	词典大多采用人工方式构建，难以及时收录网络大数据中产生的新词、新义，无法避免词语缺失的问题
基于语料库的方法	利用表达实体关系的词汇在语料库中的词的分布或是 n-grams 的分布度量实体关系之间的语义相似度	借助搜索引擎等方式比较容易获得大规模语料库，利用丰富的语料库提升关系之间相似度判断的准确率	采用与搜索引擎交互的方式，增加了关系之间相似度计算的时间开销；利用查询方式获取语料库时无法有效处理关系的歧义性

通过对相关工作的分析可以看出，计算描述关系的词汇之间的语义相似度进行关系扩充是一种比较准确、有效的方式，但是这种方法过度依赖外部语义词典或语料库，当词典中词语缺失，语料库稀疏时则无法有效工作.

3.2.2　基于嵌入学习的关系扩充方法

基于嵌入(Embedding)学习的方法[113]是一种基于能量模型的方法，这种方法通过在嵌入空间中寻找一个恰当的能量函数学习实体的嵌入表示，然后利用实体的嵌入表示表达实体关系，进而判断两个描述实体的关系是否表达同一种关系，从而实现实体关系的扩充.

给定一个实体关系，这里用一个简化的三元组表示：(h,r,t)，其中 h 和 t 分别表示关系的头部实体和尾部实体，r 表示实体 h 和 t 之间的一种关联关系. 嵌入方法需要将实体 h 和 t 映射到一个语义空间，学习其在该空间的向量表示 $\boldsymbol{h}$ 和 $\boldsymbol{t}$，并通过打分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$ 度量 (h,r,t) 在嵌入空间中的合理性. 而 r 在嵌入空间的表示 $\boldsymbol{r}$ 则通过 $\boldsymbol{h}$ 和 $\boldsymbol{t}$ 表达.

典型的基于嵌入学习的工作是 Bordes 等人[113]

提出的 TransE 模型，该模型将实体映射到一个低维的嵌入空间，利用实体在嵌入空间中的向量表示刻画实体的关系. TransE 的基本思想是将实体间的关系转化为两个实体在嵌入空间中的一个翻译，即当(h,r,t)成立时，在嵌入空间中则存在 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$ 成立，这也表明在嵌入空间中 $\boldsymbol{t}$ 应该是 $\boldsymbol{h}+\boldsymbol{r}$ 最近的邻居. TransE 的基本思想如图 6 所示.

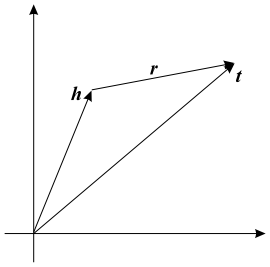


图 6 TransE 模型[113]

在 TransE 模型中，打分函数定义如下：$f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2^2$，如果 $f_r(\boldsymbol{h},\boldsymbol{t})$值较小，则说明关系$(h,r,t)$为真，否则为假. TransE 适用于 1－1 实体关系但对 N－1、1－N 和 N－N 的关系则存在问题. 以一个 1－N 关系为例：$\forall i\in\{0,\cdots,m\},(h_i,r,t)\in S$ 如果通过 TransE 映射，所有的元组都成立，这也就表明 $h_0=\cdots=h_m$，这与事实不符.

为了解决 TransE 在处理 N－1、1－N 和 N－N 关系存在的问题，Wang 等人[114]提出了 TransH 模型，TransH 保证使一个实体在涉及不同的关系时有不同的分布表示. 对于关系 r，TransH 将其建模为超平面上的向量 $\boldsymbol{r}$，其中超平面的法向量为 $\boldsymbol{w}_r$. TransH 的基本思想如图 7 所示.

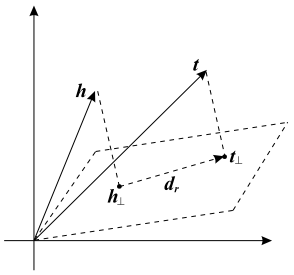


图 7 TransH 模型[114]

在 TransH 模型中，对于每一个(h,r,t)，将其对应的嵌入形式 $\boldsymbol{h}$ 和 $\boldsymbol{t}$ 利用法向量 $\boldsymbol{w}_r$向关系超平面进行投影得到 $\boldsymbol{h}_\perp$ 和 $\boldsymbol{t}_\perp$. 打分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$定义为 $f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}_\perp+\boldsymbol{r}-\boldsymbol{t}_\perp\|$，如果限制 $\|\boldsymbol{w}_r\|_2=1$，则推出 $\boldsymbol{h}_\perp=\boldsymbol{h}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r$，$\boldsymbol{t}_\perp=\boldsymbol{t}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{t}\boldsymbol{w}_r$. 通过将实体嵌入到关系超平面的方式，可以表达实体在各种关系中扮演不同的角色.

TransE 和 TransH 都是将实体和关系嵌入到同一个空间，但关系和实体是完全不同的对象，这很难用一个共同的语义空间表示它们. 虽然 TransH 利用关系超平面增强了模型的灵活性，但是它并没有完全打破这个假设的限制. 为了解决这个问题，Lin 等人[115]提出了 TransR 模型，该模型利用两个不同的语义空间建模实体和关系，即实体空间和关系空间，并在对应的关系空间中执行翻译. TransR 的基本思想如图 8 所示.

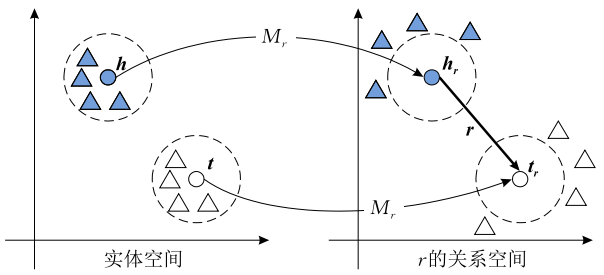


图 8 TransR 模型[115]

在 TransR 模型中，对于每一个(h,r,t)，首先将 h 和 t 嵌入到 k-维空间 $\boldsymbol{h},\boldsymbol{t}\in\mathbb{R}^k$，将 r 嵌入到 d-维空间 $\boldsymbol{r}\in\mathbb{R}^d$；对每一个关系 r 设置一个投影矩阵 $\boldsymbol{M}_r\in\mathbb{R}^{k\times d}$，然后在 $\boldsymbol{M}_r$的作用下将实体空间的实体表示映射到关系 r 对应的关系空间，映射结果分别为 $\boldsymbol{h}_r=\boldsymbol{h}\boldsymbol{M}_r$和 $\boldsymbol{t}_r=\boldsymbol{t}\boldsymbol{M}_r$，从而得到 $\boldsymbol{h}_r+\boldsymbol{r}\approx\boldsymbol{t}_r$. 打分函数定义为 $f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}_r+\boldsymbol{r}-\boldsymbol{t}_r\|_2^2$. 这种关系特定的映射可以保证具有相同关系的头部/尾部实体(深色圆圈)在嵌入空间中彼此接近，而没有关系的实体(深色三角形)在嵌入空间中则离的较远.

TransE 和 TransH 以及 TransR 为每一个关系学习一个唯一的向量表示，这种方式可能无法适合这个关系下的所有实体，因为这些关系通常是多元化的. 为了更好地建模这些关系，Lin 等人[115]在 TransR 模型的基础上提出了 CTransR 模型，该模型首先将输入实例分成若干组，即对一个特定的关系 r，将训练数据中所有与之相关的实体对(h,t)分为多个组；然后，对每一组学习一个关系向量 $\boldsymbol{r}_c$，对每一个关系学习投影矩阵 $\boldsymbol{M}_r$，定义投影实体 $\boldsymbol{h}_{r,c}=\boldsymbol{h}\boldsymbol{M}_r$和 $\boldsymbol{t}_{r,c}=\boldsymbol{t}\boldsymbol{M}_r$. 打分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$定义为

$$f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}_{r,c}+\boldsymbol{r}_c-\boldsymbol{t}_{r,c}\|_2^2+\alpha\|\boldsymbol{r}_c-\boldsymbol{r}\|_2^2,$$

其中：$\|\boldsymbol{r}_c-\boldsymbol{r}\|_2^2$的目标是确保每一个组的 $\boldsymbol{r}_c$不能与原始关系 $\boldsymbol{r}$ 离的太远；α 用于控制这一约束.

上述模型都是通过将关系 r 看作是从头部实体 h 到尾部实体 t 的翻译完成嵌入表示，然而，这些模型无法有效处理反射关系、1－N、N－1 和 N－N 关系，并且扩展性较差、学习效率较低. 为了解决上

述问题，Feng 等人[116]提出了 TransF 模型，该模型认为头部实体 h 和尾部实体 t 之间的翻译满足弹性大小，如果(h,r,t)成立，与 TransE 定义的 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$ 的假定不同，TransF 的定义方式为 $\boldsymbol{h}+\boldsymbol{r}\approx\alpha\boldsymbol{t}$，$\alpha>0$. 也就是说，TransF 只需要保证向量 $\boldsymbol{h}+\boldsymbol{r}$ 与 $\boldsymbol{t}$ 的方向，并不考虑向量本身的大小.

除此之外，Fan 等人[117]针对 TransE 无法灵活处理关系元组的多种映射性质的问题，提出了 TransM 模型，TransM 利用知识库的结构，根据每个关系的映射性质，预先计算训练数据中每个关系元组的权重. 对于每一个(h,r,t)，TransM 定义其在嵌入空间中的约束为 $f_r(\boldsymbol{h},\boldsymbol{t})=w_r\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{L_1/L_2}$. w_r表示关系 r 映射程度的权重，计算方式如下：

$$w_r=\frac{1}{\log(h_rpt_r+t_rph_r)},$$

其中，h_rpt_r和t_rph_r分别表示对关系 r 来说每个尾部实体平均关联的头部实体数以及每个头部实体平均关联的尾部实体数.

除了上述翻译模型以外，还有一些其它基于能量模型的方法，这些方法为知识库中的元组分配一个较低的能量并采用神经网络的方法学习，如语义匹配能量模型 SME[118]、非结构化模型 UM[118-119]、结构化嵌入模型 SE[120]、潜在因素模型 LFM[121-122]、张量神经网络模型 NTN[123-124]以及基于矩阵分解的模型 RESCAL[125]等.

表 4 汇总了关系扩充中的典型的基于嵌入学习的关系扩充方法.

表 4　关系扩充中的典型的基于嵌入学习方法的汇总表

方法	主要特点	优点	缺点
TransE	将实体映射到一个低维的嵌入空间，利用实体在嵌入空间中的向量表示刻画实体的关系	模型简单有效，能够直接建立实体和关系之间的复杂语义联系，模型参数较少，计算复杂度低	无法有效处理 1－N、N－1 和 N－N 复杂关系类型
TransH	在 TransE 的基础上增加关系超平面，提出让一个实体在不同的关系下拥有不同的表示	通过对实体的多样性表示学习，提升了处理 1－N、N－1 和 N－N 复杂关系类型的表示学习能力	与 TransE 一样，都假设实体和关系处于相同的语义空间，在一定程度上限制了模型的表示能力
TransR	针对不同的关系建立不同的语义空间，将实体投影到对应的关系空间，在关系空间中建立实体之间的翻译关系	通过利用不同语义空间的表示学习方式，较 TransE 和 TransH 有显著改进	较 TransE 和 TransH，模型参数急剧增加，计算复杂度大大提高；关系关联的头、尾实体共享相同的投影矩阵，没有考虑头、尾实体的类型的差异
CTransR	在 TransR 的基础上，引入聚类思想，对一个关系进行更细粒度的划分	通过将关系细分为多个子关系，实现更精确地建立实体和关系之间的投影	相比 TransR，模型的计算复杂度更高
TransF	在 TransE 的基础上，引入实体在嵌入空间中的弹性大小表示学习方式	只考虑实体和关系向量表示的方向，不考虑大小，简单、灵活，学习效率较高，计算复杂度低	与 TransE 和 TransH 一样，都采用一个相同的空间表示实体、关系，在一定程度上限制了模型的表达能力
TransM	在 TransE 的基础上，引入知识的结构信息，考虑关系的映射性质，加入关系的权重信息	通过分析关系的映射性质，引入关系的权重信息，提升具有多映射的关系表示学习问题处理的灵活性	与 TransE 和 TransH 一样，都采用一个相同的空间表示实体、关系，在一定程度上限制了模型的表达能力

这些模型仅利用实体关系本身的信息，将实体、关系映射到一个低维空间中，自动编码学习关系在该空间的特征表示，基于在低维空间学习到的向量表示度量实体关系之间的相似性，结合实体扩充方法即可实现实体关系的扩充. 基于嵌入学习的关系扩充方法可以有效解决传统的基于语义方法面临的数据稀疏问题，使关系扩充的性能得到显著提升，能够适用于从网络大数据的文本中获取的实体关系的扩充问题. 然而，现有的嵌入学习方法主要是考虑实体描述的关系的表示学习，对描述实体关系的上下文、层次类型、别名现象、时间信息等缺乏有效的表示建模，除此之外，这些方法都是针对单一关系采用面向实体表示的学习方式，无法对关系之间的语义关系进行有效表示学习.

3.3　分类扩充方法

在第 3.1 节和第 3.2 节分别介绍了构成知识库的点（实体）和边（关系）的扩充方法. 在这一节，将从知识库的结构角度介绍知识库的分类扩充方法. 分类体系作为知识库的骨架结构，是一个在知识库中用于语义分类或标注知识项的分类的集合. 由于不同的知识库可能会含有重叠或互补的数据，已经有越来越多方法开始尝试通过匹配不同知识库中的公共元素来将它们进行关联合并.

分类扩充的主要目标是将描述知识的两个分类体系进行集成，实现知识的复用和共享，其主要包含两个部分：一是分类对齐，在不同知识库中发现匹配分类体系中共同的元素；二是分类合并，根据分类体系对齐的结果，将描述知识的两个分类体系进行集成，从而完成两个分类体系的合并.

3.3.1　分类对齐方法

分类对齐的主要目标是在不同分类体系中发现对齐分类体系中共同的元素. 分类对齐的根源来自

于实体链接、重复检测或共指消解问题.此外,该问题也与模式匹配问题类似.根据 Rahm 和 Shvaiko 等人[126-127]对已有的模式匹配工作的研究分析,现有的模式匹配工作主要分为 3 类:基于相似度的方法、基于统计的方法和混合方法,这些工作的目标是试图为多个数据源建立一个共同的模式或找到不同模式之间内在的关联方式.

尽管模式匹配问题与分类对齐问题类似,但是与模式匹配问题相比,分类对齐有着自己独特的特点:(1)与数据模式相比,分类体系在定义数据时提供更高的灵活性和更明确的语义信息;(2)数据模式通常是为特定数据库定义的,而分类体系本质上是可重用共享的;(3)在分类体系中,知识表示的基本元素的数量更大、更复杂,如传递性,分类不相交性和类型检查约束等.因此模式匹配的方法无法直接适用于分类对齐问题.

除此之外,分类对齐问题特别是在本体匹配的概念下已被广泛研究.Choi 和 Shvaiko 等人[128-129]对已有的分类对齐工作进行了研究分析,现有的工作根据使用策略的不同主要分为以下 5 类:

(1)利用分类在分类体系中的指代形式(词汇表示)的策略.这种策略主要基于编辑距离或 Jaccard 系数计算分类名称之间的文本相似度判断分类之间的等价关系.这种策略简单、直接.然而,这种策略完全取决于分类的词汇表示,无法区分分类同义和多义的情况.

(2)利用语义词典(如 WordNet)的策略.这种策略主要利用语义词典的信息丰富分类体系中分类的背景信息,如 Chen 等人[42]利用 WordNet 的 Synset 信息扩展分类的信息,提出了一种结合模糊理论与形式概念分析的分类对齐方法 FFCA.这种策略受限于词典的覆盖率,当词典中词语缺失时则无法有效工作.

(3)利用分类在分类体系中的上下位关系的策略.这种策略利用分类在分类体系中的近邻结构计算两个分类之间的等价关系,如 Li 等人[130]提出了一种动态多策略框架 RiMOM,该框架基于两个评估因素:词汇相似度和结构相似度,自动选择分类对齐使用的策略.RiMOM 通过在两个分类体系关系图上采用 Similarity Flooding 技术,提高结构信息对分类对齐的影响.这种策略适用于分类结构相似程度高的分类体系之间的匹配.

(4)利用分类下包含的实例信息的策略.这种策略利用分类包含的实例的重叠率计算分类之间的等价关系.Suchanek 等人[131]提出了基于实例的概率化方法 PARIS,该方法通过不同的修剪启发式规则的稀疏表示(特别是,在每一步保持分类体系中每个元素的最大分配)来处理维护所有分类对齐的可扩展性问题.Demidova 等人[132]采用基于实例的方法将 YAGO 和 Freebase 对应的分类体系进行关联合并,形成新的 YAGO+F 知识库,丰富 Freebase 的语义信息.这种策略适用于分类下实例丰富且重叠率高的分类体系之间的匹配.

(5)利用上述信息的组合形式的混合策略.如 Ba 等人[133]利用词汇和实例信息计算分类之间的相似度,基于本体服务器设计开发了一个面向生物医学领域本体匹配的系统 ServOMap.Jiménez-Ruiz 等人[134]结合词汇相似度、语义相似度和结构相似度计算分类体系中共同的元素,设计开发了 LogMap 系统.Lacoste-Julien 等人[135]针对大规模分类体系提出了一种基于贪心的分类对齐方法 SiGMa,该方法组合词汇、属性和结构信息以贪婪的局部搜索方式的发现匹配的分类.基于贪心的方法对处理大规模的分类对齐任务来说可能是一种有效的方法.然而,由于其贪心的性质,它在决策时无法修正之前的错误.因此,基于贪心的方法不能保证为两个分类体系获得全局最优的匹配.

图 9 展示了汽车领域的两个不同的分类体系的对齐结果.

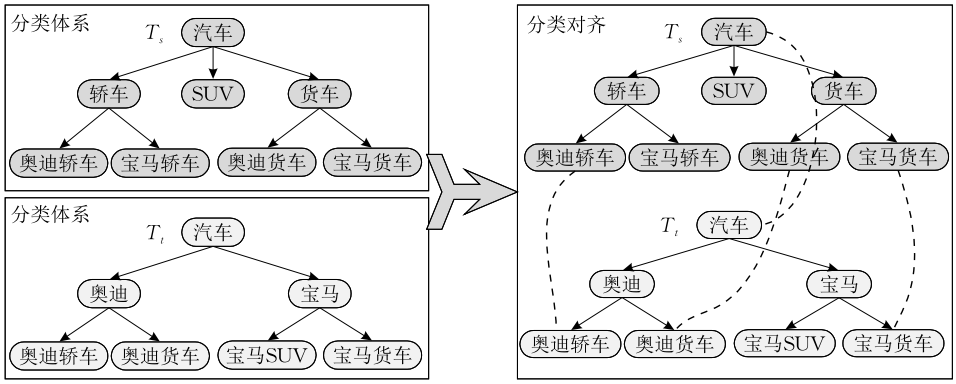


图 9 分类对齐模型

通过分析可以看出，现有的分类体系对齐方法大部分是通过计算两个分类体系之间的元素相似度来实现的. 虽然目前已经提出了很多分类体系对齐方法，但是这些方法无法有效处理大规模的分类体系[136]，主要原因在于对于分类体系中的每一个分类来说，与大规模的分类体系中的分类进行对齐时，会产生更多可能的候选选择，其对应的可能的对齐空间将会随分类体系中分类数量的增加呈现指数级增长. 不仅如此，它们中没有一个主导性的分类体系对齐方法能够在所有应用领域都表现地很好. 特别是，由于网络大数据的爆炸性增长，分类体系将变得越来越庞大和复杂. 因此，需要研究新的分类体系对齐方法以便最大程度提升分类扩充的正确率.

3.3.2　分类合并方法

分类合并的主要目标是根据分类对齐的结果，将两个分类体系中的元素进行集成，消除两个分类体系中的冗余元素，得到一个完整的分类体系，其模型如图 10 所示.

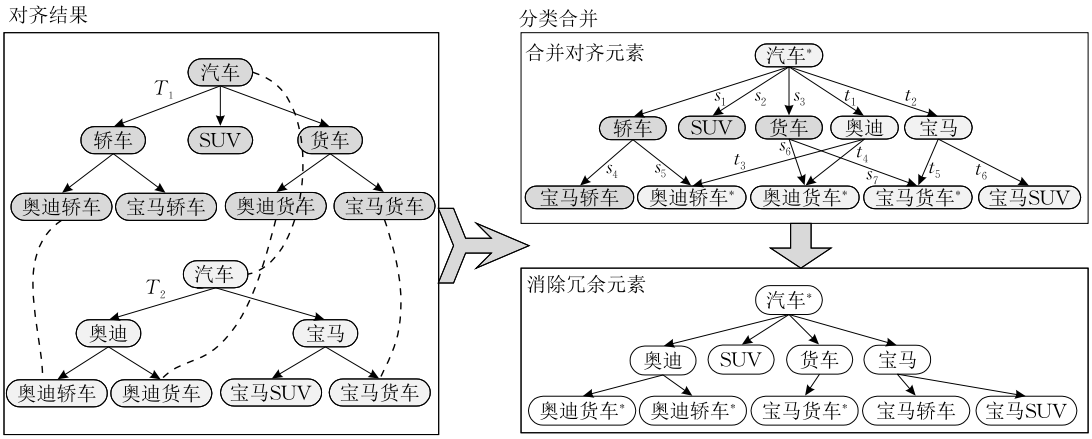


图 10　分类合并模型

针对从大规模半结构化和无结构化的网页中提取的分类体系进行合并面临噪音、冗余和信息缺失的问题，微软 Probase[14] 提出了一种基于语义的概率化局部合并方法，该方法将局部分类合并分为两种方式：水平合并和垂直合并，以分类的语义为基础，通过不断迭代执行水平合并或垂直合并，从而将分类体系关联起来. 为解决分类合并带来的不一致问题，Probase 将不一致消解问题转化为寻找图的最优多路割问题求解，保证方法的可扩展性.

水平合并[14] 主要是分析分类的种属关系，Probase 根据分类之间的 isA 关系，计算两个分类包含的孩子节点的相似度，通过对分类所属关系的语义判断，将属于同一个父类的分类进行关联，执行分类的水平合并，合并方式如图 11 所示.

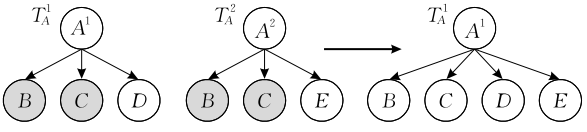
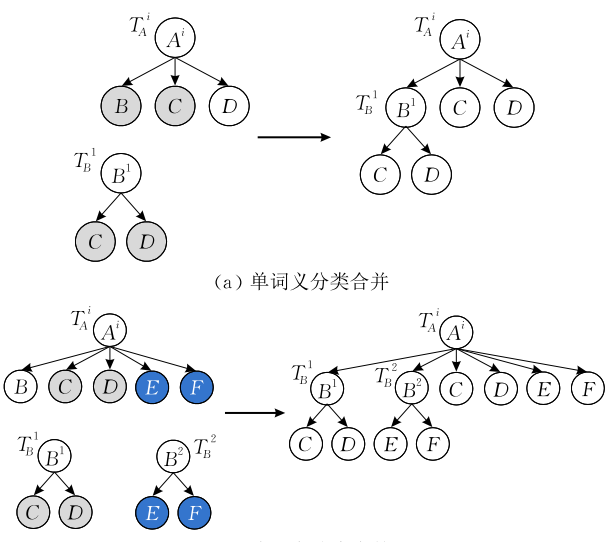


图 11　分类水平合并模型[14]

垂直合并[14] 则是充分分析分类的词义，利用其所扮演的不同语义角色，将具有相同语义角色的分类进行关联合并. Probase 根据分类之间的词义信息，计算两个分类之间的语义覆盖，执行分类的垂直合并，垂直合并方式如图 12 所示.



(a) 单词义分类合并

(b) 多词义分类合并

图 12　分类垂直合并模型[14]

除此之外，Raunich 等人[137] 提出了一种基于目标驱动的分类合并方法，该方法区分源分类体系和目标分类体系，将源分类体系和目标分类体系之间等价的分类合并成一个公共的分类，然后在这个合并的分类体系中正确地放置剩下的源分类体系中的

分类信息,在合并过程中,该方法重点维护目标分类体系的结构信息.

总体上讲,目前在分类的合并方面,已开展了多方面的研究,相关的技术和成果都有了一定程度的尝试和积累.但是,当前关于分类合并的研究工作大多针对的是一对一的分类对齐映射关系的合并,缺乏针对多对多映射的合并研究.

4 面向网络大数据的知识融合方法总体框架

前面几小节讨论了知识融合的模型思路,主要针对来自网络大数据的多源异构知识的评估以及知识库要素的扩充,包括实体扩充、关系扩充和分类扩充等.由于从网络大数据中获取的知识的多源性、动态性、多样性以及冗余、歧义等特点,在对这些知识进行融合时,需要建立度量知识质量寻求知识真值的评价体系,并能够将验证为正确的知识进行有效关联与合并,构建可动态扩展的知识库.因此,如何建立一个端到端(end-to-end)的模式框架,提供自底向上的方法将面向网络大数据的知识融合分解成易于处理的多个方面,并采用独立可行的技术.这些技术之间可建立共生的关系,互相补充互为条件来系统地实现开放网络知识的融合,保障知识库的维护和更新.

近年来,国内外在面向网络大数据的知识融合模式及框架研究方面也开展了一些研究,如微软在构建 Probase[14] 时对来自网络大数据的知识进行融合采用的模式如下:首先,从网络大数据中采集数据;其次,基于 Hearst Patterns[14] 是应用词法-句法信息从文本中识别上下位 isA 关系;接着,基于合理性和典型性指标评估知识为真概率;然后,基于知识的语义分析实现知识的关联合并,从而实现从网络大数据中获取的知识的融合,生成概率化的知识库 Probase. Google 在构建新一代知识库 Knowledge Vault[15] 对来自网络大数据的知识进行融合时,首先也是从网络大数据中采集数据;其次,基于不同的知识抽取器从网络大数据不同数据源中抽取知识;接着,基于局部封闭世界假设评价机制对不同来源获取的知识进行正确性判断;然后,基于实体链接等技术,对网络大数据中获取的知识进行关联合并,从而实现从网络大数据中获取的知识的融合.虽然 Probase 和 Knowledge Vault 都针对来自网络大数据的知识给出了融合计算的方法,但是它们主要是基于离线方式计算的,忽略了网络大数据中动态产生的新知识与知识库中已有知识的融合计算问题,难以保障知识库中知识的覆盖率和时新性.

根据前面对于知识融合问题的定义和分析,结合现有的知识融合模式及框架,可以看出在网络大数据时代,为了更好地利用网络大数据的价值,理解大数据中蕴含的知识,提高知识库的实用性,需要将网络大数据中动态产生的知识不断融合到知识库中,保持知识库随网络大数据发展的更新和演化.而开放知识网络[138](Open Knowledge Network,OpenKN)是一个异质的具有时空演化特性的网络,网络中的点和边都具有时间跨度和空间约束来定位以跟踪知识的演化过程,它能够更好地适用网络大数据环境下知识的挖掘、组织和计算.因此,我们以开放知识网络 OpenKN 作为网络大数据知识统一表示和计算的框架,总结了面向网络大数据的知识融合的模式和融合框架,如图 13 所示.

该框架自底向上主要分为数据采集、知识抽取和知识融合 3 个模块.数据采集模块的主要功能是从网络大数据中准确地识别、提取不同来源和形式的数据,高效地采集大量的信息,其关键在于采用分布式并行采集技术,并提供增量采集机制实现大规模网络数据的高效采集.知识抽取模块的主要功能是基于数据采集模块获取的海量碎片化数据中提取出组成知识库的知识要素,其关键在于采用自然语言处理技术,通过对自然语言的词法、句法的分析实现实体、关系、分类和属性等知识要素的抽取.知识融合模块则建立在知识抽取模块的基础上,负责将知识抽取模块从网络大数据不同数据源中提取的实体、关系、分类等进行对齐关联、合并计算,按序将知识组织生成知识库.

考虑到来源于网络大数据的知识可能存在错误、不一致等冲突情况,为此,知识融合模块又分为知识评估和知识扩充两个子模块.在对从网络大数据中抽取的知识进行融合时,首先利用知识评估子模块对来源于网络大数据的知识进行质量度量,解决知识冲突,寻找知识真值.其次,将验证为正确的知识,根据知识的类型,包括实体、关系、分类等,通过知识扩充子模块基于相应的知识扩充算法将其动态更新到知识库中.通过分析可以看出,知识评估是实现知识融合的先决条件,而知识扩充的结果可以作为先验知识指导知识评估.因此,知识融合包含的这两部分是互为补充互为条件,缺一不可.虽然不同类型知识的扩充的模型涵盖的内容和侧重点各异,

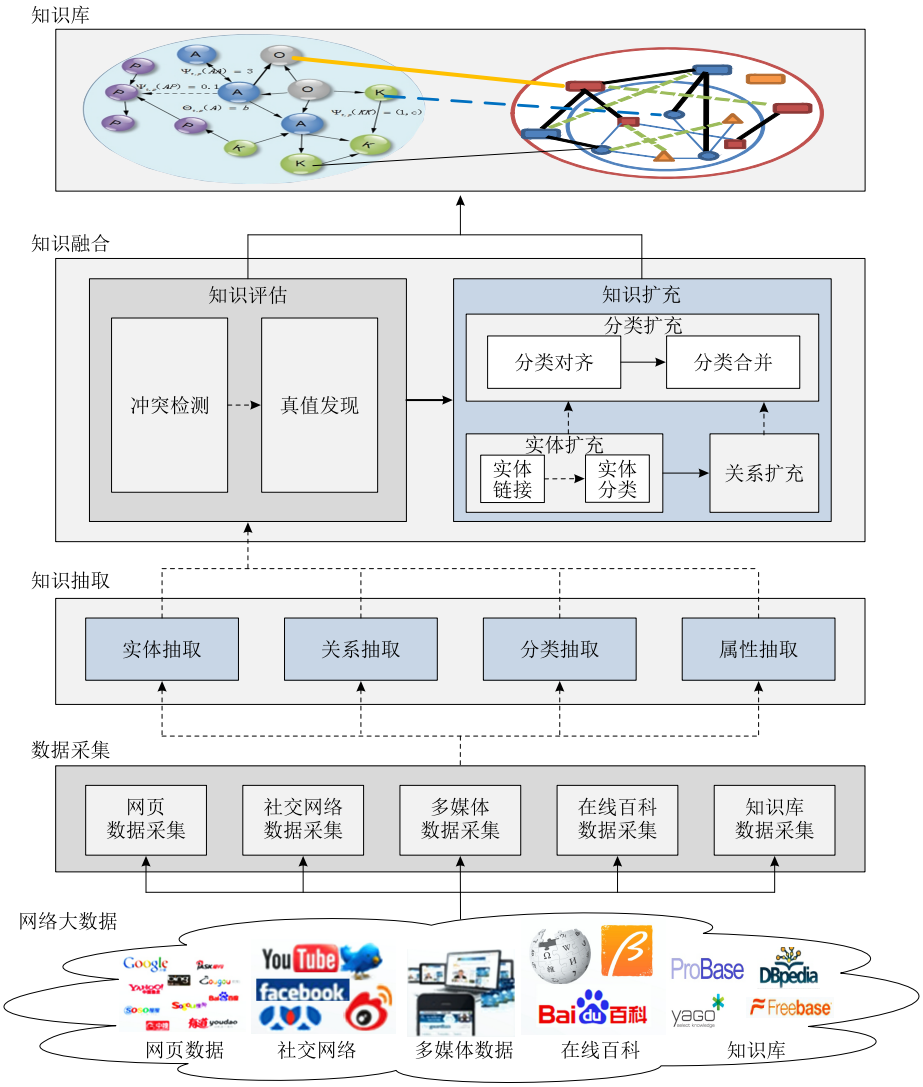


图 13　面向网络大数据知识融合的总体框架

包括实体扩充、关系扩充、分类扩充等，但是它们都与知识库的结构相关. 因此，可以考虑将它们归结到统一的知识扩充模型，这将是一个具有前景的方向. 从方法论的意义上讲，在上述模型框架中，由于实体、关系和分类刻画知识库的不同层面的结构内容，将它们的扩充以统一的方式建模是一个充满挑战的课题.

5　研究展望

面向网络大数据的知识融合为构建基于网络大数据的知识库，提高知识库的实用性，为人们深入利用网络大数据的价值提供有效的途径. 通过上述分析，我们看到知识融合已经取得了一定的成果，已有不少成熟的理论模型和方法，但无论是知识融合涉及到的知识的评估，还是知识的扩充，它的实现都还不能完全满足人们的应用需求，这意味知识融合尤其是在当前的网络大数据时代是极具挑战性的工作. 经过前几节的技术梳理发现面对网络大数据知识的融合，现存的融合技术仍然存在很多局限性，仍有大量问题需要研究和解决.

(1) 网络大数据中动态时序知识的评估. 由于网络大数据的高速变化，来源于网络大数据的知识随着网络大数据的发展，具有动态演化特性，上一时刻正确的知识，下一时刻未必为真，而上一时刻未发生的知识，下一时刻可能就变成了现实. 并且不同来源的知识更新频率不尽相同，导致知识的时效性难以达到一致. 而现有的知识评估的方法主要是针对静态知识的评估，无法直接适用于随时间动态演化的知识的评估，缺乏针对动态知识的处理方法. 不仅如此，现有的方法在进行知识评估时，缺乏对数据源之间关系的分析，缺乏对知识获取渠道和获取方式

的建模,因此难以从不可靠的知识获取方式中区分不可靠的数据源,这导致这些方法在处理网络大数据中大规模、多源异构、动态演化的知识的评估时面临准确率不高、鲁棒性较差等问题.目前,针对动态时序知识评估的研究成果还不多,主要是 YAGO 提出的基于时间一致性的评估方法[11-12],该方法主要针对维基百科数据,无法直接适用于来自网络大数据不同来源的动态时序知识的评估.所以为了提高知识评估方法处理网络大数据知识的扩展能力,这就需要开放网络知识在评估时能够充分融入知识的时间信息,跟踪这些时间信息的变化.不仅如此,结合数据源之间的关系,对知识获取渠道和获取方式进行建模,实现从不可靠的知识获取方式中区分不可靠的数据源的评估能力,提升知识评估的鲁棒性和扩展能力.

(2)实体扩充联合推断方法.通过对现有的实体扩充工作的分析发现,实体链接和实体分类都是实体扩充的重要组成部分,两者虽然作为实体扩充两种不同的处理情况,但其关键在于确定实体在知识库中的位置,两者之间互为补充,相互联系,这促使研究统一的实体扩充处理框架,实现实体链接和实体分类的一体化计算.但现有的工作基本是实体链接和实体分类独立化执行,不感知彼此的相互影响.不仅如此,现有的工作主要利用基于统计的特征(实体流行度)和词汇的特征(上下文相似度、话题连贯性),这些特征对流行的实体相对丰富但对长尾的实体却表现稀疏,导致长尾实体扩充的准确率不高,面临鲁棒性和可扩展性的挑战[33].目前,在实体扩充的联合推断方面的工作相对较少,主要是 Shen 等人提出的基于标签传播的方法[33],该方法通过将实体扩充中的实体链接和实体分类问题进行统一建模计算,借助实体之间的语义依赖关系提升实体扩充的准确率,但该方法在针对短文本中实体的背景知识稀疏时,实体扩充仍然面临准确率不高的问题.所以借助实体之间的语义依赖关系,采用联合推断这种互增益的学习机制仍然是下一步实体扩充研究的一大热点.

(3)联合多元信息的基于嵌入学习的关系扩充方法.在关系扩充方面,基于语义的方法的理论和技术的发展已相对成熟,但受限于语义词典或语料库的完备性,导致基于语义的方法的鲁棒性受到影响.近年来采用深度学习理论,提出的基于嵌入学习的关系扩充方法已经崭露头角,在面向大数据的关系扩充任务中展现了巨大的应用潜力,通过将关系的语义信息表示为低维空间中的稠密实值向量,该技术可以在低维空间中高效的计算关系之间的语义关系,有效解决数据稀疏的问题.然而,现有的嵌入学习方法主要是考虑实体描述的关系的表示学习,对描述实体关系的上下文、层次类型、别名现象、时间信息等缺乏有效的表示建模,有机表示这些信息,将显著提升嵌入学习方法的表示能力,提高关系扩充的准确率.此外,这些方法都是针对单一关系采用面向实体表示的学习方式,无法对关系之间的语义关系进行有效表示学习.目前有些研究工作已经利用卷积神经网络模型建立起了结合上下文信息的关系嵌入学习模型[139-140],这为联合多元信息的嵌入学习方法提供了技术基础.所以联合多元的与实体关系相关的其他信息的嵌入学习方法将是下一步关系扩充研究的热点问题.

(4)大规模异构分类体系的扩充方法.通过对现有的分类体系扩充工作的分析发现,分类对齐和分类合并都是分类体系扩充的重要组成部分,分类对齐是分类合并的先决条件,直接影响到分类扩充的效果,因此,分类对齐问题得到广泛的研究.然而,目前大部分工作还只能在特定领域发挥作用,而且无法有效地处理大规模的分类体系[136].导致这一问题的原因在于:不同的分类体系通常使用不同的词汇和层级结构来表示自己的分类,而且其对应的可能的匹配空间随分类体系中分类的规模的增加呈现指数级增长.特别是,随着网络大数据的发展,分类体系变得越来越庞大和复杂.目前有些研究工作已利用贪心算法解决大规模分类体系对齐的问题[136],但由于该方法贪心的性质,它在匹配决策时难以修正之前的错误.因此,该方法无法保证两个分类体系获得全局最优的匹配,导致分类体系扩充的准确率受到影响.所以考虑到现有单一的分类体系扩充方法存在领域适应性弱,规模扩展能力差等问题,这就需要集合多种分类体系扩充方法,即构建聚合型的分类体系扩充方法,通过提供统一的评价机制,对多种独立的分类体系扩充方法进行整合、控制和优化利用,将值得深入探索.

6 总 结

网络大数据是指“人、机、物”三元世界在网络空间中交互、融合所产生并在互联网上可获得的大数据.这些数据具有多源异构、时效性、社会性、突发性和高噪声等特点,其背后蕴含着丰富的知识资源,这

些知识之间复杂关联形成强大的知识库.对这些知识进行深入分析成为有效利用网络大数据价值的重要体现.然而,由于网络大数据的大规模、变化高速、多样性和正确性给知识库的实用性带来巨大的挑战.为此,本文以提高知识库的实用性为最终目标,以面向网络大数据的知识融合为技术手段,详细分析和讨论了知识融合相关模型和方法,包括开放网络知识质量的评估、开放网络知识的扩充.基于这些分析和讨论,希望能够为未来的网络大数据的知识融合的研究提供一些有益的指导和启发.

参 考 文 献

[1] Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network big data: Present and future. Chinese Journal of Computers, 2013, 36(6): 1125-1138(in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望. 计算机学报, 2013, 36(6): 1125-1138)

[2] Turner V, Gantz J F, Reinsel D, et al. The digital universe of opportunities: Rich data and the increasing value of the Internet of Things. International Data Corporation, White Paper, IDC_1672, 2014

[3] Meng Xiao-Feng, Ci Xiang. Big data management: Concepts, techniques and challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169(in Chinese)
(孟小峰, 慈祥. 大数据管理: 概念, 技术与挑战. 计算机研究与发展, 2013, 50(1): 146-169)

[4] Wen Jun. Knowledge Base System Theory and Applications. Shanghai: Fudan University Press, 1995(in Chinese)
(文君. 知识库系统原理及其应用. 上海: 复旦大学出版社, 1995)

[5] Wang Yuan-Zhuo, Jia Yan-Tao, Liu Da-Wei, et al. Open web knowledge aided information search and data mining. Journal of Computer Research and Development, 2015, 52(2): 456-474(in Chinese)
(王元卓, 贾岩涛, 刘大伟等. 基于开放网络知识的信息检索与数据挖掘. 计算机研究与发展, 2015, 52(2): 456-474)

[6] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[7] Etzioni O, Cafarella M, Downey D, et al. Web-scale information extraction in knowitall: (preliminary results)//Proceedings of the 13th International Conference on World Wide Web (WWW). New York, USA, 2004: 100-110

[8] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Vancouver, Canada, 2008: 1247-1250

[9] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a web of open data//Proceedings of the 6th Semantic Web and 2nd Asian Conference on Asian Semantic Web (ISWC-ASWC). Busan, Korea, 2007: 722-735

[10] Ponzetto S P, Navigli R. Large-scale taxonomy mapping for restructuring and integrating Wikipedia//Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI). Pasadena, USA, 2009: 2083-2088

[11] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia// Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI). Beijing, China, 2013: 3161-3165

[12] Mahdisoltani F, Biega J, Suchanek F M. YAGO3: A knowledge base from multilingual Wikipedias//Proceedings of the 7th Bennial Conference on Innovative Data Systems Research (CIDR). Asilomar, USA, 2015: 1-11

[13] Carlson A, Betteridge J, Wang R C, et al. Coupled semi-supervised learning for information extraction//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM). New York, USA, 2010: 101-110

[14] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Scottsdale, USA, 2012: 481-492

[15] Dong X, Gabrilovich E, Heitz G, et al. Knowledge Vault: A web-scale approach to probabilistic knowledge fusion//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD). New York, USA, 2014: 601-610

[16] Singhal A. Introducing the knowledge graph: Things, not strings. Official Google Blog, 2012

[17] Lu Ru-Qian, Jin Zhi. From knowledge-based software engineering to knowware-based software engineering. Science in China Series E, 2008, 38(6): 843-863(in Chinese)
(陆汝钤, 金芝. 从基于知识的软件工程到基于知件的软件工程. 中国科学: E辑, 2008, 38(6): 843-863)

[18] Niu X, Sun X, Wang H, et al. Zhishi.me-weaving Chinese linking open data//Proceedings of the 10th International Semantic Web Conference (ISWC). Bonn, Germany, 2011: 205-220

[19] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion. The VLDB Endowment, 2014, 7(10): 881-892

[20] Dong X L, Naumann F. Data fusion: Resolving data conflicts for integration. The VLDB Endowment, 2009, 2(2): 1654-1655

[21] Zhou Fang, Wang Peng-Bo, Han Li-Yan. Multi-source knowledge fusion algorithm. Journal of Beijing University of Aeronautics and Astronautics, 2013, 39(1): 109-114(in Chinese)
(周芳, 王鹏波, 韩立岩. 多源知识融合处理算法. 北京航空航天大学学报, 2013, 39(1): 109-114)

[22] Dempster A P. Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics, 1967, 38(2): 325-339

[23] Shafer G. A Mathematical Theory of Evidence. Princeton, USA: Princeton University Press, 1976

[24] Dempster A P. Upper and lower probabilities induced by a multivalued mapping. Classic Works of the Dempster-Shafer Theory of Belief Functions, 2008, 219: 57-72

[25] Zadeh L A. Fuzzy sets. Information and Control, 1965, 8(3): 338-353

[26] Abdulghafour M, Chandra T, Abidi M A. Data fusion through fuzzy logic applied to feature extraction from multi-sensory images//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA, 1993: 359-366

[27] Grabisch M, Sugeno M, Murofushi T. Fuzzy Measures and Integrals: Theory and Applications. USA: Springer-Verlag New York Inc, 2000

[28] Tahani H, Keller J M. Information fusion in computer vision using the fuzzy integral. IEEE Transactions on Systems, Man and Cybernetics, 1990, 20(3): 733-741

[29] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK, 2011: 529-539

[30] Dong X L, Gabrilovich E, Murphy K, et al. Knowledge-based trust: Estimating the trustworthiness of web sources. The VLDB Endowment, 2015, 8(9): 938-949

[31] Zhao B, Han J. A probabilistic model for estimating real-valued truth from conflicting sources//Proceedings of the 10th International Workshop on Quality in Databases (QDB). Instanbul, Turkey, 2012: 1-7

[32] Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2): 443-460

[33] Shen W, Wang J, Luo P, et al. A graph-based approach for ontology population with named entities//Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM). Maui, USA, 2012: 345-354

[34] Monge A E, Elkan C. The field matching problem: Algorithms and applications//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD). Portland, USA, 1996: 267-270

[35] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records//Proceedings of the KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation. Washington, USA, 2003, 3: 73-78

[36] Tejada S, Knoblock C A, Minton S. Learning domain-independent string transformation weights for high accuracy object identification//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Edmonton, Canada, 2002: 350-359

[37] Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Washington, USA, 2003: 39-48

[38] Budanitsky A, Hirst G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures //Proceedings of the Workshop on WordNet and Other Lexical Resources. Pittsburgh, USA, 2001: 29-34

[39] Budanitsky A, Hirst G. Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics, 2006, 32(1): 13-47

[40] Liu Qun, Li Su-Jian. Word similarity computing based on How-net. Chinese Computational Linguistics, 2002, 7(2): 59-76(in Chinese)
(刘群，李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学，2002，7(2)：59-76)

[41] Wang Bin. Automatic Alignment of Bilingual Corpus in English and Chinese[Ph. D. dissertation]. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 1999(in Chinese)
(王斌. 汉英双语语料库自动对齐研究[博士学位论文]. 中国科学院计算机技术研究所，北京，1999)

[42] Chen R C, Bau C T, Yeh C J. Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques. Applied Soft Computing, 2011, 11(2): 1908-1923

[43] Arasu A, Götz M, Kaushik R. On active learning of record matching packages//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Indianapolis, USA, 2010: 783-794

[44] Li Y, Wang C, Han F, et al. Mining evidences for named entity disambiguation//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Chicago, USA, 2013: 1070-1078

[45] Ratinov L, Roth D, Downey D, et al. Local and global algorithms for disambiguation to Wikipedia//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, USA, 2011: 1375-1384

[46] Bunescu R C, Pasca M. Using Encyclopedic knowledge for named entity disambiguation//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Trento, Italy, 2006: 9-16

[47] Ananthakrishna R, Chaudhuri S, Ganti V. Eliminating fuzzy duplicates in data warehouses//Proceedings of the 28th International Conference on Very Large Data Bases (VLDB). Hong Kong, China, 2002: 586-597

[48] Bhattacharya I, Getoor L. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-36

[49] Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data//Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, Czech Republic, 2007: 708-716

[50] Milne D, Witten I H. Learning to link with Wikipedia//Proceedings of the 17th ACM Conference on Information and Knowledge Management(CIKM). Napa Valley, USA, 2008: 509-518

[51] Cilibrasi R L, Vitanyi P M B. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

[52] Bhattacharya I, Getoor L. A latent Dirichlet model for unsupervised entity resolution//Proceedings of the 6th SIAM International Conference on Data Mining (SDM). Bethesda, USA, 2006: 47-58

[53] Sen P. Collective context-aware topic models for entity disambiguation//Proceedings of the 21st International Conference on World Wide Web (WWW). Lyon, France, 2012: 729-738

[54] Han X, Sun L, Zhao J. Collective entity linking in web text: A graph-based method//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Beijing, China, 2011: 765-774

[55] Cucerzan S. TAC entity linking by performing full-document entity extraction and disambiguation//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2011: 1-7

[56] Gottipati S, Jiang J. Linking entities to a knowledge base with query expansion//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK, 2011: 804-813

[57] Lafferty J, Zhai C. Document language models, query models, and risk minimization for information retrieval//Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). New Orleans, USA, 2001: 111-119

[58] Zhang W, Sim Y C, Su J, et al. Entity linking with effective acronym expansion, instance selection, and topic modeling//Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI). Barcelona, Spain, 2011: 1909-1914

[59] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK, 2011: 782-792

[60] Thater S, Fürstenau H, Pinkal M. Contextualizing semantic representations using syntactically enriched vector models//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Columbus, USA, 2010: 948-957

[61] Lee T, Wang Z, Wang H, et al. Web scale taxonomy cleansing. The VLDB Endowment, 2011, 4(12): 1295-1306

[62] Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships. The VLDB Endowment, 2010, 3(1-2): 1338-1347

[63] Shen W, Wang J, Luo P, et al. Linden: Linking named entities with knowledge base via semantic knowledge//Proceedings of the 21st International Conference on World Wide Web (WWW). Lyon, France, 2012: 449-458

[64] Guo Y, Qin B, Liu T, et al. Microblog entity linking by leveraging extra posts//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, USA, 2013: 863-868

[65] Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Chicago, USA, 2013: 68-76

[66] Guo S, Chang M W, Kiciman E. To link or not to link? A study on end-to-end tweet entity linking//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Atlanta, USA, 2013: 1020-1030

[67] Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 2007, 30(1): 3-26

[68] Fleischman M, Hovy E. Fine grained classification of named entities//Proceedings of the 19th International Conference on Computational Linguistics (COLING). Mumbai, India, 2002: 1-7

[69] Sekine S. NYU: Description of the Japanese NE system used for MET-2//Proceedings of the 7th Message Understanding Conference (MUC). San Diego, USA, 1998: 1-6

[70] Borthwick A, Sterling J, Agichtein E, et al. NYU: Description of the MENE named entity system as used in MUC-7//Proceedings of the 7th Message Understanding Conference (MUC). San Diego, USA, 1998: 7-12

[71] Asahara M, Matsumoto Y. Japanese named entity extraction with redundant morphological analysis//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Edmonton, Canada, 2003: 8-15

[72] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Edmonton, Canada, 2003: 188-191

[73] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics(ACL). Ann Arbor, USA, 2005: 363-370

[74] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan, 2003: 184-187

[75] Collins M, Singer Y. Unsupervised models for named entity classification//Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, USA, 1999: 100-110

[76] Collins M. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL). Philadelphia, USA, 2002: 489-496

[77] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Edmonton, Canada, 2003: 25-32

[78] Cucchiarelli A, Velardi P. Unsupervised named entity recognition using syntactic and semantic contextual evidence. Computational Linguistics, 2001, 27(1): 123-131

[79] Pasca M, Lin D, Bigham J, et al. Organizing and searching the world wide web of facts-step one: The one-million fact extraction challenge//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Boston, USA, 2006: 1400-1405

[80] Nadeau D, Turney P, Matwin S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity//Proceedings of the 19th Canadian Conference on Artificial Intelligence. Québec, Canada, 2006: 1-12

[81] Alfonseca E, Manandhar S. An unsupervised method for general named entity recognition and automated concept discovery//Proceedings of the 1st International Conference on General WordNet. Mysore, India, 2002: 34-43

[82] Evans R. A framework for named entity recognition in the open domain. Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 2003, 260: 267-274

[83] Hearst M A. Automatic acquisition of hyponyms from large text corpora//Proceedings of the 14th Conference on Computational Linguistics (COLING). Nantes, France, 1992: 539-545

[84] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence, 2005, 165(1): 91-134

[85] Rahman A, Ng V. Inducing fine-grained semantic classes via hierarchical and collective classification//Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China, 2010: 931-939

[86] Ling X, Weld D S. Fine-grained entity recognition//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Toronto, Canada, 2012: 94-100

[87] Yosef M A, Bauer S, Hoffart J, et al. HYENA: Hierarchical type classification for entity names//Proceedings of the International Conference on Computational Linguistics (COLING). Mumbai, India, 2012: 1361-1370

[88] Cimiano P, Völker J. Towards large-scale, open-domain and ontology-based named entity classification//Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP). Borovets, Bulgaria, 2005: 66-72

[89] Tanev H, Magnini B. Weakly supervised approaches for ontology population. Frontiers in Artificial Intelligence and Applications, 2008, 167: 129-143

[90] Ganti V, König A C, Vernica R. Entity categorization over large document collections//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Las Vegas, USA, 2008: 274-282

[91] Giuliano C, Gliozzo A. Instance-based ontology population exploiting named-entity substitution//Proceedings of the 22nd International Conference on Computational Linguistics (COLING). Manchster, UK, 2008: 265-272

[92] Giuliano C. Fine-grained classification of named entities exploiting latent semantic kernels//Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL). Boulder, USA, 2009: 201-209

[93] Nakashole N, Tylenda T, Weikum G. Fine-grained semantic typing of emerging entities//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria, 2013: 1488-1497

[94] Nakashole N, Weikum G, Suchanek F. PATTY: A taxonomy of relational patterns with semantic types//Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Jeju, Korea, 2012: 1135-1145

[95] Yosef M A, Bauer S, Hoffart J, et al. HYENA-live: Fine-grained online entity type classification from natural-language text//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria, 2013: 133-138

[96] Tsoumakas G, Zhang M L, Zhou Z H. Introduction to the special issue on learning from multi-label data. Machine Learning, 2012, 88(1-2): 1-4

[97] Baluja S, Seth R, Sivakumar D, et al. Video suggestion and discovery for YouTube: Taking random walks through the view graph//Proceedings of the 17th International Conference on World Wide Web (WWW). Beijing, China, 2008: 895-904

[98] Rada R, Mili H, Bicknell E, et al. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 1989, 19(1): 17-30

[99] Resnik P. Using information content to evaluate semantic similarity in a taxonomy//Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). Montréal, Canada, 1995: 448-453

[100] Banerjee S, Pedersen T. An adapted lesk algorithm for word sense disambiguation using WordNet//Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). Mexico-City, Mexico, 2002: 136-145

[101] Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone//Proceedings of the 5th International Conference on Systems Documentation (SIGDOC). Toronto, Canada, 1986: 24-26

[102] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts//Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy, 2006: 1-8

[103] Sánchez D, Batet M, Isern D. Ontology-based information content computation. Knowledge-Based Systems, 2011, 24(2): 297-303

[104] Li P, Wang H, Zhu K Q, et al. Computing term similarity by large probabilistic isA knowledge//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM). San Francisco, USA, 2013: 1401-1410

[105] Alvarez M A, Lim S J. A graph modeling of semantic similarity between words//Proceedings of the International Conference on Semantic Computing (ICSC). Irvine, USA, 2007: 355-362

[106] Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Athens, Greece, 2009: 33-41

[107] Agirre E, Cuadros M, Rigau G, et al. Exploring knowledge bases for similarity//Proceedings of the International Conference on Language Resources and Evaluation (LREC). Reykjavik, Iceland, 2010: 373-377

[108] Chen H H, Lin M S, Wei Y C. Novel association measures using web search with double checking//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL). Sydney, Australia, 2006: 1009-1016

[109] Bollegala D, Matsuo Y, Ishizuka M. A web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 977-990

[110] Radinsky K, Agichtein E, Gabrilovich E, et al. A word at a time: Computing word relatedness using temporal semantic analysis//Proceedings of the 20th International Conference on World Wide Web (WWW). Hyderabad, India, 2011: 337-346

[111] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2013: 3111-3119

[112] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Atlanta, USA, 2013: 746-751

[113] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data//Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2013: 2787-2795

[114] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes//Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI). Québec City, Canada, 2014: 1112-1119

[115] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion//Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI). Austin, USA, 2015: 2181-2187

[116] Feng J, Zhou M, Hao Y, et al. Knowledge graph embedding by flexible translation. arXiv preprint/1505.05253, 2015

[117] Fan M, Zhou Q, Chang E, et al. Transition-based knowledge graph embedding with relational mapping properties//Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation (PACLIC). Phuket, Thailand, 2014: 328-337

[118] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing//Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). La Palma, Canary Islands, 2012: 127-135

[119] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data. Machine Learning, 2014, 94(2): 233-259

[120] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases//Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI). San Francisco, USA, 2011: 301-306

[121] Sutskever I, Tenenbaum J B, Salakhutdinov R R. Modelling relational data using Bayesian clustered tensor factorization//Proceedings of the 23rd Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada, 2009: 1821-1828

[122] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data//Proceedings of the 26th Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2012: 3167-3175

[123] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion//Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2013: 926-934

[124] Chen D, Socher R, Manning C D, et al. Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. arXiv preprint/1301.3618, 2013

[125] Nickel M, Tresp V, Kriegel H P. Factorizing YAGO: scalable machine learning for linked data//Proceedings of the 21st International Conference on World Wide Web (WWW). Lyon, France, 2012: 271-280

[126] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching. The VLDB Journal, 2001, 10(4): 334-350

[127] Shvaiko P, Euzenat J. A survey of schema-based matching approaches. Journal on Data Semantics IV, 2005: 146-171

[128] Choi N, Song I Y, Han H. A survey on ontology mapping. ACM Sigmod Record, 2006, 35(3): 34-41

[129] Shvaiko P, Euzenat J. Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 158-176

[130] Li J, Tang J, Li Y, et al. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8): 1218-1232

[131] Suchanek F M, Abiteboul S, Senellart P. PARIS: Probabilistic alignment of relations, instances, and schema. The VLDB Endowment, 2011, 5(3): 157-168

[132] Demidova E, Oelze I, Nejdl W. Aligning freebase with the YAGO ontology//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM). San Francisco, USA, 2013: 579-588

[133] Ba M, Diallo G. Large-scale biomedical ontology matching with ServOMap. IRBM, 2013, 34(1): 56-59

[134] Jiménez-Ruiz E, Grau B C. LogMap: Logic-based and scalable ontology matching//Proceedings of the International Semantic Web Conference (ISWC). San Francisco, USA, 2011: 273-288

[135] Lacoste-Julien S, Palla K, Davies A, et al. SiGMa: Simple greedy matching for aligning large knowledge bases//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Chicago, USA, 2013: 572-580

[136] Grau B C, Dragisic Z, Eckert K, et al. Results of the ontology alignment evaluation initiative 2013//Proceedings of the 8th ISWC Workshop on Ontology Matching. Zurich, Switzerland, 2013: 61-100

[137] Raunich S, Rahm E. ATOM: Automatic target-driven ontology merging//Proceedings of the 27th International Conference on Data Engineering (ICDE). Hannover, Germany, 2011: 1276-1279

[138] Jia Y, Wang Y, Cheng X, et al. OpenKN: An open knowledge computational engine for network big data//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Beijing, China, 2014: 657-664

[139] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks //Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal, 2015: 1753-1762

[140] dos Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL). Beijing, China, 2015: 626-634

LIN Hai-Lun, born in 1987, Ph.D., assistant professor. Her main research interests include open knowledge network, information extraction, etc.

WANG Yuan-Zhuo, born in 1978, Ph.D., associate professor. His main research interests include social computing, open knowledge network, etc.

JIA Yan-Tao, born in 1983, Ph.D., assistant professor. His main research interests include open knowledge network, social computing, etc.

ZHANG Peng, born in 1984, Ph.D., associate professor. His current research interests include cloud computing, stream data processing, etc.

WANG Wei-Ping, born in 1975, Ph.D., professor and Ph.D. supervisor. His main research interests include big data storage and processing.

Background

This work is supported by Grants from the National Science and Technology Support Program of China (No. 2012BAH46B03), the National HeGaoJi Key Project of China (No. 2013ZX01039-002-001-001), the National Key Research and Development Program of China (No. 2016YFB1000902), and the National Natural Science Foundation of China (Nos. 61602467, 61303056, 61402442, 61402464, 61572469, 61572473, 61502478).

With the coming of the era of network big data, this has motivated the research on the construction of open network data oriented knowledge base. Although a large number of knowledge bases have been constructed, these knowledge

bases are still much limited in knowledge coverage and freshness. Therefore, in order to better support a wide range of knowledge base based applications, such as knowledge search, knowledge recommendation and so on, researching on adaptive heterogeneous knowledge fusion for open network is of great significance. However, the knowledge in network data has a few typical features, such as multi-sourced, temporal evolution and ambiguity, so the task of adaptive knowledge fusion is challenging. This has motivated the research on the network big data oriented knowledge fusion to improve the utility of knowledge bases. The authors' main objectives are to provide a complete overview on methods for knowledge fusion.

In this paper, the authors present a survey on the techniques and algorithms of knowledge fusion in decades. They indicate the most commonly knowledge evaluation methods used to judge the authenticity of knowledge in knowledge fusion, and introduce the research progress of knowledge population in detail from entity population, relation population and taxonomy population aspects. The authors also discuss the overall framework of knowledge fusion. Finally, the authors give some concluding remarks on new and challenging directions for future and potential research of knowledge fusion for network big data.