**知识融合现状研究**

知识图谱创建的信息抽取过程,实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标，然而，这些结果中可能包含大量的冗余和错误信息，数据之间的关系也是扁平化的，缺乏层次性和逻辑性，因此有必要对其进行清理和整合。通过知识融合，可以消除概念的歧义，剔除冗余和错误概念，从而确保知识的质量，帮助构建出高质量的知识图谱。

一、实体链接

实体链接(entity linking)是指对于从文本中抽取得到的实体对象, 将其链接到知识库中对应的正确实体对象的操作。它的基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象,然后通过相似度计算将指称项链接到正确的实体对象。

实体链接的一般流程是: 1)从文本中通过实体抽取得到实体指称项; 2)进行实体对齐、实体消歧，消除数据存在的实体冲突问题，判断知识库中的同名实体与之是否代表不同的含义以及知识库中是否存在其他命名实体与之表示相同的含义; 3)在确认知识库中对应的正确实体对象之后,将该实体指称项链接到知识库中对应实体。

二、实体对齐

实体对齐(entity alignment)，也称为实体匹配(entity matching)，目的是消除异构数据中存在的知识指代不明、实体冲突等问题。实体对齐本质上是在计算知识之间的相似度，通过相似度来决定知识对齐的操作。

实体对齐算法可以分为成对实体对齐、协同实体对齐以及基于表示学习的对齐算法三类。成对实体对齐主要考虑实例及其属性的相似程度，而协同实体对齐则是在成对对齐的基础上考虑不同实例之间的相互关系用以计算相似度，基于表示学习的实体对齐则主要通过对低纬度的空间映射来使用向量表示实体与关系，提供更丰富的语义信息。

1）成对实体对齐算法

成对实体对齐根据属性相似性评分来判断实体是否匹配，主要包括基于概率模型的对齐算法和基于机器学习的对齐算法。最早的基于概率的对齐算法是基于属性相似性评分方法[1]，该方法将实体匹配问题转换为分类问题(匹配、可能匹配和不匹配三种)，在此基础上建立起基于概率的实体链接模型，为每个匹配的属性分配不同的权重，提高准确性[2]。

基于机器学习的对齐算法则是将实体对齐问题转换为二分类问题进行处理。根据是否采用标注数据可以将该方法分为有监督学习和无监督学习两类。有监督学习采用一部分标注数据作为训练集来训练模型，然后将模型用于未标注数据的对齐。这一类算法集中于决策树、支持向量机、集成学习等[3][4][5][6]。无监督的学习方法则基于聚类的思想，将类似的实体聚在一起，然后通过少量的标记数据来推理实体的匹配情况[7]。

2）协同实体对齐算法

协同实体对齐算法为实体本身的属性以及与它相关的实体的属性分别设置不同的权重，并通过加权来计算总体的相似度，来实现实体的对齐。基于相似性传播的方法就是一种典型的协同实体对齐方法，主要考虑了需要匹配的两个实体及与他们直接关联的其他实体的相似性。而基于概率的协同实体对齐算法则采用 LDA模型、CRF模型、Markov逻辑网等[8][9][10]。

3）基于表示学习的实体对齐方法

表示学习方法来源于深度学习，主要思想是将知识图谱中的实体和关系映射到低维空间，学习得到实体和关系的向量表示。这种低维稠密的向量蕴涵了图谱的结构信息以及实体和关系的属性特征，具有丰富的语义信息。基于表示学习的知识图谱实体对齐算法由两部分组成：知识表示的学习和实体间映射关系的学习。首先，将待对齐知识图谱分别映射到低维空间得到对应的知识表；其次,基于新的知识表示和人工标注的实体对齐数据集来学习得到实体间的对应关系。

三、实体消歧

实体消歧(entity disambiguation)是专门用于解决同名实体产生歧义问题的技术，即经常会遇到的某个实体指称项对应于多个命名实体对象的问题。通过实体消歧，就可以根据当前的语境，准确建立实体链接。实体消歧主要采用聚类法，以实体对象为聚类中心，将所有指向同一目标实体对象的指称项聚集到以该对象为中心的类别下。聚类法消歧的关键问题是如何定义实体对象与指称项之间的相似度，常用方法模型有以下四种：

1. 空间向量模型(也称词袋模型)。典型的方法是取当前语料中实体指称项周边的词构成特征向量，然后利用向量的余弦相似度进行比较，将该指称项聚类到与之最相近的实体指称项集合中。

2. 语义模型。该模型与空间向量模型类似，区别在于特征向量的构造方法不同，语义模型的特征向量不仅包含词袋向量，而且包含一部分语义特征。

3. 社会网络模型。该模型的基本假设是物以类聚、人以群分。在建模时,首先利用实体间的关系将与之相关的指称项链接起来构成网络,然后利用社会网络分析技术计算该网络中节点之间的拓扑距离(网络中的节点即实体的指称项),以此来判定指称项之间的相似度。

4. 百科知识模型。百科类网站通常会为每个实体 (指称项) 分配一个单独页面，其中包括指向其他实体页面的超链接，百科知识模型正是利用百科类网站中实体页面的链接关系来计算实体指称项之间的相似度。

四、共指消解

共指消解(entity resolution)技术主要用于解决多个指称项对应于同一实体对象的问题。共指消解的主要解决思路分为两种：基于自然语言处理和基于统计机器学习的方法。

基于自然语言处理的共指消解是以句法分析为基础的，代表性方法是Hobbs算法和向心理论(centering theory)。Hobbs算法主要思路是基于句法分析树进行搜索。向心理论则将表达模式 (utterance) 视为语篇(discourse)的基本组成单元，通过识别表达模式中的实体，可以获得当前和后续语篇中的关注中心(实体)，根据语义的局部连贯性和显著性，就可以在语篇中跟踪受关注的实体[11]。除此之外，统计机器学习方法也被引入该领域，并进入了快速发展阶段，通过将共指消解作为聚类问题来求解，以实体指称项为中心，通过实体聚类实现指称项与实体对象的匹配。

1. Newcombe H B, Kennedy S J, Axford S J, et al. Automatic linkage of vital records [J]. Science, 1959, 130 (3381): 954-959.
2. Herzog T N, Scheuren F J, Winkler W E. Data quality and record linkage techniques [M]. Berlin: Springer, 2007.
3. Han J W, Kambe M. Data mining: Concepts and techniques [M]. San Francisco, CA: Morgan Kaufmann, 2006.
4. Vapnik V. The nature of statistical learning theory [M]. Berlin: Springer, 2000.
5. Kantardzic M. Data mining [M]. Hoboken, NJ: John Wiley & Sons, 2011.
6. Cohen W W, Richman J. Learning to match and cluster large highdimensional data sets for data integration [C]// Proc of the ACM SIGKDD Conf on Knowledge Discovery and Data Mining, New York: ACM Presser, 2002: 475-480.
7. Bhattacharya I, Getoor L. Alatent dirichlet allocation model for unsupervised entity resolution [C]// Proc of the 6th SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2006: 47-58
8. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.
9. Fisher J, Christen P, Wang Qing. Active learning based entity resolution using Markov logic [C]// Proc of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Auckland:Springer, 2016: 338–349.
10. 李春华. 基于机器学习模型与众包的知识融合方法研究 [D]. 苏州:苏州大学, 2017. (Li Chunhua. Knowledge fusion based on machine learning model and crowdsourcing [D]. Suzhou: Soochow University, 2017.)
11. Grosz B J, Weinstein S, Joshi A K.Centering:A framework for modeling the local coherence of discourse[J].Computational Linguistics, 1995, 21 (2) :203-225