高技术通讯 2022 年 第 32 卷 第 12 期:1302-1311

doi:10.3772/j.issn.1002-0470.2022.12.010

面向行业电商知识图谱应用的实体对齐算法①

陈富强②* 肖明明③* 韩凯南** 任 毅*** 王文文*** 李 克*

(*北京联合大学智慧城市学院 北京 100101)

(**中铁物贸集团有限公司 北京 102308)

(***鲁班(北京)电子商务科技有限公司 北京 102308)

摘 要 针对多源异构知识图谱数据融合中的实体对齐问题,本文面向行业电商领域电商平台真实数据,提出了一种基于领域知识的集合相似度实体对齐算法。首先,基于领域知识针对性设计数据预处理技术,如实体属性值原子化、统一术语和去除冗余等,以规范化电商底层多源异构数据、提升数据处理效率和准确性;然后,以行业电商知识图谱应用为导向,筛选实体对生成高质量候选集,优化集合相似度测量和实体对排序方法,实现实体对的高效匹配。实验结果表明,本文算法可有效提高多源异构数据融合的准确率,大幅减少人工干预,可为行业电商发展提供新思路。

关键词 多源异构数据;知识图谱;实体对齐;集合相似度;电子商务

0 引 言

随着电子商务的兴起,电子商务开始替代传统贸易模式,在各个领域内快速发展。作为一种新型的商业运营模式,电商在交易领域生产、供货和物流等方面经常面临着不同客户数据的多源异构问题。在铁路建设领域内大宗商品的交易过程也面临着同样的问题,数据的多源异构性使得交易过程中数据流转不方便、使用困难[1],对数据使用的效率产生了极大影响。传统数据管理方式已经不能满足发展的需要,而基于图数据库的知识图谱在数据存储和管理上更为灵活。知识图谱通过节点存储现实世界中的实体,通过边存储实体间的关系,同时在知识图谱里可以存储不同类型的数据以及复杂的实体关系,提升了数据流动的速度和使用的效率。一些消费类电商已将知识图谱技术应用到实际生产交易中,以激发行业活力,带动行业发展,如阿里巴巴建设的电商认知图谱“AliCoCo”可认知用户需求实现更加智能的搜索和精准推荐[2],美团构建的餐饮娱乐知识图谱“美团大脑”可充分挖掘关联的各个场景数据实现智能搜索和商圈美食的个性化推荐[3],京东建立了基于商品知识图谱的兴趣召回等。

随着知识图谱广泛的应用,各个领域都开始建设属于自己领域的知识图谱,统称为领域知识图谱。不同领域的知识图谱因为业务的不同使得其从本体到实体都存在较大的差异,而同一领域的知识图谱也因为规范不统一存在一定差异,不同来源的知识图谱无法直接建立起联系[4]。因此,需要通过知识融合将不同来源的知识图谱融合在一起,实体对齐作为知识融合的一部分,可以有效地对不同来源的实体进行匹配,在实体层面建立起联系。

中铁物贸集团供应链协同平台(简称协同平台)是其大宗铁路建设物资采购的核心电商交易平台,本文将以该平台为具体场景,尝试构建行业电商

① 国家自然科学基金(61972040),中铁物贸集团鲁班公司科技研究开发计划课题,北京市教育委员会科研计划(KM201911417010)和北京联合大学校内科研专项课题(ZB10202004)资助项目。

② 男,1998 年生,硕士生;研究方向:数据挖掘,知识图谱和机器学习;E-mail:cfq2828@163.com。

③ 通信作者,E-mail:xxtmingming@buu.edu.cn。

(收稿日期:2021-12-24)

知识图谱替代传统的数据系统,通过知识推理来满足该电商平台中跨系统的大宗物资数据的匹配等业务需求。前文所提消费类电商平台主要是面向普通用户,为用户提供搜索和推荐服务、满足用户的消费需求;而行业电商平台所面向的客户大多为公司或企业,是一个为特定用户提供信息流通和交易的平台,具有行业特殊性。

考虑到钢材类物资是协同平台的主要交易品类(占总交易额的80%以上),本文以钢材类物资数据为例。具体来说,基于领域知识图谱构建框架,研究如何从各类结构化和半结构化数据中,基于已抽取的各类实体和属性,结合领域知识建立各类实体尤其是中铁物资实体与钢厂产品实体间的匹配关系。本文基于行业内的领域知识进行数据预处理,包括数据转换、去噪、规范性校验等,还包括根据领域知识建立领域字典,以提高数据质量、提升数据处理的效率和准确性。结合钢材领域知识,过滤非配对实体,缩小实体对齐计算空间,生成高质量候选集。将集合相似度的思想应用到实体对齐的过程中,综合实体名称、属性和行业领域信息,定义实体相似度评估函数,有效评估实体对的匹配概率,形成了一种基于领域知识的集合相似度实体对齐算法。应用不同数据集的实验结果表明,本文提出的基于领域知识的集合相似度实体对齐算法对不同钢厂产品实体和中铁物资实体间的匹配具有较高的准确率。

1 相关工作

1.1 行业电商知识图谱的构建框架

知识图谱的构建方法可以分为自顶向下构建和自底向上构建2种。自顶向下的构建方法通常由领域专家设计好本体模型或者由行业内固定的知识体系转化成为本体模型,然后依照本体模型从不同来源的数据中抽取实体和关系融入到知识图谱中;自底向上的构建方法先进行实体和关系挖掘然后从实体中抽象出概念和本体模型。自顶向下和自底向上的方法的最大的不同就是本体的由来,自顶向下的方法是先人工提供本体再导入数据,自底向上的方法是先导入数据再从数据中抽象出本体。因此,这2种不同方法的适用对象也是不同的,自顶向下的方法面向的是知识体系比较成熟的领域,而自底向上的方法更多面向的是知识体系欠完备的领域,使用基于数据驱动的方式进行构建[5]。以自底向上的方法为例,其过程主要包括知识抽取、知识融合和知识加工[6]。知识抽取从各种类型的数据源中进行实体、属性和关系的抽取;知识融合通过融合实体消除矛盾和歧义,通过知识合并融合外部数据源;知识加工主要进行本体构建、知识推理。通过实体的聚合生成上层的概念,逐层抽象形成本体模型。通过知识推理挖掘语义层次的关联信息,发现新的知识。

因为协同平台所涉及的钢材类物资信息具有比较确定的上下文结构,其知识图谱可采用自顶向下的方式构建。先确定知识图谱的数据模型,再根据模型去填充具体数据,最终形成知识图谱,本文称之为行业电商知识图谱。

其构建框架如图1所示。该平台现有物资数据以结构化形式管理存储,第三方系统(主要指各钢厂的信息系统)则形式比较多样化,部分钢厂为结构化数据,但部分属性字段为半结构化和非结构化描述。实体对齐是构建该图谱的核心,是实现行业电商平台与外部系统之间数据柔性对接和顺畅流转的基础,可为平台物资采购交易和交付等业务提供底层数据支撑。

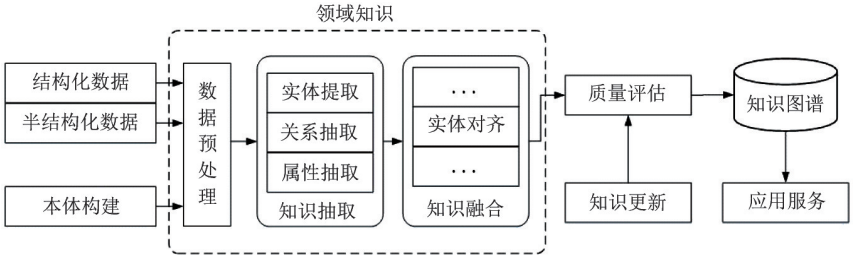

图1 领域知识图谱构建框架

1.2 实体对齐

实体对齐是挖掘不同数据源中指向现实世界中同一对象的关键技术[7]，其思想最早可追溯至 1946 年 Dunn[8]所提的记录链接，记录链接可以判断相同或者不同数据集中的 2 个实体是否指向真实世界中的同一个对象。实体对齐又可以被称为实体匹配、链接预测、对象识别等，能够高质量链接多个现有知识库，将多来源数据建立联系，并构建一个大规模统一的知识库，从而帮助机器理解底层数据，实现底层数据的有效融合。因此，实体对齐具有广泛的应用，如在电商平台中进行相似商品的推荐，在社交平台中进行好友的推荐，在搜索引擎中进行数据的搜索等。

实体对齐的常见方法可以分为 2 类，一类是基于实体表示学习的实体对齐方法，另一类是基于属性信息的实体对齐方法。基于实体表示学习的实体对齐方法通常先将实体抽象为向量表示，然后在向量空间中学习实体的对齐关系[9]。基于机器学习、深度学习的方法在进行大规模实体和关系的学习中进行了充分探索，TransE[10]作为一种有效的实体表示方法得到了广泛使用并衍生出多个模型（如 STransH、TransD、NTransGH 等）以实现复杂的实体和关系映射，一些基于图神经网络（graph neural networks，GNN）的方法[11]已被尝试应用于实体对齐以加强实体表示学习能力。基于属性信息的实体对齐方法通过属性的字符串相似性或者值的相似性判断 2 个实体是否可以对齐。基于字符串的 Jaccard 相似度和基于编辑距离的方法在实体对齐中得到了很好的应用。

基于表示学习的实体对齐方法通常需要一定的标注数据，但是由于在多个来源的知识图谱间进行标注需要较强的专业性，而这类方法对于标注数据又有很强的依赖性，在多源异构数据中学习的模型往往不能发挥很好的效果。在知识融合初期缺乏标注数据的情况下基于属性信息的实体对齐方法则具有更强的通用性。

基于 Jaccard 相似度的方法由于其效果显著、简单而被广泛应用。该方法已经在多种不同知识图谱的任务中进行了实验，其适用性、效率、性能得到了验证。文献[12]基于 Jaccard 和余弦相似度进行相似度的度量，在电影推荐系统中起到了较好效果，同时通过实验证明了基于余弦相似度的方法具有较强的鲁棒性。文献[13]基于 Jaccard 相似度和混合优化模型进行了跨领域的本体构建，提出的算法具有较高性能。文献[14]基于 Jaccard 相似度在大规模知识库中根据实体属性信息进行实体对齐，并用实验证明了该方法具有较高的精确度和效率，但是其算法依赖于邻域信息，而本文面向的物贸领域缺乏有效邻域信息，因而此方法不适用。文献[15]基于词频比改进 Jaccard 相似度，按照词频比确定交集中词的权重，在对文本分词后使用该方法计算文本的相似度，在本文中也使用了该方法进行测试。相关研究已经证明了 Jaccard 相似度具有较强的适用性，本文针对电商领域相关数据多源异构的问题提出了一种新的基于 Jaccard 方法，接下来对具体问题和所设计方法进行描述。

2 问题描述

行业电商领域涉及业务部门多、上下游产业规模大，其智能化电商平台底层数据往往来源广泛、数据标准不统一、质量参差不齐。如图 2 所示，在订单流转的过程中行业电商经常面临着数据对接的问题，不同供货方提供的数据存在较大的差异，甚至同一供货方数据也存在异构，电商平台的数据描述规范和供货方的数据规范也有很大区别，现有的方法通常需要人工进行数据匹配，从而在不同数据源间建立映射。如表 1 所示，某钢厂的原始数据存在着大量的异构问题，如包含属性字段中复合字段造成的异构、单位描述规范不统一造成的异构、属性字段冗余造成的异构等。人工进行数据的处理往往需要具备大量专业背景知识，同时人工处理大批量数据

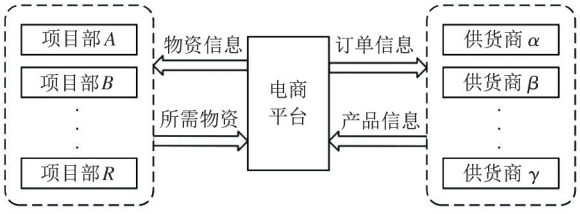

图 2 订单生成过程中数据流转示意图

表 1　某钢厂产品数据

物料编码	物料描述	品种	规格	分类
BHRB5E025003T	抗震带肋钢筋 HRB500E Φ25 * 12 000 (型钢生)	BHRB5E0	Φ25 * 12	1
JHQ235B040020001	H 型钢 Q235B H400 × 200 × 8 × 13 12M	H 型钢 Q235B	H400 × 200	2
BHRB5E0200003	抗震带肋钢筋 HRB500E Φ20 * 12 000	BHRB5E0	Φ20 * 12	1
BHRB4E0120021	六五零 抗震带肋钢筋 HRB400E Φ12 * 9000	BHRB4E0	Φ12 * 9	1
JIQ235B025B02	工 Q235B I250 × 118 × 10 × 13 9M	工字钢 Q235B	工 25B 国标	2

的效率较低,并且数据处理的准确性难以把控。

基于知识图谱的方法对数据进行处理和存储,把上述订单流转过程中的数据匹配问题转化成行业电商知识图谱中供货方和电商平台的实体匹配任务,也即实体对齐。本文以中铁物贸集团供应链协同平台为具体场景,完成协同平台中跨系统的大宗物资数据的匹配。

3　基于领域知识的实体对齐方法

3.1　算法框架

本文实体对齐旨在供应商提供的产品实体和电商交易平台的物资实体中找出指向同一个对象的实体对。算法的工作流程如图 3 所示。

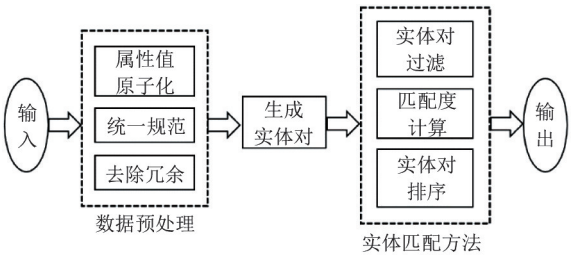

图 3　算法工作流程图

实体对齐方法主要包括以下 3 个步骤。

(1)数据预处理。针对多源异构知识库实体属性数据存在的问题进行统一处理,根据领域知识进行实体属性值的原子化分割、统一单位、规范表示及去除冗余,得到处理后的知识库。

(2)生成实体对。在供应商产品知识库的实体和电商交易平台物资知识库中的实体间计算笛卡尔积,得到实体对集合。

(3)实体匹配。遍历实体对集合,基于领域知识过滤不存在匹配关系的实体对;计算实体对相似度,根据不同的实体对之间的相似度进行排序,得到与供应商产品实体 $E^F = \{key\_F1:value\_F1, key\_F2:value\_F2, \cdots, key\_Fn:value\_Fn\}$ 最为匹配的电商平台物资实体 $E^M = \{key\_M1:value\_M1, key\_M2:value\_M2, \cdots, key\_Mm:value\_Mm\}$,输出存在匹配关系的实体对 (E^F, E^M)。

3.2　数据预处理

行业电商领域涉及业务部门多、上下游产业模大,信息存在不对称性,并且各行业术语存在特殊性,这些导致行业电商平台底层数据来源广泛且异构、数据标准不统一、质量参差不齐,因此无法直接进行实体对齐,需对数据进行预处理。以中铁物贸集团供应链协同平台为例,其数据来源有昆明钢铁厂钢产品数据、陕西钢铁厂钢产品数据、中铁钢材物资数据等。不同钢厂有各自对钢铁产品数据的描述规范,中铁也有其对钢材物资的描述规范,这必然导致协同平台底层数据的多源异构性,其存在的主要问题有:数据的量纲不统一、数据中含有的单位描述不规范以及属性字段重复造成冗余。针对上述问题,本文在数据预处理阶段,根据领域知识建立领域字典(产品名称字典 *nameDicList*、分词词典 *dicList*、停用词典 *stopDicList*、替换词典 *replaceDicList* 和删除词典 *deleteDicList*)对数据进行规范化处理,包括属性值原子化、统一单位、规范表示、去除冗余等,以提高实体对齐的准确率。具体方法如下。

(1)属性值原子化

知识库 $S = \{S_1, S_2, \cdots, S_N\}$,$S_N$ 表示电商平台知识库,$S_1, S_2, \cdots, S_{N-1}$ 为供应商知识库,每个知识库包含多个钢产品实体,知识库中对钢产品属性的划分不统一,不同知识库实体属性的描述存在较大

差异。如表 1 所示,钢厂的物料描述字段内包含多个关键参数,字段内包含的参数直接影响和这个钢产品匹配的中铁物资,如果不对其处理,实体属性值无法直接进行字符串比较,从而导致实体对齐失败。对于一个属性融合多类属性值的属性值融合异构现象,使用自然语言处理中基于规则的正向最大匹配分词技术对不同类别属性值进行划分,虽然处理的不是自然语言,但是对于这类半结构化数据可以通过引入分词词典 *dicList* 和停用词典 *stopDicList* 进行分词,将属性值原子化。然后使用替换词典 *replaceDicList* 将表示相同含义的字符进行统一替换,对类似表 1 中规格字段的值进行拆分,最后删除多余的字符。经过处理的数据如图 4 所示。

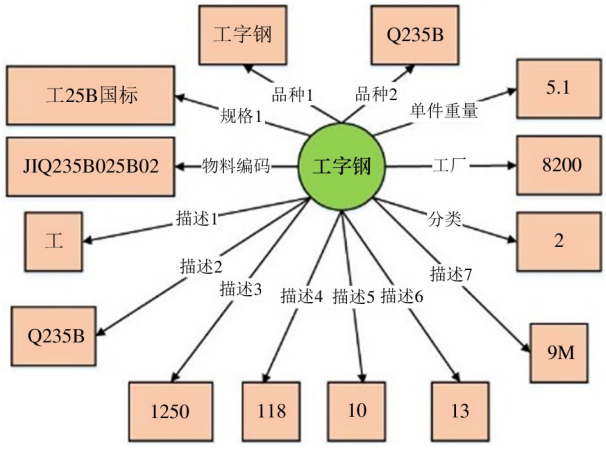

图 4　经过分词处理之后的数据

(2)统一规范

知识库 $S=\{S_1,S_2,\cdots,S_N\}$ 中数值型属性值的单位描述和书写规范均不统一。单位描述的不统一将导致数值型属性值存在较大差异,单位书写的不统一将会影响字符串的匹配结果,因此,直接进行属性匹配将会导致属性匹配的失败。对于单位不统一的数据,统一使用量纲较大的单位,长度的单位使用米、质量的单位使用千克。然后,统一去除所有属性值 *value* 中含有的单位,确保相同的数据可以进行精确匹配。

(3)去除冗余

知识库 $S=\{S_1,S_2,\cdots,S_N\}$ 中原始属性字段中存在冗余,实体属性值经过拆分后变成了一个个独立的属性值,从而出现了属性重复的现象;同时属性字段存在多个索引列,冗余的索引值在实体匹配过程中也会被作为一个重要的属性值。在实体对齐的过程中,冗余的属性会作为重要的属性参与属性值的匹配,从而影响了最终匹配结果。算法通过比较列的相似性去除具有相似内容的列,通过判断列内容的唯一性去除冗余索引列。

数据预处理具体流程为对属性值进行原子化操作,去除冗余索引列,根据钢产品实体名称将知识库进行划分,在划分之后的知识库中进行实体属性度量单位的转换,删除重复属性和冗余列,相应算法伪代码如算法 1 所示。

算法 1　数据预处理

输入:知识库实体集合 $S=\{S_1,S_2,\cdots,S_N\}$,产品名称字典 *nameDicList*、分词词典 *dicList*、停用词典 *stopDicList*、替换词典 *replaceDicList* 和删除词典 *deleteDicList*

输出:预处理后的实体集合 $\tilde{S}=\{\tilde{S}_1,\tilde{S}_2,\cdots,\tilde{S}_N\}$

1) 基于 *dicList* 和 *stopDicList* 进行分词
2) 通过 *replaceDicList* 统一 *value* 中表示
3) 对 *value* 进行字段拆分
4) 根据 *deleteDicList* 删除字符
5) 删除多余的索引属性
6) 利用 *nameDicList* 提取名称集合 *name _ set*
7) 根据 *name _ set* 将 *entitySet* 分块为 *entityBlocks*
8) for *entityBlock _ i* in *entityBlocks*
9)　　对 *value* 进行单位转换
10)　　删除重复的属性
11) end for

3.3　基于集合相似度的实体匹配

在知识库间进行笛卡尔积得到实体对集合 *entitySet*1($entitySet1=\{(E^F,E^M)\mid E^F\in\{\tilde{S}_1,\tilde{S}_2,\cdots,\tilde{S}_{N-1}\},E^M\in\tilde{S}_N\}$)。基于集合相似度的实体匹配旨在从 *entitySet*1 中找出指向同一对象的实体对,主要包括 3 个步骤,即实体对过滤、匹配度计算以及实体对排序。通过实体对过滤实现实体对筛选,缩小实体对齐规模,提高实体匹配效率。然后计算实体相似度,以此衡量不同知识库的 2 个实体是否指向了同一个对象。最后,基于实体对的实体相似度排序输出最匹配的实体对。如图 5 所示,在得到匹配的实体对之后,就可以在实体之间建立匹配关系,表示 2 个实体指向了真实世界中的同一个对象。

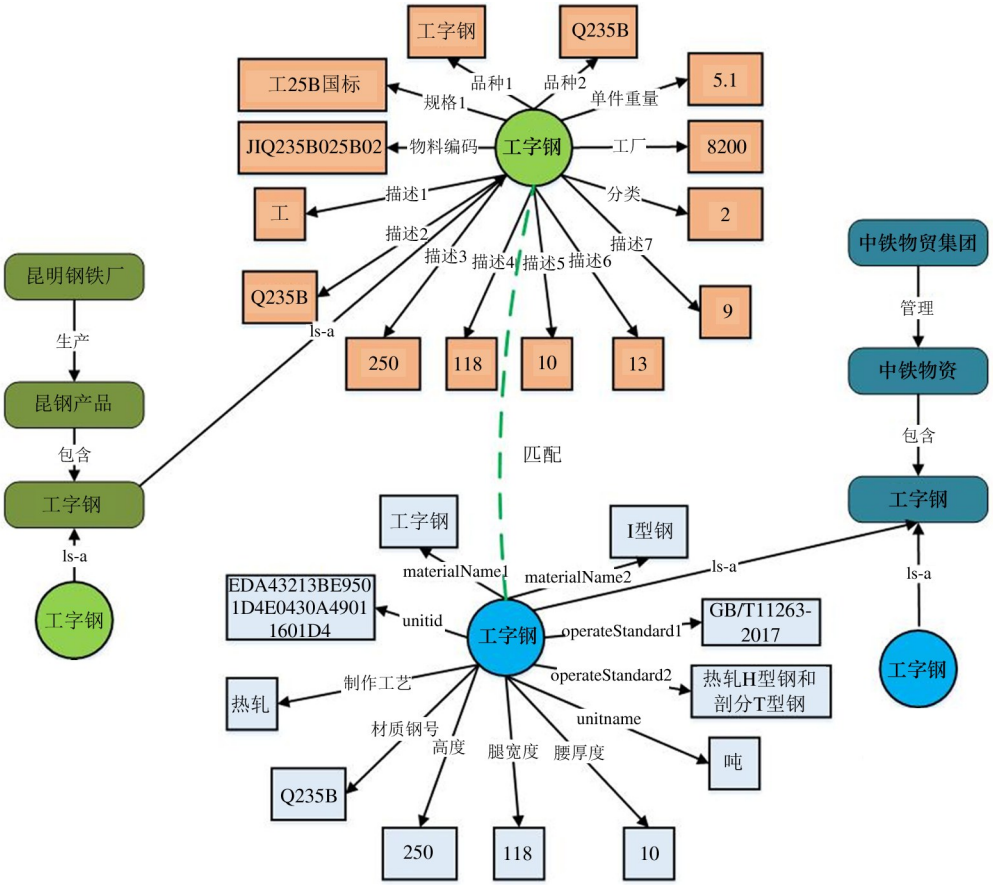

图 5　实体匹配示意图

(1)实体对过滤

为了提高实体对齐效率,可根据实体关键属性对实体进行划分,减少无效匹配,以提升匹配效率。在钢材交易领域,钢产品名称和钢号可以作为判断 2 个实体是否指向同一个对象的关键属性。因此,通过钢产品名称字典找出实体名称 *value _ name*,根据实体名称的 Jaccard 相似度进行排序,过滤掉名称相似度低于阈值的实体对。如果 2 个实体 E^F 和 E^M 满足匹配条件则记为 $E^F \cong E^M$。如果所有的实体对中没有名称相似的实体对,则进行全局实体对齐。为提高全局匹配效率,使用领域字典抽取出材质或钢号字段 *value _ grade*,根据 *value _ grade* 进行实体对过滤,得出新的实体对集合 $entitySet2(entitySet2 = \{(E^F, E^M) \mid E_t^F \cong E_l^M,\ E^F \in \{\tilde{S}_1, \tilde{S}_2, \cdots, \tilde{S}_{N-1}\},\ E^M \in \tilde{S}_N\})$。

(2)匹配度计算

匹配度计算旨在量化实体对 (E^F, E^M) 的相似度(记为 $sim(E^F, E^M)$),通过匹配度的值 $sim(E^F, E^M)$ 可以对不同实体对进行比较找出存在匹配关系的实体对。钢厂钢产品实体和中铁物资实体的属性字段具有天然的集合结构,属性的先后顺序不影响实体的匹配结果,因此可用 Jaccard 相似度(式(1))来度量 2 个实体的匹配度。

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

其中 A 和 B 是 2 个实体属性集合。由于传统 Jaccard 相似度的计算中分母会随着实体属性数量的变化而变化,算法不能直接利用 Jaccard 相似度值进行比较。因此,本文在实体匹配方法设计中,首先统计钢厂产品实体属性值与中铁物资实体属性值相同的属性个数,即 $\sum_{k=1}^{n} f(value\_Fk)$,其中 $f(value\_Fk)$ 表示 E^F 的属性值 *value _ Fk* 是否出现在 E^M 的属性值集合 Ω 中,其取值如式(3)所示,出现时取值为 1、否则为 0;然后将相同的属性个数除以钢厂钢产品属性个数(n)得到实体对匹配度 $sim(E^F, E^M)$。计算公式如式(2)所示,$sim(E^F, E^M)$ 表示实

体 E^F 和 E^M 的相似度。本算法中对相似度计算的优化可保证不同实体对评判标准的统一性,使得相似度计算变得更加合理。

$$sim(E^F,E^M)=\frac{\sum_{k=1}^{n}f(value\_Fk)}{n} \tag{2}$$

$$E^F\in\{\tilde{S}_1,\tilde{S}_2,\cdots,\tilde{S}_{N-1}\},\ E^M\in\tilde{S}_N$$

$$f(value\_Fk)=\begin{cases}1 & value\_Fk\in\Omega\\ 0 & \text{其他}\end{cases}$$

$$\Omega=\{value\_M1,\ value\_M2,\cdots,value\_Mm\} \tag{3}$$

(3)实体对排序

由于中铁物资实体和实体属性值具有多样性,2 个实体对的匹配度相同不代表这 2 个实体对同时匹配。因此,本算法在获得实体对匹配度之后,针对性设计有效实体对排序规则对实体对进行排序,对供应商任意产品实体 E^F 取排名第一的实体对 (E^F,E^M) 作为最佳匹配,然后将其保存至实体匹配集合 $entitySet3$($entitySet3=\{(E^F,E^M)\mid E^F=E^M,E^F\in\{\tilde{S}_1,\tilde{S}_2,\cdots,\tilde{S}_{N-1}\},E^M\in\tilde{S}_N\}$)。基于领域知识的实体对排序规则具体为:首先计算实体 E^F 的匹配阈值 $\frac{1}{\|E^F\|}$,$\|E^F\|$ 表示实体 E^F 的属性个数;然后选择与实体 E^F 匹配时 $sim(E^F,E^M)$ 大于或等于阈值的实体对 (E^F,E^M),通过该阈值的设定可排除明显不匹配实体对;最后对所选实体对 (E^F,E^M) 按 $sim(E^F,E^M)$ 值从大到小规则进行排序,当 $sim(E^F,E^M)$ 值相同时,按匹配对象 E^M 的属性字段长度 $\|E^M\|$ 从小到大排序;当 $\|E^M\|$ 值相同时,考虑中铁物资实体 E^M 属性值越丰富匹配越可靠,按实体 E^M 不同属性值个数从大到小排序。

基于集合相似度实体匹配的具体算法如算法 2 所示,基于算法可找出存在匹配关系的实体对,然后在实体对间建立联系,建立联系后的实体匹配示意图如图 5 所示。

算法 2 基于集合相似度的实体匹配算法

输入:实体对集合 *entitySet*1、钢号字典 *steelGradeList*、钢产品名称字典 *nameDicList*

输出:实体对集合 *entitySet*3

1) for E^F in $\{\tilde{S}_1,\tilde{S}_2,\cdots,\tilde{S}_{N-1}\}$
2) 根据 *nameDicList* 抽取 E_i^F 名称 value _ *name*
3) 根据 value _ *name* 对 *entitySet*1 进行实体对过滤,得出 *entitySet*2
4) if *entitySet*2 为空
5) 根据 *steelGradeList* 得出 E^F 属性值中钢号 *value _ grade*
6) 根据 *value _ grade* 进行实体对筛选保存到 *entitySet*2 中
7) 计算 $\sum_{i=1}^{n}f(value\_Fi)$
8) 计算实体对相似度 $sim(E^F,E^M)$
9) 选择 $sim(E^F,E^M)\geqslant 1/\|E^F\|$ 的实体对,并根据实体对排序规则进行排序,保存排名第一的实体对到 *entitySet*3 中
10) end for

4 实验分析

在本节中使用了上文提出的基于领域知识的集合 相似度实体对齐算法,在5个不同的钢铁厂的钢材数据和中铁物资数据上进行了测试。本实验使用的服务器的处理器为 Intel(R) Xeon(R) CPU E5-2680 v4@ 2.40 GHz,内存为 256 GB,操作系统为 CentOS 7,使用 Python 语言进行模型的搭建和实验。

4.1 数据集

实验使用协同平台的中铁钢材类物资数据和昆钢、南钢等 5 个钢厂的钢产品数据,数据类别和实体数量如表 2 所示。

表 2 实验数据汇总

	类别数量	实体数量
中铁钢材物资	386	48 218
昆钢钢产品	14	1089
南钢钢产品	2	48
攀钢钢产品	2	336
陕钢钢产品	34	497
韶钢钢产品	4	151

每个钢铁产品有多种不同的属性,不同数据源中描述属性的名称和属性字段长度存在较大差异。如昆钢钢产品的属性字段中存在着属性值融合的字段,同时还存在着冗余的数据。由于各个数据源的数据表述标准不统一,因此同一钢产品的描述存在

较大差异。数据中还存在着一些自定义的标识数据,从而影响了数据的质量。

4.2 性能评价指标

Hits@k 表示实体对齐的前 k 个实体中对齐到正确的实体的比例。平均倒数排名(mean reciprocal rank,MRR)为匹配到正确实体的排名的倒数的平均值,计算公式如式(4)所示,其中 $rank_i$ 表示在对齐到正确的实体的排名,N 为实体的总数。

$$MRR = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{rank_i} \tag{4}$$

一个钢产品实体如果存在对应的物资实体则说明这个钢产品实体是一个正例,如果不存在对应的物资实体说明这个钢产品实体是一个负例。在最大相似度准则下,如果一个钢产品的相似度最高的实体对的相似度值超过指定阈值并且准确匹配到了对应的物资实体则说明这个钢产品的预测结果是正例。为了验证阈值的影响,设定不同的阈值,比较不同阈值下的各项指标的变化,得出最佳的参数。根据真实标签和预测标签可以计算出查全率(recall,简称 R)、查准率(precision,简称 P)和 F1 度量。

4.3 实验结果分析

为验证本文算法的正确性和有效性,利用 5 个不同钢厂数据进行实验。对比算法包括基础算法(对数据采用属性值原子化处理后直接用 Jaccard 相似度计算实体匹配度)、Jaccard 算法(使用 3.2 节所述数据预处理方法处理数据,采用 Jaccard 相似度计算实体匹配度)、词频比算法(文献[15]中使用的算法)和本文所提算法,对比指标为 Hits@1 和 MRR。4 个对比算法中均包含实体属性值原子化处理,旨在解决实际数据中实体单个属性融合多类属性值(如表 1 所示)而导致的字符串无法直接匹配问题。

如表 3 所示,本文所提算法总体实体对齐的 Hits@1 值达到 96.4%,明显高于其他 2 种算法。在表 4 中,本文所提算法的 MRR 值也均高于其他算法,说明本文算法可以更有效地识别出存在匹配关系的实体对。文献[15]提出的基于词频比的算法在计算中考虑到了不同词出现的频率,但是在相似度计算公式中仍然使用了并集,影响了不同实体间匹配的准确率,因此也验证了本文方法的有效性。进一步看,对于本文所提算法,南钢、攀钢、韶钢 3 个钢厂实体对齐的 Hits@1 均已达到了 100%,由于昆钢和陕钢异构数据较多、数据描述复杂,Hits@1 为 93.3% 和 96.3%,该结果同样高于其他 2 个算法。此外,从韶钢的数据可以看出,使用仅含普通数据预处理方法的基础算法,Hits@1 只有 49.3%,这也说明了本文所设计数据预处理算法的有效性。

表 3 不同钢厂的实体匹配 Hits@1 对比

	基础算法	Jaccard 算法	词频比算法	本文算法
昆钢	86.5%	87.9%	86.2%	93.3%
南钢	95.7%	95.7%	95.7%	100%
攀钢	100%	100%	100%	100%
陕钢	92.5%	94.1%	94.1%	96.3%
韶钢	49.3%	96.7%	96.7%	100%
总体	86.9%	93.1%	92.4%	96.4%

表 4 不同钢厂的实体匹配 MRR 对比

	基础算法	Jaccard 算法	词频比算法	本文算法
昆钢	0.905	0.913	0.893	0.959
南钢	0.975	0.975	0.957	1.0
攀钢	1.0	1.0	1.0	1.0
陕钢	0.950	0.960	0.955	0.977
韶钢	0.652	0.982	0.967	1.0

表 3 列出了实体匹配的 Hits@1 值。如果放宽查找范围,在相似度最大的前 k 个候选实体中如果存在对应的物资实体,则视为匹配成功,由此可以得出在不同 k 值下的实体匹配的 Hits@k 值。从图 6 可以看出,随着 k 值的逐渐增大,实体匹配的准确率逐渐升高,陕钢实体匹配的 Hits@5 已经达到了 100%,昆钢实体匹配的 Hits@5 已经超过了 98%,这个结果也再次验证了本文算法的有效性。

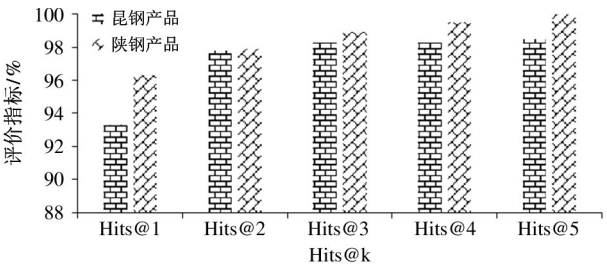

图 6 不同尺度下实体对齐查全率

进一步地,为验证使用匹配度阈值对实体对齐准确性的影响,本文比较了不同阈值下模型的各项性能指标。以下实验将针对昆钢和陕钢的数据进行分析。

根据实验的结果可得出阈值为 0.35 时本文算法在昆钢数据中表现出色,阈值为 0.28 时本文算法在陕钢的数据上表现良好。

本文实验从不同角度对模型的有效性进行了验证。首先比较了最大相似度准则下的实体对齐 Hits@1 和 MRR,然后在 Hits@1 偏低的钢厂数据中又再次比较了不同 k 值下的 Hits@k,通过不同算法对比可以看出本文所提算法达到了较高的准确率。然后从相似度值的角度进行衡量,探索阈值对于实验结果的影响。结果表明,算法的性能有所下降,原因是钢厂不同钢产品属性描述存在差异,直接使用统一的阈值进行实体对齐的衡量导致部分数据无法有效进行匹配。图 7、图 8 为钢厂数据在设定不同阈值时评价指标的变化趋势图。图 7 中的查准率折线随着阈值的升高出现了明显的波动,这也反映出基于阈值的实体对齐存在一定问题,后期将会探索更科学的实体对齐的衡量尺度,从多角度科学地衡量实体对的匹配关系,进一步提高实体对齐的准确度。

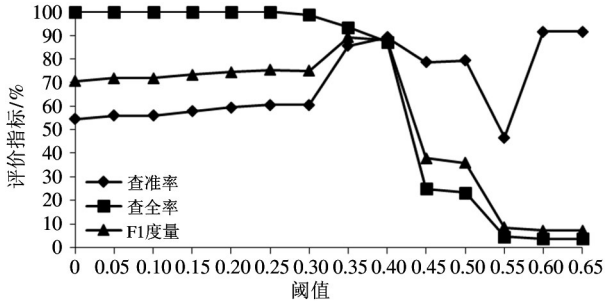

图 7 不同阈值下昆钢的实验结果

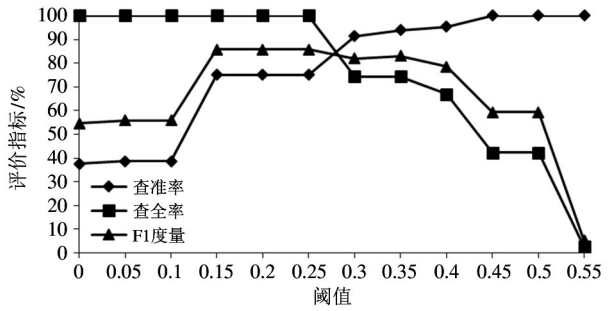

图 8 不同阈值下陕钢的实验结果

5 结 论

本文面向行业电商知识图谱应用,将电商平台底层多源异构数据中同一对象的匹配问题转化为知识图谱领域中实体对齐问题,并提出一种基于领域知识的集合相似度实体对齐算法,实现数据融合。在算法设计中,基于领域知识针对性设计数据预处理技术,以规范化电商底层多源异构数据、提升数据处理效率和准确性;结合钢材领域知识,过滤非配对实体,缩小实体对齐计算空间,生成高质量候选集;应用集合相似度思想,综合实体名称、属性和行业领域信息,定义实体相似度评估函数,有效评估实体对的匹配概率。实验结果表明,本文算法可有效提高实体匹配准确率,实现多源异构数据的高效融合。本文进一步从基于相似度值排序的度量和基于阈值的度量 2 个方面比较了实体对齐的匹配标准,探索了不同阈值对实体对齐的影响。

在未来研究中,将会融合不同标准进行实体对齐的度量,同时也会从知识图谱的拓扑结构出发,探索基于图神经网络等方法进行实体对齐,进一步提高实体对齐的准确度,减少人工干预,提高行业效率。

参考文献

[1] 李娜,金冈增,周晓旭,等. 异构网络中实体匹配算法综述[J]. 华东师范大学学报(自然科学版),2018(5):41-55

[2] LUO X S, LIU L X, YANG Y H, et al. AliCoCo: Alibaba E-commerce cognitive concept net[C]// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, Portland, USA, 2020: 313-327

[3] WANG H W, ZHANG F Z, ZHANG M D, et al. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, USA, 2019:968-977

[4] 庄严,李国良,冯建华. 知识库实体对齐技术综述[J]. 计算机研究与发展,2016,53(1):165-192

[5] 王昊奋,漆桂林,陈华钧,等. 知识图谱方法、实践与应

用[M]. 北京:电子工业出版社,2019:425-426
[6] 韩蕊. 阿里巴巴 B2B 电商算法实战[M]. 北京:机械工业出版社,2020:254-256
[7] 王凌阳,陈钦况,寿黎但,等. 多源异构数据的实体对齐方法研究[J]. 计算机工程与应用,2019,55(19):87-95,152
[8] DUNN H L. RecordLinkage[J]. *American Journal of Public Health and the Nations Health*, 1946, 36(12): 1412-1416
[9] 朱继召,乔建忠,林树宽. 表示学习知识图谱的实体对齐算法[J]. 东北大学学报(自然科学版), 2018,39(11): 1535-1539
[10] BORDES A, USUNIER N, GARCÍA-DURÁN A, et al. Translating embeddings for modeling multi-relational data [C] // Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2013: 2799-2807
[11] CAO Y X, LIU Z Y, LI C J, et al. Multi-channel graph neural network for entity alignment[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 1452-1461
[12] KERMANY N R, ALIZADEH S H. A hybrid multi-criteria recommender system using ontology and neuro-fuzzy techniques[J]. *Electronic Commerce Research and Applications*, 2017, 21: 50-64
[13] KAKAD H, DHAGE S. Cross domain-based ontology construction via Jaccard semantic similarity with hybrid optimization model[J]. *Expert Systems with Applications*, 2021, 178: 115046
[14] LACOSTE-JULIEN S, PALLA K, DAVIES A, et al. Sigma: simple greedy matching for aligning large knowledge bases[C] // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, USA, 2013: 572-580
[15] 谢红. 基于词频比的改进 Jaccard 系数文本相似度计算[J]. 内江科技,2021,42(8):27-28

Research of entity alignment in the B2B e-commerce knowledge graph

CHEN Fuqiang*, XIAO MingMing*, HAN Kainan**, REN Yi***, WANG WenWen***, LI Ke*
(*Smart City College, Beijing Union University, Beijing 100101)
(**China Railway Material Trade Group Co Ltd., Beijing 102308)
(***Luban (Beijing) Electronic Commerce Technology Co Ltd., Beijing 102308)

Abstract

Aiming at the entity alignment problem in the fusion of multi-source heterogeneous knowledge graph data, this paper is oriented to the real data of the e-commerce platform in the industry e-commerce field, and proposes an entity alignment algorithm based on domain knowledge of the set similarity. First, data pre-processing techniques, such as atomizing property value, unifying terminology, and removing redundancy, are specifically designed based on domain knowledge to normalize the multi-source heterogeneous data at the bottom of e-commerce, thus improving the accuracy of data application. Then, considering the application of B2C e-commerce knowledge graph, an effective and efficient entity matching method is proposed, which mainly consists of selecting high-quality pairs of entities and sorting them by optimizing set similarity evaluation function. The experimental results show that the proposed algorithm can effectively improve the accuracy of data fusion, reduce workload, and can provide new ideas for the development of the industry.

Key words: multi-source heterogeneous data, knowledge graph, entity alignment, set similarity, e-commerce