

面向关联数据的实体链接发现方法研究^{*}

高劲松 周习曼 梁艳琪

摘 要 随着关联数据应用的不断深入,已有众多的数据集发布在网上,但目前已发布的关联数据集之间关联很少,为数据的共享使用带来不便。本研究提出一种基于统计学习方法进行关联数据集间实体识别及链接构建的方法。首先进行数据集间的实体匹配,采用基于 K 中心点聚类算法实现属性的聚合及关系发现,对具有高相关度的属性进行匹配关系描述,降低实体匹配时的属性匹配计算次数;其次对已匹配的属性进行实体属性值的相似度比较计算,实现实体间相似度的比较,在 SILK 框架下实现实体的链接构建工作,以达到实体链接发现的目的;最后通过实验验证,这一方法能降低数据集间实体匹配计算次数,提高实体链接的正确率,具有可行性及实用性。图 12。表 4。参考文献 19。

关键词 关联数据 实体链接 数据链接 链接发现

分类号 G254

Linked Data-Oriented Method of Entity Linking Discovery

GAO Jinsong, ZHOU Ximan & LIANG Yanqi

ABSTRACT

The World Wide Web has been developed into a global data space, which links web data and database data. Linked data is one of the best tools to achieve this information evolution. Linked data publish data in a structured form to interlink resources. With the depth of linked data being deeply applied, more and more data are published on the web as linked data. The published web information also has been transformed into linked data in automatic or semi-automatic ways. Practically, there are still only a few connections between the released linked dataset, and it is inconvenient to share data. So based on the entity linking discovery, we can discover the real relation between entities, build the entity linking according to the publishing standard, realize the goal of discovering potential entity linking, enhance the interlinking between datasets, and then increase the accuracy of published linked data.

In this thesis, a statistical learning method is proposed to recognize entities and build links across different linked datasets. Before the entities comparing computation, first, the method finds class correspondences to classify related entity attributes correspondences across datasets. It gives a matching relationship description for the high correlation attributes and reduces the calculation times to match entity attributes. Second, our

^{*} 本文系国家社会科学基金一般项目“基于关联数据的知识创造中知识外化和融合机制研究”(编号:12BTQ039)的研究成果之一。(This article is an outcome of the project “The Mechanisms of Externalization and Combination of Knowledge Creation Based on Linked Data”(No. 12BTQ039) supported by National Social Science Foundation of China.)

通信作者:高劲松, Email: jsgao@mail.ccnu.edu.cn, ORCID: 0000-0003-0022-5923 (Correspondence should be addressed to GAO Jinsong, Email: jsgao@mail.ccnu.edu.cn, ORCID: 0000-0003-0022-5923)

method compares the similarity of entities based on calculating the similarity of the matched attributes, and builds entities' linking to complete the goal of linking discovery across different datasets. When to cluster the attributes correspondences, we use K-medoids clustering algorithm to discover the potential attributes correspondences. K-medoids clustering algorithm is mainly aimed at classifying property concepts and corresponding attributes that represent the same expression meanings between datasets. At last, the attributes can be compared and matched in groups. Then EDOAL language is used to define the clustered attributes and describe the correspondences relation between those attributes. According to the matching relation, we compare and calculate the similarity between entity attributes. Finally our method works out the linking under the SILK framework: mapping the property relationship to SILK scripts, building entities linking between datasets according to a preset confidence value, endowing entities with RDFs properties, and realizing entity links discovery between datasets. The thesis testifies different open linked datasets on the basis of linked data entity linking discovery method. The datasets mainly include IM@ OAIE2014(dataset Abox3) 、CKAN(dataset EUROSTAT) and GADM-RDF(dataset GADM), and data are used to cluster matched attributes and interlink entities. Through twice entity linking discovery process of experimental verification, experimental results show that K-medoids clustering algorithm calculates the similarity of entities matching between dissimilar properties can increase the number of entities links. The method already reaches the high accuracy rate and F values. So the proposed method can reduce the calculation times of matching entities across different datasets and improve the accuracy of physical links. It has high feasibility and practicability to solve this problem. 12 figs. 4 tabs. 19 refs.

KEY WORDS

Linked data. Entity linking. Data linking. Linking discovery.

0 引言

目前,万维网已经发展成为一个将网页数据及数据库数据链接起来的全球数据空间,关联数据是实现这一信息进化的最佳工具之一^[1]。关联数据是将数据以结构化的形式在网络上发布,实现资源的互联。它的本质及目标是,数据与其他数据的链接关系越多越全面,表明该数据的使用价值越全面。关联数据云(Linked Data Cloud)发展到今天已经成为一个非常大的数据集集合,从原先的12个发展到目前近600个,关联数据愈来愈获得技术人员及用户的认可与支持,其涵盖生命科学、地理信息、科学出版物、电影、音乐等众多领域^[2]。每个发布者只要基于关联数据规则进行数据的发布就可被其他用户链接及应用。这将使用户实现双

重的身份认定,数据发布者既是关联数据内容的创造者,也是其他关联数据内容的使用者。那么就存在一定的数据发布自定义内容,对于数据发布者想要描述的现实世界中的实体对象,每个人均会根据自我的认知及使用方便而规定其身份标识及属性名称。不同的数据集之间,可能对于现实世界中的同一实体对象有不同的描述内容及不一样的属性标识,这为数据的共享使用带来不便。因此,找寻有效的方法来发现这些数据集间的联系是亟待解决的问题。

在进行关联数据链接的发现工作时,因关联数据集中每个实体的信息均采用RDF三元组的形式进行描述^[3],可以保证每个实例数据均有各自的描述内容以区别于其他实例数据。同时关联数据采用URI进行实体的描述及表征链接^[4],在进行URI赋予描述实体信息时还存在

一些问题。

(1) 实体对象 URI 的定义多样性。在不同领域数据源中使用不同 URI 描述形式展现同一个实体对象时,会产生资源的重复定义,对实体对象的利用造成冗余。

(2) 实体对象名称的多重性。一个实体对象可能会拥有多个可描述的名称,而且这些名称很可能在字符表现上存在很大的差别。

本文采用一种新的思路构建数据集间的链接关系,主要进行数据集中的属性发现,将具有潜在对应关系的属性发现出来,进行关系匹配,并针对其中的属性进行函数运算,比较其相似度,以此来确定类内的实体匹配置信度,在此基础上进行实体链接的构建工作。本文提出的数据集实体链接发现方法可提高数据集的数据精度与细度,实现跨数据集的数据互联,提高这些数据集中实例数据的质量,尽可能简单快捷地发现数据之间的链接关系,并对重复数据进行校对及矫正。

1 研究综述

学者已对关联数据环境下的实体链接展开研究。国外学者主要是针对数据集或知识库之间的实体链接发现方法,以及关联数据环境下的实体识别展开研究。其中,Isle 等^[5]提出用于融合异构数据源的开源框架 SILK,它可以用于生成不同数据源的相关数据之间的链接。基于 SILK 链接发现过程有如下特点:用灵活的 SILK-链接规范语言(SILK-LSL)来指定链接规则;支持 RDF 链接生成;采用分布式系统;可用于混合不同词汇表的术语。SILK 对包含有大量数据的关联数据集的关联发现有较好的实现效果。Ngomo、Auer^[6]提出基于关联数据集之间的链接发现框架 LIMES(Link Discovery Framework for Metric Spaces),该框架主要利用空间度量算法处理链接发现中的可扩展性问题,计算实体的相似度,基于这些相似度,LIMES 可以过滤掉因用户设置不满足匹配条件的大量数据集;然

后计算剩余数据集的真正相似性,并返回匹配实体,发现数据集之间丢失的链接,提高数据检索结果的精确度,这些可通过其 Web 在线匹配工具实现。Vidal 等^[7]提出基于概率逻辑抽样方法及贝叶斯网络方法在关联数据集中进行权威实体排名算法操作,从而发现相互关联的实体间的隐藏链接,并对该方法进行广泛的实例验证。Jain 等^[8]在研究关联数据中本体映射规则的基础上,推出 BLOOM 链接查询工具,用于发现 LOD 云数据集之间的模式层级实体链接。BLOOM 可以利用维基百科和其类目结构等数据源中产生的噪声数据,计算关联数据中的本体映射对;此外,还将映射应用程序作为基础系统,用以补充原本的自举方法。Hienert 等^[9]开发了名为“Vizgr”的关联数据图形可视化工具,它提供了一种在图形可视化层面映射异构信息的有效方法,该工具能直接以图形的方式显示关联数据集中的实体对象及关系,用户可直接用手动链接的方式实现实体间的链接关系构建,不同来源的数据由用户输入,最终都可以以图表、标记云等可视化形式表达出来,数据点之间可以相互联系,也可与其他视图或网络处理器的数据联系。

国内的研究主要对关联数据环境下的链接发现过程及链接发现方法进行分析,探讨链接发现的实际应用。沈志宏等^[10]对关联数据应用下链接发现工作过程的定位、目标及复杂性进行探讨,提出链接发现过程中多任务、多路径、多步骤的特性,对关联数据链接发现方法、工具进行比较分析,提出链接发现方法需要进一步满足大关联数据的需求,发现框架需要能进一步支持多任务流程。贾丽梅等^[11]对关联数据语义相似度算法进行研究,提出在影响相似度计算的三个因素,即属性取值类型、属性权值及关联关系的基础上,进行相应的相似度计算算法。周文磊等^[12]提出基于关联开放数据的命名实体相似度计算方法,通过提取 URI 中的命名实体名称进行两个命名实体相似度计算,并进行相关验证,实现实体的动态链接发现。

综上所述,近年来针对于语义网或关联数据环境下的实体识别和链接发现的研究取得了一定的成果,但是针对关联数据环境下的实体链接构建方法研究较少。基于关联数据的数据链接特性,为我们提供了从分布式数据集中发现实体潜在关系的巨大可能,无需再对实体的来源进行复杂的操作,通过对不同来源的数据进行关联、发现、链接、合并、融合操作,可以生成一个高度集中的数据云。因此,在关联数据环境下进行实体发现不仅可以避免对实体的大量聚类计算等过程,而且可以高准确度地进行实体的识别及歧义消除,实现实体内容的语义显示。

2 面向关联数据的实体链接发现方法

2.1 实体描述模型

在关联数据环境下,实体链接更多的是表示关联数据实体对象之间的实体关联关系,对实体描述方式进行研究,可以在此基础上进行实体的关联发现及匹配。依据关联数据规则进行实体对象信息的描述,主要可以分为三个部分:实体基本信息描述,实体内容信息描述,实体链接信息描述。实体基本信息主要是指实体所属类的信息内容及实体所属资源集的内容;实体内容信息主要是指实体对象的属性信息,主要包括文本类属性及数值类属性;实体链接信息是指针对描述对象与其他实体对象所构成的链接关系,包括同一性链接关系、继承性链接关系及相关性链接关系等,如图1所示。

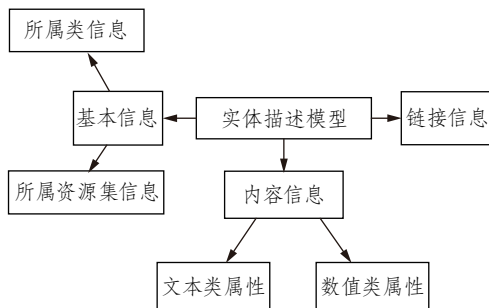


图1 关联数据实体描述模型

对实体进行描述信息分析可以发现,实体节点的主要信息分为属性描述信息和关系描述信息。在进行关系发现时对其属性进行比较分析,可以明确实体匹配的过程,在此基础上主要进行基于描述信息属性值包括文字类属性和数值类属性的计算处理及链接关系构建处理工作。

2.2 实体链接发现过程

在关联数据环境下,实体链接发现是数据资源聚合的一个过程,依据自下而上的顺序进行实体数据解析、实体匹配处理及实体链接构建的过程,如图2所示。

(1) 数据集解析

关联数据中的数据集结构多样,URI内容具有多样性。数据发布者一般只根据自身数据构成及发布需要发布相关数据集,因而对数据集进行解析就变得非常重要。在实体链接发现过程中,根据实际需要分为来源数据集(Source Dataset)及目标数据集(Target Dataset),两者处于同等地位,主要目的是实现其中实体及属性的匹配及链接处理,两者不具有绝对性。针对本文的主要研究问题,数据集解析是必须要进行的前期工作。对数据集进行解析,需要了解数据集的实体构成及相互关系,同时让数据质量得到保证。针对数据集的实际存储数据内容,要能够明确数据集中的类、实体、属性的URI描述信息内容,并且查明其中的有效类及有效属性,确保数据集中尽量不存在空类或空属性的情况,这样能够保证在对应属性发现阶段依据属性值进行属性聚类计算时的准确性,并能保证在实体链接构建时的有效性。对于本文所用到的数据集均采用RDF/XML格式进行存储,解析过程中采用SPARQL查询语句进行数据内容及格式的确认。

(2) 潜在对应属性发现过程

整个链接发现过程需要进行实体所在类及实体属性的发现及比较。对于数据集对应类内描述实体的属性进行所有属性值的遍历及比较过程,本文采用一个直观的方法,在属性概念之

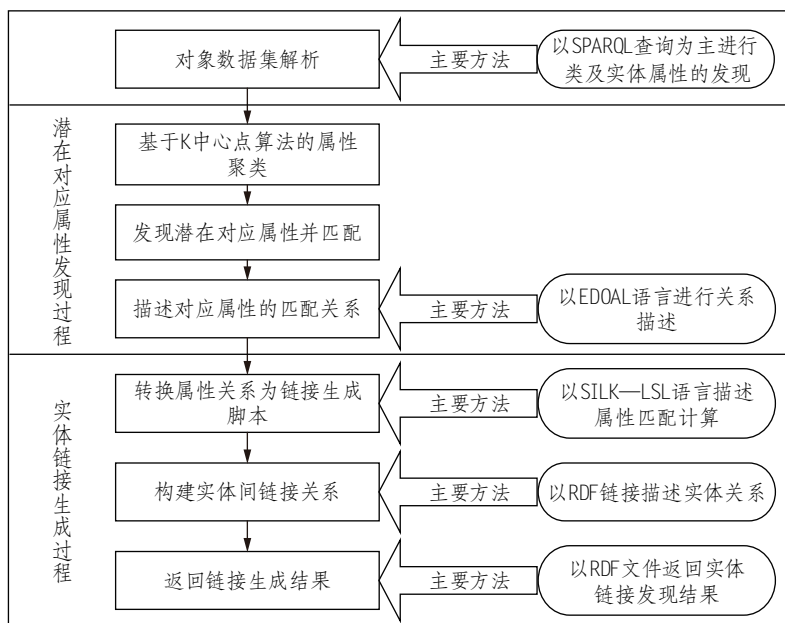


图2 实体链接发现过程

间发现其潜在的语义关系,依据其属性值的不同,对属性概念进行聚类处理,将属性进行分组,并对其属性关系依据 EDOAL 语言标准进行描述,以此来实现实体属性的匹配及计算,通过聚合比较来确定实体是否存在共指关系。

通常情况下,在数据集间发现对应属性是比较困难的,需要同时评估两个属性是否可以真正构成一个等价属性,也需要对其相似性阈值进行设置。在等价属性发现过程中,如果设置了较高的相似性阈值,就可能会有许多属性因为不满足这个阈值而被遗漏;另一方面,如果对待评估的属性设置了较低的相似性阈值,就会产生本不相关的假的等价属性对。所以在进行等价属性发现时,不仅要明确其所在类是对应关系,同时对类内的属性间关系还要进行路径确定,一般的属性路径是一组属性队列,其中以属性的关系进行访问,开始于某个属性关系,结束于另一个属性关系或属性值。在本研究中,路径里的每个属性三元组中的对象 O 均是其后关系表示中属性三元组的主语 S 。

对应属性发现主要采用的方法是,根据属

性值本身的数据类型将属性分为文字类属性和数值类属性,并在此基础上进行相应类别的属性值聚类,以属性特征值的形式表现属性本身应具有的关键含义,并在此基础上实现聚类分组的目的,实现潜在对应属性的发现。在对某个类中的属性进行具体聚类分组时,采用参照 K 中心点聚合算法进行计算,目的是在最大限度下保证所有的潜在对应属性均能够被发现,尽可能实现所有潜在对应属性的匹配关系描述,在后面进行链接比较时会保证实体属性相似度计算的完整性。

(3) 实体链接构建过程

在实现对应属性的发现之后,对属性匹配关系进行关系描述,主要目的是实现实体属性相似度比较,以此判断所有实体是否真正存在语义相关关系,并通过构建链接的形式发现其链接关系。在整个链接的构建过程中,通过进行已匹配的属性对计算,可以大大减少对所有属性对两两比较的次数,降低匹配计算所花费的时间,同时提高匹配计算的准确性。SILK 作为一个开源链接发现工具,一直被广泛用于关

联数据集的链接发现,采用此工具可以在一定的实践基础上验证本文提出的方法的可行性,并对生成的可执行链接脚本进行二次利用。

文中针对对应属性的关系描述采用 EDOAL 语言,但其不能被 SILK 工具所识别,因此在进行链接生成时,就需要实施转换,将对应属性关系转化为 SPARQL 查询语句,实现 RDF 路径访问表达,使其可被 SILK 读取。在进行链接生成脚本的编写时,需要分辨对应属性是文字类属性还是数值类属性,选择不同的匹配计算函数,并进行相应的置信度设置,使其成为可被 SILK 执行的脚本文件。SILK 工具执行的脚本文件,将会直接进行数据集中实体相似度的合成计算,并导出实际的实体匹配链接结果。

3 数据集解析方法

数据集解析的主要目的:一是保证数据集的数据架构简洁明确,不存在空类或空属性;二是保证来源数据集及目标数据集中有潜在实体链接。数据集解析的主要对象包括:数据集名称、数据集规模(实体数目)、数据集内实体对象内容结构、实体属性数目及内容、RDF 数据存储的格式及数据集 SPARQL 查询端点等。对数据集的结构内容进行解析探究,不仅可以在宏观上把握数据集的领域用途及需求目标,也可在微观上明确数据集内部的实体描述标准及主要的描述属性内容。同时,由于 RDF 存储形式多样,针对不同格式的数据集需要采用不同的工具进行处理。图 3 显示数据集解析的内容,在概念上对数据集解析任务进行归类划分。

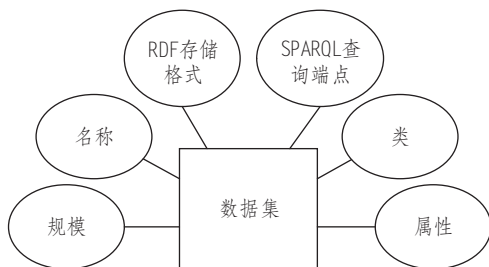


图3 数据集解析内容

通过对数据集进行解析,将关联数据集规范为符合实际程序处理的数据集对象,实现对关联数据集中的结构、实体类及属性等内容的读取,并对实体的具体属性结构进行标识化处理。根据关联数据集特性,对数据集内部数据进行类、实体及属性的区分,按实体类进行实体的分类,在同一类内拥有若干个实体对象实例,每个实体的属性定义标准均相同,但属性值不同,以此来描述实体之间的区别,在数据集层次上对实体数据进行由上到下、由粗到细的层次解析。

数据集解析过程如图 4 所示:①进行类查询,读取数据集中的有效类,确认数据集中描述的主要概念内容;②进行某个类中实体信息的查询,确认实体的数目及范围;③对某个实体进行查询,通过实体信息描述内容分析实体的属性及链接信息,对属性名称及属性值数据类型进行确认,包括文本类属性及数值类属性的属性值;④对类及属性关系进行分析,明确其所定义的实体信息内容,在进行对应类及对应属性发现时可根据语义进行属性匹配的处理工作。

通过实施完整的数据集解析流程,可以完成对数据集的名称、规模及 RDF 存储形式、数据集中可匹配的实体数目、类的形式及种类等内容的确认,通过对多个数据集进行解析可以发现潜在实体关系,为对应属性的发现提供前提条件,确保后续属性匹配及链接构建工作的顺利进行。

本文中的主要数据集对象是地理信息数据集 GADM 与 EUROSTAT 数据集中的 NUTSRegion。其中,GADM 依据 NeoGeo 本体发布全球行政区域的空间地理相关数据,包括国家名称、所包含区域、区域面积及区域间的包含关系等内容,NUTSRegion(Nomenclature of Territorial Units for Statistics,领土单位统计术语标准数据集)是欧盟统计局依据 Neo 标准发布的欧盟国家行政区域地理编码标准,主要对欧盟范围内的地理信息进行标准发布,包括区域名称、

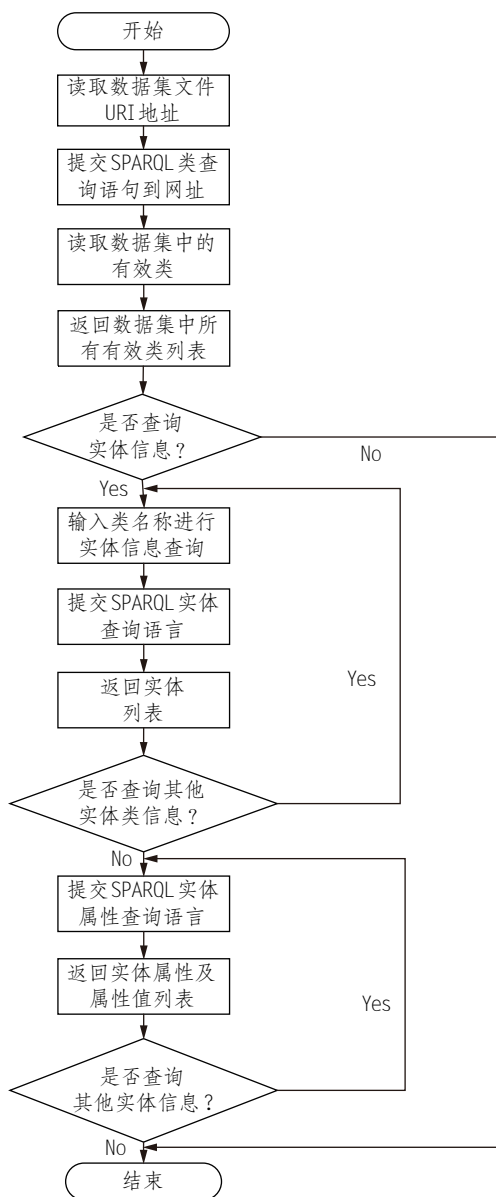


图4 数据集解析过程

名称编码及区域间包含关系。目前数据集 GADM 与数据集 NUTSRegion 均以关联数据形式进行发布,但还没有在数据层面形成链接关系,可为本文进行关联数据实体链接发现提供良好的实例,在下文的方法分析中也以这两个数据集为分析对象。

对 EUROSTAT 数据集的解析过程如下。

(1)数据集 EUROSTAT 的远程 SPARQL 查询平台地址为: <http://wifo5-03.informatik.uni-mannheim.de/eurostat/snorql>。

(2)数据集的 SPARQL 类查询语言为: `SELECT DISTINCT ?class WHERE { [] a ?class } ORDER BY ?class`,结果如图5所示,显示数据集有两个类: `eurostat:countries` 和 `eurostat:NUTSRegions`。

(3)对 EUROSTAT 中 NUTSRegion 类中的前 100个实体进行查询,SPARQL 查询语句为: `SELECT ?entity WHERE { ?entity a <classes [] > } OFFSET 10`,结果如图6所示。

(4)对类 `eurostat:NUTSRegion` 的某个实体即德国首都柏林 URI 为 `<http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Berlin>` 进行属性信息查询,SPARQL 查询语句为: `SELECT DISTINCT ?property ?hasValue WHERE { { <http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Berlin> ?property ?hasValue } } ORDER BY (!BOUND (?hasValue)) ?property ?hasValue`,结果如图7所示。实体包含的属性主要包括 `eurostat: name`, `eurostat: geocode`, `eurostat: GDP`, `eurostat:disposabl_income` 等地区基本统计信息类型及关系属性 `eurostat:hasParentCountry`,通过数据集解析流程,完成对数据集内容及结构的解析工作。

```
<binding name="class">
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/countries</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/regions</uri>
```

图5 数据集类查询结果

```

<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/%C3%85land</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/%C3%8Ele_De_France</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/%C3%96vre_Norrland</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Alentejo</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Algarve</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Andaluc%C3%ADa</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Arag%C3%B3n</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Attiki</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Auvergne</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Baden-w%C3%BCrttemberg</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Basilicata</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Basse-normandie</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Bassin_Parisien</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Bayern</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Bedfordshire_And_Hertfordshire</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Berlin</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Border_Midland_And_Western</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Bourgogne</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Brandenburg</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Brandenburg_-_Nordost</uri>
<uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/regions/Brandenburg_-_S%C3%BCdost</uri>

```

图6 数据集实体查询结果

```

<variable name="property"/> <variable name="hasValue"/> <variable name="isValueOf"/> </head> <results>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/GDP</uri>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">79169.9E0</literal>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/RnD_exp_mio_euro</uri>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">3096.146E0</literal>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/personel_percent_of_act_pop</uri>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">2.31E0</literal>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">0.01E0</literal>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/disposable_income</uri>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">14737.6E0</literal>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/electricity_consumption_GWh</uri>
<binding name="hasValue"> <literal datatype="http://www.w3.org/2001/XMLSchema#double">16684.0E0</literal>
<binding name="property"> <uri>http://wifo5-04.informatik.uni-mannheim.de/eurostat/resource/eurostat/electricity_production_capacity_MWh</uri>

```

图7 数据集实体属性查询结果

4 基于 K 中心点聚类算法的对应属性发现方法

本文中“对应类”是指包含相同或相似实体对象集共同特征的概念集合,“对应属性”是描述实体对象间相同或相似个体特征的具体名称变量。“对应类”和“对应属性”将会帮助实现实体的识别及实体间的相似度匹配计算。

对应属性发现主要采用的方法是,根据属性值本身的数据类型将属性分为文字类属性和数值类属性,并在此基础上进行相应类别的属性值聚类,以属性特征值的形式表现属性本身

应具有的关键含义,进行聚类分组,实现潜在对应属性的发现。在对某个类中的属性进行具体聚类分组时,参照 K 中心点聚类算法进行计算,目的是尽可能保证所有的潜在对应属性均能够被发现,实现所有潜在对应属性的匹配关系描述,在后面进行链接比较时会保证实体属性相似度计算的完整性。

4.1 不同类间的属性聚类

本文使用 K 中心点聚类算法的目的是对属性概念进行分类分组^[13]。利用此方法,属性概念得以分组,并能最终实现数据集间具有相同表达含义的属性的对应,最后实现其分组匹配

比较的目的。其中最重要的工作是进行样本数据的选取,计算属性概念的数字坐标。

出聚类的首次中心点对象坐标及最后生成的聚类簇数。属性特征值聚类算法如图 8 所示。

进行聚类之前,本文依据 Forgy^[14]的方法给

```
1 输入:属性样本点集合 M,样本点总数 N,迭代次数 T=30。
2 输出:属性样本聚合组。
3: C=样本点集合的初始中心点个数;
4: KMedoids(M,C) 分配样本点到最近的中心点簇内;
5: 计算样本点到中心点距离 L;
6: t=1;
7: while( t<T) do
8: Cnew=新的随机中心点个数;
9: KMedoids(M,C) 分配样本点到最近的中心点簇内;
10: 计算样本点到中心点距离 Lnew;
11: If( Lnew<L) then
12: L=Lnew;
13: C=Cnew;
14: t++;
15: while( |C|>2) do
16: 重复迭代进行 KMedoids(M,C);
17: 随机选择每个簇内新的中心点 Cnew;
18: 计算 ad;/ad 为整个样本点内中心点的距离之和/
19: C=Cnew;
20: If( ci, cj ∈ C, L( ci, cj ) <  $\frac{1}{2}$ ad) then
21: C= MergeCenter( C,  $\frac{1}{2}$ ad );/对中心点进行合并/
22: If( |C|<2) then
23: Break;
24: Else
25: Break;
26: 输出 C.
```

图 8 K 中心点属性聚类算法

聚合计算主要是进行两个类间的属性关系发现,所以最后形成的簇数目为 2。当某两个中心点之间的距离小于所有中心点的平均距离时,对这两个中心点所在的簇进行合并处理。依据样本点集合数目的大小确定初始中心点数目,当属性样品点数目 N=50 时,取其中心点数目 C=20,依据 Forgy 方法进行中心点坐标的随机选择。当两个中心点之间的坐标距离小于所有中心点坐标距离的 1/2 时,我们认为两个中心点所在的簇具有较强的相似性,合并为新的簇并进行新的簇中心点的选择,再次进行新的聚合迭代过程。

本文以 EUROSTAT 数据集和 GADM 数据集

中类 eurostat: NUTSRegion 和 gadm: AdminRegion 为例进行属性聚类计算。其中,对于 eurostat: NUTSRegion 类中的属性,前文提到其 eurostat: name、eurostat: geocode 及关系属性 eurostat: hasParentCountry 等均属于文字类属性,对于 GADM 数据集中 gadm: label、gadm: ID 及关系属性 gadm: In_Country 等依据属性值进行属性的归类聚合。

EUROSTAT 数据集中的类 eurostat: NUTSRegion 的主要属性为:

```
{ eurostat: name, eurostat: geocode, eurostat:
hasParentCountry }
```

GADM 数据集中的类 gadm: AdminRegion 的

主要属性为:

{ gadm: label, gadm: ID, gadm: In_Country }

对这两个类内的属性值进行样本数据取样并计算统计特征值,属性特征聚合值的坐标为(X,Y,Z),代表对上述文字类属性的聚合结果

矩阵,形成聚类样本点,可依据此属性特征聚合值进行属性之间的聚合计算,依据欧式空间中属性点的距离发现属性之间的相似关系,依据聚类结果进行属性的分组,并做出潜在对应属性对的匹配。其主要的聚类过程如图9所示。

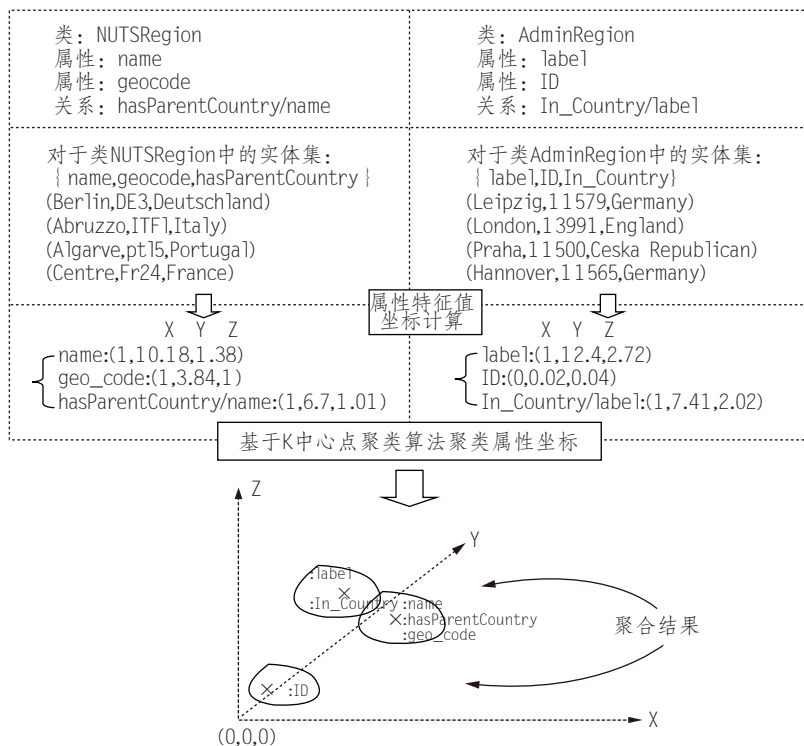


图9 属性聚类过程图

4.2 对应属性匹配规则

在对每个类中的实体属性进行聚类后,属性依据其本身的特征值被分配到不同的簇中,表明属性本身与其他属性的相关关系。这为不同的类C与C'中进行对应属性的关系构建提供了前提条件,每个在类C中被聚合的属性均可以与另外一个类C'中的属性实现对应匹配。

针对两组对应类中的对应属性匹配关系构建的主要过程是:假设在类C中有p组实体文字类属性及相应属性值,标记为 $G_1, G_2, G_3, \dots, G_p$,经计算可得每组属性的属性特征坐标为 (x_i, y_i, z_i, u_i) ,其中 $i=1, 2, \dots, p$ 。同样在类C'中有q组实体文字类属性及相应属性值,被标记为

$G'_1, G'_2, G'_3, \dots, G'_q$,同样经计算可得其属性坐标为 (x'_j, y'_j, z'_j, u'_j) ,其中 $j=1, 2, \dots, q$ 。那么依据每个属性的特征坐标值进行聚合计算,所形成的匹配对距离为 $D=(G_i, G'_j)$,那么对于两个类间的对应属性匹配集合可以表示为:

$$C_{(i,j)} = \{ \arg \min D(G_i, G'_j) \}$$

其中,当属性坐标索引编号为(i,j)时,对所形成的属性匹配对进行比较,取所有属性匹配对距离的最小值作为两个数据集属性匹配的最终结果,其本质上就代表了不同类间属性特征值坐标距离最短,表明两个属性的语义相关性最强。依据这样的取值方法,针对类C及C'内的属性,会形成一个表示所有潜在对应属性组的集合。

依据 K 中心点聚类算法将会实现对不同类间的潜在对应属性发现,对于潜在对应属性之间的关系描述采用 EDOAL^[15] 语言标准进行表示,该语言可以准确描述出属性之间的复杂关系。对于位于不同数据集中对应类下的属性分别用 P 及 P' 进行表示,依据 EDOAL 语言标准可以对属性 P 及 P' 进行关系描述为:

(1) 属性 P 与 P' 是对应关系

```
<Cell>
  <entity1>
    <edoal:Property rdf:about="P"/>
  </entity1>
  <entity2>
    <edoal:Property rdf:about="P'"/>
  </entity2>
  <relation>equivalence</relation>
  <measure>1.0</measure>
</Cell>
```

(2) 属性 P 与某属性路径下的关系组合 $R'_1, R'_2 \dots R'_n$, 和属性 P' 为对应关系, $n = [1, +\infty)$

```
<Cell>
  <entity1>
    <edoal:Property rdf:about="P"/>
  </entity1>
  <entity2>
    <edoal:Property>
      <edoal:compose rdf:parse Type="Collection">
        <edoal:Relation rdf:about="R'_1"/>...
        <edoal:Relation rdf:about="R'_n"/>
      </edoal:compose>
    <edoal:Property rdf:about="P'"/>
  </entity2>
  <relation>equivalence</relation>
  <measure>1.0</measure>
</Cell>
```

以上两种属性描述规则基本涵盖了所有在数据集中出现的属性及关系的存在形式,包括通过对属性关系进行推理,实现数据集间对应属性的真正匹配。本文对匹配关系赋予具体的

取值 $[0, 1]$, 默认所有潜在对应属性的匹配值均为 1, 两个互相对应的属性处于完全匹配的关系, 在后面进行属性值计算时会通过计算属性值的方法来判断实体是否完全匹配。

5 基于 SILK 的链接构建方法

5.1 数据集及实体属性获取方式

以匹配关系描述不同数据集间的对应属性关系,在进行实体的对比时,采用对比该属性下两组属性值的相似度的方法,这需要将实体的属性值抽取出来,进行基于对应属性关系的成对比较,根据其所包含的属性类型,选择不同的比较方法进行属性值的比较。同时需要在 EDOAL 语言规范下,实现类、属性及关系的查询表达,使得被约束的对应关系能够被 SILK-LSL^[5] 脚本“理解”,在进行链接构建过程中,可以生成实体间自动的链接。

在进行实体属性抽取之前,需要对实体数据的来源进行声明,包括数据集名称、数据集的访问地址及数据集的数据具体查询条件等。其主要表达形式为:

```
<SourceDataset dataSource="dataSource id"
var="resource variable name">
  [<RestrictTo>SPARQL restriction</RestrictTo>]
</SourceDataset>
<TargetDataset dataSource="dataSource id"
var="resource variable name">
  [<RestrictTo>SPARQL restriction</RestrictTo>]
</TargetDataset>
```

在进行数据源信息语句声明后,对数据集的访问地址及访问形式等多个内容均可进行实际的定义,然后进行属性路径的声明,基于对应属性的 RDF 路径输入表达式为“path=? x/属性名”,这种基于 SILK-LSL 的声明方式将会实现具体属性数据的抽取,从而为实体的匹配提供前提条件。

5.2 实体链接规则构建

根据 SILK 框架下的链接构建步骤总结关

联数据实体链接的构建规则^[16],主要分为三个部分:路径规则构建、比较规则构建、聚合规则构建。这三个部分的规则能约束数据集间实体链接构建的全过程语句表达及方法选择,对基于 EDOAL 语言表述的对应属性及对应类进行比较,实现实体链接的构建。

路径规则构建的主要目的是在 SILK-LSL 语言语法中,支持 XPath 用法路径输入操作。以 P 表示路径规则,则 P 中的规则内容能实现从某个节点主语对象 S 到某个属性或某个关系的访问路径。其具体的规则内容为:

$$P(u) = /u; P(u^{-1}) = \backslash u; P(R) = /u[s, <u>, o] \quad (\text{规则 1})$$

比较规则构建的主要目的是,针对已经进行转换格式操作的属性及关系,以 tr 表示转换函数,对其在统一格式条件下的相似度进行比较,以 C 表

示比较规则,则针对属性的比较规则为:

$$C(a, a', tr) = M_{t, t'}(tr(s/P(a), s'/P(a'))) \quad (\text{规则 2})$$

其中, t 表示对应类, a 表示类的相关属性, s 表示属性 a 的实体属性值, $M_{t, t'}$ 表示依据 s 的数据类型,选择具体的属性比较算法,进行属性值的比较。通过 Transform 函数转换格式,统一实体属性值格式,实现对实体属性值的方法比较。

聚合规则构建的主要目的是为了能够将两个对应属性中的属性相似度值进行聚合计算,这两个对应类或属性是属于集合 A (A 代表两个数据集中的对应属性的集合,对应类及对应属性的关系用 EDOAL 语言描述)。在本文中假设两个属性的 URI 分别为 u 和 u', 类所包含属性的相似度计算及置信值计算过程可表示如下:

$$\begin{aligned} R_{s, s'}(u, u', A) &= \text{Average}_{(a, a', =, tr) \in k(u, u', A)} C(a, a', tr) \\ R_{s, s'}(u, u', A) &= \text{Max}_{(a, a', =, tr) \in k(u, u', A)} C(a, a', tr) \\ R_{s, s'}(u, u', A) &= \text{Min}_{(a, a', =, tr) \in k(u, u', A)} C(a, a', tr) \end{aligned} \quad (\text{规则 3})$$

其中, $k(u, u', A)$ 表示对应属性对集合表示形式,当 u 代表具体的属性值 URI 值,为了构建更多链接,我们选择聚合方法 Max(取最大置信值)。

通过对实体链接规则包括路径规则、比较规则及聚合规则的方法构建,可以在 SILK 框架下实现实体链接构建的整个过程描述,包括确认数据来源、抽取实体信息、计算潜在对应属性的相似度及确认实体链接存储地址等内容,其 xml 代码如图 10 所示。

6 实验

为了验证基于 K 中心点聚类算法发现不同类间的对应属性及实现实体链接构建方法的可行性、实体链接发现的合理性及适用性,本文进行相应实验。

实验的开发环境为 Java MyEclipse,通过调用 XMLAPI 中的 dom4j 程序读取 RDF/XML 文件,通过 AlignmentAPI-4.7 实现对应属性关系命

令编写,通过 SILKAPI2.6.0 实现实体链接规则编写及链接结果导出。

6.1 对象数据集的选取

为验证关联数据实体链接的发现方法,分别对不同的开放关联数据集进行实验,主要对 IM@OAIE2014^[17](数据集 Abox3)、CKAN^[18](数据集 EUROSTAT)和 GADM-RDF(数据集 GADM)^[19]中的数据进行对应属性的聚合发现及实体链接。在 EUROSTAT 与 GADM 中,对欧洲国家和地区的地理数据进行关联数据发布,文中所举实例也是基于这两个数据集;同时对 Abox3 中两个书目信息数据集进行实验验证,进行关联数据形式的表示。两个数据集中均存有不一致的书目信息,包括书籍的作者、主题、类型及文学流派等信息。分别对数据集中的对应类及对应类的属性数目和实体数目进行统计,如表 1 所示。本文对四个数据集进行两次实验,以验证本文所提方法的可行性、适用性及实用性。

```
<?xml version=" 1.0 " encoding=" utf-8 " ?>
<Silk>
  <Prefixes>
    <Prefix id=" prefix id " namespace=" namespace URI " />.....
  </Prefixes>

  <!--确定数据集地址-->
  <DataSources>
    <DataSource id=" data source id of g " type=" data source type " >
      <Param name=" parameter name " value=" parameter value " />
    </DataSource>
    <DataSource id=" data source id of g' " type=" data source type " >
      <Param name=" parameter name " value=" parameter value " />
    </DataSource>
  </DataSources>

  <!--从两个对应类中抽取实体信息-->
  <Interlinks>
    <Interlink id=" interlink id " >
      <LinkType>owl:sameAs</LinkType>
      <SourceDatasetdataSource=" data source id of g " var=" resource variable name " >
        <RestrictTo> SPARQL query on instances in class C </RestrictTo>
      </SourceDataset>
      <TargetDatasetdataSource=" data source id of g' " var=" resource variable name " >
        <RestrictTo> SPARQL query on instances in class C' </RestrictTo>
      </TargetDataset>
    </Interlink>
  </Interlinks>

  <!--对潜在对应属性进行相似度计算-->
  <LinkageRule>
    <Aggregate type=" max " >
      <Compare metric=" similarity metric " >
        <Input path=" RDF path of  $A_1$  in  $AC_1$  " />
        <Input path=" RDF path of  $A'_1$  in  $AC'_1$  " />
      </Compare>.....
      <Compare metric=" similarity metric " >
        <Input path=" RDF path of  $A_i$  in  $AC_i$  " />
        <Input path=" RDF path of  $A'_i$  in  $AC'_i$  " />
      </Compare>
    </Aggregate>
  </LinkageRule>

  <!--确定实体链接存储地址-->
  <Outputs>
    <Output type=" output type " >
      <Param name=" parameter name " value=" parameter value " />
    </Output>
  </Outputs>
</Interlink>
</Interlinks>
</Silk>
```

图 10 基于 SILK 的实体链接构建过程 xml 代码

表 1 实验数据集

数据集	对应类	属性数目	实体数目
EUROSTAT 和 GADM	eurostat:NUTSRegion	4	1 200
	gadm:AdminiRegion	6	37 933
Abox3	owl:NamedIndividual	6	173
	owl:NamedIndividual	6	172

6.2 实验分析

6.2.1 对应属性发现实验分析

分别对本次试验发现的潜在对应属性对,及对应属性对占总属性对的比例进行统计,结果如表

表 2 发现的对应属性对所占比例

数据集	潜在对应属性对占总属性对的比例(%)	对应属性对占总属性对的比例(%)
EUROSTAT 和 GADM	25	6
Abox03	49	8

比例计算过程依据是:假设源数据集中某类有 n 个属性,目标数据集的对应类中有 n' 属性,进行基于 K 中心点的聚类过程,那么可进行匹配属性对数目为 $n * n'$,在完成属性聚类后,假设所发现的潜在对应属性对数目为 A ,那么发现潜在对应属性对占总属性对的数目是 $|A|/n * n'$,在对属性对进行相似性匹配计算时,选择 Max 聚合方法进行实体属性相似度的聚合运算,此类对应属性对数目为 B ,则对应属性对占总属性对的比例是 $|B|/n * n'$ 。

对表 2 进行分析可以发现,潜在对应属性对比例均大于对应属性对比例,说明在进行 K 中心点聚类之后,实体属性可以满足属性匹配关系的构建,可降低 50% 的属性错误匹配及计算,其中大部分的属性对不能实现默认匹配。 K 中心点聚类算法减少了实体链接构建阶段的相似度遍历计算次数,同时不存在属性匹配缺失的现象,能够最大限度保证所有的实体属性均能实现关系发现及匹配计算。

在对应属性发现的同时,也可对不同类之间的聚合关系进行发现。在两个数据集中依据属性特征标准即字符串单词个数及字符串字

2 所示。其中对应属性对是指在构建实体链接后,两个实体属性值为等价关系,其表示的实体属性内容的含义相同;潜在对应属性对为根据 K 中心点聚类算法发现具有相关关系的实体属性。

总长度为坐标取值标准,对两个类进行聚合,聚合的效果如图 11、图 12 所示。可以看到,在图 11 中,对于数据集 EUROSTAT 和 GADM 有两个聚合中心点(1.00,2.00)和(2.50,15.00),这表明对于类 eurostat:NUTSRegion 和类 gadm:Admin-Region,实体属性值字符串单词个数及字符串总长度在聚合中心点附近的实体属性为潜在对应属性,对于实体属性的分组能达到一个较好的效果。在图 12 中,对于两个类 owl:NamedIndividual

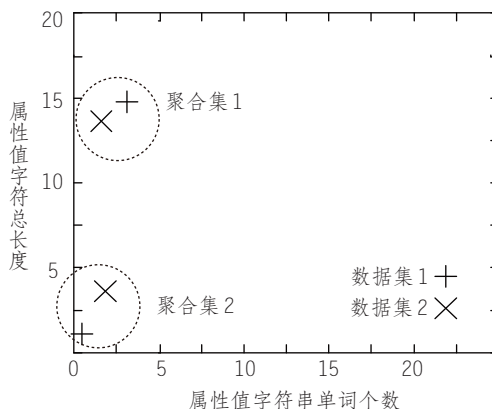


图 11 对类 eurostat:NUTSRegion 和类 gadm:AdminRegion 的聚合效果

和类 owl:NamedIndividual 进行属性值聚合,其中心点为(8.50,52.00),表明对于属性值字符串单词个数及字符总长度在聚合中心点附近的实体属性为潜在对应属性,能进行有效的发现及匹配。

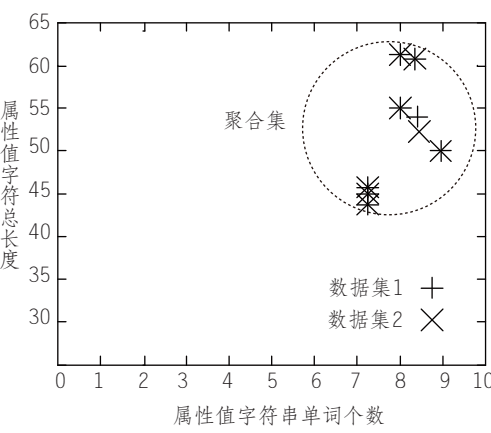


图 12 对类 owl:NamedIndividual 和类 owl:NamedIndividual 的聚合效果

6.2.2 实体链接构建实验分析

对对象数据集中的属性进行聚合分析后,

通过 SILK 框架进行实体属性相似度计算,并赋予 RDF 链接:owl:sameAs,实现实体的链接。在实体链接阶段分别对以上三类共四个关联数据集进行链接发现实验,在实验过程中采用不同置信度对实体匹配结果进行统计。

其中 eurostat:NUTSRegion 类中共有1 200个实体,与 GADM 数据集中对应类 gadm:Admin-Region 中的实体进行相似度计算,采用 Max 方法取置信度最大值进行链接构建,然后统计链接数目,同时实体总数参照源数据集 NUTSRegion 类下的实体个数1 200进行计算。具体的统计结果如表 3 所示。

其中 F 值计算采用信息检索中的查全率与查准率标准,见下列公式。

查全率 $R = \frac{\text{正确实体链接数目}}{\text{实体总数目}}$

查准率 $p = \frac{\text{正确实体链接数目}}{\text{实体链接发现数目}}$

$$F = \frac{2PR}{P+R}$$

表 3 数据集 EUROSTAT 和 GADM 实体链接发现结果统计

源数据集 实体总数	目标数据集 实体总数	实体链接 发现数目	实体链接 失败数目	正确实体 链接数目	F 值
1 200	37 933	965	235	935	0.878 9

对数据集 Abox03 中的两个书本信息数据集进行实体链接构建实验,其中源数据集的类 owl:NamedIndividual 与目标数据集的类 owl:NamedIndividual 为对应类,包含的对应属性有 rdf:label、oeai:subject、oeai:author、oeai:publisher、oeai:genre、oeai:literaryGenre,进行实体属性值比

较,构建实体链接。源数据集中有 173 个实体,目标数据集中有 172 个实体,进行实体链接总数计算时,以源数据集的实体总数 173 为参照,实体相似度计算选择采用 Max 取置信度最大值进行实体链接数目的统计。具体的统计结果如表 4 所示。

表 4 数据集 Abox03 实体链接发现结果统计

源数据集 实体总数	目标数据集 实体总数	实体链接 发现数目	实体链接 失败数目	正确实体 链接数据	F 值
173	172	163	10	158	0.940 5

实验结果表明:通过两次数据集的实体链接实验发现,采用基于 K 中心点聚类算法发现

不同类之间的对应属性进行实体匹配计算的方法,可提高实体链接的数目,具有较高的准确率及F值。采用本文的方法进行实体链接发现具有可行性及实用性。

7 结语

目前,万维网已经发展成一个将网页数据及数据库数据链接起来的全球数据空间,关联数据是实现这一信息进化的最佳技术之一。利用关联数据发布标准,通过对相互关联的信息进行标准化的链接描述来表达语义,同时还将数据标识作为唯一的并且可寻址的方式,将全球各种领域的信息连接起来,成为最大的互联

数据库。因此,针对关联数据集进行实体发现及链接构建具有非常重要的研究价值。

本文从链接发现角度阐述了在数据集中进行实体属性匹配及实体相似度的计算方法,并对所存在的问题进行分析。在此基础上,提出基于统计学习方法的实体属性聚类方法及链接构建规则设计,并进行实验验证及方法评估。本文针对关联数据环境下的数据链接问题提出的方法及实现流程,可为解决实体匹配难题提供新思路和新方法,为解决数据集层次的语义互联问题奠定基础。在下一步的研究工作中,我们拟进行关联数据集的互联模式与关联方法研究,改善关联数据的集成效果,进一步提高关联数据语义查询的效率。

参考文献

- [1] Bizer C. The emerging web of linked data[J]. IEEE Intelligent Systems, 2009(5): 87-92.
- [2] Schmachtenberg M, Bizer C, Paulheim H. State of the LOD Cloud 2014[EB/OL]. [2016-02-20]. <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state>.
- [3] Bizer C, Cyganiak R, Heath T. How to publish linked data on the web[EB/OL]. [2016-03-08]. <http://www4.wiwiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial>.
- [4] Berners-Lee T. Linked data[EB/OL]. [2016-01-20]. <http://www.w3.org/DessignIssues/LinkedData.html>.
- [5] Isele R, Jentzsch A, Bizer C, et al. Silk tools controduction[EB/OL]. [2016-04-10]. <http://www4.wiwiwiss.fu-berlin.de/bizer/silk>.
- [6] Ngomo A C N, Auer S. A time-efficient hybrid approach to link discovery[C]//Volume 814 of CEUR Workshop Proceedings, Bonn, Germany, 2011.
- [7] Vidal M E, Rivera J C, Ibáñez L D, et al. An authority-flow based ranking approach to discover potential novel associations between linked data[J]. Semantic Web, 2014, 5(1): 23-46.
- [8] Jain P, Hitzler P, Sheth A P, et al. Ontology alignment for linked open data[J]. Semantic Web, Proceedings, 2010, 6496: 402-417.
- [9] Hienert D, Zapilko B, Schaer P, et al. Vizgr linking data in visualizations[R]. Web Information Systems and Technologies, 2012(101): 177-191.
- [10] 沈志宏, 黎建辉, 张晓林. 面向 LOD 的关联发现过程的定位、目标与复杂性分析[J]. 中国图书馆学报, 2013(6): 1-8. (Shen Zhihong, Li Jianhui, Zhang Xiaolin. Insights into link discovery process for linked open data: positioning, goals and complexity[J]. Journal of Library Science in China, 2013(6): 1-8.)
- [11] 贾丽梅, 郑志蕴, 李钝, 等. 基于动态权值的关联数据语义相似度算法研究[J]. 计算机科学, 2014, 41(8): 263-273. (Jia Limei, Zheng Zhiyun, Li Dun, et al. Research on semantic similarity algorithm of linked data base

- on dynamic weight [J]. Computer Science, 2014, 41(8): 263-273.)
- [12] Zhou W, Wang H, Chao J, et al. LODDO: using linked open data description overlap to measure semantic relatedness between named entities[C]//Joint International Semantic Technology Conference, JIST 2011, Hangzhou, China, December 4-7, 2011.
- [13] 吴文亮. 聚类分析中 K 均值与 K 中心点算法的研究[D]. 广州: 华南理工大学, 2011. (Wu Wenliang. Study of K-means and K-medoids algorithms in clustering analysis [D]. Guangzhou: South China University of Technology, 2011.)
- [14] Forgy E W. Cluster analysis of multivariate data: efficiency versus interpretability of classification[J]. Biometrics, 1965, 21(3): 768-769.
- [15] Scharffe F. EDOAL: expressive and declarative ontology alignment language[EB/OL]. [2016-03-30]. <http://alignapi.gforge.inria.fr/edoal.html>.
- [16] Jentzsch A, Isele R, Bizer C. Silk-generating RDF links while publishing or consuming linked data[C]//International Semantic Web Conference (ISWC2010), Shanghai, November 2010.
- [17] OAEI. Ontology alignment evaluation initiative 2014 campaign[EB/OL]. [2015-09-30]. http://islab.di.unimi.it/im_oaei_2014/index.html.
- [18] CKAN. Datahub[EB/OL]. [2016-04-15]. <http://datahub.io>.
- [19] GADM-RDFproject. An RDF spatial representation of all the administrative regions in the world[EB/OL]. [2016-04-20]. <http://gadm.geovocab.org>.

高劲松 华中师范大学信息管理学院教授, 博士生导师。湖北 武汉 430079。

周习曼 华中师范大学信息管理学院硕士研究生。湖北 武汉 430079。

梁艳琪 华中师范大学信息管理学院博士研究生。湖北 武汉 430079。

(收稿日期: 2016-07-22; 修回日期: 2016-09-05)