基于属性权重的链接数据共指关系构建

张晓辉 蒋海华 邸瑞华
(北京工业大学计算机学院 北京 100124)

摘 要 来自异构数据源的语义数据集之间关联的缺失严重影响了数据网的构建和发展。语义数据集中,实例数据之间共指关系的发现和构建能够丰富数据集之间的关联,从而有助于在数据集之间进行推理和查询。在基于相似度分析的共指关系构建的过程中,实例属性的权重及属性值的相似度对实例相似度具有重要作用。提出一种新的基于数据集统计信息计算属性权重的模型,并从概率统计的角度证明其合理性。同时分析了这种权重计算模型相对于传统的权重计算方法的优势。基于新的权重计算方法,实现了共指关系构建系统,并利用开放的语义数据集验证了其正确性。

关键词 链接数据,共指关系,属性权重
中图法分类号 TP393 **文献标识码** A

Property Weight Based Co-reference Resolution for Linked Data

ZHANG Xiao-hui JIANG Hai-hua DI Rui-hua
(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract The construction and development of the Web of data are affected seriously by the loss of links amongsemantic datasets fromheterogeneousdata sources. The construction of co-reference between instances in semantic datasets will be helpful for enriching the links among datasets, which will improve the reasoning and query across datasets. The property weights and the similarities of property values play an important role in the co-reference resolution based on similarity analysis. This paper proposed a new model for obtaining the weights of properties from the statistical information of datasets, and proved its rationality theoretically. Furthermore, the advantages of this model compared with traditional method were analyzed. Based on the new method proposed by this paper, a system for building co-reference among semantic datasets was implemented, and the performance was verified with some open datasets.

Keywords Linked data, Co-reference, Property weight

1 引言

链接数据(Linked Data)的概念由 Time Berners-Lee 在2006年7月首次提出[1],它是一种基于语义技术在互联网上发布和关联结构化数据的方法。链接数据使用可访问的HTTP URI作为任何事物的标识名称,并以RDF形式为访问者提供被访问URI的相关信息的链接。RDF提出了一个简单的二元关系表达模型来表示任意类型的数据,它用主体(subject)、谓词或属性(predicate 或 property)、客体或属性值(object 或 property value)所构成的三元组来描述资源的元数据[2]。RDF是语义网数据表达的核心规范,由于其灵活性,许多领域已经将它作为表达元数据的基本方法,因此RDF已经成为知识表达的通用形式。

随着语义网的成熟以及链接数据技术的发展,越来越多的组织和个人开始在互联网上以链接数据的形式发布数据[3,4],这导致互联网上语义数据的快速增长。然而这些来自众多不同数据源的语义数据集是由不同领域的人员独立创建和发布的,由于互联网的分布性和动态性,使得数据创建者遵循统一的命名规范并不现实。因此不同的数据集会使用不同的URI来表示同一个现实世界中的实体,这就导致数据集在创建的过程中没有或者很少引用其他数据集中的URI,从而造成来自不同数据集的数据之间的关联存在不同程度的缺失。这种关联的缺失会严重影响计算机对语义数据的理解,使计算机无法综合分析和关联不同数据源中描述同一实体的数据,从而导致在数据分析和知识发现的过程中遗漏一些重要信息,甚至产生错误的推理和结论。

目前,在不同数据集之间建立数据链接的主要方法是基于共指关系来建立URI之间的关联。共指关系是表示同一实体的两个不同URI之间的关系,通常利用owl:sameAs来表达这种关系[5]。Owl:sameAs是本体描述语言OWL定义的用于表示等价语义的词汇。不同数据集之间共指关系的发现和建立能够丰富不同数据集之间的数据链接,使来自不同数据源的数据集形成一个有机联系的数据网。数据网中的链接越多,计算机就能在数据分析和知识发现的过程中关联越

到稿日期:2012-06-30 返修日期:2012-08-02 本文受中国教育科研网格二期建设项目(ChinaGrid 2)资助。
张晓辉(1986—),男,博士生,主要研究方向为分布式系统、语义网,E-mail:xh-zh@msn.cn;邸瑞华(1947—),女,教授,主要研究方向为网络分布计算环境、分布式系统、软件工程等。

多的相关知识，从而挖掘出数据更多的价值。

2 相关工作

如何在大规模的语义数据集中快速、准确地发现共指关系成为自动构建高度发达的数据网的关键问题。目前在共指关系发现方面的研究主要分为两个方面。

(1)基于 OWL 语义进行推理

文献[6,7]利用一些属性的反函数性质进行推理。我们把一个具有反函数性质的函数设为 IFP(Inverse Functional Property)，OWL 规定两个不同实体的同一个 IFP 的值不可能相同(例如身份证号是人的一个 IFP，两个不同人的身份证号就不可能相同)。假设有 I_1 和 I_2 表示两个实体，它们都有一个 IFP 属性 P_1，对应的值分别为 V_1 和 V_2，当 $V_1=V_2$ 时，可以判定 I_1 和 I_2 表示同一个实体。这种方法要求目标数据集中必须定义完备的语义，而现实中的数据集往往不符合这样的要求，因此这种方法并不具有通用性。

(2)基于相似度理论进行分析

假设待匹配的两个实体分别为 I_a 和 I_b，它们有相同的属性 $P_i(i=1,2,\cdots,n)$，I_a 和 I_b 的 P_i 属性对应的属性值分别为 V_{ai} 和 V_{bi}。这种方法的基本思路就是首先计算 V_{ai} 和 V_{bi} 的相似度 $\mathrm{Sim}(V_{ai},V_{bi})$，然后基于 $\mathrm{Sim}(V_{ai},V_{bi})$ 得出两个实体的相似度 $\mathrm{Sim}(I_a,I_b)$。如果 $\mathrm{Sim}(I_a,I_b)$ 大于预先设定的阈值，则判定 I_a 和 I_b 具有共指关系。

文献[8]首先利用字符串相似度算法得到两个 URI 共有属性的属性值的相似度，然后求出各个属性值相似度的算术平均值作为两个 URI 所指向实体的相似度。这种将算术平均值作为实体相似度的做法并不科学，因为实体的各个属性在区分两个实体是否相同时所起的作用是不一样的，即在综合分析各个属性值相似度的过程中，应该为每个属性赋予不同的权重。

文献[9]提供了一种声明式语言 Silk-LSL(Silk-Link Specification Language)来帮助用户指定需要比较的属性值、属性值相似度的计算方法以及相关的阈值。Silk-LSL 虽然定义了为属性指定权重的标签，但是每个属性的权值必须人为指定。这种人为指定的权值往往因人而异，并不能真实地反映属性在区分实体中的重要程度。

文献[10]提出根据待匹配的数据集中相关属性的统计信息来决定属性的权重。它的基本思想是针对特定数量的以属性 P 为谓词的三元组，假如 P 的不同属性值的数量越多，那么它在判别实体共指关系中的作用就越大。因此该文提出将不同属性值的数量 $Num(VP)$ 与以 P 为谓词的三元组的数量 $Num(TP)$ 的比值 $Num(VP)/Num(TP)$ 作为属性 P 的权重。这种算法只有在特定情况下才能较为准确地反映出属性真正的权值。本文利用概率论的方法对该算法进行了数学推导，并分析了其成立的条件。

基于上述方法的不足，本文通过对大量语义数据集的分析和总结，发现属性的权重不仅与数据集中属性的取值数量有关，还与属性值的分布有关。因此，本文将属性值的分布情况作为决定属性权重的一个重要因素，基于概率模型提出一种更加通用和准确的权重计算方法。

3 基于属性权重的链接数据共指关系构建

大量的语义数据集中往往包含了很多隐含的信息，通过综合分析数据集中的各项统计信息，就能得到一些有助于实现数据链接的规律性知识，例如可以利用数据集中各个属性及其值的相关统计信息来计算出各个属性在计算实例相似度中的权重。

在基于实例属性相似度的相似实例检测方法中，属性的权重能够增加相似实例检测的准确度。由于通过人工设定的权重主观性较强，在很多情况下并不能真正反映属性在判别实例相似程度中的重要程度。因此从对大量数据的统计分析中得出属性权重，成为一种主要的属性权重判定方法。

3.1 属性权重计算的概率模型及推导

属性权重反映了属于某类实例的各个属性在判别实例相似度中的重要程度。当两个实例的同一属性 P 的值相等时，这两个实例越有可能表示同一个实体，属性 P 的权重就越大。本文利用概率模型对上述结论进行说明。

假设实例 I_1 和 I_2 同属于类 C，分别都有属性 P_1 和 P_2。实例 I_1 的属性值分别为 $V_{I_1}^{P_1}$ 和 $V_{I_1}^{P_2}$，实例 I_2 的属性值分别为 $V_{I_2}^{P_1}$ 和 $V_{I_2}^{P_2}$，$I_1=I_2$ 表示实例 I_1 和 I_2 代表同一个实体，即具有共指关系。当实例 I_1 和 I_2 在属性 P 的值相等的条件下代表同一个实体的条件概率 $P(I_1=I_2|V_{I_1}^{P}=V_{I_2}^{P})$ 越大，属性 P 的权重 W_p 就越大，即

$$P(I_1=I_2|V_{I_1}^{P_1}=V_{I_2}^{P_2})>P(I_1=I_2|V_{I_1}^{P_2}=V_{I_2}^{P_2})\rightarrow W_{P_1}>W_{P_2} \tag{1}$$

由式(1)可知，属性权重 W_p 的大小可以通过条件概率 $P(I_1=I_2|V_{I_1}^{P}=V_{I_2}^{P})$ 来间接判定。

通过分析总结，我们发现当同属某类的两个实例的某个属性 P 的值相同的概率很小时，在给定属性 P 的值相同的条件下，两个实例为同一个实体的概率就会很大。例如对于人的姓名和性别两个属性，人的姓名相同的概率要远远小于性别相同的概率；在分别给定姓名相同和性别相同两个已知条件的情况下，姓名相同的两个实例更有可能表示同一个人。设在 I_1 和 I_2 代表不同的实体的条件下，其属性 P 的值相等的概率为 $P(V_{I_1}^{P}=V_{I_2}^{P}|I_1\neq I_2)$。利用概率公式的推导，得出条件概率 $P(I_1=I_2|V_{I_1}^{P}=V_{I_2}^{P})$ 与 $P(V_{I_1}^{P}=V_{I_2}^{P}|I_1\neq I_2)$ 之间存在如下关系：

$$P(V_{I_1}^{P_1}=V_{I_2}^{P_1}|I_1\neq I_2)>P(V_{I_1}^{P_2}=V_{I_2}^{P_2}|I_1\neq I_2)\rightarrow P(I_1=I_2|V_{I_1}^{P_1}=V_{I_2}^{P_1})<P(I_1=I_2|V_{I_1}^{P_2}=V_{I_2}^{P_2}) \tag{2}$$

式(2)表示在已知不等式 $P(V_{I_1}^{P_1}=V_{I_2}^{P_1}|I_1\neq I_2)>P(V_{I_1}^{P_2}=V_{I_2}^{P_2}|I_1\neq I_2)$ 成立的条件下，对不等式 $P(I_1=I_2|V_{I_1}^{P_1}=V_{I_2}^{P_1})>P(I_1=I_2|V_{I_1}^{P_1}=V_{I_2}^{P_2})$ 的证明。下面通过概率推导的方式对式(2)进行证明。

为便于推导，我们假设事件 A 表示 $I_1=I_2$，$\neg A$ 表示 $I_1\neq I_2$；事件 B 表示 $V_{I_1}^{P_1}=V_{I_2}^{P_1}$，$\neg B$ 表示 $V_{I_1}^{P_1}\neq V_{I_2}^{P_1}$；事件 C 表示 $V_{I_1}^{P_2}=V_{I_2}^{P_2}$，$\neg C$ 表示 $V_{I_1}^{P_2}\neq V_{I_2}^{P_2}$。因此式(2)可以表示为：

$$P(B|\neg A)>P(C|\neg A)\rightarrow P(A|B)<P(A|C) \tag{3}$$

由贝叶斯公式可得 $P(A|B)=\dfrac{P(B|A)P(A)}{P(B)}$，$P(A|C)=\dfrac{P(C|A)P(A)}{P(C)}$，且由于当两个实例表示同一个实体时，它们的相同属性的值相等成为必然事件，即 $P(B|A)=1$，$P(C|A)=1$，因此，对 $P(A|B)<P(A|C)$ 的证明可以转换为对 $P(B)>$

$P(C)$的证明。

由全概率公式可得 $P(B)=P(B|A)P(A)+P(B|\neg A)P(\neg)A=P(A)+P(B|\neg A)P(\neg)A$，$P(C)=P(A)+P(C|\neg A)P(\neg)A$。已知 $P(B|\neg A)>P(C|\neg A)$，且由于概率的非负性，可得 $P(B)>P(C)$。因此，式(3)得证。

由式(2)可知，通过确定不同属性的条件概率 $P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)$，就可以得出条件概率 $P(I_1=I_2|V_{I_1}^p=V_{I_2}^p)$，进而可以确定各个属性的权重。而条件概率 $P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)$，则可以通过语义数据集中属性值的统计信息来近似得出。本文基于数据集中实例属性及其值的具体分布，推导出条件概率$P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)$的计算公式。

假设有语义数据集 D，D 中所有类的集合为 $C=\{c_i|i=1,2,3,\cdots,n\}$，与属于类 c_i 的实例 Ic_i 相关的属性集为 $Pro(c_i)=\{P_j|j=1,2,3,\cdots,m\}$。设数据集中属性 P_j 共出现了 N 次，属性 P_j 的不同值共有 m 个，每个属性值 v 会被 n 个实例的属性 P_j 取到，具体分布如表 1 所列。

表 1　属性 P_j 的值分布

属性值	v_1	v_2	v_3	v_4	……	v_m
属性出现次数	n_1	n_2	n_3	n_4	……	n_m

由于来自单一数据源的数据集中出现共指关系实例的概率很小，因此在面向包含大量实例数据的语义数据集时，可以近似认为数据集中的实例之间不存在共指关系。

根据以上分布，从 N 个 P_j 中任取两个，其值相等的概率为 $P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)=\frac{\sum_{i=1}^{m}n_i(n_i-1)}{N(N-1)}$；又由于$\sum_{i=1}^{m}n_i=N$，可得式(4)。

$$P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)=\frac{\sum_{i=1}^{m}n_i^2-N}{N(N-1)} \tag{4}$$

由式(2)可知，$P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)$与 $P(I_1=I_2|V_{I_1}^p=V_{I_2}^p)$存在反比关系，而 $P(I_1=I_2|V_{I_1}^p=V_{I_2}^p)$又与属性权重有相同的变化趋势，因此可以将 $P(V_{I_1}^p=V_{I_2}^p|I_1\neq I_2)$的倒数作为计算属性权重的基本公式，即：

$$W'_{p_j}=\frac{N(N-1)}{\sum_{i=1}^{m}n_i^2-N} \tag{5}$$

由于根据式(5)计算 W_p 得出的值变化过大，不利于后期相似度的计算，为了使 W_p 的变化较为平缓，可以对式(5)取对数，经过平滑化处理和修正后，可得式(6)。

$$W'_{p_j}=\log\left(\frac{N(N-1)}{\sum_{i=1}^{m}n_i^2-N+1}+1\right) \tag{6}$$

为了便于理解，我们可以将权重的值限定在[0,1]的范围之内。通过求出与一个类相关的所有属性权重的最大值，记为 $W_{p\max}$，再利用式(7)可以得到最终的属性权重。

$$W_{p_j}=W'_{pj}/W_{p\max} \tag{7}$$

文献[10]仅仅使用 m/N 作为权重的计算公式，在很多情况下并不能准确地反映出属性的权重。

(1)针对某个类 C 的属性 P，当 N 很小时，假设 m 接近于 N，根据文献[10]的计算公式可得属性 P 的权重接近于 1。但是由于属性 P 所关联的三元组数量很少，不能从中总结规律性的知识，因此其在判别实例共指关系中所占的作用应该很小。

假设类 C 共有 1000 个实例，其中只有 10 个实例有属性 P_1，1000 个实例都有属性 P_2，且属性值都各不相同。那么根据文献[10]中的计算公式得出 P_1 和 P_2 的权重都为 1，但是由于 P_1 所关联的三元组数量较大，因此计算出的权重并不能真正反映其实际的权重，所以 P_1 在判别实例是否具有共指关系中的作用应该小于 P_2。利用本文提出的计算公式，P_1 和 P_2 的权重分别为 $W'_{p_1}=1.95$，$W'_{p_2}=5.99$。

(2)当有两个属性 P_1 和 P_2 的数量和值的数量都分别为 N 和 m 时，m 个值的分布并不相同，文献[10]提出的方法不能很好地区分出这两个属性权重的大小。

假设类 C 共有 1000 个实例，属性 P_1 和 P_2 都各有 100 个不同的值。P_1 的值是均匀分布，即每 10 个实例取相同的值；P_2 的值是非均匀分布，假设最坏的情况是 901 个实例的属性 P_2 取同一个值，其余 99 个实例的属性 P_2 的值各不相同。根据文献[10]的计算公式可得属性 P_1 和 P_2 的权重都为 0.1，而根据本文提出的计算公式可得 $W'_{p_1}=2.04$，$W'_{p_2}=0.35$。

3.2　共指关系的构建

共指关系的构建基于两个实例数据的相似度 $\mathrm{Sim}(I_1,I_2)$ 进行判断，如果 $\mathrm{Sim}(I_1,I_2)$大于设定的阈值 t，则判定 I_1 和 I_2 具有共指关系。而 $\mathrm{Sim}(I_1,I_2)$主要基于实例的 m 个共有属性的值相似度 $\mathrm{Sim}(V_{I_1}^{p_i},V_{I_2}^{p_i})$与各个属性的权值 W_{p_i} 计算得出，计算公式如式(8)所示：

$$\mathrm{Sim}(I_1,I_2)=\sum_{i=1}^{m}\mathrm{Sim}(V_{I_1}^{p_i},V_{I_2}^{p_i})\times W_{p_i}/\sum_{i=1}^{m}W_{p_i} \tag{8}$$

属性值相似度 $\mathrm{Sim}(V_{I_1}^{p_i},V_{I_2}^{p_i})$的计算主要基于字符串相似度算法，例如基于编辑距离的相似度算法、基于语义距离的相似度算法等。为了提高相似度计算的准确性，可以采用多种相似度算法综合计算的方法得出 $\mathrm{Sim}(V_{I_1}^{p_i},V_{I_2}^{p_i})$。本文针对普通字符串主要采用 jaroWinklerSimilarity 算法[12]来计算其相似度；针对 URI 则只判断其是否相同，即当两个 URI 相同时，返回其相似度为 1，否则其相似度为 0。

4　实现及验证

本文基于开源的语义网开发工具包 Jena 对上述权重计算算法以及共指关系的构建方法进行了实现，并利用开放的语义数据集进行了测试和验证。

4.1　测试数据集

本文采用两种类型的数据集进行测试，一种是由研究机构整理发布的数据集，包括由 RKB Explorer[13]管理的 DBLP 和 ACM。这两个数据集是由研究机构从传统的数据转换而来，格式和内容都比较统一，易于实现相似实例数据的发现和链接。文献[10]中使用的测试数据集也包括这两个。DBLP 中包含 78625 个 Publication 实例，而 ACM 中包含 290401 个 Publication 实例。

本文分别从这两个数据集中提取一个 Publication 实例，并计算其属性〈http://www.aktors.org/ontology/portal#has-title〉的值的相似度，如果大于 0.5，则形成一个待匹配的实例对。然后利用本文提出的链接构建方法对待匹配的实例对进行精确匹配，从而最终确定其是否具有共指关系。

另一种是利用链接数据爬虫工具 LDSpider[14]从网络获取的 FOAF[15]数据。由于这类数据是互联网用户手动配置和生成的，即使采用了统一的 FOAF 模式，其内容也不统一，

因而实例的属性也都不尽相同，甚至会包含错误的格式和内容。本文利用 LDSpider 从互联网上获取了 188MB 的 FOAF 数据，其中包含 Person 实例 84804 个。通过属性〈http://xmlns.com/foaf/0.1/name〉的值的相似度形成带匹配的实例对，然后对其进行精确匹配。

4.2 数据集的存储和统计

本文采用 Jena TDB 实现数据集的存储和查询。TDB 使用三元组的形式对 RDF 数据提供持久性存储，以提供高性能的查询引擎。同时采用 Jena Fuseki 作用 TDB 的远程访问代理服务器。Jena Fuseki 是一个 Sparql 服务器，基于 HTTP 协议提供了一系列针对 TDB 的数据操作接口，包括查询和更新等。通过 Fuseki 可以实现对 TDB 中存储数据的远程访问。

本文使用 Sparql 查询语言对 TDB 中的数据集进行查询，获取相关的统计信息，并计算相应的属性权重。本文设计了一套 XML Schema 来对统计信息及权重进行存储，见图 1。XML Schema 中包含三级统计信息。

```
<dataset>
  <dataset-name>  </dataset-name>
  <dataset-uri>  </dataset-uri>
  <class-number>  </class-number>
  <triple-number-dataset> </triple-number-dataset>
  <class>
    <class-uri> </class-uri>
    <triple-number-class> </triple-number-class>
    <property-number> </property-number>
    <instance-number> </instance-number>
    <property>
      <property-uri>  </property-uri>
      <triple-number-property>    </triple-number-property>
      <value-number>  </value-number>
      <property-weight1>  </property-weight1>
      <property-weight2>  </property-weight2>
    </property>
  </class>
</dataset>
```

图 1 数据集统计信息模型

数据集：包含数据集的名字、URI、三元组的数量、类的数量；

类：包含类的 URI、类中三元组和实例的数量、相关的属性的个数；

属性：包含属性的 URI、属性相关的三元组数量、不同值的数量、使用不同算法计算出的多个权重。

4.3 链接结果的验证

本文主要通过人工核对的方法来验证生成的共指实例对的正确性。

针对 DBLP 和 ACM，本文用提出的权重计算方法对从待匹配的实例对中随机抽取的 200 个进行精确匹配。在给定阈值 $t=0.8$ 的条件下，共生成了 126 个共指实例对。对这些共指实例对的验证主要通过 RKB 提供的一项名为 Consistent Reference Service(CRS)的在线服务来进行实现。通过 CRS 可以查询与给定 URI 具有共指关系的相关 URI。将共指实例对中的一个实例的 URI 作为参数进行查询，如果在给出的结果中包含另一个实例的 URI，则说明这两个实例之间的共指关系确实存在。通过人工核对，共有 118 个共指实例对被确认存在共指关系。因此，针对这两个数据集，本文提出的链接构建方法的准确率达到了 93.6%。而基于文献[10]提出的权重计算方法进行精确匹配，其准确率为 89%。

针对 FOAF 数据集，本文采用同样的方法得到 200 个待匹配的实例对，并分别基于两种权重计算方法进行精确匹配，分别得到 119 和 91 个共指实例对。经过人工确认之后的共指实例对分别为 98 和 71 个，准确率分别为 82.3% 和 78%。

FOAF 数据集中的数据具有较大的随意性，内容和格式的规范性不强，这些都影响了共指关系构建的准确率。

结束语 基于属性值分布的权重计算方法从大量语义数据中分析总结出属性的权重，其能够更加准确地反映属性在共指关系构建中的真实作用。实验结果表明，本文提出的共指关系构建方法能够较为准确地从大量语义数据中找出具有共指关系的实例，但也存在以下不足：

(1)属性值相似度的计算较为简单，应该根据属性值的类型使用不同的相似度计算方法，以提高共指关系构建的准确率。

(2)准确性验证不够充分，需要在多种具有不同数据特征的数据集上进行验证。

(3)本文提出的方法在处理更大规模的数据集时，性能需要进一步提高和优化。

今后，需要在以上几个方面进行更加深入的研究和实验，提高共指关系构建的准确率和性能。

参 考 文 献

[1] Berners-Lee T. Linked Data- Design Issues[OL]. http://www.w3.org/DesignIssues/LinkedData.html

[2] Manola F, Miller E. RDF Primer. W3C [OL]. http://www.w3c.org/TR/rdf-primer/, February 2004

[3] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space[M]. Synthesisi Lectures on the Semantic Web: Theory and Technology, 2011

[4] 沈志宏，张晓林. 关联数据及其应用现状综述[J]. 现代图书情报技术，2010(11)：1-9

[5] Bizer C, Heath T, Berners-Lee T. Linked data- the story so far [J]. Int. J. Semantic Web Inf. Syst., 2009, 5(3): 1-22

[6] Hogan A, Harth A, Decker S. Performing Object Consolidation on the Semantic Web Data Graph[C]// I3. 2007

[7] Sleeman J, Finin T. Computing FOAF Co-reference Relations with Deductionand Machine Learning[C]// Proceedings of the Third International Workshop onSocial Data on the Web. November 2010

[8] Raimond Y, Sutton C, Sandler M. Automatic Interlinking of Music Datasets on the Semantic Web[C]// Linked Data on the Web Workshop (LDOW2008). 2008

[9] Volz J, Bizer C, Gaedke M, et al. Silk - A Link Discovery Framework forthe Web of Data[C]// Proceedings of the 2nd Workshop on Linked Data on the Web(LDOW2009). 2009

[10] Song D, Heflin J. Domain-independent entity coreference in RDF graphs[C]// Huang J, Koudas N, Jones G J F, et al., eds. 'CIKM'. ACM, 2010: 1821-1824

[11] 白海燕，朱礼军. 关联数据的自动关联构建研究[J]. 现代图书情报技术，2010(2)：44-49

[12] Winkler W. The state record linkage and current research problems[R]. Technical report. Statistics of Income Division, Internal Revenue Service Publication, 1999

[13] Glaser H, Millard I, Jaffri A. Rkbexplorer.com: A knowledge driven infrastructure for linked data providers[C]// ESWC. 2008: 797-801

[14] Isele R, Umbrich J, Bizer C, et al. Dspider: An Open-source Crawling Framework for the Web of Linked Data[C]// Axel Polleres&Huajun Chen, ed. 'ISWC Posters&Demos', CEUR-WS.org, 2010

[15] Brickley D, Miller L. Foaf vocabulary specification[OL]. http://xmlns.com/foaf/0.1/, 2003