软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(7):1729−1744 [doi: 10.3724/SP.J.1001.2012.04215]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

语义 Web 中对象共指的消解研究*

胡 伟[1,2+], 柏文阳[1,2], 瞿裕忠[1,2]

[1](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210093)

[2](南京大学 计算机科学与技术系,江苏 南京 210093)

Research on Resolving Object Coreference on the Semantic Web

HU Wei[1,2+], BAI Wen-Yang[1,2], QU Yu-Zhong[1,2]

[1](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210093, China)

[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: whu@nju.edu.cn

Hu W, Bai WY, Qu YZ. Research on resolving object coreference on the semantic Web. *Journal of Software*, 2012,23(7):1729−1744 (in Chinese). http://www.jos.org.cn/1000-9825/4215.htm

Abstract: Semantic Web data proliferates with the rapid growth of the semantic Web. An object on the semantic Web is likely to be denoted with many identifiers (e.g., URIs) by different parties. Resolving an object coreference on the semantic Web is to identify different identifiers for the same object and eliminate the inconsistency between their involved RDF (resource description framework) data, which is important for semantic Web data fusion, search, browsing, etc. In this paper, the problem of resolving object coreference on the semantic Web is first formalized. Next, the state of the art of works are surveyed and categorized into five aspects: The used characteristics for coreference identification, the mechanisms for data conflict resolution, the applicable scopes of current approaches, prototypes, and benchmarks. Finally, open research issues are discussed and possible future research directions are also pointed out.

Key words: object coreference; coreference resolution; instance matching; semantic Web; data fusion

摘 要: 随着语义 Web 的快速发展,语义 Web 数据大幅增长.在语义 Web 中,单个对象很可能由多个不同的标识符(例如 URI)指称.语义 Web 中,对象共指的消解是识别语义 Web 中指称相同对象的不同标识符,并消除描述这些标识符的 RDF(resource description framework)数据之间不一致性的过程,它对于语义 Web 数据的融合、搜索、浏览等具有重要作用.首先,形式化定义了语义 Web 中对象共指的消解问题;然后,从对象共指识别使用的特征、数据冲突的消解方式、对象共指消解方法的适用范围、现有原型系统和基准测试集这 5 个方面调研了最新的研究进展;最后,讨论了尚存的挑战,并展望未来可能的研究发展方向.

关键词: 对象共指;共指消解;实例匹配;语义 Web;数据融合

中图法分类号: TP182 **文献标识码**: A

* 基金项目: 国家自然科学基金(61003018, 61100040, 61021062); 国家社会科学基金(11AZD121); 国家教育部博士点基金(20100091120041); 江苏省自然科学基金(BK2011189)

收稿时间: 2011-05-20; 定稿时间: 2012-04-01; jos 在线出版时间: 2012-04-19

CNKI 网络优先出版: 2012-04-19 11:30, http://www.cnki.net/kcms/detail/11.2560.TP.20120419.1130.001.html

语义 Web(semantic Web)是 Web(万维网)的一个重要发展方向,它提供了一个通用框架,使得数据的共享和重用可以跨越应用系统、企业和社区的边界.在原始 Web 上,只有文档的交换和共享.语义 Web 以 RDF(resource description framework)为基础,而 RDF 以 URI(uniform resource identifier)作为标识机制、以 XML 作为语法,能够将各种不同应用中的数据和服务容易地集成起来.本体(ontology)在语义 Web 中扮演着重要的角色,语义 Web 本体一般是指使用 RDFS(RDF schema)或者 OWL(Web ontology language)等语言描述的本体,其中定义了类(class)、属性(property)和实例(instance).而类、属性和实例又可统称为实体(entity).近年来,有关 RDF 数据查询的 SPARQL 和有关规则表示的 RIF(rule interchange format)等技术也日趋成熟,标志着语义 Web 的数据模型、本体语言、规则语言和数据存取等技术基础已经奠定.

随着语义 Web 的快速发展,特别是链接开放数据项目(linking open data project)[1]的大力推动,语义 Web 的数据量已经达到了一个相当大的规模,覆盖的范围包括了社会网络、生物医学、政府数据、地理信息和图书音乐等众多领域,正逐步形成一个“数据之网(Web of data)”.目前,我们研发的语义 Web 搜索系统 Falcons[2]已在 Web 上发现了约 1.6 千万个语义 Web 文档(包含语义 Web 数据的 Web 文档).通过对采集的语义 Web 文档分析后发现,2008 年 9 月~2011 年 4 月期间,标识语义 Web 实体的 URI 数量已从 7.6 千万增长到 3.1 亿,增长幅度超过了 3 倍,其中,90%以上的 URI 标识的是语义 Web 实例.

由于语义 Web 中的任何机构和个人都可以自由发布语义 Web 数据,导致语义 Web 数据具有多样性和异构性(heterogeneity).语义 Web 数据的大量涌现,常会造成多个不同的标识符(例如 URI)指称真实世界中的相同对象,称为对象共指(object coreference).例如,关于万维网之父兼语义 Web 的倡导者 Berners-Lee 先生,就已经发现数百个不同的 URI 指称他.语义 Web 中普遍存在的对象共指现象阻碍了语义 Web 数据的共享和集成,不利于网络效应的发挥,造成了知识重用的“困局”[3].

语义 Web 中,对象共指的消解是识别语义 Web 中指称真实世界相同对象的不同标识符,并消除描述这些标识符的 RDF 数据之间不一致性的过程(如图 1 所示).这里,对象共指的识别和数据冲突的消解并不是相互独立的两个过程,冲突消解后的新数据可以被用来促进识别,而新识别的对象共指又需要进一步消解可能的冲突.另外,在整个消解过程中还可能引入人工参与.

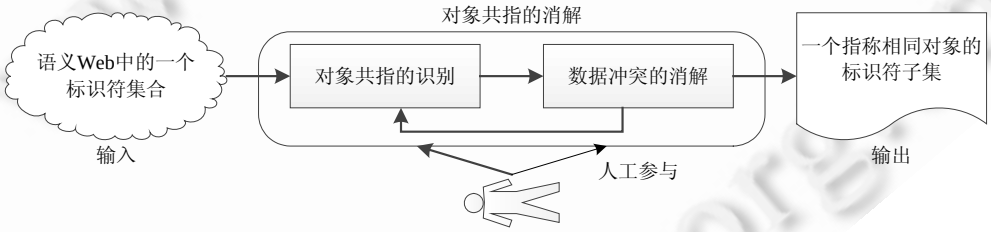

Fig.1 A basic process for object coreference resolution on the semantic Web

图 1 语义 Web 中对象共指消解的基本过程

语义 Web 中,对象共指的消解是语义 Web 研究中的一个核心问题,对于实现数据之网至关重要,是未来语义 Web 数据集成及应用的关键.目前的发展趋势表明,存储在 Web 上的海量数据将逐渐由纯文本文档以及存储在数据库中的数据(称为 deep Web)向基于 RDF 的数据模型转变,语义 Web 的应用将大幅度快速增长.对于使用异构数据的语义 Web 应用而言,对象共指的消解是消除数据之间语义异构性的一种有效途径,可以为应用程序间的交互建立一种互操作性(interoperability).这种互操作性使得这些应用程序之间可以实现诸如语义 Web 环境下的数据集成与共享、分布式查询处理、服务组合、多 Agent 通信和语义 Web 搜索与浏览等功能[4].

在语义 Web 领域,对于语义 Web 中对象共指的消解研究已经取得了部分成果.然而,语义 Web 数据呈现出的规模快速增长和异构性强等特点,导致现有方法愈显不足;而且 RDF 数据模型与传统数据模型(例如关系模型、非结构化的数据模型)之间存在较大差异,致使数据库和自然语言处理等领域的相关研究成果不能被直接应用.因此,语义 Web 中对象共指的消解问题仍有待进一步加以研究.本文对现有的语义 Web 中对象共指的消解

工作进行总结和分析,为读者提供进一步研究的方向和基础.首先,形式化地给出语义 Web 中对象共指消解的定义,并分析当前面临的主要挑战;然后,从对象共指识别使用的特征、数据冲突的消解方式、对象共指消解方法的适用范围、现有原型系统和基准测试集这 5 个方面介绍最新的研究进展,并进行归纳和比较;最后,总结目前工作的不足,并展望未来可能的研究发展方向.

1 问题描述

本节形式化地给出语义 Web 中对象共指消解问题的相关定义,并分析解决该问题所面临的主要挑战.限于篇幅,本文假设读者已经对语义 Web 的基本概念有所了解.

1.1 语义Web中对象共指的消解

给定一个 URI 引用(URI reference)的集合 $\boldsymbol{U}$、一个匿名结点(blank node)的集合 $\boldsymbol{B}$ 和一个字面量(literal)的集合 $\boldsymbol{L}$,三元组$\langle s,p,o\rangle\in(\boldsymbol{U}\cup\boldsymbol{B})\times\boldsymbol{U}\times(\boldsymbol{U}\cup\boldsymbol{B}\cup\boldsymbol{L})$被称为是一个 RDF 三元组,其中,$s$ 被称为是该三元组的主语,p 被称为是谓语,而 o 被称为是宾语.一个 RDF 图是一个 RDF 三元组的集合,而一个 RDF 文档是对一个 RDF 图的序列化(serialization)[5].

对于某个 RDF 图,一个标识符 $u\in(\boldsymbol{U}\cup\boldsymbol{B})$是一个类(或属性)当且仅当它在该 RDF 图中能够推理出一个 RDF 三元组$\langle u$,rdf:type,rdfs:Class$\rangle$($\langle u$,rdf:type,rdf:Property$\rangle$).例如:假设一个 RDF 图中存在这样一个 RDF 三元组,它的主语是 u,谓语是 rdfs:subClassOf,则可以推理出 u 是一个类.类似地,假设存在一个 RDF 三元组,它的谓语是 owl:onProperty,宾语是 u,则可以推理出 u 是一个属性.如果某个标识符 u 既不是一个类也不是一个属性,则它被认为是一个实例[2].参考 OWL DL 规范[6],假设类、属性和实例之间是不相交的,但在实际的 RDF 文档解析过程中还经常设定一些启发式规则来处理定义不一致等情况.

语义 Web 中,关于对象共指消解的研究相对较多.针对不同研究目标,存在多种不同的定义.本文从链接数据(linked data)的角度出发[7],给出语义 Web 中对象共指消解的一种定义.

定义 1(对象共指的消解). 给定一个语义 Web 实例标识符的候选集合 $\boldsymbol{D}_o=\{I_1,I_2,\ldots,I_n\}$,可能指称真实世界中的某个对象 o.对 $\boldsymbol{D}_o$ 的对象共指消解被定义为一个函数 Resolve: $D_o\rightarrow 2^{D_o\times(0,1]}$. 其中, $2^{D_o\times(0,1]}$ 表示 $\boldsymbol{D}_o\times(0,1)$的幂集.对于 $\boldsymbol{D}_o$ 的某个子集 $\boldsymbol{d}_i$,满足其确信程度 $v_i\in(0,1)$.

该定义相较于其他定义,比如文献[8]以及数据库领域中许多研究工作的定义[9],最大的不同之处在于,它不限定输入为两个待消解的实例标识符,而是允许输入为一个实例标识符的集合,这更符合语义 Web,尤其是链接数据的实际情况.即,需要在一个开放的 Web 环境中同时考虑多个实例标识符之间是否共同指称相同的对象.事实上,输入为两个实例标识符的情况可以看作是本文定义的一个特例.

这里,还有必要区别语义 Web 中对象共指的消解和语义 Web 本体的匹配(ontology matching).虽然本体包含了类、属性和实例,但是语义 Web 本体的匹配主要针对两个本体,目标是发现不同本体中类或属性之间的映射(mapping),在这一方面已有多个研究综述[4,10−14];而语义 Web 中,对象共指的消解则特别针对实例,目前,国内外高质量的相关综述还很少[15,16],因此有必要对其进行详细调研.但是这两类研究也不能完全割裂,在本体匹配方面,就有研究基于对象共指(也称为实例匹配)来匹配类或属性[17];而在对象共指消解方面,也有一些工作通过匹配类或属性来提高对象共指消解的准确度[18].

另外,本文在上下文清晰的情况下也将实例标识符简称为标识符,而将对象共指的消解简称为共指消解.

1.2 研究难点

数据库领域中,对象共指的消解常被称为记录链接(record linkage)、重复检测(duplicate detection)或记录匹配(record matching)[19−21];在自然语言处理和信息检索领域,常称之为共指消解(coreference resolution)[22−24],属于指代消解(anaphora resolution)中的一类工作;而在语义 Web 领域,也常称之为引用调和(reference reconciliation)以及对象合并(object consolidation).本质上,这些工作同属于异构数据的集成问题,但是由于语义 Web 数据具有许多不同的特点,它与其他领域中的共指消解研究还有所区别,体现在对象共指识别使用的特征、数据冲

突的消解方式、消解方法的适用范围、原型系统和基准测试集等方面.在现有的一些工作中,已经对语义 Web 中对象共指消解所面临的挑战有了部分阐述[15].本文认为,研究难点主要包括以下 3 个方面:

首先,语义 Web 数据具有明确的语义,而数据库记录和自然语言文本的语义相对较弱,甚至可能含混不清.因此,针对语义 Web 中对象共指的消解,需要充分考虑如何合理、有效地利用语义.第一,在对象共指的识别过程中可以利用 owl:sameAs、反函数型属性(inverse functional property,对于单个对象,其反函数型属性的值唯一)等构建一个规模较小但准确度高的集合.而数据库和自然语言处理领域均不具有这一特性,它们通常采用属性值相似度计算的方式来识别对象共指;第二,在数据冲突的消解方面,语义 Web 中可以考虑使用逻辑推理的方式调试(debug)或诊断(diagnose)所涉及的 RDF 数据之间的一致性,而数据库领域中,或不考虑冲突消解、或采用简单的方法实施消解,例如选择冲突数据中的最大(最小、平均)值或最近更新数据等[21];第三,在原型系统开发或基准测试集构建时,也需要考虑本体语义.

其次是应用场景不同.数据库通常仅为有限的几个应用程序服务,因此经常由应用程序的开发者独自创建和管理维护;而本体是对某一领域中公认知识的建模,所以本体模型和具体应用经常是分开的.这导致了语义 Web 数据的规模更大,异构性更强.在数据库领域,通常是针对两个数据集开展记录链接;而在语义 Web 领域则针对的是多个数据源,甚至可能涉及整个语义 Web.因而,语义 Web 中对象共指的消解方法需要考虑多个实例标识符之间的共指关系,并且原型系统可能以搜索的形式出现.另外,在基准测试集方面,真实环境下一些大型数据集只允许以 SPARQL 查询的方式在线访问,不允许直接全部下载到本地,这与其他领域的测试也存在差异.

再次,语义 Web 实例通常使用 URI 标识,因此具有 Web 可访问性(accessibility).即,可以通过标识实例的 URI 来获取实例的权威描述.这个过程被称为解引(dereference)[25].而数据库中的记录或自然语言文本均不具有这一特点.利用 Web 可访问性,不但可以获取到关于实例的更多“权威”RDF 数据,实例之间也被链接成一个更广泛的有向图结构.另外,还可以考虑利用数据源的可信度来消解数据冲突.

2 对象共指的消解方法分类

在语义 Web 发展的初期,Google 公司的 Guha 等人就指出,语义搜索中的研究型搜索(research search)主要针对的是 1~2 个对象,并给出了一个包含对象共指的搜索范例[26].近年来,随着链接开放数据项目的不断开展,众多领域中的数据通过 RDF 的形式发布和链接,语义 Web 数据量激增,语义 Web 中对象共指现象日益严重.自 2007 年起,每年国际万维网会议(WWW)都有涉及该主题的研讨会召开,例如 2007 年的“身份、标识符与鉴别(Identity,Identifiers,Identification)”研讨会和 2008 年至今每年一届的“链接 Web 数据(Linked Data on the Web)”研讨会.而每年国际语义 Web 会议(ISWC)和语义 Web 扩展会议(ESWC,原名欧洲语义 Web 会议)等高水平学术会议上也有不少相关文章发表.

现有工作从多个方面研究了语义 Web 中对象共指的消解问题.例如,设计对象共指消解的框架流程[15,27]、提出具体的共指消解方法、定义表达对象共指的语法语义[28]以及具体应用[29].共指消解方法是整个语义 Web 中对象共指消解问题的核心,结合对象共指消解的基本流程(如图 1 所示)以及第 1.2 节所述的问题难点,我们将着重从 3 个方面对已有的语义 Web 中对象共指的消解方法进行介绍和归纳,具体分类方法如图 2 所示.

首先是基于识别对象共指所使用的特征分类.既可以利用语义 Web 数据包含的“等价”语义,也可以使用基于属性值相似度计算的方法,还可以是这两种方法的不同组合.对于每类方法还可以进一步细分.具体分析可参见第 2.1 节.

其次,根据消除具有共指关系的实例标识符涉及的 RDF 数据之间不一致性的方式分类.相对简单的方式是忽略或预先避免数据之间的不一致.而消解冲突的方式可以分为基于语义一致性检测的方法和基于数据源可信度评估的方法.具体分析可参见第 2.2 节.

最后是针对对象共指消解方法的适用范围分类.对于不同的应用场景,需要选择不同的消解方法.例如在开放的语义 Web 环境中,应当选择全自动的、可以实施多个数据源之间对象共指消解的方法及工具;而针对某些封闭的特定领域,还可以有其他解决方案.具体内容请参见第 2.3 节.

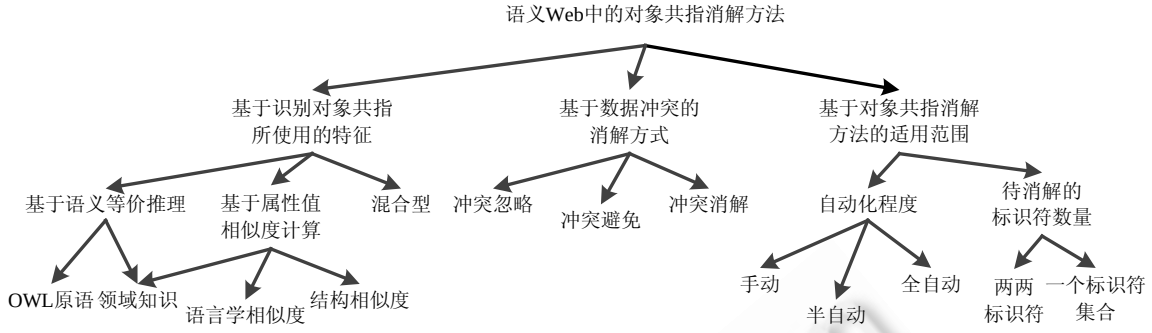

Fig.2 Classification of existing approaches for object coreference resolution

图 2 语义 Web 中现有对象共指消解方法的分类

2.1 基于识别对象共指所使用的特征分类

2.1.1 基于语义等价推理的方法

OWL 语言定义了一组原语,可以用来定义或推理不同实例标识符间的对象共指关系.其中,owl:sameAs 原语直接定义,所有使用该原语链接的标识符可以表示为⟨*s*,owl:sameAs,*o*⟩的 RDF 三元组形式,拥有相同的身份(identity),即 *s* 和 *o* 应是同样的资源[6].文献[30]正是基于 owl:sameAs 原语为指称相同对象的 URI 之间建立等价传递闭包(称为 bundle),并实现了一个对象共指查询服务(网址:http://www.rkbexplorer.com/sameAs/).

也可以使用 OWL 语言提供的反函数型属性(owl:InverseFunctionalProperty)间接推理出对象共指关系.一个反函数型属性的语义保证了对于单个对象,其反函数型属性的值唯一.例如,FOAF(friend of a friend)本体定义电子邮件地址(foaf:mbox)为一个反函数型属性,表明如果两个实例标识符拥有相同的邮件地址,则这两个标识符指称相同的对象.文献[31,32]分别在其开发的语义 Web 搜索系统 SWSE 和 Sindice 中利用反函数型属性发现具有对象共指关系的标识符,并考虑了匿名结点的对象共指,以此扩展搜索结果.爱尔兰 DERI 研究所最新开发的 Sig.ma 系统(网址:http://sig.ma/)利用语义 Web 搜索系统 Sindice 采集的反函数型属性来寻找对象共指,对以关键词指称的对象的属性值进行聚合(mash-up),提供对象的浏览服务[33].

与反函数型属性相似,还可以使用 OWL 语言中的函数型属性(owl:FunctionalProperty)和(最大)基数(owl:cardinality 或 owl:maxCardinality)来发现对象共指.其中,(最大)基数把一个约束类(restriction class)与一个数值 *N* 绑定,描述了一个类的所有实例对于某个属性而言恰好(最多)包含 *N* 个不同的取值.当 *N*=1 时,它的语义类似于函数型属性,但其作用域仅限于该类.通过分析语义 Web 搜索系统 Falcons 采集的约 6 亿条 RDF 三元组,我们发现了近 800 万条 owl:sameAs 三元组,而函数型属性和(最大)基数数量较少,这也造成了现有研究较少使用函数型属性、(最大)基数以及其他 OWL 语义等价推理规则来识别对象共指.

除了 OWL 语言定义的可以用于识别对象共指的原语外,还存在一些其他本体定义的属性也可以被用来推理对象共指关系,例如 SKOS 规范(http://www.w3.org/2004/02/skos/)中的 exactMatch 属性、UMBEL 通用本体(http://umbel.org/umbel)中的 isLike 属性以及 Gene 本体(http://www.geneontology.org/formats/oboInOwl)中的 hasExactSynonym 属性.使用这些“外部”属性的对象共指识别方法的不足之处在于,这些属性可能尚不具备普遍接受的语义,在实际使用中可能存在误用的情况.

综合使用多种原语,文献[34]设计了一种新型的对象共指数据模型 RDFS+,它在传统 RDFS 模型中加入了 OWL 语言的 owl:disjointWith、函数型和反函数型属性以及语义 Web 规则语言(SWRL)的可判别型属性.文献[35]则考虑了 owl:sameAs、owl:differentFrom、函数型和反函数型属性这 4 种原语.而英国 Southampton 大学最新研发的 SameAs.org 系统既使用了 owl:sameAs 原语,也手工指定了其他本体中 6 种可用于发现对象共指的属性,以此开发了一个在线的对象共指查询服务(网址:http://sameas.org/).

总体而言,基于等价语义推理的方法利用 OWL 语言中定义的特殊原语或其他本体属性来识别对象共指,

其准确度较高,但通常数量较少;并且也无法根据本体的语用情况发现标识符之间隐式的对象共指关系,适应性较差.最近,文献[36,37]宏观分析了链接开放数据中 owl:sameAs 原语的使用情况,发现“等价”语义的传递可能会导致错误,这也对基于语义等价推理的对象共指识别方法提出了警告.

2.1.2 基于属性值相似度计算的方法

根据 RDF 图结构,每个实例标识符一般包含一组描述它的属性及属性值,因此可以通过比较描述标识符的多个属性和属性值之间的异同来识别对象共指.基于属性值相似度计算的方法通常包含两个假设[38]:一是共指的实例标识符应该具有某些共同的特征,体现在具有相同或相似的属性和属性值;二是不同的属性及属性值对于识别对象共指具有不同的确信度,需要加以区分.语义 Web 领域中,基于属性值相似度计算的对象共指识别方法也常被称为实例匹配.

文献[7,39]分别利用本体匹配工具 HMatch 和 RiMOM,通过计算描述不同实例标识符的多个属性值之间的相似度来识别对象共指.HMatch 和 RiMOM 均集成了多种匹配算法,例如基于字符串的编辑距离、向量空间模型等,充分利用了实例的语言学描述、实例在 RDF 图中的链接结构等信息.文献[40]则以关键词作为输入,首先根据同义词和语义 Web 搜索系统 Watson 对关键词进行扩展,获得候选实例标识符集合;再为每个标识符抽取上下文(即本体模块),并计算不同上下文之间的相似度;最后,使用一种层次化聚类算法来生成多个对象共指的集合.文献[8]在假设不同的实例具有相同的本体模式的前提下,采用整数线性规划的方法来求解对象共指关系,其中综合利用了实例的语言学特征和本体模式的结构特征.此外,还基于不精确图匹配算法提出了一种近似算法来提高方法的可扩展性.文献[41]为减小实例间两两相似度计算的时间复杂度,提出了一种基于相似度度量空间的三角不等式约束的优化方法,可以过滤掉大量不满足相似度度量阈值的候选实例对,在过滤过程中所需要的阈值可以通过主动学习和遗传算法获得.

还有一些工作利用外部背景知识来辅助相似度的计算.文献[42]从 Web 2.0 服务中抽取出语义社会图结构(semantic social graphs),根据随机游走(random walk)算法计算语义社会图与描述实例标识符的 RDF 图之间的距离,并将具有对象共指关系的标识符聚类在一起,其本质是一种相似度传播算法.文献[38]在一个较大规模的数据集上(包含了约两千万个 RDF 三元组)统计了属性和属性值的使用频次,然后挑选出具有高可区分度的“伪键”属性(quasi-key property)来识别对象共指.

对于特定领域,文献[43]鉴别人名、文献[44]区分地名、文献[45]识别冗余的音乐数据、而 LinkedMDB 项目链接重复的电影纪录[46],它们在属性值相似度计算的基础上,均不同程度地引入了领域背景知识,例如人工选定有助于共指识别的特定属性(如文献[43]选用论文的标题、摘要和合作者等属性),用于提高识别的准确度.

此外,部分本体匹配方法通常先要计算实例之间的相似度,因此与基于属性值相似度计算的对象共指识别方法关系紧密.GLUE 利用实例的文本描述(例如名字、标签)训练朴素贝叶斯分类器,通过交叉分类来获得类之间的匹配[47];文献[48]首先使用字符串匹配算法获得实例匹配,然后使用 Jaccard 集合相似度和信息熵计算统计意义下的概念匹配;而文献[17]改进了文献[48]的方法,使用马尔可夫随机场(Markov random field)进一步区分描述实例的多个属性的不同重要程度.后两种方法均需要两两匹配所有实例,因此效率较低.另有研究关注概念匹配和实例匹配的相互促进,例如,文献[49]提出了一种名为 ILIADS 的匹配算法来同时匹配概念和实例,而文献[18]则通过概念匹配提高实例匹配的准确度.

另外,对于数据库领域中的重复记录检测,早在 1969 年,Fellegi 和 Sunter 就基于指称相同对象的不同记录应具有某些共性这一假设,提出了一种链接记录的方法[9],数据库领域的后续研究也大都遵循这一假设.现有研究主要基于属性值比较的思路,包括两大类方法:一类方法强调简单和高效,能够处理大规模的记录;而另一类方法则采用机器学习和概率统计等相对复杂的方法.具体请参见文献[20],在此不再赘述.

总体上,基于属性值相似度计算的方法需要比较描述实例标识符的多个属性和属性值之间的异同,当属性或属性值之间差异较大时,很难选取一个合适的确信度阈值(threshold),影响识别的准确度.而且,当实例标识符的数目很多时,成对比较这些标识符的属性和属性值效率很低.值得注意的是,数据库领域中有许多基于属性值相似度计算的研究工作可供参考.

2.1.3 混合型方法

正是由于基于语义等价推理的方法和基于属性值相似度计算的方法都存在某些不足,许多研究关注如何有效集成这两类方法,以取得更好的对象共指识别效果.其中的一种方法是先分别使用上述两类方法识别对象共指,再把识别结果合并起来,这种方法称为平行型(parallel)集成策略[4].

与平行型集成策略不同,文献[50]提出了一种基于自我训练(self-training)模型的方法,它利用语义等价推理构建一个初始训练集,随后基于这个训练集不断学习,自举式地识别对象共指,其中的核心技术是从训练集中学习出最适合识别对象共指的属性及属性值.与文献[50]类似,文献[41]也是基于一个扩展式的框架,它们的不同之处在于,文献[41]是基于 WordNet 构建初始训练集,并且仅使用实例标识符的本地名(local name)和 rdfs:label 属性的值进行扩展,没有用到语义等价推理.文献[35,51]也同样基于语义等价推理构建一个对象共指的初始集合,所不同的是,它们采用相似度传播的方法来发现新的对象共指,这类方法被称为序列型(sequential)集成策略.

虽然已有一些工作开始着手研究如何有效集成基于语义等价推理的方法和基于属性值相似度计算的方法,但是这些研究目前还处在初级阶段,还存在不少亟待解决的技术问题,例如如何根据应用场景灵活选取合适的集成策略.但是,也应该看到,这种混合型方法不仅能够发现更多的对象共指,还具有更好的适应性和灵活度,因此将成为未来对象共指识别研究的发展方向.

2.2 基于数据冲突消解方式的分类

在语义 Web 数据的发布过程中,由于数据源质量层次不齐、信息抽取和转换方法不完善等原因,常会产生不一致的 RDF 数据.对于共同指称同一对象的一组语义 Web 实例标识符,消除描述这些标识符的 RDF 数据之间不一致性的过程被称为冲突消解.

关于对象共指的 RDF 数据冲突大致可以分为两类:一类是数据层面的冲突,即描述不同实例标识符的相同属性的属性值不同.这又可细分为两种情况[19]:一种情况是由于属性或属性值的缺失,称为不确定(uncertainty).例如,对于共同指称同一对象的两个标识符,其中一个标识符使用了某一属性,而另一个则没有使用;另一种情况是对于相同属性的属性值不同,称为矛盾(contradiction).处理这种情况比处理属性或属性值缺失的情况难度更大.介于语义 Web 本体基于开放世界假设(open world assumption,区别于数据库中经常采用的封闭世界假设),属性或属性值的缺失或矛盾并不一定意味着发生冲突,需要仔细辨别;第 2 类数据冲突是语义层面的冲突,即,如果将描述不同实例的 RDF 数据融合到一起,不能推理出一个逻辑一致的模型,而这个推理过程通常要用到本体模式信息.

对于冲突处理,主要可以分为 3 种方式:冲突忽略(ignorance)、冲突避免(avoidance)和冲突消解(resolution).冲突忽略允许所有可能的不一致的 RDF 数据而不加处理;而冲突避免和冲突消解则意识到冲突的存在,并处理冲突[19].其中,冲突避免可以通过人工指定等方式预先设定冲突处理规则,当冲突发生时,总是按照该规则处理冲突.例如,对于发生属性值缺失的情况,一种冲突避免方法可以是始终选择存在的属性值;冲突消解则根据具体的冲突数据和其他相关信息来决定如何处理冲突,更为灵活也更加复杂.目前,对于语义 Web 中数据冲突的消解研究,除了采用简单的基于统计的方法,例如基于少数服从多数的选举策略、选择冲突数据中的最近更新数据等,还可以采用基于语义一致性推理和数据源信任度评估的方法.

基于语义一致性推理的方法既可以用来发现冲突,也可以用来消解冲突,具体可以分为两种方法:第 1 种方法是假设共同指称同一对象的多个实例标识符及其 RDF 数据(包含本体模式)已经合并为一个可能具有不一致性的本体,然后寻找这个不一致本体的某个一致的子集作为冲突消解后的结果[52,53];第 2 种方法则明确考虑对象共指消解的不准确性,认为产生冲突的主要原因是共指消解结果不够好,所以需要修补[54−56].一种启发式的消解方法是,对于冲突的共指消解结果,保留确信度高的对象共指,而删除确信度低的.对于前一种方法,在 RDF 数据合并之前,如果需要对象共指,那么一般使用对象共指识别的结果对实例标识符重命名;对于后一种方法,可以赋予对象共指和本体模式一种分布式语义,用分布式描述逻辑(distributed description logics)来解释它们[54].基于语义一致性推理方法的不足之处在于,逻辑推理的计算复杂度高且可伸缩性差,而且依赖于高质量的不相交关系(disjointness),但现实中仅存在少量这种关系.文献[57]给出了一种不相交关系的受监督学习方法.

在发现数据冲突的基础上,由于语义 Web 实例标识符通常具有 Web 可访问性,也可以考虑基于数据源的信任度来消解数据冲突.文献[58]在 Web 内容整合(syndication)中使用了基于数据源的信任度和更新时间的冲突消解方法.文献[59]基于因子图结构(factor graph)为已经发现的对象共指加入数据源的信任度,并设计了一种面向二部图(bipartite graph)的分布式算法来迭代传播数据源的信任度.以上研究均假设数据源的信任度是事先给定的.文献[60]详细调研了计算机科学以及语义 Web 领域中的信任度评价方法,供有兴趣的读者进一步阅读.

根据我们的调研,目前集成对象共指消解的语义 Web 应用系统还很少考虑数据冲突的消解.ObjectCoref[50]基于语义等价关系是否可以解引以及实例标识符在不同 RDF 文档中的出现次数,对共同指称同一对象的实例标识符进行排序,将可信度高的标识符排在前面,但是还未考虑数据冲突的消解.而以 Sig.ma 和 Potluck[61]为代表的语义 Web 对象聚合系统也仅简单地将描述对象共指的相同属性的属性值堆砌在一起,只提供手工方式过滤属性值之间的冲突.

在数据库领域中,早在 1983 年,Dayal 就提出了一种冲突消解的方法[62].由于数据库能够提供的信息较少,目前针对数据库中数据冲突的消解方法较为简单,一般仅利用数据本身或相关元数据(例如更新时间),从数据层面检测和消解冲突.文献[19,21]罗列了 9 种数据库中常用的冲突消解机制,并且指出,目前针对不确定型冲突的消解研究较多,而针对矛盾型冲突的研究较少.

2.3 基于对象共指消解方法的适用范围分类

针对不同的应用场景,需要选择不同的消解方法以适应需求.例如在生物医学领域,由于测量、采集、组织和管理数据的难度较大,所以只存在少数含有大量实例的本体.因为这些本体本身的规模过于巨大(例如医学领域的著名本体 LinkedCT[63]中存在大约 80 万个实例),完全依靠手工的方式构建对象共指是不现实的.同时,因为医学领域高度关注安全性,共指消解方法最好能够保证消解的准确度,而对运行速度可以适当放宽要求,因此需要尽可能地采用半自动的消解方法.对于开放式的语义 Web 环境,由于存在大量可能的对象共指,而且数据还存在频繁的变更(例如,有新的 RDF 数据发布或原来的 RDF 数据失效),所以需要同时考虑多个实例之间是否共同指称相同的对象.另外,由于时常采用搜索驱动的服务形式,一般采取全自动的消解方法.

根据上述例子,可以从两个维度分析归纳目前已有的消解方法:一是从对象共指消解方法的自动化程度上分类(手动、半自动、全自动);二是从消解方法针对的待消解的实例标识符数量上分类.即,同时考虑多个标识符,还是只考虑标识符的两两消解.

从语义 Web 中对象共指消解方法的自动化程度上看,目前已有的方法主要采用全自动的方法,而手动和半自动的方法相对较少,仅发现 Sig.ma 系统[33]允许用户对共指消解结果作确认和删除.造成这一现象的主要原因是,虽然手动或半自动的方法有时能够发现隐式的对象共指,但是对于大规模的语义 Web 数据,非常费时费力,并且发现的对象共指在很大程度上受到用户交互质量的影响.

而目前,全自动方法的不足之处在于,很难在所有领域上都取得很好的共指消解效果.换句话说,在某些特定领域上实施对象共指消解很可能会失效,原因在于针对不同领域很难选取一个统一的集成策略以及相关阈值.例如,对于生物领域实例标识符的本地名可以很好地作为消解使用的特征属性[64],但是如果使用本地名识别人,由于重名的人较多,准确度则会有所降低.针对上述不足,一些改进方法针对不同领域,或是设定某些背景知识和启发式规则用于提高消解的准确度[7],或是通过一些机器学习的方法,在小规模已标记的数据集上进行训练,再进行共指消解[50].

目前的发展状况表明,对于全自动的对象共指消解方法已有不少研究.但是人工参与的重要性也不可忽视,未来如何有效设计人机交互模式和人机交互界面来提高共指消解的效果,是一个值得深入研究的问题.文献[65]在本体匹配过程中利用了主动学习模型来提高匹配的准确度,这一工作对语义 Web 中对象共指的消解有启发作用.

从消解方法针对的待消解的实例标识符数量上看,基于语义等价推理的方法[30,50]普遍能够识别多个标识符之间的对象共指关系,其原因在于,这些方法通常假设“等价”关系具有传递性,利用这种传递性可以很容易地推理出多个实例标识符之间的对象共指.

而对于基于属性相似度计算的方法,既有针对两个实例标识符的,也有针对多个标识符的.针对两个标识符的比较,需要计算这两个标识符涉及的多个属性和属性值之间的相似度,再将这些相似度综合起来,构成这两个标识符之间的共指确信度[7,39].进而可以采用聚类的方法,将相似度较高的标识符聚类在一起,形成多个标识符之间的对象共指关系[42].另外,也可采用基于特征搜索或模式匹配的方法.例如,选取某一个标识符的某个属性及属性值作为特征,在整个数据集甚至语义 Web 中查找具有相同属性和属性值的其他标识符[41,50].

总体上看,针对多个实例标识符共指关系的消解方法更符合语义 Web 和链接数据的发展方向,也更能利用多个标识符之间的不同特征,因此在未来研究中需要多加关注.

3 原型系统与基准测试集

上一节从 3 个不同的角度对现有的语义 Web 中对象共指的消解方法进行了分类和归纳.本节首先简要介绍 8 个具有代表性的系统工具,并对它们作进一步的比较和分析;然后介绍现有的基准测试数据集.

3.1 原型系统

3.1.1 系统简介

SameAs.org 是由英国 Southampton 大学开发的一个在线的对象共指消解系统(网址:http://sameas.org/),它的前身是 RKBExplorer 系统中的 sameAs 模块[30].SameAs.org 系统基于语义 Web 搜索引擎 Sindice 提供的数据集,目前已包含 4 700 多万个实例标识符.它主要采用语义等价推理的方法来识别对象共指,其中既使用了 owl:sameAs 原语,也使用了其他本体中定义的用于表示“等价”语义的属性,例如 skos:exactMatch 等.系统的输入有两种:一种是指称某个对象的 URI,对应的系统输出是一个与该 URI 具有对象共指关系的 URI 集合(出于自反性,包括输入 URI 自身);另一种输入为关键词,其输出被划分为多个对象共指集合,每个集合中的 URI 指称真实世界中的同一个对象.该系统的特点主要在于能够准确地发现对象共指,但是针对每个输入该系统能够找到的对象共指数量相对较少,并且没有给出识别对象共指的证据.

Sig.ma[33]主要由爱尔兰 DERI 研究所设计开发,提供了一个在线的语义 Web 数据聚合(mash-up)服务(网址:http://sig.ma/),其原始数据主要来源于 RDF 数据和网页中的 RDFa 或 Microformat 格式的数据.Sig.ma 以关键词作为输入,输出为聚合后的语义 Web 数据.严格意义上说,Sig.ma 并不完全是对象共指的消解系统,但是它提供的数据浏览能力使用户能够查询对象共指.技术实现上,Sig.ma 主要使用了 owl:sameAs 和反函数型属性以及查询 OKKAM 系统来发现对象共指.另外,Sig.ma 系统提供了精心设计的用户界面,允许用户通过交互反馈来过滤错误或不一致的 RDF 数据和数据源,以提高共指消解的准确度.

OKKAM[27]是由欧盟委员会资助的第七框架项目(FP7)下的一个大规模集成项目,其基本理念是,根据 14 世纪的“奥卡姆剃刀(Occam’s razor)”原则,提倡如果没有必要则不增加实体的标识符.OKKAM 为内容创建者、编辑和开发人员等提供一个全球性的基础设施,称为实体命名系统(entity name system,简称 ENS),用于帮助人们便捷地查找相关实体的公用标识符.其中,实体命名系统包含了一种基于特征的实例匹配方法 FBEM[66],通过集成两个实例标识符的多种不同特征属性及其属性值之间的相似度,识别可能的对象共指.例如,FBEM 使用了基于 Levenstein 编辑距离的方法来比较实例标识符的本地名.

Silk[28]主要由德国 Freie Universität Berlin 大学开发,是一个用于发现两个不同链接数据源中数据之间关系的工具.Silk 定义了一种描述性的链接规范语言(link specification language),可以用来指定数据源之间待发现的链接类型(主要是共指关系)以及相互链接所需满足的条件.Silk 集成了多种相似度计算及组合方法,用于发现共指关系,例如 Levenshtein 距离、Jaro-Winkler 距离、字符串相等,或不等判断、Jaccard 距离等,而且还针对浮点数、时间类型数据或地理经纬度数据提供了特殊的计算方法,相似度组合方法上允许选择最大值、最小值或平均值等.针对不同的运行环境,Silk 开发了 3 个不同的版本,例如,可以部署到 MapReduce 计算集群上.

LN2R[34,51]主要由法国 INRIA 研究所设计开发,系统包含一个基于逻辑的对象共指消解模块 L2R 和一个基于数值的对象共指消解模块 N2R,两个模块既允许单独使用也可以组合使用.其中,L2R 方法将本体模式和数据中的语义转换为 Horn 规则,并且通过这些规则推理出对象共指或不共指关系;而 N2R 方法则将本体模式的语义

转换成不严格的相似度指标,并通过一个迭代过程传播对象共指的相似度.另外,LN2R 还设计了一种对象共指的表达语言 RDFS+.

KnoFuss[18,35]是英国 Open 大学开发的一个语义数据融合系统,其主要目标是为基于本体标注的语义数据集的融合或互联解决数据集成问题,其中重点考虑对象共指的消解和知识库的更新问题.针对对象共指的消解,KnoFuss 特别关注本体类和属性的匹配与对象共指消解之间的相互促进.具体地,KnoFuss 通过对象共指来发现类和属性之间的匹配,再利用这些类和属性的匹配来进一步提高对象共指消解的准确度以及扩大其覆盖面.其中,对象共指的识别用到了多种 OWL 语言定义的原语和信念传播(belief propagation)算法.

RiMOM[39,67]是清华大学研发的一种集成了多种本体匹配方法的多策略本体匹配系统,其中也包含了多种实例匹配方法.针对实例匹配,RiMOM 将每个实例所含信息分为 6 类:URL、元信息、名称、字符串类型信息、非字符串类型信息和邻居信息.通过基于编辑距离的方法和向量空间模型,计算实例所含各种信息之间的相似度,并使用元信息和非字符串类型信息进一步过滤,最后通过多种策略将各种相似度集成起来用于发现对象共指.RiMOM 系统参加了多届本体匹配工具评测(OAEI),在各个测试数据集上均取得了不错的效果.

ObjectCoref[50]是南京大学研发的一种基于自我训练模型的对象共指消解系统(在线网址:http://ws.nju.edu.cn/objectcoref/).它基于语义 Web 搜索系统 Falcons 提供的数据集,目前已经包含 7 300 多万个实例标识符.ObjectCoref 首先利用语义等价推理,包括 owl:sameAs、函数型或反函数型属性以及基数或最大基数限制,构建出一个初始训练集;随后,基于这个训练集不断学习,自举式地识别对象共指,其中的关键技术是从训练集中学习出最适合识别对象共指关系的属性及属性值.该系统还考虑了频繁属性组合,同时使用两个属性识别对象共指(例如经度和纬度、姓和名),进一步提高消解的准确度.另外,还基于语义等价关系是否可以解引以及实例标识符在不同 RDF 文档中的出现次数等,对共同指称同一对象的实例标识符进行排序.ObjectCoref 提出了一种新的语义等价推理与相似度计算相集成的体系结构,能够较为全面地识别对象共指,但是训练集中的错误共指关系可能会导致学习过程中的错误积累,使得识别的准确性降低.

3.1.2 分析与比较

除了以上提到的 8 个具有代表性的原型系统外,还有许多各具特色的系统和工具,例如 ASMOV[56],CODI[8],DSSim[68],HMatch[7]和 Zhishi.links[69].一般而言,语义 Web 中对象共指消解原型系统之间的主要区别在于:对象共指识别使用的特征、数据冲突的消解方式以及能够处理的实例标识符的数量.

根据对象共指识别使用的特征、数据冲突的消解方式和能够处理的实例标识符数量这 3 个方面,现将上述 8 个原型系统进行比较总结,详见表 1.表中给出的原型系统名称表示该系统使用相应的对象共指识别特征(列)以及每次能够处理的实例标识符数量(行),而系统名称后的方括号内给出的是本文认为的该原型系统所具有的突出特点.

Table 1 Comparison on the prototypes

表 1 原型系统比较

待消解的标识符数量	对象共指识别使用的特征		
	基于语义的	基于相似度的	混合型
两两标识符		OKKAM [特征匹配、ENS] Silk [链接规范语言] RiMOM [相似度组合]	LN2R [逻辑推理和数值相似度传播]
一个标识符集合	SameAs.org [OWL 原语和领域知识] Sig.ma [手动冲突消解、用户友好]		ObjectCoref [自我训练模型] KnoFuss [本体匹配与共指消解互动]

表 1 表明,针对两个实例标识符的对象共指消解系统和针对一组实例标识符的共指消解系统数量相等.这是由于各个原型系统所针对的应用场景不同,致使所采用的消解方法有所差异.对于使用到属性相似度计算的原型系统而言,主要考虑处理两个数据源内实例标识符之间的共指关系;而针对链接开放数据项目的原型系统,则通常考虑多个标识符之间的对象共指消解.本文认为,随着语义 Web 数据量的不断增加,能够消解多个标识符之间对象共指关系的方法和系统将逐渐增多.

其次,采用基于属性相似度计算和混合型方法的原型系统稍多(各 3 个),并且这些原型系统均为全自动的;而单独使用基于语义等价推理的系统略少.这个情况说明,越来越多的方法和工具都意识到,任何一种对象共指识别方法都无法应对所有的应用场景,必须考虑多种识别方法的集成问题,从而保证在大多数情况下都能获得较为理想的结果,既保证对象共指识别的准确性,也尽量多地发现对象共指.

再者,除了 Sig.ma 支持手工的数据冲突消解以外,其他原型系统还均未考虑这一问题.虽然关于数据冲突消解的理论重要性已被研究人员普遍认同,但在实际应用中还需要考虑用户交互模式和界面设计等问题.有关资料表明,ObjectCoref 正朝着这一方向努力.

3.2 基准测试集

基准测试集给研究人员提供一个公平的测试平台,例如,已举办多年的内容自动抽取评测 ACE(automatic content extraction)为自然语言处理中的指代消解提供了基准测试集.基准测试集的有无及是否科学,从另一个角度反映出目前该领域的研究是否成熟.语义 Web 中对象共指的消解研究还没有严格意义的基准测试集,但已经出现了一些简单的测试集.

3.2.1 测试集简介

本体匹配工具评测 OAEI(ontology alignment evaluation initiative,网址:http://oaei.ontologymatching.org/)是一个自 2004 年起每年一届的国际性比赛,旨在建立一个基准测试平台,以评估和比较本体匹配工具.OAEI 可被看作对现有本体匹配工具性能的一个较全面的考察,其中包含的测试集覆盖了真实世界的许多领域,且不同测试集之间也存在显著差异,体现在测试集的规模、评测方法等方面.在 OAEI 2009~OAEI 2011 中,评测组织者共提供了 5 个包含标准结果(gold standard)的测试集,用于评估和比较各个实例匹配/对象共指消解系统.评测方法主要采用了传统信息检索领域的精度(precision)、召回率(recall)和 F-Measure 这 3 项指标.相关测试集介绍如下:

IIMB(ISLab instance matching benchmark)测试集由意大利 Milan 大学构建,数据来源于 OKKAM 项目,包含了关于演员、运动员和企业公司的 RDF 数据.根据不同的修改策略,该测试集被划分成 37 个子文档,每个测试子文档含有约 200 个实例标识符及 RDF 数据,需与一个特定 RDF 文档进行匹配.测试子文档按照 5 种策略划分:第 1 类(01)仅改变实例的标识符,不改变其他任何数据;第 2 类(002~010)进行了属性值的变换,例如对每个属性包含的属性值随机修改,同时对修改的程度也进行了控制,在一些例子上修改较多而在另一些例子上修改较少;第 3 类(011~019)进行了结构修改,例如删除某些属性值、将某些数据类型属性(datatype property)改为对象属性(object property)以及将某个属性拆分为多个属性等;第 4 类(020~029)对数据作了逻辑变换,例如将某个实例定义为同一个类的不同子类的实例;而最后一类(030~037)综合运用了上述几种修改方法.IIMB 测试集数据量适中,可以从不同角度较为全面地分析评价各对象共指消解系统的优缺点,但是由于该测试集是按照某些模式人为构建的,因此可能与真实情况存在差异.

PR(persons-restaurants)测试集包含了 3 对规模适中的数据集,其中两对是关于人的,另一对是关于餐馆的,每个数据集包含数百个到几千个实例标识符.该测试集与 IIMB 测试集类似,也是根据某些修改策略对数据集作了不同程度的改变.但与 IIMB 不同,PR 测试集的修改主要是添加或删除一些属性,而对具体的属性值则很少改动.对于对象共指消解系统而言,可以通过识别关键的特征属性值来发现对象共指,比如对于餐馆而言,餐馆的电话号码以及地址可以作为识别对象共指的特征.

DI(data interlinking)测试集是一个规模较大的数据集,主要包含了多个生物医学领域的大型本体,还涉及了关于电影的大型本体 LinkedMDB 以及 DBpedia.其中,部分数据集以 RDF 文件的形式下载,例如 DrugBank 和 Diseasome;而另一部分本体由于版权等问题,需要通过 SPARQL 查询方式获取,例如 STITCH.该测试集强调全自动的对象共指消解方法,并且不应使用先验知识及实例数据背后的本体模式.由于这些本体的规模很大,同时部分数据集需要通过网络在线访问,因此给目前的对象共指消解系统带来了很大的挑战.在 OAEI 2010 中,仅有 RiMOM 和 ObjectCoref 两个系统参与了测试,并且效果都不理想.

NYT(interlinking New-York times data)测试集也是一个规模较大的数据集,数据来源于纽约时报的开放链接数据,主要包含人、组织和地点 3 个领域.测试包括两个任务:一个是重建纽约时报数据集内部的共指关系;另

一个是与 Dbpedia,Geonames 以及 Freebase 数据集之间寻找共指关系.由于测试集的规模很大,评测组织者建议使用搜索和解引的方式来发现对象共指关系,而不将数据集下载到本地后再作处理.

A-R-S(eprints-rexa-SwetoDBLP)测试集和 T-S-D(TAP-SwetoTestbed-DBpedia)测试集分别包含了 3 个大型数据集,对它们需要两两匹配.其中,A-R-S 测试集的实例数据来源于文献出版领域,而 T-S-D 测试集中的实例覆盖了多个不同领域.

3.2.2 分析与比较

现有的测试集主要来源于本体匹配工具评测 OAEI,测试数据覆盖了多个领域,对于语义 Web 中对象共指的消解方法研究和系统构建起到了推动作用.但是应该看到,目前这些测试集主要还是要求以两两匹配的形式消解对象共指,无法扩展到开放的语义 Web 环境,使得一些适用于多个实例标识符间对象共指消解的方法和系统无法完全发挥其能力.因此,进一步完善现有测试集或者构建更好的基准测试集将是未来一个重要的课题.

本文认为,近年来语义 Web 挑战赛(semantic Web challenge)中发布的 BTC(billion triple challenge)数据集能够提供某种尝试:首先,BTC 数据集从多个语义 Web 搜索引擎(例如 Sindice 和 Falcons)以及 DBpedia 本体中抽取了超过 10 亿个 RDF 三元组,覆盖了链接开放数据项目中的众多真实领域;其次,该数据集已经被用于语义搜索评测等比赛中,受到了较为广泛的认可.因此,可以考虑使用 BTC 数据集以查询驱动的方式来评估对象共指的消解方法,但是构建全面而合理的测试输入集合是一个难点.

4 总结与展望

本文形式化地定义了语义 Web 中对象共指的消解问题,并分析了主要的研究难点.从对象共指识别使用的特征、数据冲突的消解方式以及对象共指消解方法的适用范围这 3 个角度,详细介绍了语义 Web 中对象共指消解问题的基本特征、常用的解决途径和最新研究进展.还分别介绍并比较了 8 个具有代表性的对象共指消解原型系统和 5 个常用的基准测试集.通过调研可以看出,虽然语义 Web 中对象共指的消解是一个新兴的研究课题,但是也有相关领域的研究成果可供借鉴,例如重用数据库和自然语言处理领域的一些方法.同时,语义 Web 中对象共指的消解问题本身也具有特殊性,仍然存在着许多亟待应对的挑战.概括起来,主要有以下 3 个方面:

- 对于语义 Web 中对象共指消解方法的改进,目前虽然已经提出了不少方法,但是绝大部分方法还不完善.首先,针对对象共指识别的方法较多,但是考虑数据冲突消解的较少,并且很少有工作将对象共指识别和数据冲突消解有机地结合起来;其次,现有的部分方法仅考虑实例标识符之间的两两共指关系,而针对多个实例标识符之间的消解研究还不够深入,没有完全体现出链接开放数据环境下对象共指消解的新特点.此外,伴随语义 Web 的不断发展,多语言问题将逐渐显现出来,面向多语言或跨语言的对象共指消解方法也需要受到进一步关注.所以在今后的研究中,有必要提出一些新的消解方法,同时对于现有方法也需要进一步加以改进;
- 对于语义 Web 中对象共指消解原型系统的研发,现有系统的功能都非常有限.这些系统仅仅集成了一种或很少几种消解方法,并且对于消解结果的集成也不够灵活,还尚未真正创造出新型的体系架构.而且从 OAEI 结果不难看出,目前的原型系统处理大规模数据的能力还很缺乏,因此在实际应用中还存在局限性.另外,通过使用可以感觉到,目前的系统无论在功能还是在稳定性、易用性等方面都与实际使用还有一段距离.未来需要的是一个具有良好可伸缩性的、整合的、能够完成多种消解任务的实用系统.因此,如何对各种不同方法或者算法进行有效集成、构建出更好的对象共指消解系统,是未来的一个工作方向;
- 对于基准测试集及对象共指消解结果的评价,目前已有的 OAEI 测试集主要针对实例标识符之间的两两匹配,并没有涉及多个标识符之间的共指消解.而且,其中部分数据集是人为构造的,并不能完全反映出不同方法或原型系统在真实环境下的性能.在传统的自然语言处理及语义搜索等领域中,都已经构建出一些来自真实世界的大规模测试集,例如 ACE 测试集和 BTC 测试集.对于语义 Web 中对象共指的消解研究,也迫切需要建立一个适用于链接数据环境的基准测试集,使开发人员能够统一地评价消解

结果,促进语义 Web 中对象共指消解方法和工具的进一步发展.并且在这种大规模开放环境中,仅采用传统的基于精度、召回率的评价方法可能不够,还需要设计出新的定量指标以更全面、合理地评价消解结果.

总之,语义 Web 中对象共指的消解是语义 Web 研究中的一个重要问题.国际上,在语义 Web 对象共指消解方面的研究很活跃,而国内也有相关研究,并且取得了一定进展.可以预见,随着语义 Web 数据量的不断增加,将会有更多针对语义 Web 中对象共指消解的方法、系统和测试集涌现出来.

致谢 感谢东南大学计算机科学与工程学院漆桂林教授为本文第 2.2 节的写作提供了部分素材.感谢荷兰阿姆斯特丹自由大学计算机系王胜惠博士和南京大学计算机科学与技术系程龚博士的修改意见.

References:

[1] Bizer C, Heath T, Berners-Lee T. Linked data—The story so far. Int'l Journal on Semantic Web and Information Systems, 2009, 5(3):1−22. [doi: 10.4018/jswis.2009081901]

[2] Cheng G, Qu YZ. Searching linked objects with Falcons: Approach, implementation and evaluation. Int'l Journal on Semantic Web and Information Systems, 2009,5(3):49−70. [doi: 10.4018/jswis.2009081903]

[3] Alani H, Brewster C. Ontology ranking based on the analysis of concept structures. In: Clark P, Schreiber G, eds. Proc. of the 3rd Int'l Conf. on Knowledge Capture. Banff: ACM, 2005. 51−58. [doi: 10.1145/1088622.1088633]

[4] Euzenat J, Shvaiko P. Ontology Matching. Heidelberg: Springer-Verlag, 2007.

[5] Klyne G, Carroll JJ. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation 10 February 2004. Latest version. http://www.w3.org/TR/rdf-concepts/

[6] Patel-Schneider PF, Hayes P, Horrocks I. OWL Web ontology language semantics and abstract syntax. W3C Recommendation 10 February 2004. Latest version: http://www.w3.org/TR/owl-semantics/

[7] Ferrara A, Lorusso D, Montanelli S. Automatic identity recognition in the semantic Web. In: Bouquet P, Halpin H, Stoermer H, Tummarello G, eds. Proc. of the 1st Int'l Workshop on Identity and Reference on the Semantic Web. Tenerife, 2008. http://ceur-ws.org/Vol-422/irsw2008-submission-2.pdf

[8] Noessner J, Niepert M, Meilicke C, Stuckenschmidt H. Leveraging terminological structure for object reconciliation. In: Aroyo L, Antoniou G, Hyvönen E, ten Teije A, Stuckenschmidt H, Cabral L, Tudorache T, eds. Proc. of the 7th Extended Semantic Web Conf. LNCS 6088, Heidelberg: Springer-Verlag, 2010. 334−348.

[9] Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Society, 1969,64(328):1183−1210.

[10] Shvaiko P, Euzenat J. Ten challenges for ontology matching. In: Meersman R, Tari Z, eds. Proc. of the Move to Meaningful Systems. LNCS 5332, Heidelberg: Springer-Verlag, 2008. 1164−1182. [doi: 10.1007/978-3-540-88873-4_18]

[11] Choi N, Song I, Han H. A survey on ontology mapping. SIGMOD Record, 2006,35(3):34−41. [doi: 10.1145/1168092.1168097]

[12] Noy NF. Semantic integration: A survey of ontology-based approaches. SIGMOD Record, 2004,33(4):65−70. [doi: 10.1145/1041410.1041421]

[13] Kalfoglou Y, Schorlemmer M. Ontology mapping: The state of the art. The Knowledge Engineering Review, 2003,18(1):1−31. [doi: 10.1017/S0269888903000651]

[14] Qu YZ, Hu W, Zheng DD, Zhong XY. Mapping between relational database schemas and ontologies: The state of the art. Journal of Computer Research and Development, 2008,45(2):300−309 (in Chinese with English abstract).

[15] Glaser H, Lewy T, Millard I, Dowling B. On coreference and the semantic Web. Technical Report, 15245, Southampton: University of Southampton, 2007. http://eprints.soton.ac.uk/265245/

[16] Morris A, Velegrakis Y, Bouquet P. Entity identification on the semantic Web. In: Gangemi A, Keizer J, Presutti V, Stoermer H, eds. Proc. of the 5th Workshop on Semantic Web Applications and Perspectives. Rome, 2008. http://disi.unitn.it/~velgias/docs/MorrisVB08.pdf

[17] Wang S, Englebienne G, Schlobach S. Learning concept mappings from instance similarity. In: Sheth A, Staab S, Dean M, Paolucci M, Maynard D, Finin T, Thirunarayan K, eds. Proc. of the 7th Int'l Semantic Web Conf. LNCS 5318, Heidelberg: Springer-Verlag, 2008. 339−355. [doi: 10.1007/978-3-540-88564-1_22]

[18] Nikolov A, Uren V, Motta E, de Roeck A. Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: Gomez-Perez A, Yu Y, Ding Y, eds. Proc. of the 4th Asian Semantic Web Conf. LNCS 5926,

Heidelberg: Springer-Verlag, 2009. 332−346. [doi: 10.1007/978-3-642-10871-6_23]

[19] Bleiholder J, Naumann F. Data fusion. ACM Computing Surveys, 2008,41(1):1−41.

[20] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans. on Knowledge and Data Engineering, 2007,19(1):1−16. [doi: 10.1109/TKDE.2007.250581]

[21] Naumann F, Bilke A, Bleiholder J, Weis M. Data fusion in three steps: Resolving inconsistencies at schema-, tuple-, and value-level. IEEE Data Engineering Bulletin, 2006,29(2):21−31.

[22] Mitkov R. Anaphora Resolution. Edinburgh: Longman, 2002.

[23] Wang HF. Survey: Computational models and technologies in anaphora resolution. Journal of Chinese Information Processing, 2002,16(6):3−9 (in Chinese with English abstract).

[24] Zhao J. A survey on named entity recognition, disambiguation and cross-lingual coreference resolution. Journal of Chinese Information Processing, 2009,23(2):4−17 (in Chinese with English abstract).

[25] Jacobs I, Walsh N. Architecture of the World Wide Web. Vol.1. W3C Recommendation 15 December 2004. Lastest version: http://www.w3.org/TR/webarch/

[26] Guha R, McCool R, Miller E. Semantic search. In: Proc. of the 12th Int'l World Wide Web Conf. Budapest, 2003. 700−709. http://dl.acm.org/citation.cfm?id=775250

[27] Bouquet P, Stoermer H, Bazzanella B. An entity name system (ENS) for the semantic Web. In: Bechhofer S, Hauswirth M, Hoffmann J, Koubarakis M, eds. Proc. of the 5th European Semantic Web Conf. LNCS 5021, Heidelberg: Springer-Verlag, 2008. 258−272.

[28] Volz J, Bizer C, Gaedke M, Kobilarov G. Discovering and maintaining links on the Web of data. In: Bernstein A, Karger DR, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K, eds. Proc. of the 8th Int'l Semantic Web Conf. LNCS 5823, Heidelberg: Springer-Verlag, 2009. 650−665. [doi: 10.1007/978-3-642-04930-9_41]

[29] Chen JX, Zhu ZM, Liu YT, Wu DL. Study on the applications of contributor identifiers. Journal of Intelligence, 2010,29(12): 134−140 (in Chinese with English abstract).

[30] Glaser H, Jaffri A, Millard IC. Managing co-reference on the semantic Web. In: Bizer C, Heath T, Berners-Lee T, Idehen K, eds. Proc. of the 2nd Workshop on Linked Data on the Web. Madrid, 2009. http://eprints.soton.ac.uk/267587/1/ldow2009_paper11.pdf

[31] Hogan A, Harth A, Decker S. Performing object consolidation on the semantic Web data graph. In: Bouquet P, Stoermer H, Tummarello G, Halpin H, eds. Proc. of the 1st Workshop on I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web. Banff, 2007. http://www2007.org/workshops/paper_135.pdf

[32] Oren E, Delbru R, Catasta M, Cyganiak R, Stenzhorn H, Tummarello G. Sindice.com: A document-oriented lookup index for open linked data. Int'l Journal of Metadata, Semantics and Ontologies, 2008,3(1):37−52. [doi: 10.1504/IJMSO.2008.021204]

[33] Tummarello G, Cyganiak R, Catasta M, Danielczyk S, Delbru R, Decker S. Sig.ma: Live views on the web of data. Journal of Web Semantics, 2010,8(4):355−364. [doi: 10.1016/j.websem.2010.08.003]

[34] Saïs F, Pernelle N, Rousset M. L2R: A logical method for reference reconciliation. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2007. 329−334. http://membres-liglab.imag.fr/rousset/publis/L2R-aaai07.pdf

[35] Nikolov A, Uren V, Motta E, de Roeck A. Refining instance coreferencing results using belief propagation. In: Domingue J, Anutariya C, eds. Proc. of the 3rd Asian Semantic Web Conf. LNCS 5367, Heidelberg: Springer-Verlag, 2008. 405−419. [doi: 10.1007/978-3-540-89704-0_28]

[36] Ding L, Shinavier J, Shangguan Z, McGuinness DL. SameAs networks and beyond: Analyzing deployment status and implications of owl:sameAs in linked data. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, Zhang L, Pan JZ, Horrocks I, Glimm B, eds. Proc. of the 9th Int'l Semantic Web Conf. LNCS 6496, Heidelberg: Springer-Verlag, 2010. 145−160.

[37] Halpin P, Hayes PJ, McCusker JP, McGuinness DL, Thompson HS. When owl:sameAs isn't the same: An analysis of identity in linked data. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, Zhang L, Pan JZ, Horrocks I, Glimm B, eds. Proc. of the 9th Int'l Semantic Web Conf. LNCS 6496, Heidelberg: Springer-Verlag, 2010. 305−320. [doi: 10.1007/978-3-642-17746-0_20]

[38] Hogan A, Polleres A, Umbrich J, Zimmermann A. Some entities are more equal than others: Statistical methods to consolidate linked data. In: Ceri S, Della Valle E, Hendler J, Huang ZS, eds. Proc. of the 4th Workshop on New Forms of Reasoning for the Semantic Web: Scalable & Dynamic. Heraklion, 2010. http://wasp.cs.vu.nl/larkc/nefors10/paper/nefors10_paper_4.pdf

[39] Li JZ, Tang J, Li Y, Luo Q. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Trans. on Knowledge and Data Engineering, 2009,21(8):1218−1232. [doi: 10.1109/TKDE.2008.202]

[40] Gracia J, d'Aquin M, Mena E. Large scale integration of senses for the semantic Web. In: Quemada J, León G, Maarek YS, Nejdl W, eds. Proc. of the 18th Int'l Conf. on World Wide Web. Madrid: ACM, 2009. 611−620. [doi: 10.1145/1526709.1526792]

[41] Ngomo AN, Auer S. LIMES—A time-efficient approach for large-scale link discovery on the Web of data. In: Walsh T, ed. Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence. Barcelona: AAAI Press, 2011. 2312−2317. http://ijcai.org/papers11/Papers/IJCAI11-385.pdf

[42] Rowe M. Applying semantic social graphs to disambiguate identity reference. In: Aroyo L, Traverso P, Ciravegna F, Cimiano P, Heath T, Hyvönen E, Mizoguchi R, Oren E, Sabou M, Simperl EPB, eds. Proc. of the 6th European Semantic Web Conf. LNCS 5554, Heidelberg: Springer-Verlag, 2009. 461−475. [doi: 10.1007/978-3-642-02121-3_35]

[43] Aswani N, Bontcheva K, Cunningham H. Mining information for instance unification. In: Cruz I, Decker S, Allemang D, Preist C, Schwabe D, Mika P, Uschold M, Arogo L, eds. Proc. of the 5th Int'l Semantic Web Conf. LNCS 4273, Heidelberg: Springer-Verlag, 2006. 329−342. [doi: 10.1007/11926078_24]

[44] Volz R, Kleb J, Mueller W. Towards ontology-based disambiguation of geographical identifiers. In: Bouquet P, Stoermer H, Tummarello G, Halpin H, eds. Proc. of the 1st Workshop on I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web. Banff, 2007. http://www2007.org/workshops/paper_132.pdf

[45] Raimond Y, Sutton C, Sandler M. Automatic interlinking of music datasets on the semantic Web. In: Bizer C, Heath T, Idehen K, Berners-Lee T, eds. Proc. of the 1st Workshop on Linked Data on the Web. Beijing, 2008. http://events.linkeddata.org/ldow2008/papers/18-raimond-sutton-automatic-interlinking.pdf

[46] Hassanzadeh O, Consens M. Linked movie data base. In: Bizer C, Heath T, Berners-Lee T, Idehen K, eds. Proc. of the 2nd Workshop on Linked Data on the Web. Madrid, 2009. http://ceur-ws.org/Vol-538/ldow2009_paper12.pdf

[47] Doan A, Madhavan J, Domingos P, Halevy A. Learning to map between ontologies on the semantic Web. In: Proc. of the 11th Int'l World Wide Web Conf. Honolulu: ACM, 2002. 662−673. [doi: 10.1145/511446.511532]

[48] Issac A, van der Meij L, Schlobach S, Wang S. An empirical study of instance-based ontology matching. In: Aberer K, Choi K, Noy N, Allemang D, Lee K, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux P, eds. Proc. of the 6th Int'l Semantic Web Conf. and 2nd Asian Semantic Web Conf. LNCS 4825, Heidelberg: Springer-Verlag, 2007. 253−266.

[49] Udrea O, Getoor L, Miller RJ. Leveraging data and structure in ontology integration. In: Chan CY, Ooi BC, Zhou A, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Beijing: ACM, 2007. 449−460. http://dl.acm.org/citation.cfm?id=1247531

[50] Hu W, Chen JF, Qu YZ. A self-training approach for resolving object coreference on the semantic Web. In: Srinivasan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R, eds. Proc. of the 20th Int'l Conf. on World Wide Web. Hyderabad: ACM, 2011. 87−96. [doi: 10.1145/1963405.1963421]

[51] Saïs F, Pernelle N, Rousset M. Combining a logical and a numerical method for data reconciliation. Journal on Data Semantics, 2009,XII:66−94. [doi: 10.1007/978-3-642-00685-2_3]

[52] Huang Z, van Harmelen F, ten Teije A. Reasoning with inconsistent ontologies. In: Kaelbling LP, Saffiotti A, eds. Proc. of the 19th Int'l Joint Conf. on Artificial Intelligence. Edinburgh: AAAI Press, 2005. 454−459.

[53] Schlobach S. Diagnosing terminologies. In: Veloso MM, Kambhampati S, eds. Proc. of the 20th National Conf. on Artificial Intelligence. Pittsburgh: AAAI Press, 2005. 670−675.

[54] Meilicke C, Stuckenschmidt H, Tamilin A. Repairing ontology mappings. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2007. 1408−1413. https://dkm.fbk.eu/tamilin/publications/2007/aaai/paper.pdf [doi: 10.1007/978-3-540-87696-0_11]

[55] Qi G, Ji Q, Haase P. A conflict-based operator for mapping revision: Theory and implementation. In: Bernstein A, Karger DR, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K, eds. Proc. of the 8th Int'l Semantic Web Conf. LNCS 5823, Heidelberg: Springer-Verlag, 2009. 521−536. [doi: 10.1007/978-3-642-04930-9_33]

[56] Jean-Mary YR, Shironoshita E, Kabuka MR. Ontology matching with semantic verification. Journal of Web Semantics, 2009,7(3): 235−251. [doi: 10.1016/j.websem.2009.04.001]

[57] Meilicke C, Völker J, Stuckenschmidt H. Learning disjointness for debugging mappings between lightweight ontologies. In: Gangemi A, Euzenat J, eds. Proc. of the 16th Int'l Conf. on Knowledge Engineering: Practice and Patterns. LNCS 5268, Heidelberg: Springer-Verlag, 2008. 93−108. [doi: 10.1007/978-3-540-87696-0_11]

[58] Golbeck J, Halaschek-Wiener C. Trust-Based revision for expressive Web syndication. Journal of Logic and Computation, 2009, 19(5):771−790. [doi: 10.1093/logcom/exn045]

[59] Cudré-Mauroux P, Haghani P, Jost M, Aberer K, de Meer H. idMesh: Graph-Based disambiguation of linked data. In: Quemada J, León G, Maarek YS, Nejdl W, eds. Proc. of the 18th Int'l Conf. on World Wide Web. Madrid: ACM, 2009. 591−600.

[60] Artz D, Gil Y. A survey of trust in computer science and the semantic Web. Journal of Web Semantics, 2007,5(2):58−71. [doi: 10.1016/j.websem.2007.03.002]

[61] Huynh DF, Miller RC, Karger DR. Potluck: Data mash-up tool for casual users. In: Aberer K, Choi K, Noy N, Allemang D, Lee K, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux P, eds. Proc. of the 6th Int'l Semantic Web Conf. and the 2nd Asian Semantic Web Conf. LNCS 4825, Heidelberg: Springer-Verlag, 2007. 239−252. [doi: 10.1016/j.websem.2008.09.005]

[62] Dayal U. Processing queries over generalization hierarchies in a multidatabase system. In: Schkolnick M, Thanos C, eds. Proc. of the 9th Int'l Conf. on Very Large Data Bases. Florence: Morgan Kaufmann Publishers, 1983. 342−353.

[63] Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A linked data space for clinical trials. Technical Report, CoRR abs/0908.0567, Toronto: University of Toronto, 2009. http://arxiv.org/abs/0908.0567

[64] Ghazvinian A, Noy NF, Jonquet C, Shah N, Musen MA. What four million mappings can tell you about two hundred ontologies. In: Bernstein A, Karger DR, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K, eds. Proc. of the 8th Int'l Semantic Web Conf. LNCS 5823, Heidelberg: Springer-Verlag, 2009. 229−242. [doi: 10.1007/978-3-642-04930-9_15]

[65] Shi F, Li JZ, Tang J, Xie GT, Li HY. Actively learning ontology matching via user interaction. In: Bernstein A, Karger DR, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K, eds. Proc. of the 8th Int'l Semantic Web Conf. LNCS 5823, Heidelberg: Springer-Verlag, 2009. 585−600. [doi: 10.1007/978-3-642-04930-9_37]

[66] Stoermer H, Bouquet P. A novel approach for entity linkage. In: Proc. of the IEEE Int'l Conf. on Information Reuse and Integration. Las Vegas: IEEE, 2009. 151−156. [doi: 10.1109/IRI.2009.5211542]

[67] Tang J, Liang BY, Li JZ, Wang KH. Automatic ontology mapping in semantic Web. Chinese Journal of Computers, 2006,29(11): 1956−1976 (in Chinese with English abstract).

[68] Nagy M, Vargas-Vera M, Stolarski P. DSSim results for OAEI 2009. In: Shvaiko P, Euzenat J, Giunchiglia F, Stuckenschmidt H, Noy NF, Rosenthal A, eds. Proc. of the 4th Int'l Workshop on Ontology Matching. Chantilly, 2009. 160−169. http://disi.unitn.it/~p2p/OM-2009/oaei09_paper5.pdf

[69] Niu X, Wang HF, Wu G, Qi GL, Yu Y. Evaluating the stability and credibility of ontology matching methods. In: Antoniou G, Grobelnik M, Simperl EPB, Parsia B, Plexousakis D, Leenheer PD, Pan JZ, eds. Proc. of the 8th Extended Semantic Web Conf. LNCS 6643, Heidelberg: Springer-Verlag, 2011. 275−289.

附中文参考文献:

[14] 瞿裕忠,胡伟,郑东栋,仲新宇.关系数据库模式和本体间映射的研究综述.计算机研究与发展,2008,45(2):300−309.

[23] 王厚峰.指代消解的基本方法和实现技术.中文信息学报,2002,16(6):3−9.

[24] 赵军.命名实体识别、排歧和跨语言关联.中文信息学报,2009,23(2):4−17.

[29] 陈金星,祝忠明,刘玉婷,吴登禄.责任者唯一标识符应用研究.情报杂志,2010,29(12):134−140.

[67] 唐杰,梁邦勇,李涓子,王克宏.语义 Web 中的本体自动映射.计算机学报,2006,29(11):1956−1976.

胡伟(1982－),男,江苏南京人,博士,讲师,主要研究领域为语义 Web,本体工程,数据融合.

瞿裕忠(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 科学,语义 Web,软件方法学.

柏文阳(1967－),男,副教授,CCF 高级会员,主要研究领域为数据库,数据管理,数据挖掘.