关联数据冲突消解方法研究

何绯娟[1],刘文强[2],缪相林[1],许大炜[1]
(1. 西安交通大学城市学院,陕西 西安 710018;
2. 西安交通大学 电子与信息工程学院,陕西 西安 710049)

摘 要:关联数据源的独立自治特点导致不同数据源对真实世界相同实体可能提供冲突的描述,限制了关联数据的可用性。对关联数据进行冲突消解成为近年来语义网研究领域的热点问题。针对冲突在 RDF 三元组中出现的位置,对主语、谓词、宾语的冲突消解研究现状进行了综述。其中,对于主语冲突消解(也称实体共指消解),分析了基于语义等价推理与基于属性值相似度两类典型方法;对于谓词冲突消解(也称本体匹配),介绍了基于相似度、基于结构匹配以及基于实例三类方法;对于宾语冲突消解,介绍了冲突避免和真值发现两类方法。目前,国内外在关联数据主语与谓词的冲突消解问题已开展了大量研究,但对宾语冲突消解的研究还比较薄弱。文中总结了宾语冲突消解在多值冲突、时变数据、数据拷贝等方面的挑战,并对未来的研究工作进行了展望。

关键词:关联数据;资源描述框架;冲突消解;共指消解;本体匹配
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2018)11-0111-04
doi:10.3969/j.issn.1673-629X.2018.11.025

Research on Conflict Resolution for Linked Data

HE Fei-juan[1],LIU Wen-qiang[2],MIAO Xiang-lin[1],XU Da-wei[1]
(1. Xi'an Jiaotong University City College,Xi'an 710018,China;
2. School of Electronic and Information Engineering,Xi'an Jiaotong University,Xi'an 710049,China)

Abstract:Different linked data sources may provide conflicting descriptions for the same real-world entities due to the autonomy of the sources,which limits the availability of linked data. Therefore,to resolve the conflicts in linked data has received extensive attention in semantic web research. According to the three elements of RDF (resource description framework) triples,we survey the research work on conflict resolution of subject,predicate,and object in linked data respectively. For the subject conflicts resolution (also called resolving entity coreference),we analyze two typical methods which are based on the semantic inference and the similarity of attribute values respectively. For the predicate conflicts resolution (also called ontology alignment),we introduce the similarity-based,structural matching-based and instance-based methods. For the object conflicts resolution,we present the conflict-ignoring methods and the truth discovery methods. At present,extensive research has been conducted on the conflict resolution of subject and predicate,but the research on object conflict resolution is still weak. In this paper,we summarize the challenges of object conflict resolution,including multi-value conflicts,time-varying data,data replication,and discuss the future research work.

Key words:linked data;resource description framework;conflict resolution;coreference resolution;ontology alignment

0 引 言

关联数据是语义网的一种实现形式,采用资源描述框架(resource description framework,RDF)模型对信息实体及其关系进行描述、发布与部署,从而为互联网上智能化的应用提供支撑。关联数据的核心是 RDF 模型,它是一种由主语(subject)、谓词(predicate)、客体(object)构成的三元组形式,其中主语采用统一资源标识符(uniform resource identifier,URI)描述 Web 上的信息实体,谓词表示实体的属性,客体则是属性对应的值。例如,RDF 三元组<http://dbpedia.org/resource/China, dbo: capital, http://dbpedia.org/resource/Beijing>表示中国的首都是北京。近年来,由于关联开放数据项目(linked open data,LOD)的大力推动,关联数据源的数量及数据规模都快速增长,截至

收稿日期:2018-01-15 **修回日期**:2018-05-17 **网络出版时间**:2018-09-18
基金项目:陕西省教育科研计划项目(17JK1012)
作者简介:何绯娟(1977-),女,讲师,硕士,研究方向为知识管理、数据挖掘。

2018 年,已集成了 1 200 多个数据源,RDF 三元组规模达到了百亿级,内容涵盖地理、生命科学、出版物、社交网络等领域,并在数字图书馆、生物医学、教育等领域得到应用。

关联数据具有发布自由、独立自治等特点,导致多个独立维护的数据源对真实世界相同实体可能提供冲突的描述[1]。冲突的原因主要包括两个方面:一是在一些数据源构建中,采用了机器学习、自然语言处理等算法自动实现关联数据生成,容易引入脏数据;二是对数据更新时差的不同也会导致冲突问题。冲突问题严重影响关联数据的可用性。

根据冲突出现位置的不同,可将 RDF 数据的冲突问题分为主语冲突、谓词冲突和宾语冲突[2],分别对应 RDF 三元组的三个元素。其中,主语冲突或谓词冲突是指不同 RDF 数据集为真实世界相同实体(主语)或属性(谓语)提供不同标识符。宾语冲突是指不同数据集为同一个主语的相同谓词提供不同的值。例如,不同数据源对“Java 语言的设计者”提供不同的值。DBpedia 数据源认为 Java 语言的设计者是 Sun 公司和 James Gosling,而 Freebase 认为只是 James Gosling。

近年来,人们对关联数据的冲突消解问题开展了一系列研究,文中将从主语、谓词、宾语三个方面对冲突消解研究工作进行分析、总结与展望。

1 主语冲突消解

主语冲突消解也称为实体共指消解(resolving entity coreference),旨在消除描述关联数据中相同实体的标识符不一致问题。例如,DBpedia 数据源的标识符“Beijing”和 Freebase 的“m. 01914”都指代“北京”。目前,主语冲突消解主要有基于语义等价推理与基于属性值相似度两类方法[3]。

基于语义等价推理的方法主要利用网络本体语言(ontology web language,OWL)推理不同实体标识符间的对象共指关系。OWL 语言定义一组原语来描述关联数据中类及类间关系。典型的原语包括:

(1) owl:sameAs 原语。该原语通过 < s, owl:sameAs,p>三元组形式直接定义 s 和 p 指代相同的实体。然而实证分析表明,关联数据中仅有 51% 的 sameAs 被正确使用[4];因而,单纯基于 sameAs 有很大的局限性。

(2)反函数属性(owl:InverseFunctionalProperty)原语。如果两个实体具有相同的反函数属性,如电子邮件地址(foaf:mbox),则指代相同的实体对象。例如,Nikolov 等综合 owl:sameAs、owl:differentFrom 等原语提出了一种共指消解模型[5]。

基于语义等价推理的方法能够利用 OWL 的精确原语识别共指关系,准确率较高,但由于关联数据中大量缺失这些 OWL 原语,导致召回率较低。

基于属性值相似度计算方法通过比较实体标识符的属性和属性值来识别对象实现主语冲突消解,主要依据是相同的实体通常具有相同或者高度相似的属性值对。例如,Wang 等提出一种基于马尔可夫随机场的主语冲突消解方法[6],该方法通过计算描述不同标识符的多个属性值间的相似度来实现主语冲突消解。基于属性值相似度计算方法具有很强的适应性,但是难以确定合适的相似度阈值,且计算复杂度较高。

目前也有综合两种方法实现主语冲突消解的研究工作。例如,Hu 等提出一种自训练的共指消解识别方法[7]。该方法首先利用 OWL 语言识别出部分候选训练集,进而基于相似度计算的方法扩展该训练集,并过滤掉低可信度的训练集,直至得出高可信度的共指关系集。基于两者结合的方法有助于提高主语冲突消解的准确率和召回率,但随着关联数据量的增加,计算开销也急剧增加。

2 谓词冲突消解

谓词冲突消解又称为本体匹配(ontology alignment),主要解决关联数据源对同一实体的相同属性采用不一致标识符的问题。例如,DBpedia 数据源中表示“人口”为 populationTotal,而 Geoname 数据源中则是 population。目前,谓词冲突消解方法主要包括基于相似度、基于结构匹配以及基于实例三类方法。

基于相似度的方法利用两个谓词之间的文本相似度(如编辑距离、N - gram 距离等)或词义相似度(WordNet 距离、层次距离等)挖掘两个谓词间关联,实现谓词冲突消解。例如,Schadd 与 Roos 利用词汇在分类体系中的层次距离提出了一种词汇相似性指标,并设计了一种基于该指标的谓词冲突消解方法[8]。潘有能等提出了一种利用谓词在 WordNet 中父子概念的相似度进行本体匹配的方法[9]。总体上,基于相似度的方法简单、直接,但对谓词表达形式依赖性大;此外,相似度计算函数的选择也影响此类方法的性能。

基于结构匹配的方法利用关联数据本身或者本体的拓扑结构来实现谓词冲突消解。例如,Xiang 等利用关联数据图结构,提出了一种基于随机游走的相似度传播算法实现谓词冲突消解[10]。王颖等提出一种利用 RDF 图结构相似性进行本体匹配的方法[11]。清华大学也研制了谓词匹配系统 RiMOM[12],能够综合利用多种策略以及结构、文本相似性进行谓词冲突消解。基于结构匹配的消解方法是相似度计算方法的扩展,通过拓扑结构能够有效提升匹配效率;但该类方法对结构信息过于依赖,适应性较差。

基于实例的方法主要采用机器学习方法自动实现谓词冲突消解。例如,Wang 等在谓词相似度、值的匹配程度等特征的基础上,提出了一种基于分类的谓词冲突消解方法,并在知识图谱融合中取得了较好的效果[13]。蒋湛等从实例、结构等维度计算每项特征的置信度,并提出基于特征自适应的本体映射方法[14]。这类方法适应性强,但是需要大量的人工标注数据。

3　宾语冲突消解

宾语冲突消解用于消除不同关联数据源相同实体的同一属性的属性值不一致的过程。例如,北京的总人口数在关联数据源 Freebase 和 DBpedia 上的数值分别是“20 180 000”和“21 516 000”。目前,宾语冲突消解可分为冲突避免和真值发现(truth discovery)两类方法。

冲突避免方法采用人工预设规则避免冲突。例如,Mendes 等提出了一种关联数据质量评估框架 Sieve[15],该框架指定特定数据源的值是可信值,以此解决属性值的不一致问题。部分研究也采用少数服从多数的投票策略,即把出现次数最多的属性值作为最可信的值。这类方法的主要缺陷是认为每个数据源的权威值是相同且固定的,这与实际不符。

真值发现方法根据数据特点、拓扑结构等特征识别出特定实体特定属性最可能的数值,据此消除宾语冲突。真值发现进一步可分为基于迭代、基于最优化以及基于概率图三类方法。

基于迭代的方法主要利用数据源权威性与数据可信性相互依赖的特点,迭代推导出实体属性的真值。例如,Dong 等采用类似 Authority-Hub 迭代机制,提出了一种基于贝叶斯推断的真值发现算法[16]。马如霞等提出一种基于数据源分类可信性的真值发现算法,该算法采用基于贝叶斯的方法迭代计算数据源分类可靠性和属性值准确性[17]。

基于最优化的方法主要是通过优化损失函数,逐步缩短冲突值与真值间的距离,进而发现真值。例如,Li 等提出了一种真值发现优化框架,该框架把数据可信性与数据源权威性作为优化函数的两个变量,提高真值发现的准确性[18]。陈超等提出了一种基于距离的异构数据联合真值发现算法,该算法采用最优化策略更新数据可信度和数据源的类簇内可靠性[19]。

基于概率图的方法把数据源权威性与数据可信性看作 0 到 1 之间的概率,利用概率图模型推断实体属性的真值。例如,Zhao 等提出了一种基于概率图的真值发现算法[20],该算法把数据源的权威性定义为敏感度与特异性两个变量,并以此构建了真值发现概率图模型框架。

基于迭代的方法适应性强,但需要较多的迭代次数,时间复杂度较高。基于最优化的方法过于依赖训练集,且由于关联数据更新速度快,易引发训练集和测试集间的分布不一致问题,造成模型的欠拟合。基于概率图的方法正确率高,但计算复杂。

4　研究展望

目前,在关联数据冲突消解方面,对于主语与谓词的冲突消解问题,国内外已开展了大量研究,已具有较成熟的算法与软件系统。但是,对于宾语的冲突消解问题,目前仍然面临多值冲突、时变数据、数据拷贝等挑战性问题。后续的研究方向主要包括:

(1)关联数据中的多值冲突问题。关联数据的某些属性本身就有很多正确值,比如图书的作者。现有宾语冲突消解算法只考虑了关联数据中单值冲突问题。未来一个有前景的研究方向是借助概率图模型或者深度学习算法研究多值冲突消解问题。

(2)关联数据的拷贝问题。数据源之间存在大量的隐含拷贝关系,这类关系妨碍了对数据源可信性的判别,如何识别并利用拷贝关系是提高宾语冲突消解算法性能的关键。

(3)关联数据的表示学习模型。以深度学习为代表的表示学习技术旨在将关联数据中的实体和属性表示成低维稠密的向量。但是由于不同的关联数据源采用不同的标识符来表示相同的实体和属性,这给关联数据的表示学习带来了挑战。

5　结束语

随着关联数据源的数量及数据规模的快速增长,关联数据的冲突问题愈发严重,成为制约其可用性的关键因素。文中根据关联数据的 RDF 三元组结构,从主语、谓词、宾语三个方面对近年来在关联数据冲突消解方面的研究工作进行了归类分析,系统总结了各类冲突消解方法的优缺点。最后,重点总结了宾语冲突消解存在的问题,给出了关联数据的多值冲突、拷贝、表示学习三个具有挑战性的研究方向。

参考文献:

[1] LIAN Xiang, CHEN Lei, WANG Guoren. Quality-aware subgraph matching over inconsistent probabilistic graph databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(6): 1560-1574.

[2] LIU Wenqiang, LIU Jun, DUAN Haimeng, et al. Exploiting source-object networks to resolve object conflicts in linked data[C]//European semantic web conference. [s. l.]: Springer, 2017: 53-67.

[3] 胡　伟,柏文阳,瞿裕忠. 语义 Web 中对象共指的消解研

究[J]. 软件学报,2012,23(7):1729-1744.
[4] HALPIN H,HAYES P J,MCCUSKER J P,et al. When owl:sameas isn't the same:an analysis of identity in linked data [C]//Proceedings of the 9th international semantic web conference. [s. l.]:Springer,2010:305-320.
[5] NIKOLOV A,UREN V,MOTTA E,et al. Refining instance coreferencing results using belief propagation[C]//Asian semantic web conference. Berlin:Springer,2008:405-419.
[6] WANG Shenghui, ENGLEBIENNE G, SCHLOBACH S. Learning concept mappings from instance similarity[C]// Proceedings of the 7th international conference on the semantic web. Karlsruhe,Germany:Springer,2008:339-355.
[7] HU Wei,CHEN Jianfeng,QU Yuzhong. A self-training approach for resolving object coreference on the semantic web [C]//Proceedings of the 20th international conference on world wide web. Hyderabad,India:ACM,2011:87-96.
[8] SCHADD F C,ROOS N. Word-sense disambiguation for ontology mapping:concept disambiguation using virtual documents and information retrieval techniques[J]. Journal on Data Semantics,2015,4(3):167-186.
[9] 潘有能,刘朝霞. 基于 WordNet 的关联数据本体映射研究[J]. 情报杂志,2013,32(2):99-102.
[10] XIANG Chuncheng,CHANG Baobao,SUI Zhifang. An ontology matching approach based on affinity-preserving random walks[C]//Proceedings of the 24th international conference on artificial intelligence. Buenos Aires, Argentina: AAAI Press,2015:1471-1478.
[11] 王 颖,刘 群,王慧强,等. 一种基于 RDF 图的本体匹配方法[J]. 计算机应用,2008,28(2):460-462.
[12] LI Juanzi,TANG Jie,LI Yi,et al. RiMOM:a dynamic multistrategy ontology alignment framework[J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(8):1218-1232.
[13] WANG H, FANG Z, ZHANG L,et al. Effective online knowledge graph fusion[C]//International semantic web conference. [s. l.]:Springer,2015:286-302.
[14] 蒋 湛,姚晓明,林兰芬. 基于特征自适应的本体映射方法[J]. 浙江大学学报:工学版,2014,48(1):76-84.
[15] NOLLE A,MEILICKE C,CHEKOL M W,et al. Schema-based debugging of federated data sources[C]//European conference on artificial intelligence. [s. l.]:Springer,2016:381-389.
[16] DONG X L, BERTI-EQUILLE L, SRIVASTAVA D. Integrating conflicting data:the role of source dependence[J]. Proceedings of the VLDB Endowment,2009,2(1):550-561.
[17] 马如霞,孟小峰. 基于数据源分类可信性的真值发现方法研究[J]. 计算机研究与发展,2015,52(9):1931-1940.
[18] LI Qi,LI Yaliang, GAO Jing, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]//Proceedings of the 2014 ACM SIGMOD international conference on management of data. Snowbird, Utah,USA:ACM,2014:1187-1198.
[19] 陈 超,申德荣,寇 月,等. 异构数据联合式的真值发现算法[J]. 东北大学学报:自然科学版,2017,38(10):1373-1376.
[20] ZHAO Bo, RUBINSTEIN B I P, GEMMELL J, et al. A Bayesian approach to discovering truth from conflicting sources for data integration[J]. Proceedings of the VLDB Endowment,2012,5(6): 550-561.

(上接第 110 页)

tual polarity in phrase-level sentiment analysis[C]//Proceedings of the conference on human language technology and empirical methods in natural language processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics,2005:347-354.
[6] 戴 敏,王荣洋,李寿山,等. 基于句法特征的评价对象抽取方法研究[J]. 中文信息学报,2014,28(4):92-97.
[7] 陈耀东,王 挺,陈火旺. 浅层语义分析研究[J]. 计算机研究与发展,2008,45:321-325.
[8] KIM S M,HOVY E. Extracting opinions, opinion holders, and topics expressed in online news media text[C]//Proceedings of the workshop on sentiment and subjectivity in text. Sydney, Australia: Association for Computational Linguistics,2006:1-8.
[9] 徐 冰,赵铁军,王山雨,等. 基于浅层句法特征的评价对象抽取研究[J]. 自动化学报,2011,37(10):1241-1247.
[10] TURNEY P D. Thumbs up or thumbs down?:semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics,2002:417-424.
[11] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of the empirical methods in natural language processing. [s. l.]: Association for Computational Linguistics,2004:79-86.
[12] MULLEN T,COLLIER N. Sentiment analysis using support vector machines with diverse information sources[C]//Proceedings of the empirical methods in natural language processing. [s. l.]:[s. n.],2004:412-418.
[13] 韦 航,王永恒. 基于主题的中文微博情感分析[J]. 计算机工程,2015,41(9):238-244.
[14] 周国光. 试论语义指向分析的原则和方法[J]. 语言科学,2006,5(4):41-49.
[15] 陆俭明. 关于语义指向分析[M]//中国语言学论丛. 北京:北京语言文化大学,1999:34-47.