毕业设计（论文）报告

题目　基于多知识库的表格实体链接系统

计算机科学与工程 院（系）计算机科学与技术 专　业

学　　号　09013119

学生姓名　严晟嘉

指导教师　李慧颖

顾问老师　漆桂林

起讫日期　2017.02.20 —2017.05.22

设计地点　东南大学九龙湖校区

摘　要

在当今的万维网上包含着超大规模的蕴含丰富价值的关系型数据，而这些关系型数据大多以 HTML 表格 (也就是 Web 表格) 的形式呈现。抽取 Web 表格中的语义来制造机器可以理解的知识如今已经成为了一个热门的研究领域。从 Web 表格中抽取出丰富且高质量的语义信息是一个充满挑战和价值的任务，其中关键一步就是实体链接 (Entity Linking)，其目的在于建立表格中的字符串指称 (Mention) 到知识库中的相应实体 (Entity) 的链接。

目前大部分的实体链接都是使用单个特定的跨领域知识库来作为实体数据的来源。但由于单个知识库中的实体数量有限且对实体的覆盖范围不广，从而导致单知识库实体链接中实体缺失和链接错误的问题常有发生。在本文中，我会介绍我的本科毕业设计“基于多知识库的表格实体链接系统”的设计思路和具体实现，这个系统中包含了 2 种表格实体链接的新方法。与先前的工作不同的是，我的方法使用多知识库而不是单知识库作为实体数据的来源，来提升表格实体链接的质量。

方法一是一个“两步走”的方法，我首先将一个基于通用的概率图模型的随机游走算法应用于单知识库的表格实体链接，然后使用多知识库间已有的和新学习到的“sameAs”关系来提升前一步中实体链接结果的质量。方法二融合了方法一中的两步，我修改了方法一中概率图模型中的节点定义，直接将多知识库间已有的和新学习到的“sameAs”关系融合进了概率图模型，从而实现了一步到位的效果。

本文中的所有实验都是在基于 Zhishi.me 人工标注的 Web 表格上进行的，Zhishi.me 涵盖了三个最大的中文百科知识库：百度百科，互动百科和中文维基百科。实验结果表明我的方法在一些评估指标上胜过当今最好的表格实体链接方法。除此之外，系统在实体消岐时引入了更多的有效的特征，还增加了一些有意义的功能，比如实体链接和“sameAs”关系的迭代学习，在表格实体链接完成之后进行表格的关系抽取。

关键词： 实体链接，Web 表格，多知识库

Abstract

The World-Wide Web contains a huge scale of valuable relational data, which are embedded in HTML tables (i.e. Web tables). It's a challenging and valuable task that extracting abundant and high-quality semantics of Web tables is. A key step in extracting the semantics of Web tables is Entity Linking (EL), which aims to map the string mentions in table cells to their referent entities in a Knowledge Base (KB).

At present most entity linking researches use single cross-domain KBs as source of entities. However, the quantity of entities and the coverage in a single KB is limited. Thus, the issues of entity absence and linking error often appear in the process of EL in a single KB. In this paper, I will introduce the design ideas and concrete implementations of my undergraduate graduate project "Entity Linking System in Web Tables with Multiple Linked Knowledge Bases", which includes two new approaches for EL in Web tables. Different from previous work, the proposed approaches replace a single KB with multiple linked KBs as the sources of entities to improve the quality of EL.

Approach One is a two-step method. I first apply a random walk algorithm based on a general probabilistic graphical model to EL in Web tables with each single KB. Then, I leverage the existing and newly learned "sameAs" relations between the entities from different KBs to help improve the results of EL in the first step. Approach Two merges two steps from Approach One. I modify node definitions in the probabilistic graphical model in Approach One and directly add the "sameAs" relations between the entities from different KBs into the model.

All experiments are conducted on the sampled Web tables with Zhishi.me, which consists of three biggest linked Chinese encyclopedic KBs: Baidubaike, Hudongbaike and Zhwiki. The experimental results show that my approaches outperform the state-of-the-art table's EL methods in different evaluation metrics. Besides, the system adds some meaningful functions, such as iterative learning between EL and "sameAs" relations and table's relation extraction after EL in Web tables is accomplished.

KEY WORDS: Entity Linking, Web Tables, Multiple Linked Knowledge Bases

目 录

第一章　绪论

1.1 研究动机

如今 Web 上的内容每天都在以指数级的速度增长[1]，这也使得 Web 近年来成为世界上最大的数据集散地之一。据估计 Web 上有超过 141 亿张表格，其中 1.54 亿张表格包含关系型数据并且仅 Wikipedia 就是大约 160 万关系型表格的来源。可见 Web 表格，换句话说是 Web 上的 HTML 表格，是关系型数据的一个重要来源和信息抽取 (Information Extraction) 系统的一个重要输入。与普通文本不同，单张关系型表格就包含一系列高质量的关系实体并且在表格的列头中包含与实体相关的元数据。Web 上关系型表格中蕴含的巨大财富和价值使得表格的语义解释 (Semantic Interpretation)，也就是将 Web 表格转换成机器能够理解的知识这一任务成为热门的研究领域。

另一方面，诸如维基百科等知识共享社区的蓬勃发展和信息抽取技术的进步已经促成了大规模机器可读知识库的自动化建设。目前世界上已经出现了上百个领域不同、规模不一的知识库并且它们的规模每天都在飞速增长。知识库中包含着整个世界的实体，实体的语义类别及其相互关系的丰富信息。这样典型的例子包括 YAGO[2]，DBpedia[3] 和 Freebase[4]。在中文知识库中较有影响力的就是由上海交通大学和东南大学共同建立的 zhishi.me[5]。

Web 上虽然蕴含着许多有价值的数据，但更多的却是各种原始并且充斥着噪声的数据，其中有些甚至是错误的。这些数据大都是以自然语言的形式存在，然而由于自然语言表达的多样性与歧义性，使得它们很难被计算机直接处理或者理解。面对如此规模庞大而嘈杂的数据，信息过载的现象每天都在发生。信息过载意味着在找到想要的有用的信息之前，需要处理大量的无用数据，计算机获取有效信息的效率常常受到限制。为了减轻信息过载带来的负面影响以及处理自然语言的多样性问题和歧义性问题，语义网 (Semantic Web) 的概念应运而生，旨在对现有万维网上的文档进行元数据 (Meta Data) 标注，使计算机能够理解词语和概念以及它们之间的逻辑关系。将 Web 数据与知识库链接起来是非常有利于标注 Web 上的大规模数据的，并且有助于实现语义网的愿景[6]。许多为了解释说明 Web 表格内含的语义的研究工作[7][1][8][9][10][11] 是将其内容标注成 RDF 三元组。这种标注的关键一步就是实体链接 (Entity Linking)，将出现在 Web 表格单元格中的命名实体指称链接到其在给定知识库中对应的实体。

实体链接技术的发展可以带动许多不同的应用的发展，比如知识库补全，自然语言问答系统和语义搜索系统。随着社会的发展，新的知识被创造出来并以数据的形式表现在 Web 上。因此，利用这些新知识扩充现有的知识库显得越发重要。然而，为了将这些

新知识插入到现有知识库中，会不可避免地需要一个系统，来将命名实体指称，也就是与已抽取出的知识相关联的指称，链接到知识库中的相应实体。例如，自然语言问答系统依靠它们支持的知识库来回答用户的问题。为了回答“苹果公司创始人史蒂夫·乔布斯的诞生日期”的问题，该系统应首先利用实体链接技术将查询语句中的“史蒂夫·乔布斯”映射到美国企业家，而不是美国传记电影，然后从知识库中直接取回名为“史蒂夫·乔布斯”的出生日期。除此之外，实体链接对数据集成很有帮助，可以将不同页面、文档和站点上的实体信息进行集成。可见，Web 表格上的实体链接技术很有价值并且拥有广阔的应用前景。

1.2 研究现状

在本节中，我会回顾了有关 Web 表格上语义注释的一些相关工作，通常会解决三个任务：实体链接（EL），列类型推断以及同一行中的实体之间的关系提取。Cafarella 等人 [6] 报道说，有超过 1.5 亿个 Web 表嵌入了高质量的关系数据，许多研究人员意识到 Web 表是可用于许多应用程序的重要来源，如信息提取和结构化数据搜索。因此，出现了关于 Web 表的语义注释的各种工作。

ignette 等 [9] 提出了一种聚合方法，用于在给定本体中使用词汇表来注释 Web 表单的内容。它首先注释单元格，然后注释列，最后关联这些列。同样，Syed 等 [19] 还提出了一种管道方法，其首先推断列的类型，然后将单元格值链接到给定的 KB 中的实体，最后选择列之间的适当关系。Zhang [22] 设计了一个名为 TableMiner 的工具来注释 Web 表。TableMiner 仅关注列类型推理和 EL，并且不能从 Web 表中提取关系。之后，Zhang [21] 也提出了一些改进 TableMiner 的策略。Limaye 等人 [10] 和 Mulwad 等人 [11] 描述了可以分别联合模拟 Web 表的 EL，列类型推断和关系提取任务的两种方法。我们的方法和这些工作之间的主要区别是我们不使用任何特定的信息来完成 EL 的任务，例如 Web 表的列标题和标题，KB 中的实体类型，网页中的语义标记等。

在没有 EL 步骤的情况下，在 Web 表格上的语义注释的具体方案中也存在一些工作。在 Venetis 等人的工作中 [20]，他们的方法削弱了 EL 的影响，直接推断了列的类型，并且通过大规模的关系数据库和关系数据库来确定关系，它们都是由网页构建的，但通常不可用大部分研究人员。此外，Mun oz et al。[12] 提出了一种从维基百科表中挖掘 RDF 三元组的方法。在这项工作中，他们可以通过内部链接和文章标题直接识别维基百科中的实体。

Shen 等人最接近我们的做法 [16] 和 Bhagavatula 等人 [2]。Shen et al。[16] 尝试将列表类 Web 表（多行与一列）中的字符串提交到给定 KB 中的实体。Bhagavatula 等 [2] 提出了 TabEL，一个表实体链接系统，它使用集体分类技术来集体消除给定 Web 表中的所有提及。这两个工作都不使用 EL 的任何特定信息，并且可以应用于任何 KB。在这里，为了提高 Web 表格中的 EL 的质量，我们专注于具有多个链接 KB 而不是单个 KB 的 EL。

1.3 本文贡献

第二章　基于多知识库的表格实体链接

2.1 任务描述

2.2 关键挑战

2.3 链接流程

2.4 本章小结

第三章　系统描述

3.1 概率图模型

3.2 随机游走

3.3 方法一: 两步走

3.3.1 单知识库表格实体链接算法

3.3.2 多知识库表格实体链接算法

3.4 方法二: 融合

3.5 本章小结

第四章　系统实现

4.1 表格语料库

4.2 知识库实体

4.3 人工标注指称来源

第五章　实验与评估

第六章　总结与展望

致　谢

感谢……

参考文献

[1] Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships[J]. PVLDB, 2010, 3(1-2):1338–1347.

[2] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]. In: WWW. 2007. 697–706.

[3] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[C]. In: ISWC. 2007. 722–735.

[4] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. In: SIGMOD. 2008. 1247–1250.

[5] Niu X, Sun X, Wang H, et al. Zhishi. me-weaving chinese linking open data[C]. In: ISWC. 2011. 205–220.

[6] T. Berners-Lee O L, J. Hendler. The semantic web[J]. Scientific American, 2001.

[7] Hignette G, Buche P, Dibie-Barthélemy J, et al. Fuzzy annotation of web data tables driven by a domain ontology[C]. In: ESWC. 2009. 638–653.

[8] Mulwad V, Finin T, Joshi A. Semantic message passing for generating linked data from tables[C]. In: ISWC. 2013. 363–378.

[9] Muñoz E, Hogan A, Mileo A. Using linked data to mine RDF from wikipedia's tables[C]. In: WSDM. 2014. 533–542.

[10] Syed Z, Finin T, Mulwad V, et al. Exploiting a web of semantic data for interpreting tables[C]. In: WebSci. 2010. 5.

[11] Venetis P, Halevy A, Madhavan J, et al. Recovering semantics of tables on the web[J]. PVLDB, 2011, 4(9):528–538.

附录 A　第一个附录

……

作者简介 (包括论文和成果清单)

作者简介