学校代码: 10286
分 类 号: TP31
密 级: 公开
U D C: 004.4
学 号: 159110

东南大学
SOUTHEAST UNIVERSITY

博士学位论文

面向社交站点的双语知识图谱构建方法的研究

研究生姓名：吴天星
导 师 姓 名：漆桂林 教授

申请学位类别 工学博士 学位授予单位 东南大学
一级学科名称 软件工程 论文答辩日期 2018 年 8 月 25 日
二级学科名称 无 学位授予日期
答辩委员会主席 曲维光 教授 评 阅 人 B1800811, B1800812
B1800813

2018 年 9 月 2 日

东南大学

博士学位论文

面向社交站点的双语知识图谱构建方法的研究

专 业 名 称：软 件 工 程

研究生姓名：吴天星

导 师 姓 名：漆桂林　教授

RESEARCH ON APPROACHES TO BILINGUAL KNOWLEDGE GRAPH CONSTRUCTION FROM SOCIAL WEB SITES

A Dissertation submitted to

Southeast University

For the Academic Degree of Doctor of Engineering

BY

Wu Tianxing

Supervised by:

Prof. Qi Guilin

School of Computer Science and Engineering

Southeast University

September 2, 2018

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____日期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆、《中国学术期刊（光盘版）》电子杂志社有限公司、万方数据电子出版社、北京万方数据股份有限公司有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名：_____导师签名：_____日期：_____

摘　要

随着语义网的不断发展，由数百亿 RDF 三元组构成的相互链接的不同类型的数据集在万维网中发布，这些数据集又称为知识图谱，是辅助语义搜索、问答系统、情报分析等众多智能应用的重要基础资源。因此，构建知识图谱已成为学术界与工业界共同关注的研究课题。

目前已存在较多关于知识图谱构建方法的研究工作，但是这些工作并未全面关注一种非常重要的知识挖掘来源，即万维网中不同类型的社交站点，包括电子商务、百科、问答、博客、游戏、旅行等站点。与此同时，随着信息全球化的发展，跨语言知识对齐已成为支撑众多跨语言应用（如跨语言信息检索、跨语言语义标注等）的关键技术。然而，由于英文是世界上使用国家数最多的语言，所以在现有的多语言知识图谱中，英文知识（包括概念、实例、三元组）的数量始终占绝对主导地位，而其他语言的知识相对较少是跨语言知识对齐的主要障碍之一。因此，如何针对任意给定的两种语言有效地构建双语知识图谱，即构建每种语言对应的知识图谱并进行跨语言知识对齐是亟需探索的研究方向，而现有的相关工作也仅在百科站点中研究如何构建双语知识图谱。

基于上述讨论，本文选择研究面向社交站点的双语知识图谱构建的方法。由于社交站点中存在大量的由分类构成的层次分类体系及标签构成的分众分类系统，且这些分类与标签均表示概念，所以采用自顶向下地从模式层到实例层的双语知识图谱的构建方式，即首先尝试挖掘社交站点中概念之间的关系，该任务在本文中又称为模式知识挖掘，而已有的方法依赖于特定语言的特征与规则，不具有语言通用性。考虑到跨语言知识对齐是双语知识图谱构建的重点工作之一，本文将跨语言概念匹配作为第二项任务，但现有的方法严重依赖于翻译后的字符串相似度与特定的领域信息，导致其不具有领域通用性且匹配效果往往不佳。本文的第三项任务旨在利用实例类别推断技术，为双语知识图谱引入实例知识，而现有工作同样依赖于特定语言的规则，所以也不具备语言通用性。因此，为了克服上述三项任务的问题，本文主要进行如下研究：

1) 在模式知识挖掘方面，提出一种新的结合机器学习与规则的方法，其中将规则嵌入到机器学习的过程中。该方法不涉及任何特定语言的特征与规则，从而达成各语言通用的目标。在实验中，将该方法分别应用于中英文社交站点中的模式知识挖掘，其在测试数据集上的查准率、查全率、F1 值均优于其他基准对比方法，并且能够生成大规模、高质量的中英文模式知识。

2) 在跨语言概念匹配方面，提出一种新的基于双语主题模型的方法，其中包含两种新的双语主题模型，利用任意一个模型均可学习得到不同语言概念的向量表示，最终通过向量相似度决定不同语言概念之间的相似程度。该方法不涉及任何特定的领域信息，从而达成各领域通用的目标。实验结果表明，此方法在两种中英文层次分类体系上的查准率@1 与 MRR 均优于其他基准对比方法。

3) 在实例类别推断方面，提出一种新的基于随机游走模型的方法，在抽取得到的实例、属性、概念组成的图上进行随机游走以计算某个概念是给定实例的类别的概率。该方法不涉及任何特定语言的规则，从而达成各语言通用的目标。在实验中，将该方法分别应用于中英文维基百科中

的实例类别推断，不仅其在测试数据集上的查准率、查全率、F1 值均优于现有工作，而且能够生成大规模、高质量的中英文实例类别知识。

关键词： 双语知识图谱，社交站点，模式知识挖掘，跨语言概念匹配，实例类别推断，语义网

Abstract

With continuous development of Semantic Web, different kinds of interlinked datasets consisting of tens of billions RDF triples have been published in World Wide Web (WWW). These datasets are also called knowledge graphs, which are fundamental resources to support semantic search, question answering, information analysis and other intelligent applications. Thus, constructing knowledge graphs has been an important research topic in both academia and industry.

There already exist many works on the methods of knowledge graph construction, but they do not pay full attention to a very important source in WWW, i.e. social web sites, including the sites of electronic commerce, encyclopedia, question answering, blog, game, travelling and etc. Meanwhile, with the development of information globalization, cross-lingual knowledge alignment has become the key technology to many cross-lingual applications (e.g. cross-lingual information retrieval and cross-lingual semantic annotation). However, since English is the most spoken language in the countries all over the world, the number of English knowledge (including concepts, instances and triples) always plays a dominant role in existing multilingual knowledge graphs. The relatively small number of other language's knowledge is one of the main obstacles for cross-lingual knowledge alignment. Thus, how to effectively construct a bilingual knowledge graph when given any two languages, i.e. construct monolingual a knowledge graph for each language and perform cross-lingual knowledge alignment, is a research direction which urgently needs to explore. Existing related works only study how to construct a bilingual knowledge graph from online encyclopedias.

Based on the above discussion, this dissertation chooses to study the approaches on bilingual knowledge graph construction. Since social web sites contain a large number of categories denoting concepts in taxonomies and tags also representing concepts in folksonomies, this dissertation chooses to apply a top-down way to constructing a bilingual knowledge graph from the schema level to the instance level. It first tries to mine relations among the concepts in social Web sites. This task here is called schema knowledge mining, and its existing methods relies on language-specific features and rules so that they are not general to any language. Considering cross-lingual knowledge alignment is one of the key tasks of bilingual knowledge graph construction, cross-lingual concept matching is chosen as the second task in this dissertation. However, the existing methods strongly depends on string similarities after translation and domain-specific information, causing that it is not general to any domain and often has unsatisfied matching performance. The third task aims at utilizing the technique of instance type inference to introduce instance knowledge to the constructed bilingual knowledge graph, but the existing works also relies on language-specific rules so that they are also not general to any language. Hence, to overcome the problems in the above three tasks, this dissertation provides the following solutions:

1) With respect to schema knowledge mining, a new method combining machine learning with rules is proposed. Rules are embedded into the machine learning process. This method does not contain any

language-specific feature and rule, so that it is general to any language. In experiments, this method is applied to schema knowledge mining in English social web sites and the Chinese ones, and the precision, recall and F1-score of the proposed method are all better than those of other baselines on the benchmark. Besides, it can help generate large-scale and high-quality English and Chinese schema knowledge.

2) With respect to cross-lingual concept matching, a novel method based on bilingual topic models is presented. This method contains two new bilingual topic models, each of which can be used to learn vector representations of the concepts in different languages. The similar degree between the two given concepts in different languages is decided by their corresponding vector similarity. This method does not leverage any domain-specific information, so that it can be applied to any domain. Experimental results show that the proposed method outperforms other baselines in both precision@1 and MRR on two different kinds of English-Chinese taxonomies.

3) With respect to instance type inference, a new method based on random walk is proposed. It performs random walks on the constructed graph consisting of extracted instances, attributes and concepts, and computes the probability of some concept being the given instance's type. This method does not contain any language-specific rule, so that it is general to any language. In experiments, this method is applied to instance type inference in English Wikipedia and the Chinese one, and the precision, recall and F1-score of the proposed method are all better than those of other baselines on the benchmark. Besides, it can help generate large-scale and high-quality English and Chinese type information.

Keywords: Bilingual Knowledge Graph, Social Web Sites, Schema Knowledge Mining, Cross-Lingual Concept Matching, Instance Type Inference, Semantic Web

目 录

插图目录

表格目录

缩略词	英文全称	中文全称
WWW	World Wide Web	万维网
W3C	World Wide Web Consortium	国际万维网联盟
RDF	Resource Description Framework	资源描述框架
BiBTM	Bilingual Biterm Topic Model	双语双词主题模型
CC-BiBTM	Concept Corrleation based Bilingual Biterm Topic Model	基于概念关联关系的双语双词主题模型
ESA	Explicit Semantic Analysis	显式语义分析
BiLDA	Bilingual Latent Dirichlet Allocation	双语隐含狄利克雷分布
BTM	Biterm Topic Model	双词主题模型
IT-ER	Infobox Template based Extraction Rule	基于信息框模板的抽取规则
TDH-ER	Top-Down Hierarchy-based Extraction Rule	自顶向下的基于层次结构的抽取规则
BUH-ER	Bottom-Up Hierarchy-based Extraction Rule	自底向上的基于层次结构的抽取规则
ARW	Attribute-Driven Random Walk	属性驱动的随机游走

缩略词	英文全称	中文全称
HR	Heuristic Rules	启发式规则
WE	Word Embedding	词嵌入
AS	Attribute Similarity	属性相似度
S-ARW	Simplified Version of Attribute-Driven Random Walk	属性驱动的随机游走的简单版

第一章　绪论

1.1 研究背景

随着万维网（World Wide Web, WWW）的不断发展，大规模爆发式增长的信息资源以 Web 网页的形式在全球范围内进行发布与共享，从而改变了数十亿个人用户获取信息与交流的方式。基于 Web 网页的内容虽便于人工阅读，但计算机却难以理解并整合海量的网页内容，这便导致用户难以通过自动化、智能化的方式准确地获取所需的信息。针对这个问题，图灵奖获得者 Tim Berners-Lee 于 2001 年正式提出语义网 [1]（Semantic Web）的概念，其实质是从万维网网页内容中构建出以数据为中心的网络，即通过国际万维网联盟（World Wide Web Consortium, W3C）制定的一系列关于数据格式与语言的标准，为 Web 网页中的内容提供语义。作为万维网的重要扩展，语义网不仅可以使计算机能够自动地处理、整合网页内容，而且能够支持不同应用中数据间的语义互操作性。

链接数据（Linked Data）作为一种实现语义网的重要技术 [2] 在近十年来得到了极大的关注，其目的是将结构化数据在万维网中发布并互连，它并不是单纯地通过超链接将网页内容互相关联，而是使用符合 W3C 标准的资源描述框架（Resource Description Framework, RDF）来将世界中任意的事物进行链接。这种构想的结果构成了对于万维网中的事物有着更加精确描述的数据万维网。通过发布链接数据，大量个人与组织均可以为构建数据万维网做出自己的贡献，这无形中降低了众多不同分布的异构数据源中的数据重用、整合与应用的门槛。开放链接数据项目 [1]（Linking Open Data Project）是目前最大的依据链接数据的思想对结构化数据进行发布和互连的社区工作。如图 1.1所示，截止 2017 年 8 月，开放链接数据中共有 1163 个数据集[2]提供约 400 亿不同的 RDF 三元组[3], 这些数据集覆盖各类不同的领域，如：地理、政府、出版物、生命科学等。

开放链接数据中的数据集又可称为“知识图谱”，此概念诞生自 Google 公司于 2012 年提出的“Knowledge Graph”项目。知识图谱 [4] 旨在以图结构描述客观世界中的概念、实例及其之间的关系，概念与实例在图结构中为节点，关系为边。概念指人类对于客观世界具体事物的抽象性表示，如“动物”、“组织”、“技术”等均为概念。实例即表示客观世界的具体事物，如演员“成龙”、互联网企业“Google 公司”等。关系则描述了概念与实例中客观存在的关联关系，如“演员”与“中国演员”之间存在概念间的上下位关系，而“导师”描述了某位高校老师与其指导的学生之间的关系。近年来，

[1] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[2] http://lod-cloud.net/

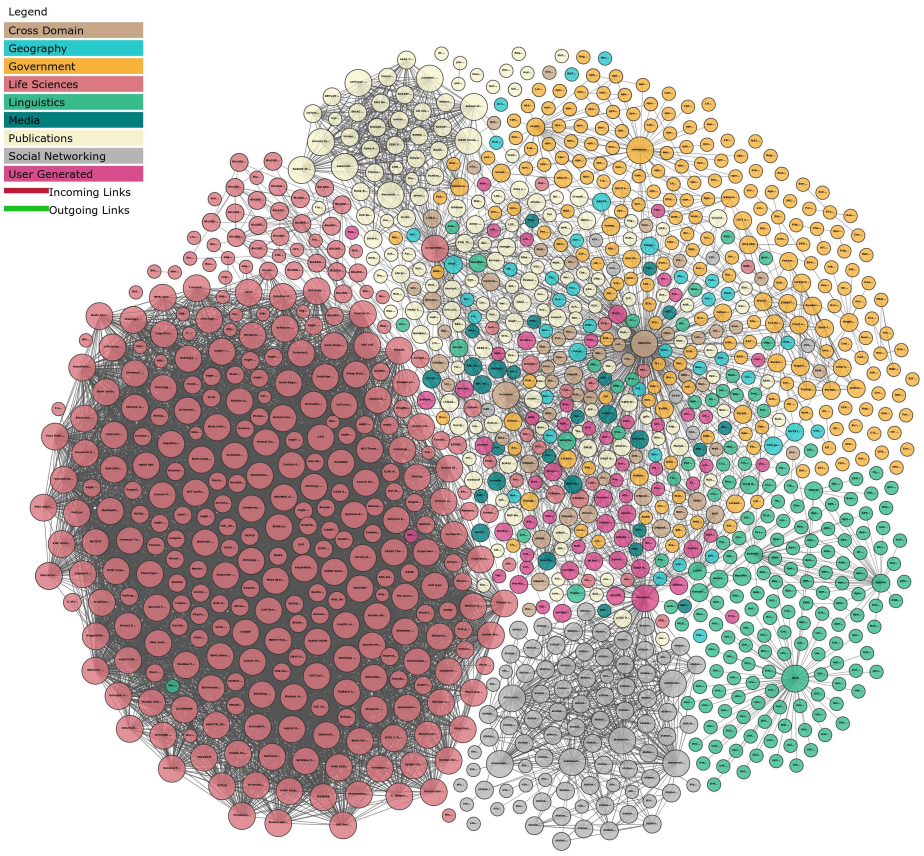


图 1.1 开放链接数据云图

知识图谱已经成为推动人工智能的发展重要驱动力之一，其重要作用在众多智能应用中得以体现，比如：

- **语义搜索：** 搜索引擎主要从以下两个方面利用知识图谱改善语义搜索的质量，一是辅助识别查询中涉及到的实体、概念、关系，真正做到理解用户的查询意图[5]；二是标注现有的 Web 网页内容，即为其添加语义[6]。从搜索的结果来看，当前的搜索引擎如 Google、Bing、百度、搜狗等不再仅仅是返回网页内容，还会以知识卡片（Knowledge Card）的形式提供知识图谱中的结构化信息。

- **问答系统：** 基于知识图谱的问答通常将用户问题解析成结构化查询语句，然后在存储知识图谱的数据库中执行查询并返回结果作为用户问题的答案。目前，许多智能问答系统均引入了知识图谱[7–9]，从而为问答提供更多的背景知识，进一步改善问答体验，比如 IBM Watson 系统、苹果公司的 Siri 语音助手、微软小娜（Cortana）、Google Now 以及百度的小度机器人。

- **情报分析：** 知识图谱通过图结构表达概念、实例之间复杂的关系，这是一种将万维网中海量的信息表示成更加接近于人类认知客观世界的方式，从而提供了组织、管理、理解万维网信息的能力。这种能力天然地有助于各领域的情报分析与决策，比如在金融领域的反欺诈情报分析中，针对借款人信息构建知识图谱，若不同地域的

借款人所填写的电话号码相同，则可能存在欺诈行为，这种异常可通过知识图谱中的不一致检测技术进行识别。此外，金融投研情报分析与公安刑侦情报分析也开始使用知识图谱辅助相关决策工作 [4]。

综上所述，知识图谱一方面已经成为实现语义网愿景的关键基础，另一方面在各领域智能应用中也已凸显出其实用的价值，因此如何从当前的万维网信息资源中构建大规模高质量的知识图谱是具有重要意义与价值的研究方向。

1.2 研究现状

本节将介绍当前学术界与工业界应用广泛的典型知识图谱与其构建方法，并以此为基础给出本文工作的主要动机。

人工编辑一直是传统知识图谱构建的主要方式之一，以这一方式进行构建的知识图谱主要包括：

- **WordNet** [10]：该知识图谱由英文词汇构成，包括名词、动词、形容词、副词，具有相同含义的词构成一个集合称为 Synset，其内在结构是由 Synset 之间的上下位关系构建而成的层次分类体系（Taxonomy）。目前，WordNet 共有约 15 万单词，11 万 Synset 以及 71 万三元组，并支持数据集的完全下载。

- **Cyc** [11]：该知识图谱起源于 1984 年的同名项目 Cyc，其目的是以专家手工编辑的方式将知识与规则编码成机器可处理的方式，从而支持人工智能的相关应用进行类人方式的推理。截至 2017，Cyc 共有约 150 万条术语（包括概念与实例），2450 万条常识规则，支撑了超过 100 家公司的相关应用。曾经 Cyc 开放了其部分数据，即 OpenCyc，但于 2017 年关闭了 OpenCyc 数据的所有公开的访问方式，目前只有通过商业合作才可使用 Cyc。

- **Schema.org** [6]：Schema.org 是 Bing、Google、Yahoo! 于 2011 年联合发布的一组通用英文模式，其实质是一个通用本体。网站管理员可以在网页中以 Microdata、JSON-LD（JavaScript Object Notation for Linked Data）等格式嵌入 Schema.org 中的结构化数据，通过这种标记方式，使得搜索引擎真正理解网页的语义，从而给用户返回更高质量的搜索结果。最新版本的 Schema.org 中，概念数为 599，属性数为 861，其数据集可完全下载。

- **Freebase** [12]：Freebase 是一个基于群体编辑的大规模多语言知识图谱，所有条目以结构化的方式呈现，其中的内容由社区成员人工编辑贡献。Freebase 定义三层结构，第一层为 topic，每个实例可视为一个 topic; 第二层为 type，实质为一类 topic 的集合，每个 type 对应一套属性；第三层为 domain，所有相关的 type 对应一个 domain。Freebase 是“Google Knowledge Graph”的重要组成部分，拥有约 6800 万实例，24 亿三元组。虽然 Freebase 项目于 2016 年完全终止，但其数据集依旧可以完全下载。

- **cnSchema**[3]：cnSchema 是 Schema.org 的中文扩展版本，由清华大学、浙江大学、复旦大学、南京大学、东南大学等高校与微软、海知智能、狗尾草科技等企业联合研制并维护，为中文领域的开放知识图谱、聊天机器人、搜索引擎优化等提供可供参考和扩展的数据描述和接口定义标准，目前拥有数千种概念、属性、关系等，并支持数据集的完全下载。

- **HowNet** [13]：该知识图谱描述中文常识，揭示概念间以及概念的属性间的关系。HowNet 包含概念间的上下位关系、同义关系、反义关系、部件 - 整体关系、相关关系、时间 - 事件关系、场所 -事件关系、事件 -角色关系等。目前包含信息结构模式 271 个，句法分布式 49 个，句法结构式 58 个，11,000 个词语。用户需填写非商业使用申请或以商业合作的方式才能获取 HowNet 数据。

- **同义词词林** [14]：该知识图谱于上世纪 80 年代编纂而成，其目的是通过总结同义词语帮助翻译等工作，其中不仅包含大量的同义词还包含相当数量的相关词。近年来，哈尔滨工业大学对同义词词林进行了扩展，剔除了罕用词与非常用词，并且借鉴了 WordNet 的表示方式表示同义词集合间的上下位关系，从而构建层次分类体系，共包含 77,343 条词语，但是该扩展版本并未提供公开的下载方式。

21 世纪初，涌现出众多的基于群体编辑的在线百科站点，如多语言的维基百科[4]，中文的百度百科[5]、互动百科[6] 等，在这些站点中，大规模的知识以半结构化的方式发布出来。基于此，许多知识图谱的构建以抽取在线百科站点中的半结构化数据为基础，其中的代表性知识图谱有：

- **DBpedia** [15]：该知识图谱拥有大规模的多语言百科知识，可视为维基百科的结构化版本。DBpedia 使用固定的模式对维基百科中的实例信息进行抽取，包括摘要（abstract）、信息框（infobox）、分类（category）和页面链接（page link）等信息，并手工构建了 DBpedia 本体。DBpedia 目前拥有 127 种语言的超过两千八百万实例与约 95 亿个 RDF 三元组，并且作为开放链接数据的核心，与许多数据集均存在实例映射关系。而根据抽样评测[16]，DBpedia 中三元组的正确率达 88%。DBpedia 支持数据集的完全下载。

- **Yago** [17]：Yago 是一个整合了维基百科、WordNet、GeoNames[7] 的大规模知识图谱。Yago 首先制定一些固定的规则对英文维基百科中每个实例的 infobox 进行指定关系的抽取，然后利用维基百科的分类信息进行实例类别推断，从而获得了概念与实例之间的上下位关系。之后，将维基百科的分类与 WordNet 中的 Synset（一个 Synset 表示一个概念）进行映射，以便于利用 WordNet 严格定义的层次分类体系完成本体构建。此外，Yago 的开发人员为 RDF 三元组增加了时间与空间信息，并以相同的

[3] http://cnschema.org/
[4] https://www.wikipedia.org/
[5] https://baike.baidu.com/
[6] http://www.baike.com/
[7] http://www.geonames.org/

方法对不同语言维基百科的进行抽取。目前，Yago 拥有 10 种语言约 459 万个实例，2400 万个三元组，三元组正确率约为 95%。该知识图谱支持数据集的完全下载。

- **Wikidata** [18]：该多语言知识图谱由维基基金会发起，首先将维基百科、维基文库、维基导游等项目中的半结构化知识进行抽取、存储、关联，然后通过群体编辑的方式，对该知识图谱不断更新补充。Wikidata 支持超过 350 种语言，拥有近 2500 万个实例与 7000 万三元组，其数据集可完全下载。

- **BabelNet** [19]：BabelNet 是当前世界范围内最大的多语言同义词典，其可视为一个由概念、实体、关系构成的语义网络。BabelNet 由 WordNet 中的英文 Synset 与维基百科页面进行映射，再利用维基百科中的跨语言页面链接以及翻译系统，从而得到 BabelNet 的初始版本。然后，BabelNet 又整合了 Wikidata、GeoNames 等多种现有知识图谱。目前共拥有 271 个语言版本，约 1400 万个 Synset。BabelNet 中的错误来源主要在于维基百科与 WordNet 之间的映射，而映射目前的正确率大约为 91%。数据使用仅支持 HTTP API 调用，而数据集的完全下载需经过非商用认证。

- **Zhishi.me** [20]：Zhishi.me 是第一份中文大规模开放链接百科知识图谱。该知识图谱采用与 DBpedia 类似的方式从中文维基百科、百度百科、互动百科中抽取知识，然后通过固定的规则将不同百科间表达相同含义的实例进行链接。Zhishi.me 中实例数超过 1200 万，三元组数量超过 1.2 亿。数据集可完全下载。

- **XLORE** [21]：XLORE 是一个大规模中英文双语百科知识图谱。它通过抽取英文维基百科、中文维基百科、百度百科、互动百科中的半结构化数据，然后挖掘中英文实例间的等价关系构建而成。目前，XLore 约有 66 万概念，5 万属性，1000 万实例，所有数据可以通过在线 SPARQL 端口查询得到。

- **CN-DBpedia** [22]：CN-DBpedia 也是中文大规模百科知识图谱，其抽取来源与 Zhishi.me 相同，为中文维基百科、百度百科、互动百科。与 Zhishi.me 不同的是，CN-DBpedia 着重于将不同百科的知识进行融合，从而形成一个知识图谱，而不是互相链接的三个知识图谱。此外，CN-DBpedia 还提出了多种知识补充与知识更新的机制，从而尽可能地为知识图谱添加新的知识。CN-DBpedia 目前共有约 1000 万实例与 8800 万三元组，并提供数据集的完全下载。

近年来，随着自然语言处理技术的不断发展，还有一类知识图谱以非结构化文本的自动抽取为基础进行构建，主要包括：

- **NELL** [23]：NELL 本质是一个英文知识学习系统，它通过预先定义的本体与抽取规则，从大规模的万维网文本中抽取三元组知识，在此过程中，不断学习新的抽取规则并将其应用到抽取的过程中，不停迭代。NELL 自 2010 年开始运行至今，已经积累了约 8000 万的三元组，每个三元组均对应一个成立的置信度。NELL 支持数据集的完全下载。

- **Microsoft Concept Graph** [24]： Microsoft Concept Graph 是一个大规模的英文层次分类体系，其中主要包含的是概念间以及概念实例间的上下位关系。Microsoft Concept

Graph 的前身是 Probase，它自动化地从数十亿网页与搜索引擎查询记录中抽取上下位关系，其中每一个上下位关系均对应一个成立的概率值。目前，Microsoft Concept Graph 拥有约 540 万个概念，1250 万个实例以及 8700 万个上下位关系。关于数据集的使用，Microsoft Concept Graph 支持数据集的完全下载，但仅能用于学术研究。

- **ConceptNet** [25]： ConceptNet 是一个大规模的多语言常识知识图谱，其本质是一个以自然语言的方式描述人类常识的大型语义网络。ConceptNet 起源于一个收集包含常识知识的语料的项目 Open Mind Common Sense，然后通过文本抽取技术获得常识知识。在该知识图谱发展的近 20 年时间里，在文本抽取的基础上，还通过设计游戏获取常识知识并融合众多其他知识图谱（如 DBpedia 与 OpenCyc）中的常识知识。目前，ConceptNet 拥有 304 种不同语言的版本，超过 350 万个概念，2800 万个三元组，并支持数据集的完全下载。

以上利用不同类型的方式构建的知识图谱各有优劣。基于人工编辑的知识图谱质量较高，但需耗费大量的人力资源与时间。基于在线百科站点中的半结构化数据抽取的知识图谱的质量往往低于基于人工编辑的知识图谱，但是这种构建方式中的人工干预较少，大大节省了人力与时间。与这两种类型的知识图谱相比，基于非结构化文本抽取的知识图谱的质量最低，其主要原因在于自然语言处理技术还不够成熟，特别是在开放域的情况下，知识抽取技术还不够实用，但是此种构建方式的自动化程度最高。

总结研究现状，发现上述三种类型的知识图谱构建方式并未全面关注一种非常重要的知识挖掘来源，即社交站点（Social Web Sites）。当前的社交万维网（Social Web）中存在大量各种类型的社交站点（见图 1.2），如电子商务、百科、问答、博客、游戏、旅行等等，并不局限于社交网络（Social Network）站点，而这些站点中的层次分类体系与由标签（Tag）构成的分众分类（Folksonomy）系统（如图 1.3 所示）可视为知识图谱构建的重要资源。基于在线百科站点中的的半结构化数据抽取的知识图谱构建仅仅考虑了社交站点中的百科站点，知识来源单一，而另外两种知识图谱构建的方式则完全忽略了社交站点。此外，一些基于分众分类的知识学习的工作[26, 27] 一度被研究人员所关注，此类工作重点利用标签之间的共现关系与标签所标注的文本信息自动化学习标签间的上下位关系，但是此类方法的实用性不佳，难以推广到实际的知识图谱构建的工作中。目前，基于社交站点中的分类与标签的知识挖掘工作已有初步探索 [28], 但是挖掘方法依赖于特定语言规则，并不通用。

与此同时，随着信息全球化的不断发展，跨语言知识对齐已经成为支撑众多应用（如跨语言信息检索 [29] 与跨语言语义标注 [30]）的关键技术。然而，由于英文是世界上使用国家数最多的语言，所以在现有的多语言知识图谱（如 DBpedia、Yago、Wikidata、BabelNet 等）中，英文知识（包括概念、实例、三元组）的数量始终占绝对主导地位，而其他语言知识相对较少是跨语言知识对齐的主要障碍之一。因此，如何针对任意给定的两种语言有效地构建双语知识图谱，即构建每种语言对应的知识图谱并进行跨语言知识对齐是亟需探索的研究方向。目前，清华大学构建的双语知识图谱 XLORE 是解决中英文跨语言知识分享的问题的重要工作，但其知识挖掘的来源依旧仅为百科站点。

针对上述讨论，本文专注研究面向社交站点的双语知识图谱构建的方法。需要说明

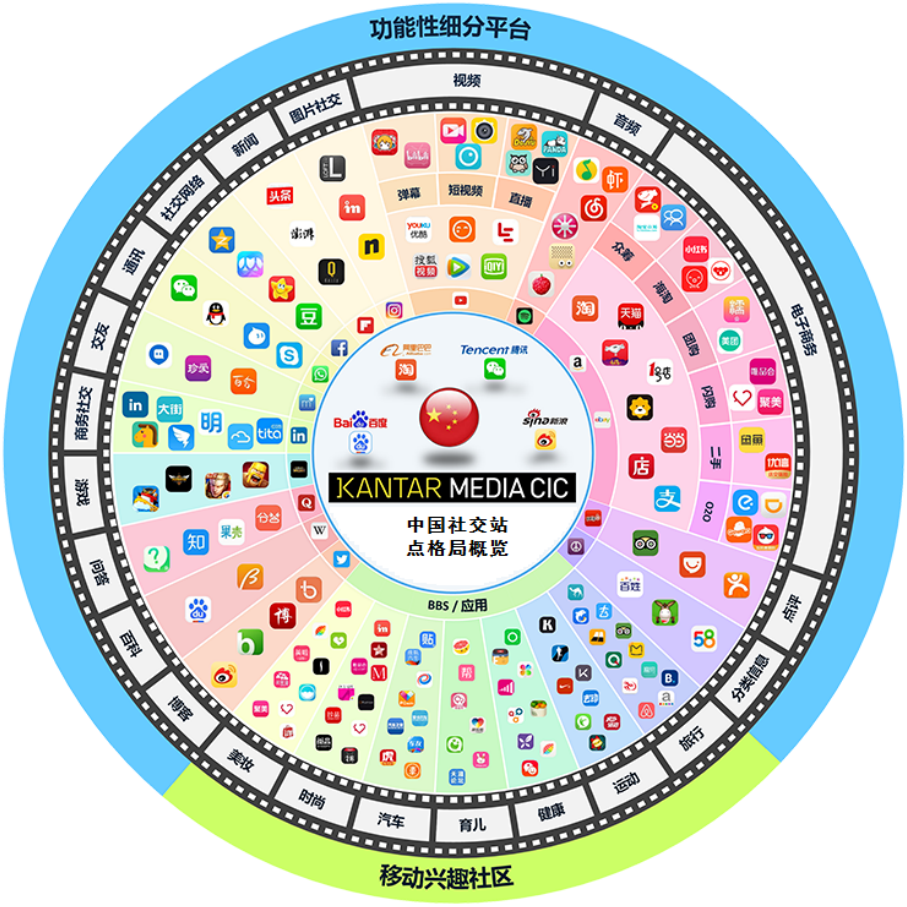


图 1.2　中国社交站点格局概览图

图 1.3　层次分类体系与分众分类系统样例

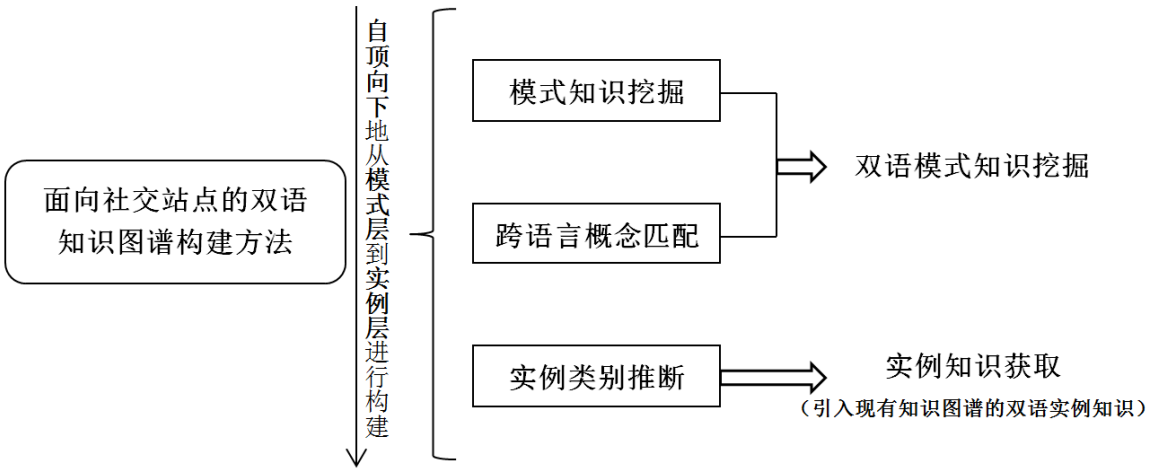


图 1.4 本文研究框架的示意图

的是，1）双语知识图谱构建指在构建两种不同语言的知识图谱之外，还需进行跨语言知识对齐；2）本文的研究实验针对中英双语进行，但方法本身并不局限于这两种语言。

1.3 本文工作

1.3.1 研究内容

由于社交站点中的层次分类体系中的分类与分众分类系统中的标签均表示概念，因此本文选择自顶向下地从模式层[8]到实例层[9]构建双语知识图谱，即首先在不同语言的环境中挖掘社交站点中概念之间的关系，称为模式知识挖掘，这是本文的第一项研究工作。此外，考虑到跨语言知识对齐是双语知识图谱构建的重点工作之一，本文将跨语言概念匹配作为第二项研究工作。以上两项工作实质是双语模式知识挖掘，而在双语知识图谱构建的过程中，实例知识是不可忽略的，由于许多多语言知识图谱（如 DBpedia 与 Yago）包含的实例均来源于维基百科，所以本文期望通过实例类别推断（Type Inference）将这些实例与社交站点中的概念进行链接，即发现这些概念与实例间的上下位关系，从而达到为本文所构建的双语知识图谱获取实例知识（源于现有的多语言知识图谱）的目的，此可作为本文的第三项研究工作。基于上述分析，图 1.4给出了本文的研究框架，而本文主要研究内容如下：

1) **研究面向社交站点的模式知识挖掘的方法。**该研究旨在挖掘社交站点中概念间的关系，概念的来源为社交站点中的层次分类体系中的分类与分众分类系统中的标签。目前此场景下的概念间关系的挖掘工作 [28] 依赖于特定语言的规则与特征，而本文提出的模式知识挖掘方法需要在不同的语言环境下进行应用，所以本文方法中用于识别两个概念间关系的特征或规则须是各语言通用的，此为该研究中的主要挑战。

[8]知识图谱模式层中的知识，即模式知识，主要包括概念间关系、概念属性、属性定义域与值域等。
[9]知识图谱实例层中的知识，即实例知识，一般指实例间关系、实例属性等。

2) **研究面向社交站点的跨语言概念匹配的方法。**跨语言概念匹配任务通常在跨语言层次分类体系对齐（Taxonomy Alignment）中完成，现有的方法 [31–33] 依赖于翻译后的字符串相似度与特定的领域信息。但由于存在大量的翻译结果不准确与词汇失配（Vocabulary Mismatch）的现象，即使不同语言的概念拥有相同的含义，翻译后的字符串往往也截然不同，易导致匹配失效。而特定的领域信息不是任意概念均可获取，所以此信息在该研究中不予考虑。综上，研究一种各领域通用的且不仅仅依赖于字符串相似度的方法是跨语言概念匹配工作的主要挑战。

3) **研究面向社交站点的实例类别推断的方法。**该研究的目的是通过类别推断技术得到存在上下位关系 (可表示为 *TypeOf* 关系) 的社交站点中的概念与维基百科中的实例，从而将相应概念与实例进行链接。由于现有的维基百科中的实例类别推断方法 [34, 35] 依赖于特定语言的规则，所以与模式知识挖掘所面临的的挑战类似，实例类别推断的主要挑战为所提出的方法须是各语言通用的。

1.3.2 主要贡献

针对上述研究内容，本文均提出了相应的解决方案，主要贡献如下：

1) **提出了一种结合机器学习与规则的模式知识挖掘的方法。**该方法首先使用一种分块（Blocking）机制生成待匹配的概念对（Concept Pairs），然后利用一种自动化策略从待匹配的概念对中生成标注数据，最后提出了一种融合了规则的半监督学习方法挖掘概念间的 *equal*、*subClassOf*、*relate*三种关系。整个方法不涉及任何特定语言的特征与规则，从而满足各语言通用的条件。在实验中，该方法分别应用于中英文社交站点中的模式知识挖掘，其在测试数据集上的查准率、查全率、F1 值均优于其他基准对比方法，并且能够生成大规模、高质量的中英文模式知识。

2) **提出了一种基于双语主题模型的跨语言概念匹配的方法。**该方法选择在社交站点中的不同语言的层次分类体系中进行跨语言概念匹配。该方法首先提出一种跨语言字符串相似度识别待匹配的双语概念对，然后利用搜索引擎与机器翻译获取每个概念对应的双语文本上下文，之后提出两种新的双语主题模型：双语双词主题模型（Bilingual Biterm Topic Model，BiBTM）与基于概念关联关系的双语双词主题模型（Concept Corrleation based Bilingual Biterm Topic Model，CC-BiBTM），这两种模型均可以生成每个概念的向量表示，最终跨语言概念匹配的结果取决于概念间的向量表示相似度。上述方法不涉及任何特定的领域信息，且综合考虑了字符串相似度与向量表示相似度。实验结果表明，本文方法在两种中英文层次分类体系上的查准率@1 与 MRR 均优于其他基准对比方法，当使用 CC-BiBTM 时，跨语言概念匹配的效果最佳。

3) **提出了一种基于随机游走模型的实例类别推断的方法。**由于多语言的维基百科中的概念是社交站点中概念的重要组成部分，且维基百科中的概念本身存在于实例所在的页面中，即可视为二者存在某种语义关联，所以该方法选择将维基百科中的概念

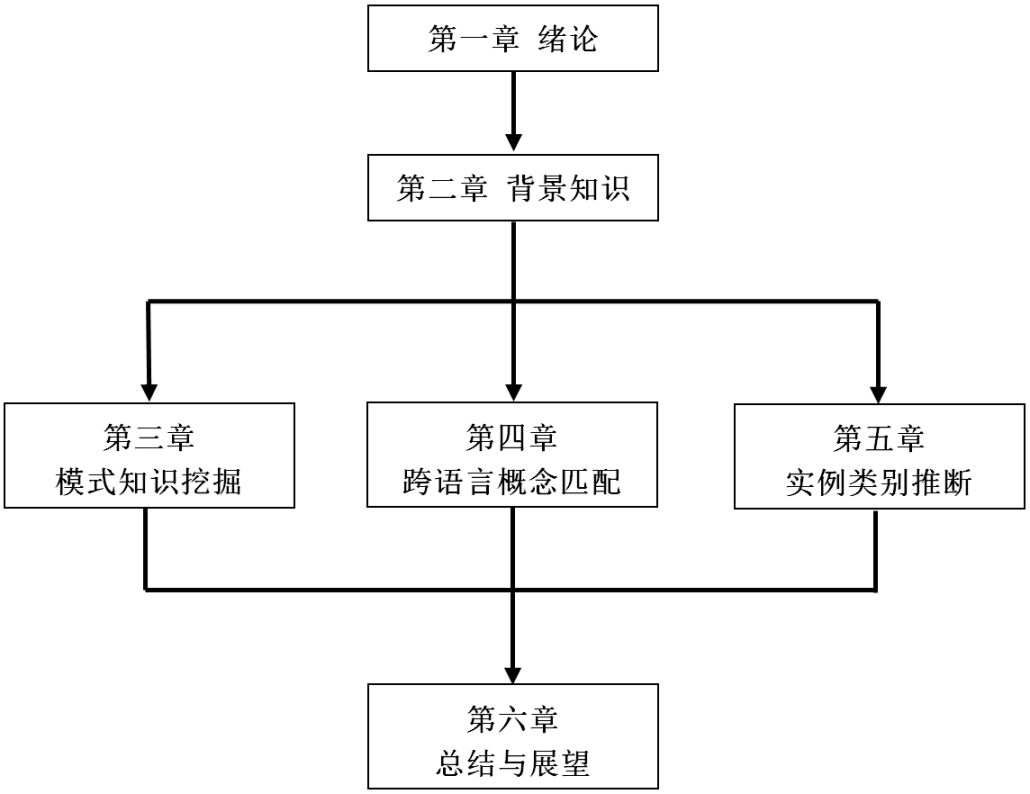


图 1.5 论文结构框图

作为候选类别进行实例类别推断。方法的基本假设为：当一个实例拥有在其页面出现的概念的具有代表性的属性时，此概念可能为给定实例的类别。对于给定的实例与概念，该方法首先抽取实例属性、概念属性、相似实例及其属性，然后构建包含上述所有抽取元素的图，最后通过图上的随机游走模型计算概念是给定实例的类别的概率。整个方法不涉及任何特定语言的规则，满足各语言通用的条件。在实验中，该方法分别应用于中英文维基百科中的实例类别推断，不仅其在测试数据集上的查准率、查全率、F1 值均优于现有工作，而且能够生成大规模、高质量的中英文实例类别知识。

1.4 论文结构

本文围绕面向社交站点的双语知识图谱构建的方法展开讨论，如图 1.5所示，具体内容组织如下：

第一章是绪论部分。该章主要介绍本文的研究背景、研究现状、研究内容及贡献。

第二章是背景知识部分。该章从知识图谱概述、术语定义与解释、论文中各方法使用的评测指标三个方面介绍了本文的基础知识，为下文各方法的设计、实验等细节的详细展开提供了良好的支撑。

第三章提出了一种结合机器学习与规则的模式知识挖掘的方法。该章重点介绍如何设计各语言通用的特征与规则，并将规则融入到机器学习模型中以挖掘概念间的 *equal*、

subClassOf、*relate*关系，以及给出所提出方法在中英双语环境中的实验结果。

第四章提出了一种基于双语主题模型的跨语言概念匹配的方法。该章重点介绍如何设计双语主题模型以学习不同语言概念的向量表示，从而辅助跨语言概念匹配，以及给出所提出方法在中英文概念匹配任务上的评测结果。

第五章提出了一种基于随机游走模型的实例类别推断的方法。该章重点介绍如何利用属性设计随机游走模型以进行不同语言实例的类别推断，以及给出所提出方法的在中英文维基百科中的实验结果。

第六章是论文的全面总结和展望。该章总结全文，并为未来的研究工作提出规划。

第二章　背景知识

本章从三个方面介绍整个论文的背景知识，分别是知识图谱概述、术语定义与解释、论文中各方法使用的评测指标。

2.1 知识图谱概述

2.1.1 发展历史

知识图谱这一概念于 2012 年由 Google 公司正式提出，其目的是汇集大规模的知识以辅助实现更加智能的搜索。知识图谱从结构上来看实质是一种语义网络 [36]（Semantic Network），所以可视为一种具有有向图结构的知识库，其中图的节点可以是概念、实例、字面量（Literal），而边可以是连接概念与概念、概念与实例、实例与实例、实例与字面量之间的各种关系。语义网络于 1956 年正式提出，是一种基于图的用于知识存储的数据结构，可方便地表达与存储自然语言句子，多用于机器翻译与自然语言理解等各项任务。1970 年开始，一些工作开始 [37, 38] 研究语义网络与谓词逻辑之间的关系，其主要目的是将二者互相表示、转化。1980 年开始，语义网络的理论趋于完善，相关研究逐步转向具有严格逻辑语义的表示和推理，重点在于提出术语逻辑 [39]（Terminological Logic）及描述逻辑 [40]，从而有效地对概念间关系进行建模并推理。2000 年以来，语义网络出现了一个新的应用场景，即语义网 [1]，它是由万维网创始人、图灵奖获得者 Tim Berners-Lee 提出的概念。语义网研究领域通过推动 RDF 标准的使用进一步促进了知识的图表示的发展，在这个阶段，知识一般被组织构成模式（Schema）或本体。之后，随着链接数据 [2] 的出现，大量的数据集在万维网中发布并互相链接，形成了一个超大规模的全球化的知识库，成为开放链接数据云（Linked Open Data Cloud）。在知识图谱（2012 年）这一概念被正式提出后，“知识图谱”一词日渐普及，目前它既可指语义网研究领域中构建的知识库（如 DBpedia、Yago 等），也可以从更加广义的层面出发，指代任意由图表示的知识集合 [41]，比如任意的 RDF 数据集或逻辑本体。

2.1.2 构建技术简介

本节从知识抽取、知识集成、知识补全三个方面简要介绍知识图谱构建技术。

2.1.2.1 知识抽取

知识抽取依据抽取源的不同可分为：面向非结构化数据的知识抽取、面向半结构化数据的知识抽取、面向结构化数据的知识抽取。

面向非结构化数据的知识抽取的研究主要是面向文本进行抽取，其首要步骤是进行实体识别。在这一步骤中，可选择两种技术，第一种是实体链接，即将候选实体与给定的知识库中的实体进行链接，从而完成文本中的实体识别任务。实体链接目前存在大量的研究工作 [42]，这些研究将实体链接的过程归纳为两个阶段：候选实体生成与候选实体消歧。在候选实体生成阶段，一般采用的是轻量级的匹配方法，如基于字符或基于字典的匹配方法。而候选实体消歧阶段通常是研究的重点，核心思想是对文本与知识库中的上下文进行有效地建模，从而能够准确地对候选实体进行排序。实体识别的第二种可选择的技术是命名实体识别，通常在无给定知识库的情况下使用。命名实体识别 [43] 是自然语言处理领域长期研究的问题，一般通过统计学方法进行实体边界识别并确定实体类别，如人名、地名、机构名等。当获得实体后，需识别句子中实体之间的关系，即关系抽取，通常是利用统计学方法 [44–47] 将该任务转换为关系分类，将各种词法、句法、词向量、位置等特征编码进机器学习模型以完成关系分类。

面向半结构化数据的知识抽取主要是利用包装器 [48]（Wrapper）归纳半结构化数据的抽取规则。包装器归纳是否有效取决于待抽取的数据是否具有大量重复性结构，一般采用的算法为序列覆盖 [49]，即通过少量的标注学习出规则，进而在整个站点下使用规则对同类型或者符合某种关系的数据进行抽取。此外，针对万维网表格（Web Table）这一特殊的半结构化数据，存在一些专门对其进行知识抽取的研究工作 [50, 51]，这些工作首先针对表格中的每个单元格中的字符串进行实体链接，与给定单元格同行同列的单元格中的信息可作为上下文辅助候选实体消歧。然后，依据实体链接结果，可将位于同一列单元格中的所有实体划分为知识库中已存在的某个类别。最后，任意给定两列单元格，若其中某同一行单元格中的两个实体在知识库中已存在一种关系，则可能这两列中任意同一行的实体之间均具有此关系。

面向结构化数据的知识抽取通常利用 D2RQ [52] 平台与 R2RML 语言将关系型数据库转换成 RDF 知识。此外，还有一些从关系型数据库抽取本体的工作 [53–55]，其实质是构建一些转换规则，从而将数据库模式转换本体知识。

在本文的研究工作中，模式知识挖掘（第 3章）与实例类别推断（第 5章）属于知识抽取范畴。由于社交站点中的概念存在一定的结构，所以本文的模式知识挖掘实质是一项面向特殊的半结构化数据的知识抽取任务。而实例类别推断中所用到的概念与实例均来源于维基百科，且二者之间存在固定的主题相关关系，所以本文的实例类别推断实质也是一项面向半结构化数据的知识抽取任务。

2.1.2.2 知识集成

知识集成分为两个步骤：1）知识匹配，包括本体匹配与实例匹配；2）知识融合。

本体匹配一直是语义网领域的研究热点，旨在发现源自不同本体的概念（或属性）间的关系，多年来大量的本体匹配工具 [56] 得以开发，如 Falcon-AO [57]、BLOOMS [58]、

PARIS [59]、LogMap [60]、AML [61] 等。与此同时，实例匹配也存在大量的研究工作 [62–65]，其目的主要在于发现源自不同知识库中的等价实例。实例的匹配结果可用于概念匹配，而概念、属性的匹配结果也可用于实例匹配，所以本体匹配与实例匹配之间可互相影响。匹配的基础是相似性度量，如基于字符的相似度、基于结构的相似度、基于外部知识的相似度等，此外还可以设计特定的匹配规则，综合所设计的特征与规则，从而最终决定是否匹配。目前的跨语言匹配研究主要是利用翻译工具与多语言的维基百科辅助进行匹配。一些研究 [31–33] 将不同语言的概念转换为同一语言再进行匹配。另一类研究 [66, 67] 则是将不同语言的实例、属性映射到相同的向量空间，再进行基于向量的相似度计算。还有一些工作 [68–70] 选择利用多语言维基百科中的关联关系，直接构造特征并利用机器学习方法完成跨语言匹配任务。

知识融合是在匹配的基础上对知识进行整合的任务，其中最为重要的问题是如何解决融合过程中产生的冲突。解决这种冲突最为简单的方法是采用多数投票机制，即通过多个不同匹配方法得出的匹配结果进行投票。另一类研究较多的知识融合方法 [71–73] 主要通过知识间的关系与外部知识构建图模型以预测不同来源知识为真的概率。

在本文的研究工作中，跨语言概念匹配（第 4章）隶属于知识集成范畴。由于本文的匹配场景是不同语言的层次分类体系中的跨语言概念匹配，所以该工作可视为一种特殊的本体匹配任务。此外，本文的跨语言概念匹配不涉及知识融合，仅针对匹配的两个概念建立链接关系。

2.1.2.3 知识补全

知识补全一般由逻辑推理或统计推理方法完成。逻辑推理在知识图谱中不仅能够运用规则推理出新的概念间、实例间、属性间的关系，还可以进行基于逻辑的冲突检测。为了使语义网络具备形式化语义并支持推理，一些研究人员针对概念描述提出描述逻辑，并以此为基础提出了许多推理系统，其中的代表性工作 Horrocks 等人提出的 FaCT 系统 [74]。但是随着数据的爆炸式增长，基于描述逻辑的推理系统难以在短时间内完成推理任务，此后的研究就着重于利用并行技术开发推理系统，其中的代表性工作有基于高性能计算平台的大规模本体推理 [75] 系统、基于 Peer-To-Peer 的分布式框架的 RDF 数据推理 [76] 系统、基于 MapReduce 开源框架的大规模本体推理系统 [76] 等等。

统计推理指运用统计学方法对知识图谱进行链接预测，即在现有的知识图谱中推断隐含的实体之间的关系。基于翻译模型的方法是近年来比较热门的研究工作，如 TranE [77]、TransH [78]、TransR [79] 等。这些方法将实体与关系映射至低维向量空间中，且认为关系向量承载了头实体向量翻译至尾实体向量的潜在特征。因此，通过比对向量空间中存在类似潜在特征的实体向量对，即可推断得到知识图谱中潜在的实体间关系。另一类方法则是基于图特征模型的方法，旨在利用现有知识图谱中三元组的关系特征预测其他隐含的关系，包括基于规则的归纳逻辑编程（Inductive Logic Programming）方法 [80] 与关联规则挖掘的方法 [81]，以及基于实体间路径特征的路径排序的方法 [82] 等。

需要说明的是，本文的研究工作并未涉及知识补全。

2.2 术语定义与解释

本小节首先对本文旨在构建的双语知识图谱给出形式化定义如下：

定义 2.1 ***双语知识图谱****: 给定语言 s 与 t 各自对应的知识图谱 KG_s 与 KG_t，其中 $KG_s = (V_s, E_s)$，$KG_t = (V_t, E_t)$，V_s 与 E_s 分别是 KG_s 中的节点集合与有向边集合，而 V_t 与 E_t 则分别是 KG_t 中的节点集合与有向边集合。基于此，定义关于 KG_s 与 KG_t 构成的双语知识图谱 $BKG = (KG_s, KG_t, R)$，其中 R 表示 V_s 中节点与 V_t 中节点之间的有向边集合。*

本文在探讨面向社交站点的双语知识图谱构建方法的过程中将使用若干专业术语，下面对其进行详细解释：

- **概念（Concept）**: 概念表示世界上任意一组具有相同特性的个体的集合，比如由“*迈克尔 · 乔丹*”、“*科比 · 布莱恩特*”、“*姚明*”等个体构成的集合即可对应概念“*篮球运动员*”。

- **分类（Category/ Class）**：分类是概念的同义词，在本文中特指存在于层级结构（即下文介绍的层次分类体系）中的概念，至少拥有一个父类（父概念）或一个子类（子概念）。

- **层次分类体系（Taxonomy）**：层次分类体系指一种由概念（即分类）构成的具有层级结构的分类系统，如商品导航目录、站点导航目录、百科分类系统等都是层次分类体系。

- **标签（Tag）**：标签是概念的同义词，在本文中特指不存在于层级结构中的概念，一般用于对博客、微博、新闻进行总结、标注。

- **分众分类系统（Folksonomy）**：分众分类系统由大量用户自发性定义的平面非层次结构型概念（即标签）组成，这种分类方式较为随意，不够严谨，但较层次分类体系而言，分众分类系统更为灵活、方便、不受限制。

- **实例（Instance）**：实例即为个体，是概念的具现化表示，比如“*姚明*”是概念“*篮球运动员*”的一个实例。

- **类别（Type）**：类别实质是概念，但概念不一定是类别，只有当一个概念与一个实例之间构成*上下位关系*（即 *TypeOf* 关系）时，此概念则可称为给定实例的类别。比如，概念“*篮球运动员*”可称为实例“*姚明*”的类别。

- **词汇中心词（Lexical Head）**: 词汇中心词指决定给定短语句法结构的词，也称为在给定短语中被修饰语所修饰、限制的中心成分，一般来说一个概念的中心词为为名词。比如“*篮球运动员*”的词汇中心词即为“*运动员*”。

2.3 评测指标

本小节针对本文所提出的双语知识图谱构建方法中使用的评测指标进行详细介绍。

查准率与查全率：研究人员通常只对运用所提出方法得到的结果中的正例感兴趣，正例指期望得到的结果，这种情况下一般使用的评测指标为查准率与查全率。查准率又称精确率，用于评测所提出方法判定为正例的正确性；查全率又称召回率，用于计算在所提出方法判定为正例的结果中实际正例的比例。此处假设所提出方法预测为正例的结果中，实际为正例的数量是 TP，实际为负例的数量为 FP；而在所提出方法预测为负例的结果中，实际为正例的数量是 FN，实际为负例的数量为 TN。在此情况下，定义查准率（Precision）与查全率（Recall）如下：

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

F1 值：虽然查准率与查全率在理论上并不相关，但是在实际应用中，高查准率往往意味着需要牺牲查全率，而高查全率则意味着低查准率，此时需使用一个评估标准 F1 值以综合考虑查准率与查全率，此处定义 F1 值（F1-score）如下：

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2.3}$$

不难看出 F1 值实际是查准率与查全率的调和平均值。

查准率@K：查准率@K 是用于排序任务（如搜索、推荐等）的评测指标。本文将跨语言概念匹配（具体见第 4章）视为一个排序任务，即在给定两种不同语言的层次分类体系的情况下，针对语言为 s 的层次分类体系中的每个概念，按照相关性高低对语言为 t 的层次分类体系中的所有概念排序并取前 K 个结果。假设给定语言为 s 的层次分类体系中的 N 个概念，在其中第 j 个概念 c_j 对应的排序后的前 K 个结果中，共有 K_j 个概念与 c_j 相关，则可计算得分 $score_j$ 如下：

$$score_j = \frac{K_j}{K} \tag{2.4}$$

然后对给定的 N 个概念的得分取均值，即可得到 $Precision@K$ 如下：

$$Precision@K = \frac{1}{N} \sum_{j=1}^{N} score_j \tag{2.5}$$

MRR（Mean Reciprocal Rank）：给定语言为 s 的层次分类体系中的 N 个概念中的第 j 个概念 c_j，由于查准率@K 只能考察方法返回的前 K 个概念（源自语言为 t 的层次分类体系）与 c_j 相关的比例，无法得知与 c_j 最相关的源自语言为 t 的层次分类体系中的概念 c_j^* 的位置信息，所以需使用一种新的评测指标 MRR。首先计算 c_j^* 的位置排名得分 $rank_j$ 如下：

$$rank_j = \frac{1}{R_j} \tag{2.6}$$

其中 R_j 是 c_j^* 的排名，然后针对所有 N 个给定概念计算 MRR 如下：

$$MRR = \frac{1}{N}\sum_{j=1}^{N} rank_j \tag{2.7}$$

2.4 本章小结

本章介绍了本文的背景知识，首先简述了知识图谱的发展历史及构建技术，覆盖了知识抽取、知识集成、知识补全三个方面。然后就本文中的“双语知识图谱”、“层次分类体系”、“概念”、“词汇中心词”等专用术语给出了定义与解释。最后介绍了本文所提出方法中用到的各种评测指标，包含查准率、查全率、F1 值、查准率@K、MRR。

第三章　模式知识挖掘

在数据集成、推理、问答等实际应用中，模式知识有着非常重要的作用。本章以社交站点中层次分类体系（如商品目录）中的分类与分众分类系统（如微博标签系统）中的标签为知识挖掘来源，重点研究模式知识挖掘的问题，旨在以一种各语言通用的方法挖掘由分类或标签表示的概念之间的关系。本章的具体内容安排如下，首先在章节 3.1中概述现有工作的问题、本章的解决方案、实验结果，然后在章节 3.2 中介绍相关工作，之后针对现有工作的问题，在章节 3.3 中详细介绍所提出的一种新的各语言通用的方法以挖掘社交站点中概念间的 *equal*、*subClassOf*、*relate* 关系，并在章节 3.4 中讨论所提出的方法在中英文社交站点上的实验结果，最后于章节 3.5进行小结。

3.1 概述

文献 [28] 对社交站点中的模式知识挖掘工作已有初步探索，该工作的目的是在层次分类体系中的分类与分众分类系统中的标签中挖掘不同的语义关系，包括 *equal*、*subClassOf*、*relate*三种关系。然而，该工作所提出的方法存在三个主要问题如下：

- 所提出的方法依赖于特定语言的规则与特征，仅能在中文社交站点中使用。这导致该方法无法应用到不同语言的模式知识挖掘的工作中。
- 所提出的方法在利用机器学习技术时，依赖手工标注数据，导致在应用该方法前需要进行大量的人工操作。
- 所提出的方法在挖掘由分类或标签表示的概念之间的关系时，分开独立使用规则与机器学习技术，即该方法未考虑综合利用规则与机器学习的优势以获取更好的结果。

为了解决上述问题，本章设计了一种新的各语言通用的方法在社交站点中挖掘概念（由分类与标签表示）之间的关系。本章所提出的新方法不仅能够自动化地为机器学习模型生成标注数据，而且将规则嵌入到机器学习的过程中以获得更好的学习结果。具体而言，所提出的方法首先使用了一种分块（Blocking）机制以减少待判定的概念对（每个概念对由两个分类或两个标签或一个分类与一个标签组成）的数量，从而尽可能保证方法能够在大规模的场景下进行应用。然后，利用了一种自动化的策略从给定的概念对中生成标注数据。最后，提出了一种半监督学习方法检测给定的两个概念之间是否存在 *equal*、*subClassOf*、*relate*关系，其中包含一个基于各语言通用的规则的后处理步骤，该步骤在迭代学习的过程中用于修正错误分类的结果。

实验结果表明本章所提出的方法在测试数据集上的查准率、查全率、F1 值三项评测指标上均优于其他所设计的对比方法。在将所提出方法分别应用于大量中英文社交站点后，评测结果表明所提出的方法能够挖掘得到高质量的中英文模式知识，与 DBpedia、Yago、BabelNet、Freebase 等知识图谱相比，通过本章方法得到的中英文模式知识均覆盖了大规模的概念且包含最大数量的 *subClassOf* 关系。

3.2 相关工作

与本章研究相关的工作可分为两类：本体学习与本体匹配，下文将进行具体介绍。

3.2.1 本体学习

本体学习，特别是层次分类体系学习在众多研究领域中均获得了极大的关注。层次分类体系学习可分为基于百科的层次分类体系学习与基于万维网的层次分类体系学习。基于百科的方法主要专注于从维基百科抽取层次分类体系。WikiTaxonomy [83] 是一个基于英文维基百科的分类系统构建而成的层次分类体系，它包含 105,000 个正确率为 88% 的 *subclassOf* 关系。Kylin 本体生成器 [84] 使用马尔科夫逻辑网推断维基百科信息框中的分类之间的包含关系。Yago [17] 将维基百科中的分类与 WordNet 中的 Synset 进行链接，整合后的层次分类体系中的叶子节点为维基百科分类，其他节点则为 WordNet 本身层次分类体系中的 Synset。目前 Yago 层次分类体系共有 350,000 个分类与 450,000 个正确率为 96% 的 *subclassOf* 关系。

关于基于万维网的层次分类体系学习，Hearst 模式 [85]（pattern）是最为常用的方法。最近的一个研究工作是 Microsoft Concept Graph [24]，它从万维网网页文本中抽取 *IsA*关系以构建大规模层次分类体系，目前共有超过 500 万个概念与 8,000 万个 *subclassOf* 关系。这样的 *IsA*关系并不区分 *subclassOf* 关系与 *instanceOf* 关系。Zhou 等人 [86] 提出一种非监督模型自动化地从社交标注信息中导出层次结构信息。Tang 等人 [26] 提出了一种主题学习模型，旨在从分众分类系统中学习层次结构语义。Lin 等人 [87] 使用关联规则挖掘与 WordNet 从分众分类系统中学习本体。此外，综述 [27] 对比了不同的从社交标签中挖掘语义的方法，它们主要关注如何从分众分类系统中抽取层次化的语义结构，即层次分类体系。

由于本章从众多社交站点中的分类与标签中挖掘三种不同的语义关系，所以本章所提出的方法应属于基于万维网的层次分类体系学习的方法。与上述工作相比，本章工作不仅学习层次分类体系，而且挖掘概念（包含分类与标签）间的 *equal*与 *relate*关系，更为重要的是本章设计了一种融合不同特征与规则的各语言通用的半监督学习方法以解决概念间关系挖掘的任务。

3.2.2 本体匹配

本体匹配是在语义网与数据库领域非常热门的一项研究。近十年来，许多本体匹配工具 [56] 得以开发与应用。下文将介绍一些具有代表性的例子。Falcon-AO [57] 是一

个利用语言学特征与图结构进行自动化异构本体匹配的系统。但是它仅能挖掘 *equal*关系，故不能解决本章面对的问题。BLOOMS [58] 是另一个本体匹配工具，它利用外部的知识图谱 WordNet 或维基百科为给定本体中的每个概念构建一个 BLOOMS 森林，然后定义一个函数计算给定的两个 BLOOMS 森林中每两个树之间的重合度，最后基于多个计算得到的重合度决定匹配是否成立。假设在本章的场景下使用 BLOOMS，如果选择维基百科为辅助匹配的知识图谱，那么在处理大规模的概念对时，BLOOMS 会多次调用维基百科，造成大量的时间消耗。如果选择 WordNet 为辅助匹配的知识图谱，由于 BLOOMS 仅能考虑社交站点中每个概念的字符串而并不分析概念的上下文结构，从而使得匹配结果极少。PARIS [59] 不仅匹配概念而且匹配关系与实例，它基于概率估计模型度量每种元素间的匹配程度，且不需要进行任何的参数调节。需要注意的是，PARIS 进行概念匹配的前提是先完成实例匹配，这导致它无法在本章的场景使用，因为本章的数据源只有概念（分类与标签），没有实例。本章需要处理的是现有层次分类体系中的分类与无显式结构的标签，而现有的本体匹配工具难以在这些分类与标签中挖掘不同的语义关系。

3.3 方法设计

3.3.1 问题定义

输入：给定某种语言的一组社交站点 $WS = \{ws_1, ws_2, ..., ws_n\}$，其中每个站点 ws 可能包含一组分类 $CA_{ws} = \{ca_1, ca_2, ..., ca_m\}$ 与一组标签 $TA_{ws} = \{ta_1, ta_2, ..., ta_o\}$。分类以层次结构的方式进行组织。在一个分类层次结构中，一个分类可能与零个或多个父类与子类相关联。如图 3.1所示，在 Google 商品层次分类体系中，分类“*Uniforms*”拥有父类“*Clothing*”及子类“*School Uniforms*”与“*Sports Uniforms*”等。而标签则是以平面的方式进行组织，并不存在任何预先定义的层次结构。图 3.2给出了一个关于标签的示例，标签“*nlp*”、“*terminology*”、“*semantics*”、“*semantic-web*”组成了一个用于标注 Stackoverflow[1]中某个问题的标签组。本章定义从社交站点收集得到的分类与标签均表示概念。由于任意一个分类 ca_i 是被网站预先定义且不可随意修改，所以定义其表示一个静态概念（Static Concept）。而由于任意一个标签 ta_j 是由万维网用户随意创建且可以根据用户需求随时更改，所以定义其表示一个动态概念（Dynamic Concept）。

输出：本章旨在生成任意给定语言的大规模模式知识。此处的模式知识包含三种概念间的语义关系，即 *equal*、*subClassOf*、*relate*关系，它们分别沿用 `owl:equivalentClass`[2]、`rdfs:subClassOf`[3]、`skos:related`[4]的定义。两个概念间存在 *equal*关系当且仅当它们包含完全相同的实例集合。一个概念是另一个概念的子概念（即前后两个概念间存在 *subClassOf*关系）当且仅当前者包含的所有实例均是后者的实例。两个概念间存在 *relate*关系当且仅当二者存在非 *equal*、非 *subClassOf*的关联关系。

[1] https://stackoverflow.com/
[2] https://www.w3.org/TR/owl-ref/#equivalentClass-def
[3] https://www.w3.org/TR/rdf-schema/#ch_subclassof
[4] https://www.w3.org/TR/swbp-skos-core-guide/#secassociative

图 3.1 分类的示例：Google 商品层次分类体系中的分类

图 3.2 标签的示例：Stackoverflow 中的标签

在这三种关系中，*relate*关系的语义最弱。多个社交站点中概念间的 *subClassOf* 关系构成了一个整合的层次分类体系，而所有抽取得到的关系组成了一个大规模的语义网络。

3.3.2 方法流程

本节主要介绍所提出方法的流程以及方法中各组件的交互方式。如图 3.3所示，本章所提出的方法主要包含四个组件：分块器、已标注数据生成器、半监督学习器、后处理器。分块器的输入是从不同社交站点中收集得到的概念（即分类与标签）。分块器将所有的概念划分到不同的块（Block）中，并且仅从划分到相同块的概念中生成未标注概念对。与从所有概念中取两个概念的组合数相比，分块器输出相对较少数量的概念对以供后续处理，这从一定程度上提高了方法的效率。已标注数据生成器通过使用 BabelNet 中已存在的 *equal*关系与 Yago 中已存在的 *subClassOf* 关系自动化地生成已标注数据。然后获取所有已标注与未标注概念对的各语言通用的特征，包括词汇特征、基于维基百科的特征、基于 BabelNet 的特征、基于网页的特征。这些特征从不同的角度衡量概念间的相关性。随后，使用半监督学习器发现概念间的 *equal*、*subClassOf*、*relate*关系。半监督学习器实质是一个迭代更新的分类器，它在每次迭代中增加具有高置信度的新的标注数据从而完成分类器的更新。与此同时，在每次迭代中，均使用一个后处理器，它利用两个各语言通用的规则纠正先前的误分类结果。方法最终的输出是由多个社交站点中的概念之间的 *equal*、*subClassOf*、*relate*关系组成的关于给定语言的大规模模式知识。

3.3.3 分块机制

从社交站点收集到的概念中枚举所有的概念对用于语义关系检测是不切实际的。分块旨在将概念划分到不同的块中，其中每个块仅包含相似的概念用于后续处理。本章假设 1）当两个概念拥有一个相同的特征时，则认为二者相似；2）仅存在于相同块中的任意两个概念之间可能存在一种语义关系。本节介绍的分块机制中的特征获取方法如下：首先抽取给定概念的词汇中心词在 BabelNet 中对应的“*sense*”（BabelNet 中的一

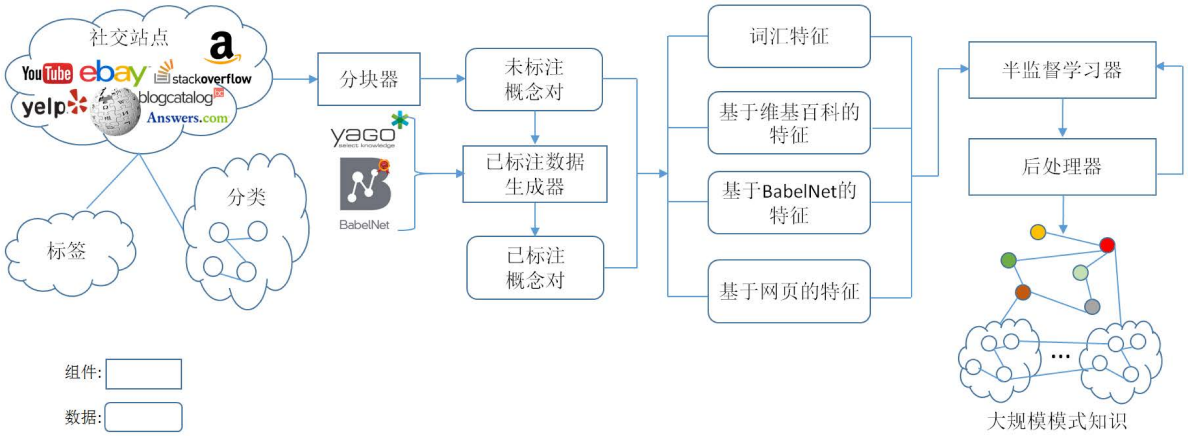


图 3.3 模式知识挖掘方法的流程示意图

种特殊结构），然后将该“*sense*”提交 BabelNet 进行搜索，最后将所有返回的上位词与下位词作为给定概念的特征。如果一个词汇中心词在给定的语言上拥有多个“*sense*”，则直接将多个“*sense*”对应的多个特征集合合并为一个集合，而不是从多个“*sense*”中选择最准确的一个，因为分块更加关注查全率而不是查准率。分块机制中产生的噪音可由后续步骤进行过滤。

基于上述两种假设，本节设计的分块机制由如下步骤构成。首先，每个概念由一些具有代表性的特征进行表示；然后，如果两个概念拥有一个相同的特征，则将它们划分到同一个块中。最终，建立倒排索引，从而使每个特征对应一个块。具体示例如下：给定四个概念 w、x、y、z，其各自对应的特征集合为 $w=\{A\}$、$x=\{A,B\}$、$y=\{B,C\}$、$z=\{C\}$，其中 A、B、C 为特征，然后建立如下的倒排索引：$A=\{w,x\}$、$B=\{x,y\}$、$C=\{z\}$，即可获得三个生成的块：$\{w,x\}$、$\{x,y\}$、$\{z\}$，最后仅在每个生成的块中抽取概念对。此外，为了去除现有层次分类体系中不正确的 *subClassOf* 关系，本章抽取所有层次分类体系中拥有父子关系的两个概念构成概念对，以便后续检测。例如，给定一个层次分类体系中一个一部分:“*Chinese Physicists*”是“*Physicists*”的孩子节点，“*Physicists*”是“*Scientists*”的孩子节点，所以可得到概念对如下：$\{(Chinese\ Physicists, Physicists), (Physicists, Scientists)\}$。

3.3.4 已标注数据生成

本章基于 Yago 和 BabelNet 自动化地生成已标注数据，为后续半监督学习提供必需的训练数据。在众多开放链接数据中的知识图谱中，Yago 包含规模最大的由 *subClassOf* 关系构成的高质量层次分类体系[88–90]。与此同时，BabelNet 是当前最大的在线同义词典并且拥有大规模的与本章定义的 *relate* 关系类似的 `Semantically_Related` 关系。因此，本章选择这两个知识图谱辅助自动化生成已标注数据。具体而言，先随机选择 200 个未标注的概念对，其中每个概念对中的两个概念在 BabelNet 中已持有 *equal* 关系；然后随机选择 800 个未标注的概念对，其中每个概念对中的两个概念在 Yago 中已持有 *subClassOf* 关系; 最后随机选择 3,000 个未标注的概念对，其中每个概念对中

的两个概念在 BabelNet 中已持有 `Semantically_Related`（即 *relate*）关系，且不在 BabelNet 中持有 *equal*关系，也不在 Yago 中持有 *subClassOf* 关系。上述 $4,000$ 个概念对作为后续半监督学习输入的已标注数据。

3.3.5 特征工程

为了从不同的角度度量概念之间的相关性，本节定义两组共六个特征。第一组称为基本特征，其中包含一个词汇特征与一个基于维基百科的特征。另外一组称为语义特征，包括两个基于 BabelNet 的特征和两个基于网页的特征。所有概念对的上述特征将被输入到半监督学习器中，从而对每个概念对进行三分类（三个分类分别为 *equal*、*subClassOf*、*relate*关系）。

3.3.5.1 基础特征

a）词汇特征：为了获取概念之间的语言学相关性，本节设计了一种基于最长公共子序列的非对称的相似度作为词汇特征。一个概念 c 的字符串表示为 l_c，词序列表示为 $seq(l_c)$。本节定义任意两个概念 c_1 与 c_2 之间的概念字符串相似度（Concept String Similarity, CSSim）如下：

$$\text{CSSim}(c_1, c_2) = \frac{|LCS(seq(l_{c_1}), seq(l_{c_2}))|}{|seq(l_{c_1})|} \tag{3.1}$$

其中 $|.|$ 表示一个词序列的长度，而 LCS 是计算两个概念字符串之间的最长公共子序列的函数。

b）基于维基百科的特征：受显式语义分析 [91]（Explicit Semantic Analysis, ESA）启发，将给定的概念与维基百科中的分类建立映射，即可使用这些分类表示给定的概念。这种表示方法有三重益处：1）丰富了概念的表示，从仅仅是概念的字符串表示扩展到由一组维基百科分类进行表示；2）维基百科分类的维度通常远远小于文本特征的维度，由此可避免维数灾难并支持更有效率的处理；3）相比于文本，维基百科分类拥有更高的质量，原因是歧义现象在分类中较少。

基于 ESA，每个概念 c 均对应一个 ESA 向量，即 $ESA_c = \langle wc_1(c), wc_2(c), ..., wc_n(c)\rangle$，其中 wc_i 是一个维基百科分类，$wc_i(c)$ 是该分类对应的权重，该权重表示概念 c 与维基百科分类 wc_i 之间的相关性。本节定义任意两个概念 c_1 与 c_2 之间的 ESA 向量相似度（ESA Vector Similarity, ESAVSim）如下：

$$\text{ESAVSim}(c_1, c_2) = \frac{\sum_{wc} wc(c_1) \cdot wc(c_2)}{\sqrt{\sum_{wc} wc(c_1)^2 \cdot \sum_{wc} wc(c_2)^2}} \tag{3.2}$$

$\text{ESAVSim}(c_1, c_2)$ 实际上是两个给定概念的 ESA 向量的余弦相似度。

3.3.5.2 语义特征

由于基础特征仅由概念的字符串本身带来的信息计算所得，并未考虑给定概念的语义。因此，本节提出若干获取概念间语义相关性的语义特征。

c）基于 BabelNet 的特征：任意给定两个概念 c_1、c_2，首先通过完全字符串匹配将它们与 BabelNet 进行映射，然后计算它们之间的 Wu & Palmer (WUP) 相似度 [92]，该相

似度（表示为 $WUPSim(c_1, c_2)$）的计算利用了给定的两个概念各自在 BabelNet 层次分类体系中的深度以及这两个概念的最低公共祖先节点的深度，具体定义如下：

$$\text{WUPSim}(c_1, c_2) = \frac{2 * depth(LCA(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{3.3}$$

其中 $LCA(c_1, c_2)$ 表示 c_1 与 c_2 在 BabelNet 层次分类体系中的最低公共祖先节点，而 $depth$ 则是计算给定概念在 BabelNet 层次分类体系中对应节点的深度的函数。需要注意的是，由于概念可能存在歧义，所以一个概念可能对应 BabelNet 层次分类体系中不同的节点，所以对于给定的两个概念可能存在不止一个 WUP 相似度。此处仅选择最高的 WUP 相似度，这不仅能够降低歧义对于概念相关性度量的负面影响，还尽可能保证了语义关系挖掘的查全率。

此外，本节还定义了另一个非对称的基于 BabelNet 的特征，即相对深度相似度（Relative Depth Similarity, RDSim），该相似度用于度量两个给定概念在 BabelNet 层次分类体系中的相对深度差异，有助于非对称关系的学习。给定任意两个概念 c_1 与 c_2，二者各自可能对应多个 BabelNet 层次分类体系中的节点，此处选择能够获得最高 WUP 相似度的两个节点用于计算 c_1 与 c_2 间的相对深度相似度如下：

$$\text{RDSim}(c_1, c_2) = \frac{depth(LCA(c_1, c_2))}{depth(c_2)} - \frac{depth(LCA(c_1, c_2))}{depth(c_1)} \tag{3.4}$$

如果 c_1 的深度高于 c_2，则表示 c_2 更有可能是 c_1 的祖先概念。比如，给定两个概念“*Product*”与“*Stroller*”，他们的最低公共祖先节点是“*Artifact*”。“*Stroller*”的深度是 10，“*Product*”的深度是 8，则存在一定的概率“*Stroller*”是“*Product*”的子孙概念。

d）基于网页的特征：每个静态概念（即分类）的父概念与动态概念（即标签）的存在于相同标签组的其他概念均可视为给定概念的上下文信息，这些信息有助于揭示每个概念的真正含义。例如，在 eBay 层次分类体系中，概念“*Sports*”是“*Shopping and Services*”的子节点，则“*Sports*”表示体育用品；而在维基百科中，“*Sports*”是“*Recreation*”的子节点，则“*Sports*”表示各类体育活动。然而，对静态与动态概念而言，上述的上下文信息是异构且有限的，所以仅利用这样的上下文信息难以合理地度量概念之间的相关性，因此选择利用搜索引擎 Google 对万维网进行查询以获得更多的概念上下文信息。

为了正确获得由 Google 返回的每个分类（即静态概念）的上下文信息，将给定分类 ca 与其父类 pca 共同提交给 Google。然而，不同层次分类体系中的根类（Root Categories）却没有父类，但又由于根类通常是没有歧义的，否则用户在自顶向下浏览层次分类体系时易产生混乱。因此，对于根类，仅将其单独提交给 Google 以获取其上下文信息。与分类不同，给定一个标签（即动态概念）ta，将随机选择一个与 ta 在相同标签组共现的标签 ota，然后将 ta 与 ota 共同提交给 Google。

在每次查询后由 Google 返回的前二十个网页片段（Snippets）中，抽取除 pca（或 ota）以外的与 ca（或 ta）在相同句子中共现的词语。此处排除 pca（或 ota）的原因是它们是查询本身的一部分，所以必然在返回的网页片段中多次出现。在过滤了停用词与词频低于 3 的词语后，采用 TF-IDF [93]（Term Frequency-Inverse Term Frequency）计

算每个词语的权重，所以每个概念（分类或标签）c 均可以由一个 n 维的上下文向量进行表示，即 $CV(c) = \langle w_1(c), w_2(c), ..., w_n(c) \rangle$，其中第 i 项 $CV(c)_i$ 的权重为 $w_i(c)$，而 n 指所有概念对应抽取得到的所有词语的数量。如果一个词语 w 没有与 c 共现，那么 w 在 $CV(c)$ 中对应的值则为 0。任意给定两个概念 c_1 与 c_2，它们之间的上下文向量相似度（Context Vector Similarity, CVSim）的定义如下：

$$\text{CVSim}(c_1, c_2) = \frac{\sum_{i=1}^{n} CV(c_1)_i \cdot CV(c_2)_i}{\sqrt{\sum_{i=1}^{n} CV(c_1)_i^2 \cdot \sum_{i=1}^{n} CV(c_2)_i^2}} \tag{3.5}$$

其中 $\text{CVSim}(c_1, c_2)$ 实际上是向量 $CV(c_1)$ 与 $CV(c_2)$ 之间的余弦相似度。除此之外，概念 c 还可以通过一组上下文集合进行表示，即 $CS(c) = \{w_1, w_2, ..., w_m\}$，其中 w_i 是第 i 个关于 c 的过滤后保留的词语，m 是关于 c 的所有的过滤后保留的词语的总数。为了进一步提升非对称关系的学习效果，依据上述表示方法，本节定义了另外一种非对称的基于网页的特征，称为相对上下文集合相似度（Relative Context Set Similarity, RCSSim），该相似度用于度量任意给定的两个概念 c_1 与 c_2 各自对应的网页上下文集合 $CS(c_1)$ 与 $CS(c_2)$ 之间的相对覆盖率差异，具体定义如下：

$$\text{RCSSim}(c_1, c_2) = \frac{|CS(c_1) \cap CS(c_2)|}{|CS(c_1)|} - \frac{|CS(c_1) \cap CS(c_2)|}{|CS(c_2)|} \tag{3.6}$$

该公式基于一个基本假设：如果大部分关于 c_1 返回的网页与关于 c_2 返回的网页相类似，但仅有一部分关于 c_2 返回的网页与 c_1 返回的网页相似，则 c_1 很有可能为 c_2 的子孙概念。相似的假设于工作 [94] 首先提出。此处使用从网页片段中抽取得到的上下文词语集合表示关于 c_1 与 c_2 返回的网页。

3.3.6 半监督学习

虽然本章利用了 Yago 与 BabelNet 自动化地生成了已标注数据，但是与分块后得到的未标注的概念对的数量相比，已标注数据的数量（仅 4,000）远远不足，所以一种自然的研究思路是利用某种半监督学习算法直接在未标注概念对中预测关系。

目前存在许多可以在当前场景应用的算法，如标签传播等，此处选择最简单且效率最高的一种算法，即 Self-Training [95]。标准的 Self-Training 算法在每次迭代中首先将已标注数据作为训练数据，并以此为基础学习得到一个分类器；然后将该分类器于未标注数据上进行应用并将具有高置信度的已分类结果添加到已标注数据中，从而为下一次算法迭代中训练新的分类器做好数据准备。整个算法在遇到如下情况即会终止：1）在连续两次迭代中，概念对分类结果的差异低于一定的阈值；2）达到预先设定的最大迭代次数。

需要注意的是，在 Self-Training 算法的每次迭代中，本章使用 SVM [96]（Support Vector Machine）模型训练一个三分类（即 *equal*、*subClassOf*、*relate*三类）的分类器，SVM 是目前已知的效果最佳的分类器之一 [97]。此外，有别于上述标准的 Self-Training

算法，本章并不直接将具有高置信度的已分类概念对加入到已标注数据中，而是提出一个后处理步骤，其中使用规则过滤出误分类的概念对，具体细节将于下一节介绍。

3.3.7 后处理

为了保证由半监督学习器学习得到的模式知识的质量，本章在学习的过程中添加了一个后处理步骤，其目的是在 Self-Training 算法的每次迭代中利用规则过滤出 SVM 模型误分类的概念对。本节提出两个通用且有效的规则如下：

规则 1：给定一个概念对 (c_1, c_2)，如果 c_1 与 c_2 具有不同的概念字符串，且 c_1 的词汇中心词 $h(c_1)$ 的字符串与 c_2 本身的字符串相同，则判定 c_1 *subClassOf* c_2。

规则 2：给定一个概念对 (c_1, c_2)，如果 c_2 与 c_1 已经在某个社交站点的层次分类体系中拥有父子关系，且它们具有相同的词汇中心词，则判定 c_1 *subClassOf* c_2。

当上述两个规则进行应用时，需要忽略某些语言中（如英文）的大小写与单复数差异。***规则 1***与 ***规则 2***作为通用的规则，可以在不同的语言中进行应用。比如，在英文中，概念“*Chinese Physicists*”的词汇中心词“*Physicists*”与概念“*Physicists*”的字符串相同，则可基于***规则 1*** 判定概念“*Chinese Physicists*” *subClassOf* 概念“*Physicists*”。在中文中，概念“*江苏学校*”与概念“*中国学校*”拥有相同的词汇中心词“*学校*”，且二者在维基百科的层次分类体系中已经具有父子关系，所以可依据***规则 2***判定概念“*江苏学校*” *subClassOf* 概念“*中国学校*”。

虽然上述规则对于任意语言均是通用的，但是在不同的语言中，关于抽取概念中的词汇中心词的方法是不同的。由于本文主要关注英文与中文，以下仅介绍如何在这两种语言的概念中抽取词汇中心词。

面向英文：此处使用与工作 [98] 相同的抽取英文概念的词汇中心词的方法。首先使用 Standford parser [99] 解析给定概念的字符串，然后限制词汇中心词发现算法 [100] 的输出为名词或多个并列的名词，比如对于概念“*Buildings and Infrastructures in Japan*”而言，将输出两个词汇中心词：“*Buildings*”与“*Infrastructures*”。

面向中文：此处使用与工作 [101] 相同的抽取中文概念的词汇中心词的方法。首先使用 FudanNLP [102] 对给定概念的字符串进行词性标注，然后将位于最后的名词作为词汇中心词输出。例如，对概念“*中国足球运动员*”进行词性标注，输出为：“*中国/LOC 足球/NN 运动员/NN*”，其中“*LOC*”表示地点，“*NN*”表示名词，所以“*运动员*”为词汇中心词。

3.4 实验分析

3.4.1 站点信息统计

本章选用 21 个英文流行社交站点与 55 个中文流行社交站点以对所提出的方法进行实验评测与分析，所有的数据于 2015 年 9 月抽取，具体每个英文站点的信息见表 3.1，每个中文站点的信息见表 3.2。这两个表格中记录了每个站点的名称、URL、类型、分

表 3.1 21 个英文社交站点的信息统计

站点名	URL	站点类型	分类数量	标签数量	层次分类体系平均深度
Amazon	http://www.amazon.com	电子商务	3,047	/	2.23
Yahoo Answer	https://answers.yahoo.com	问题回答	977	/	2.54
Youtube	http://www.youtube.com	视频分享	127	/	2.50
English Wikipedia	http://en.wikipedia.org	在线百科	103,476	/	5.00
MSN	http://www.msn.com	即时通讯	75	/	1.90
Foursquare	http://foursquare.com	社交服务	360	/	2.75
Ebay	http://www.ebay.com	电子商务	10,536	/	4.40
Expedia	http://www.expedia.com	在线百科	46	/	2.00
Answers	http://wiki.answers.com	问题回答	8,535	/	5.77
Thisnext	http://www.thisnext.com	电子商务	3,177	/	2.63
Match	http://www.match.com	婚恋交友	211	/	3.00
Yelp	http://www.yelp.com	消费评论	277	/	2.12
Epinions	http://www.epinions.com	消费评论	656	/	3.31
Bigboards	http://www.big-boards.com	论坛	771	/	3.92
Slideshare	http://www.slideshare.net	文档分享	39	/	1.00
Blogcatalog	http://www.blogcatalog.com	博客	353	/	2.02
Craiglist	http://www.craigslist.org	生活分类	96,443	/	5.00
Groupon	http://www.groupon.com	电子商务	73	/	1.75
Zynga	http://zynga.com	社交游戏	5	/	2.00
Instagram	http://instagram.com/	社交服务	/	9,519	/
Stackoverflow	http://stackoverflow.com/	问题回答	/	3,600	/

类数量、标签数量、层次分类体系的平均深度。如果一个站点不包含分类或标签，则使用符号“/”表示单元格的值为空。与存在于层次分类体系中的分类相比，标签本身的语义较为不稳定，且更改的频率较高，所以本章并未将社交站点中的所有标签作为知识挖掘的来源，而是仅在相应社交站点中选用于 2015 年 8 月期间最为流行的标签。最终，共抽取 229,184 个英文分类与 13,119 个英文标签，以及 328,248 个中文分类与 71,605 个中文标签。

3.4.2 方法评测

本节将从两个不同的角度评测所提出方法的有效性：1）分析所提出方法中的特征与规则各自的有效性；2）在中英文站点中应用所提出方法后，评测所得到的中英文知识的正确率。

3.4.2.1 特征与规则的贡献分析

为了分析所设计特征与规则对预测不同的语义关系预测的有效性，本章在两组利用 BabelNet 和 Yago 自动化生成的已标注数据中进行评测（生成过程于章节 3.3.4中介绍）。一组为英文已标注数据，一组为中文已标注数据。每组均包含 200 个被标注为“*equal*”关系的概念对，800 个被标注为“*subClassOf*”关系的概念对，以及 3,000 个被标注为“*relate*”关系的概念对。针对中英文已标注数据，分别采用三种基于不同特征与规则组合的方法进行测试。第一种方法（称基础方法）在半监督学习的过程中仅使用基于基础特征（即词汇特征与基于维基百科的特征）的 SVM 分类器。第二种方法（称基础 + 语义方法）使用基于基础特征与语义特征（即基于 BabelNet 的特征与基于网页的特征）的 SVM 分类器。第三种方法（称基础 + 语义 + 规则方法）即为本章所提出方法的完整

表 3.2 51 个中文社交站点的信息统计

站点名	URL	站点类型	分类数量	标签数量	层次分类体系平均深度
360 手机助手	http://sj.360.cn/	手机应用市场	49	/	1.69
91 手机助手	http://zs.91.com/	手机应用市场	76	/	1.55
亚马逊	http://www.amazon.cn/	电子商务	3,310	/	3.65
安卓市场	http://apk.hiapk.com/	手机应用市场	279	/	2.56
苹果应用市场	http://www.apple.com/cn/	手机应用市场	90	/	1.69
百度百科	http://baike.baidu.com/	在线百科	11,743	/	2.67
百度贴吧	http://tieba.baidu.com/	论坛	213	/	1.57
百度文库	http://wenku.baidu.com/	文档分享	298	/	1.87
百度知道	http://zhidao.baidu.com/	问题回答	2,117	/	3.24
百姓网	http://www.baixing.com/	生活分类	55,179	/	4.08
当当网	http://www.dangdang.com/	电子商务	6,847	/	2.59
点点	http://www.diandian.com/	轻博客	/	8,105	/
丁丁地图	http://www.ddmap.com/	消费评论	26,993	/	2.50
豆丁网	http://www.docin.com/	文档分享	734	/	1.60
豆瓣	http://www.douban.com/	社交网络	13,168	/	4.04
饭统网	http://www.fantong.com/	消费评论	3,842	/	2.61
鲜果	http://xianguo.com/	信息聚合	36	/	1.62
赶集网	http://www.ganji.com/	生活分类	25,274	/	3.81
逛	http://guang.com/	社交电子商务	293	/	2.61
互动百科	http://www.baike.com/	在线百科	32,293	/	5.72
江南情缘	http://www.88999.com/	婚恋交友	153	/	2.02
世纪佳缘	http://www.jiayuan.com/	婚恋交友	77	/	1.83
京东商城	http://www.jd.com/	电子商务	31,140	/	3.59
开心网	http://www.kaixin001.com/	社交网络	124	/	2.45
驴评	http://www.lvping.com/	在线旅游	40,475	/	3.57
美丽说	http://www.meilishuo.com/	社交电子商务	316	/	2.57
猫扑	http://www.mop.com/	论坛	22	/	1.55
PPS	http://www.pps.tv/	视频分享	288	/	1.50
切客	http://www.qieke.com/	移动服务	6,224	/	3.51
穷游网	http://www.qyer.com/	在线旅游	107	7,400	1.68
人和网	http://www.renhe.cn/	商务社交	249	/	2.55
人人网	http://www.renren.com/	社交网络	118	/	1.98
人人游戏	http://wan.renren.com/	社交游戏	43	/	1.70
人人小站	http://zhan.renren.com/	轻博客	/	7,038	/
若邻网	http://www.wealink.com/	商务社交	62	/	1.56
新浪爱问	http://iask.sina.com.cn/	问题回答	5,247	/	3.24
新浪博客	http://blog.sina.com.cn/	博客	27	16,190	1.56
新浪游戏	http://games.sina.com.cn/	社交游戏	52	/	1.67
新浪共享	http://ishare.sina.com.cn/	文档共享	234	/	1.57
新浪微博	http://weibo.com/	微博	183	/	2.66
淘宝网	http://www.taobao.com/	电子商务	1,843	/	3.34
腾讯博客	http://blog.qq.com/	博客	23	/	1.65
腾讯微博	http://t.qq.com/	微博	15	/	1.00
天涯	http://www.tianya.cn/	论坛	1,706	/	3.18
土豆	http://www.tudou.com/	视频分享	755	/	1.64
推他网	http://www.tuita.com/	轻博客	/	3,135	/
网易博客	http://blog.163.com/	博客	19	/	1.60
网易微博	http://t.163.com/	微博	/	29,737	/
网易阅读	http://yuedu.163.com/	信息聚合	46	/	1.83
中文维基百科	http://zh.wikipedia.org/	在线百科	55,122	/	3.71
优酷	http://www.youku.com/	视频分享	744	/	1.62

版本，使用了第二种方法中的分类器与后处理步骤中的规则。

此处采用五重交叉验证训练分类器，评测指标为查准率、查全率、F1 值。如表 3.3与表 3.4所示，本章所提出方法的完整版本在英文与中文的已标注数据中的评测结

表 3.3 各方法在英文已标注数据中的评测结果

关系	方法	查准率	查全率	F1 值
relate	基础方法	0.865	0.883	0.874
	基础 + 语义方法	0.878	0.911	0.894
	基础 + 语义 + 规则方法	**0.888**	**0.929**	**0.908**
subClassOf	基础方法	0.654	0.620	0.637
	基础 + 语义方法	0.806	0.710	0.755
	基础 + 语义 + 规则方法	**0.910**	**0.750**	**0.822**
equal	基础方法	0.860	0.770	0.813
	基础 + 语义方法	0.907	0.806	0.864
	基础 + 语义 + 规则方法	**0.930**	**0.935**	**0.932**

表 3.4 各方法在中文已标注数据中的评测结果

关系	方法	查准率	查全率	F1 值
relate	基础方法	0.832	0.876	0.853
	基础 + 语义方法	0.836	0.886	0.860
	基础 + 语义 + 规则方法	**0.882**	**0.922**	**0.902**
subClassOf	基础方法	0.711	0.590	0.645
	基础 + 语义方法	0.769	0.623	0.688
	基础 + 语义 + 规则方法	**0.879**	**0.724**	**0.794**
equal	基础方法	0.876	0.775	0.822
	基础 + 语义方法	0.914	0.795	0.850
	基础 + 语义 + 规则方法	**0.922**	**0.940**	**0.931**

果均最佳，这反映了所提出的基础特征、语义特征、规则对预测概念间的语义关系均有积极的作用。与此同时，在为半监督学习添加了使用规则的后处理步骤后，各评测指标的数值大大提升，这也突出了本章所提出的简单却通用的规则的重要性。

3.4.2.2 知识正确率评测

在中英文站点中应用所提出方法后，即可得到大规模的中英文知识如下：25,474 个英文 *equal*关系，1,047,801 个英文 *subClassOf* 关系，1,327,631 个英文 *relate*关系，11,095 个中文 *equal*关系, 947,645 个中文 *subClassOf* 关系，217,881 个中文 *relate*关系。由于并不存在针对所有给定中英文社交站点中概念间关系的已标注数据，所以只能进行手工评测。但是又因为所得到的语义关系的规模庞大，所以手工评测所有关系是否正确是不现实的。因此，先随机选择一组关系（即样本），随机选择的样本可以反映整个数据集的分布，然后对样本进行手工标注以评测样本的正确率，最后使用样本的正确率近似估计整个数据集所有关系的正确率。

本章采用与 Yago 相类似的标注方法，共四位从事知识图谱研究的硕士研究生参与

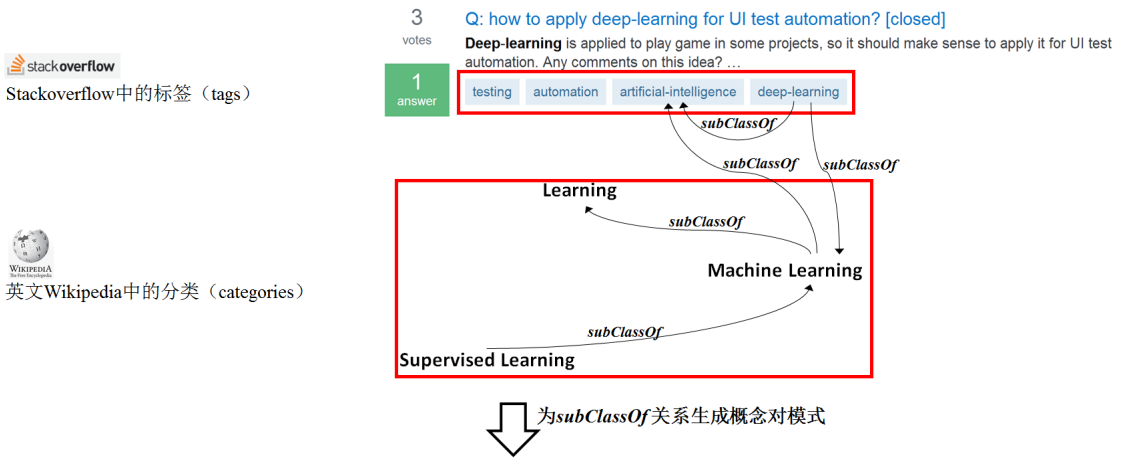


图 3.4　示例：为给定的 *subClassOf* 关系生成不同的概念对模式

标注工作。每位标注人员在标注每个样本时，有三个选择：“*同意*”、“*不同意*”、“*不清楚*”。在每位标注人员标注完所有样本后，则计算平均正确率。然后，利用威尔逊区间 [103] 并在显著性水平 $\alpha = 5\%$ 时，将样本的平均正确率泛化到整个数据集上。

按照上述策略对所有的英文关系进行评测，具体结果如下：

- 随机选择 500 个 *equal*关系，在标注后，平均“*同意*”的票数为 483，通过计算得所有英文 *equal*关系的正确率为 $96.24\% \pm 1.62\%$。
- 随机选择 500 个 *relate*关系，在标注后，平均“*同意*”的票数为 448，通过计算得所有英文 *relate*关系的正确率为 $89.29\% \pm 2.68\%$。
- 随机选择 500 个 *subClassOf*关系，在标注后，平均“*同意*”的票数为 427，通过计算得所有英文 *subClassOf*关系的正确率为 $85.13\% \pm 3.09\%$。

类似地对所有中文关系进行评测，具体结果如下：

- 随机选择 500 个 *equal*关系，在标注后，平均“*同意*”的票数为 474，通过计算得所有中文 *equal*关系的正确率为 $94.45\% \pm 1.96\%$。
- 随机选择 500 个 *relate*关系，在标注后，平均“*同意*”的票数为 461，通过计算得所有中文 *relate*关系的正确率为 $91.87\% \pm 2.36\%$。
- 随机选择 500 个 *subClassOf*关系，在标注后，平均“*同意*”的票数为 440，通过计算得所有中文 *subClassOf*关系的正确率为 $87.71\% \pm 2.85\%$。

3.4.3 知识分布分析

一个概念对的构成可能是两个分类（表示为 `category+category`），或一个分类与一个标签（表示为 `category+tag`），或两个标签（表示为 `tag+tag`）。对于一个分类与

表 3.5 关于英文关系的概念对模式分布

概念对模式	*equal*	*subClassOf*	*relate*
`category+category`	0.564	0.755	0.909
`category+tag`	0.373	0.141	0.071
`tag+category`	/	0.080	/
`tag+tag`	0.063	0.024	0.020

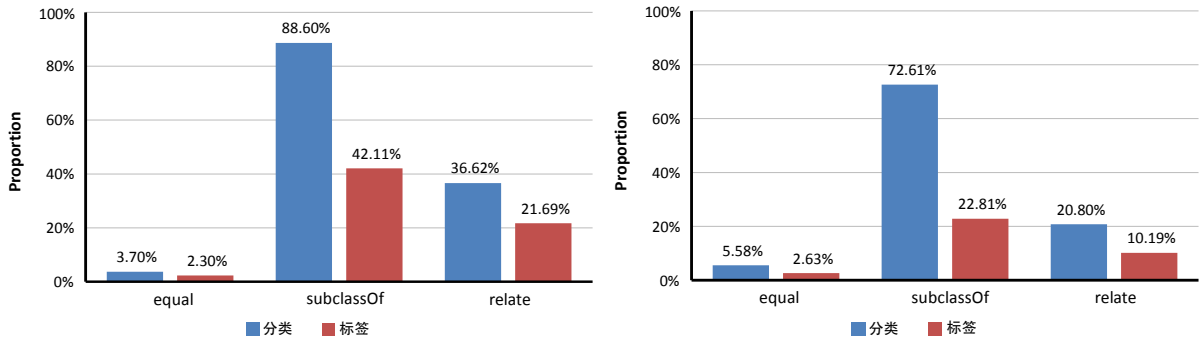
表 3.6 关于中文关系的概念对模式分布

概念对模式	*equal*	*subClassOf*	*relate*
`category+category`	0.769	0.865	0.883
`category+tag`	0.201	0.086	0.102
`tag+category`	/	0.041	/
`tag+tag`	0.030	0.008	0.015

一个标签之间的非对称的 *subClassOf* 关系而言，本章使用 `tag+category`表示一个标签是一个分类的子概念，而 `category+tag`则表示一个分类是一个标签的子概念。上述的表示方法又称为概念对模式，图 3.4给出了为给定的 *subClassOf* 关系生成不同的概念对模式的示例。表 3.5与表 3.6分别统计了英文与中文的概念对模式分布。从这两个表中不难看出 `category+category`模式对于中英两种语言的任意一种关系的贡献比例均最大（大于 0.56），原因在于 1）本章抽取的分类的数量远远大于标签的数量；2）分类的语义稳定性要高于标签。与此同时，相对较少数量的标签依旧贡献了大量概念间的语义关系，这对于仅存在于分类之间的模式知识是良好的补充。

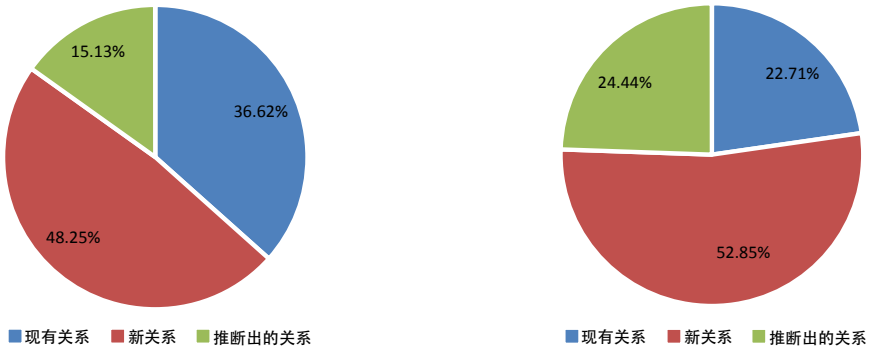
如图 3.5(a)所示，对于所有的英文分类与标签而言，88.60% 的分类与 42.11% 的标签构成了英文 *subClassOf* 关系，36.62% 的分类与 21.69% 的标签构成了英文 *relate*关系，3.70% 的分类与 2.30 的标签构成了英文 *equal*关系。从图 3.5(b)可以看出，对于所有的中文分类与标签而言，72.61% 的分类与 22.81% 的标签构成了中文 *subClassOf* 关系，20.80% 的分类与 10.19% 的标签构成了中文 *relate*关系，5.58% 的分类与 2.63% 的标签构成了中文 *equal*关系。高比例的分类与标签所构成的 *subClassOf* 关系实质上说明挖掘得到的中英文知识各自组成了一个大规模的层次分类体系。而较低比例的分类与标签构成了 *equal*关系，这从一方面说明 *equal*关系是本章所挖掘的最严格的语义关系，因此相同块中的概念之间的等价关系较少。

在现有的层次分类体系中，许多分类之间的父子关系也是正确的 *subClassOf* 关系，因此本章统计了这一类 *subClassOf* 关系的比例。如图 3.6(a)所示，对所有的英文 *subClassOf* 关系而言，已存在于现有层次分类体系中的比例为 36.62%，可以从现有层次分类体系中的路径推断得到的比例为 15.13%，而 48.25% 则为新的 *subClassOf* 关系。而从图 3.6(b)可以看出，对所有的中文 *subClassOf* 关系而言，已存在于现有层次分类体系中的比例为 22.71%，可以从现有层次分类体系中的路径推断得到的比例为 22.44%，而 52.85% 则为新的 *subClassOf* 关系。中英文较高比例的新的 *subClassOf* 关系也从侧面



(a) 英文关系　　(b) 中文关系

图 3.5　英文与中文各自的语义关系中分类与标签的比例



(a) 英文 *subClassOf* 关系　　(b) 中文 *subClassOf* 关系

图 3.6　英文与中文各自的 *subClassOf* 关系的分布

说明了所提出方法的有效性。

3.4.4 与其他知识图谱的对比

首先将通过所提出方法获得的中英文模式知识与开放链接数据中知名的多语言知识图谱（包括 DBpedia、Yago、BabelNet、Freebase）中的模式知识进行对比分析，重点考察概念与 *subClassOf* 关系的差异。如表 3.7与表 3.8所示，就概念数量而言，本章方法得到的英文模式知识中的概念数仅高于 Freebase，而所得到的中文模式知识中的概念数高于 DBpedia、Yago、Freebase。此外，本章方法得到的中英文模式知识中的概念与其他知识图谱中的概念的交集较小，这说明社交站点中的确存在较多新的概念，而这些新的概念也是对当前已有知识图谱的良好补充。关于 *subClassOf* 关系，本章方法得到的中英文模式知识均包含最大数量的 *subClassOf* 关系（是用于对比的其他知识图谱的两倍以上），而关于 *subClassOf* 关系的交集也较小，这也说明从社交站点挖掘得到的模式知识与现有知识图谱中的模式知识具有很强的互补性，如何将它们进行整合也是未来值得研究的问题。

关于所提出方法获得的中英文 *equal*关系的分析如下。在 25,474 个英文 *equal*关系中，14,849 个 *equal*关系所连接的两个概念拥有不同的字符串；而在 11,095 个中文

表 3.7 本章所得英文模式知识与其他知识图谱中的模式知识的比较

	本章的英文模式知识	DBpedia	Yago	BabelNet	Freebase
概念数量	242,303	1,213,462	408,467	1,042,196	133,075
重合概念数量	/	102,083	47,814	46,034	8,003
*subClassOf*关系数量	**1,047,801**	684	458,242	89,074	23,878
重合的 *subClassOf*关系数量	/	40	53,168	10,759	70

表 3.8 本章所得中文模式知识与其他知识图谱中的模式知识的比较

	本章的中文模式知识	DBpedia	Yago	BabelNet	Freebase
概念数量	399,853	142,411	48,621	463,485	2,035
重合概念数量	/	86,981	24,200	33,714	447
*subClassOf*关系数量	**947,645**	3	40,937	57,407	1,092
重合的 *subClassOf*关系数量	/	2	8,608	2,612	36

*equal*关系中，2,139 个 *equal*关系所连接的两个概念拥有不同的字符串。由于 BabelNet 是在当前的开放链接数据中规模最大的多语言同义词典，所以本章检测了所提出方法获得的中英文 *equal*关系被 BabelNet 覆盖的数量。在 BabelNet 中，每个概念由一个 Synset 表示，一个 Synset 由多个具有不同字符串与不同语言的同义词组成。此处不对连接两个拥有相同字符串的概念的 *equal*关系计数。在这种情况下，共有 10,160 个英文 *equal*关系与 1,013 个中文 *equal*关系已存在于 BabelNet 中。由于 *equal*关系的交集与上文提到的 *subClassOf*关系的交集均较小，所以又印证了本章获得的模式知识与 BabelNet 中的模式知识具有很强的互补性。此外，还利用 BabelNet 中的多语言同义词检测本章获得的中英文模式知识中的跨语言同义概念对（即由一个英文概念与一个中文概念构成，二者含义相同）的数量，共有 23,204 个跨语言同义概念对，仅占本章获得的模式知识中的中英文概念的较低比例，这充分说明面向社交站点的跨语言概念匹配任务具有较大的研究空间。

3.5 本章小结

本章主要介绍了一种新的各语言通用的半监督学习方法，该方法将规则嵌入到机器学习的过程中，且在不同的社交站点中的分类与标签中挖掘 *equal*、*subClassOf*、*relate*关系，即模式知识。在大规模中英文社交站点中进行实验，结果不仅表明了本章所提出方法、特征、规则的有效性，还体现了挖掘得到的中英文模式知识[5]的高正确率(大于 82%)。与开放链接数据中的大规模知识图谱 DBpedia、Yago、BabelNet、Freebase 相比，通过本章方法得到的中英文模式知识不仅拥有大规模的中英文概念，而且包含最大规模的中英文 *subClassOf*关系。

本章工作发表于 CCF B 类期刊 *Journal of Web Semantics*，论文题目为“On Building and Publishing Linked Open Schema from Social Web Sites”。

[5] http://los.linkingopenschema.info/

第四章　跨语言概念匹配

考虑到跨语言知识对齐是双语知识图谱构建的重点工作之一，且第 3章已介绍社交站点中存在大量的概念，所以如何匹配不同语言的概念成为了本章的重点研究工作。由于社交站点中的大量概念构成了不同种类的层次分类体系（如商品目录、导航站点目录等），所以本章专注于研究在不同语言的层次分类体系中进行跨语言概念匹配。本章的具体内容安排如下：章节 4.1概述了现有工作的问题、本章的解决方案、实验结果，章节 4.2介绍相关工作。之后针对现有工作的问题，章节 4.3 详细介绍所提出的一种各领域通用的基于双语主题模型跨语言概念匹配方法，并在章节 4.4中讨论、分析相应的实验结果。最后在章节 4.5中进行小结。

4.1 概述

不同语言的层次分类体系中的跨语言概念匹配指将语言为 s 的层次分类体系中的每个概念与其最相关的源自语言为 t 的层次分类体系中的概念建立映射关系，此处的“*最相关*”关系包含等价关系但不一定是等价关系，原因在于给定语言的层次分类体系中的每个概念并不一定在另一个语言的层次分类体系中存在其等价概念。跨语言概念匹配的关键在于如何度量不同语言概念之间的相似性，由于传统的单一语言相似度度量方式在跨语言场景中并不适用，所以已有一些研究 [31–33] 提出了相应的解决方案，但是依旧存在如下问题：

- 这些工作仅依赖翻译后的字符串相似度，但由于存在大量的翻译结果不准确与词汇失配的现象，即使不同语言的概念拥有相同的含义，翻译后的字符串往往也截然不同，易导致匹配失效。比如，京东商品目录中的概念“*户外/运动服*”在 eBay 商品目录中的最相关概念为“*Athletic Apparel*”，但通过 Google 翻译 [1] 获得“*户外/运动服*”对应的英文为“*Outdoor/Sportswear*”，显然其与“*Athletic Apparel*”各自对应的字符串完全不同。

- 这些工作均依赖于特定的领域信息，但是许多跨领域的层次分类体系中的概念并不拥有这样特定的信息。比如，文献 [32, 33] 利用图书实例辅助图书概念匹配，然而在许多层次分类体系中，并不存在实例信息，故这些方法并不通用。

为了解决上述问题，本章提出了一种新的各领域通用的面向不同语言层次分类体系的跨语言概念匹配方法。该方法不仅利用了字符串相似度的优势，还将不同语言的概

[1] http://translate.google.com/

念转换成位于同一空间的向量表示进而使用基于向量的相似度，即综合考虑字符串相似度与向量相似度以完成跨语言概念匹配的任务。所提出方法首先利用一种跨语言字符串相似度为给定语言为 s 的层次分类体系中的每个概念识别其在另一种语言 t 的层次分类体系中的候选匹配概念。然后提出一种利用 Google 搜索引擎与 Google 翻译获取每个概念的双语文本上下文的机制。之后提出两种新的主题模型：双语双词主题模型（Bilingual Biterm Topic Model，BiBTM）与基于概念关联关系的双语双词主题模型（Concept Corrleation based Bilingual Biterm Topic Model，CC-BiBTM），二者均可在语言为 s 与 t 的两个层次分类体系中所有概念对应的双语文本上下文中训练得到每个概念的主题向量表示。最后通过不同语言概念之间的向量余弦相似度得到匹配结果。

实验结果表明本章所提出的方法在测试数据集上的查准率@1 与 MRR 均优于其他方法，其中当使用考虑了多种概念间关联关系的 CC-BiBTM 时，匹配效果最佳。此外，通过控制阈值可发现高质量的不同数据集中的不同语言的概念间的等价关系。

4.2 相关工作

本节主要从三个方面介绍本章内容的相关工作，分别是模式（Schema）匹配、多语言知识对齐、主题模型。

4.2.1 模式匹配

模式匹配旨在识别给定两个模式之间的语义对应关系，此处的模式包括数据库模式与本体 [104]。文献 [105–107] 介绍了不同的数据库模式匹配的方法。然而，由于规模与结构的不同，数据库模式匹配与异构的层次分类体系匹配存在较大的差异。本体匹配指发现源自不同本体的同类元素（包括概念、属性等）之间的关系，如等价关系、上下位关系等。目前已存在大量的匹配不同类型本体的工作 [56]。一些方法或系统 [60, 61, 108] 可用于跨语言本体匹配，它们均依赖翻译后的字符串相似度，但由于词汇失配与翻译结果不准确的问题，匹配的效果常常不令人满意。与本体不同，层次分类体系通常没有严格的逻辑结构以及形式化定义的属性、实例、公理以帮助解决概念匹配的问题。所以一些工作 [31–33] 专门设计了针对不同语言的层次分类体系中的跨语言概念匹配的方法。这些工作专注于特定领域的不同语言的层次分类体系中的跨语言概念匹配，主要依赖于翻译后的字符串相似度与特定的领域信息。而本章工作不局限于翻译后的字符串相似度，并尝试利用概念的向量表示间的相似度以解决跨语言概念匹配的任务。此外，本章工作可用于跨领域跨语言的层次分类体系中的跨语言概念匹配，且不依赖于任何特定的领域信息。

4.2.2 多语言知识对齐

多语言知识对齐是另一类与本章相关的工作，目前已存在一些研究，比如文献 [30, 68, 69]。Wang 等人 [68] 提出了一个链接因子图模型将英文维基百科的文章与中文百度百科的文章进行链接。在此研究的基础上，他们进一步提出了一个概念标注方法与

基于回归的学习模型以迭代预测新的跨语言文章链接 [69]。这两个研究工作是著名中英双语知识图谱 XLORE [21] 的构建技术中的重要组成部分。Zhang 等人 [30] 提出了一种语义标注系统，该系统利用维基百科与开放链接数据中的多语言资源对不同语言的文档网页进行标注，其实质即为跨语言实体链接。上述研究与本章工作有极大的区别，上述研究专注于实例层面的多语言知识对齐，而本章工作则是面向模式层面的跨语言概念对齐。

4.2.3 主题模型

主题模型如概率化潜在语义分析 [109]（Probabilistic Latent Semantic Analysis, PLSA）与隐含狄利克雷分布 [110]（Latent Dirichlet Allocation, LDA）以及基于这二者的众多变种模型是长期被研究的用于分析文本隐含语义的生成模型。与本章工作相关的一种主题模型是双语隐含狄利克雷分布 [111]（Bilingual Latent Dirichlet Allocation, BiLDA），它针对成对的双语文档进行建模，每对双语文档由具有相同内容但不同语言的两个文档构成，如由维基百科跨语言链接关联的不同语言的文章或一种语言的文档与其对应的翻译后的另一种语言的文档。由于 BiLDA 在针对短文档建模时易受数据稀疏问题 [112] 的影响，从而导致主题建模效果不佳，所以该模型在本章场景中并不适用，这也是本章提出 BiBTM 模型的动机之一。另一类相关的主题模型是基于元数据的主题模型。为了同时对文本及其对应的元数据（比如：作者与标签）进行建模，一系列的基于元数据的主题模型被提出，如作者主题模型 [113]（Author Topic Model）、已标注的 LDA [114]（Labeled LDA）、带权重的标签主题模型 [115]Tag-Weighted Topic Model）、带权重的标签狄利克雷分布 [116]（Tag-Weighted Dirichlet Allocation）等。这些研究将一个元数据表示为主题或词的混合物，但无法应用于双语环境下的文本挖掘。本章所提出的 CC-BiBTM 是第一份关于双语元数据主题模型的工作，并且在跨语言概念匹配中体现了其价值。

4.3 方法设计

本节将详细介绍所提出的在不同语言的层次分类体系中进行跨语言概念匹配的方法。如图 4.1所示，输入为两个不同语言的层次分类体系，输出为对齐后的层次分类体系，而方法本身共包含三个组件，分别是候选匹配概念识别、双语文本上下文抽取、精确匹配。候选匹配概念识别利用一种跨语言字符串相似度为一种语言的层次分类体系中的每个概念识别其在另一种语言的层次分类体系中的候选匹配概念，此处输出为候选匹配概念对。与此同时，双语本文上下文抽取基于 Google 搜索引擎与 Google 翻译为两个不同语言的层次分类体系中的每个概念获取双语本文上下文。然后在精确匹配中，于所有获得的双语文本上下文上训练双语主题模型以得到每个概念对应的向量表示，最后计算候选匹配的两个概念之间的向量相似度从而得到匹配结果，即对齐后的不同语言的层次分类体系。

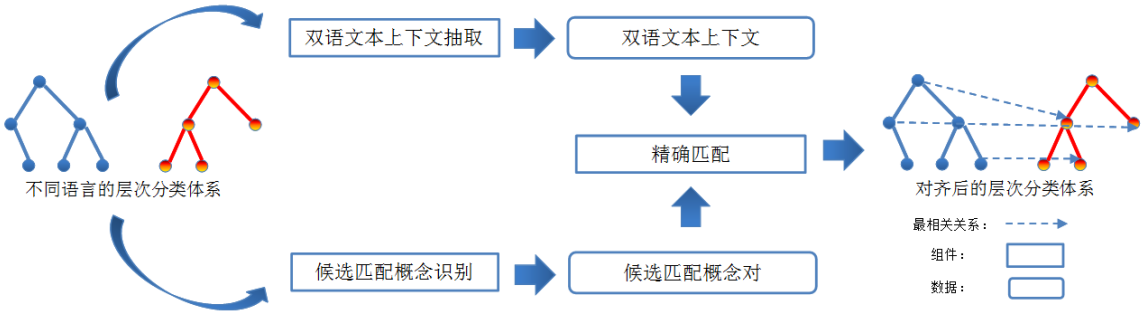


图 4.1 跨语言概念匹配方法的流程示意图

4.3.1 候选匹配概念识别

为了避免源自两个不同层次分类体系的概念之间的不必要的比较，本章所提出方法首先试图识别所有可能的候选匹配概念对。该步骤的输出作为精确匹配的输入之一。

最简单的表示一个概念的方法即使用其本身的字符串进行表示，但是该表示方法往往对于概念匹配任务效果不佳，其原因在于具有相同含义的概念可能拥有完全不通的字符串表示。例如“*Sports Clothing*”与“*Athletic Apparel*”之间便不存在任何相同的词，而该现象在不同语言的同义概念中更为普遍。此外，由于不同的语言习惯与不恰当的翻译，所以直接比较翻译后的概念字符串（相同语言环境）也存在较大的局限性。

为了解决上述问题，利用 BabelNet 计算一种词级别的概念间的跨语言字符串相似度，其关键思想在于：“*不同语言的两个概念可能相关当且仅当二者拥有相同或同义的词*”。给定语言为 s 的层次分类体系中的一个概念 c^s 与语言为 t 的层次分类体系中的一个概念 c^t，经过分词，每个概念均各自对应一组词。在去除停用词之后，c^s 与 c^t 各自对应一个词集合 $W_{c^s}=\{w_i^s\}_{i=1}^m$ 与 $W_{c^t}=\{w_j^t\}_{j=1}^n$。针对每个 w_i^s，之后通过 BabelNet 获取其对应的语言为 s 与 t 的同义词，并将它们加入到 W_{c^s} 中，而针对 w_j^t 也进行同样的操作并获得更新后的 W_{c^t}。此处定义 c^s 与 c^t 之间的跨语言字符串相似度（Cross-Lingual String Similarity, CSS）如下：

$$CSS(c^s,c^t)=\begin{cases}1, & \text{if } W_{c^s}\cap W_{c^t}\neq\emptyset\\ 0, & \text{otherwise}\end{cases} \tag{4.1}$$

如果 $CSS(c^s,c^t)$ 等于 1，那么 c^t 即可作为 c^s 的一个候选匹配概念。图 4.2给出了一个关于候选匹配概念识别的示例，其中说明了为什么 eBay 商品目录中的“*Athletic Apparel*”会是京东商品目录中的“*户外/运动服*”的一个候选匹配概念。

4.3.2 双语文本上下文抽取

概念并没有用于描述其自身的文本信息，而为了获得每个概念对应的向量表示并用于跨语言概念匹配，本节提出了一种基于 Google 搜索引擎与 Google 翻译的双语文本上下文抽取机制，抽取所得的双语文本上下文用于训练每个概念对应的向量表示。

万维网网页可能存在与给定概念相关联的文本信息，但手工寻找并不实际，所以选择利用 Google 搜索引擎对万维网进行查询从而获取每个概念对应的文本上下文。在

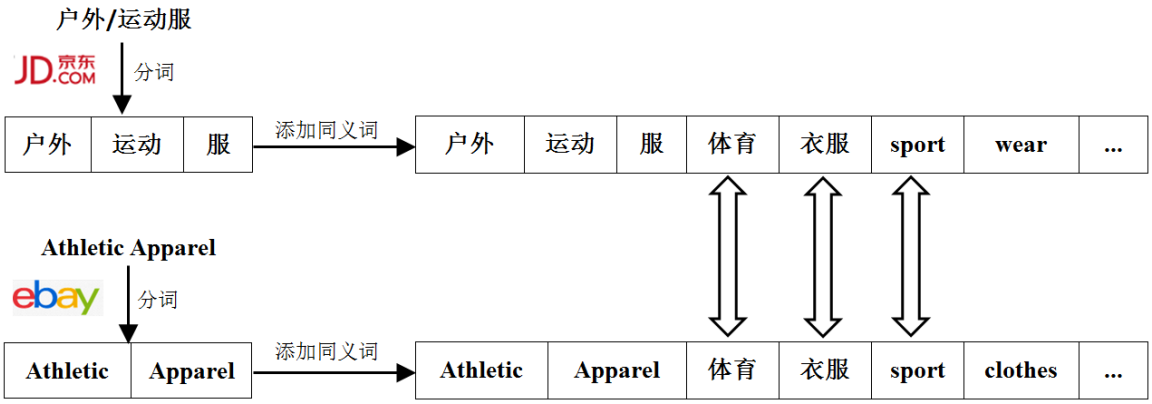


图 4.2 示例：候选匹配概念识别

不同的结构中，拥有相同字符串的概念的含义可能不同。举例而言，概念“*Sports*”在雅虎导航目录中共出现两次，其一是概念“*Shopping and Services*”的子概念，此处“*Sports*”的含义为体育用品；其二为概念“*Recreation*”的子概念，指体育活动。然而，若仅将上述两个概念的字符串（即 Sports）提交 Google，则返回结果中的标题、片段（Snippets）、网址等均相同，即无法对两个不同含义的概念进行区分。所以，为了通过 Google 正确获得与每个概念相关的文本上下文，选择将每个概念 c 与其父概念 pc 各自对应的字符串共同作为查询提交给 Google。比如，在雅虎导航目录中，将表示体育用品的概念“*Sports*”与其父概念“*Shopping and Services*”各自对应的字符串共同提交给 Google 进行查询。在每个返回的网页片段中，抽取除 pc 外与概念 c 在同一句子中共现的词语，这些词语即作为给定概念的文本上下文。由于 pc 是查询的一部分，所以返回的网页片段中必然大量出现 pc，所以 pc 不纳入文本上下文抽取的考虑。需要注意的是位于层次分类体系根节点的概念并没有父概念，但是它们通常没有歧义，否则用户难以以自上而下的方式浏览整个层次分类体系。因此选择仅将此类概念的字符串单独提交 Google 进行查询。

每个概念对应的文本上下文实际上抽取自 Google 搜索引擎返回的一组网页片段，可视为一个短文档（Short Documents）的集合。给定一个语言为 s 的短文档 d^s，可通过 Google 翻译获得语言为 t 的短文档 d^t，从而可构建一对双语短文档 (d^s, d^t)。同理针对语言为 t 的短文档 d^t，可通过 Google 翻译获得语言为 s 的短文档 d^s，并构建一对双语短文档。综上可知，每个概念对应的双语文本上下文可由多对双语短文档构成的集合 $\{(d_i^s, d_i^t)\}_{i=1}^{N_d}$ 表示，其中 N_d 为集合中元素的数量。

4.3.3 精确匹配

在候选匹配概念识别之后，旨在使用基于向量的相似度完成给定概念与其每个候选匹配概念之间精确匹配。基于每个概念对应的双语文本上下文的数据特点，首先提出双语双词主题模型（Bilingual Biterm Topic Model，BiBTM）以学习每个概念的向量表示，由此即可计算概念间的向量相似度。然而，考虑到存在多种不同概念关联关系，故将概念关联关系纳入双语双词主题模型建模的过程中，即提出了基于概念关联关系的双

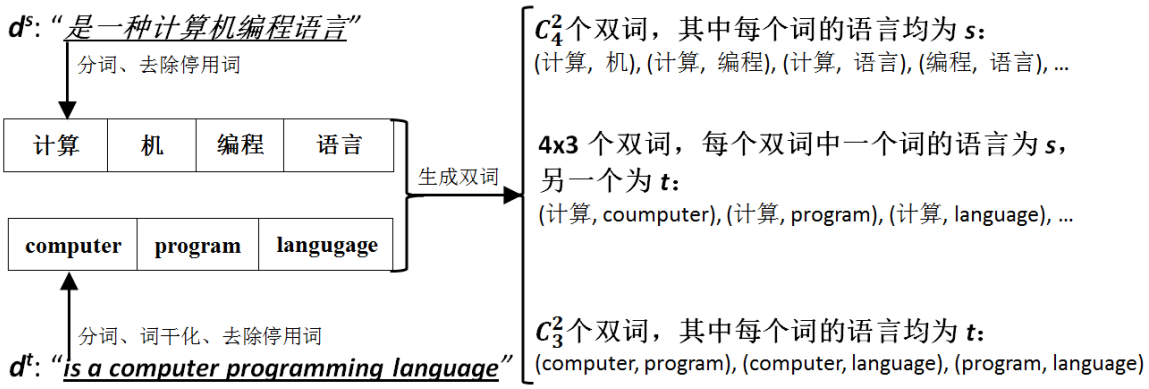


图 4.3 示例：BiBTM 中的双词生成

语双词主题模型（Concept Corrleation based Bilingual Biterm Topic Model，CC-BiBTM），从而获得对于跨语言概念匹配任务更加有效的概念向量表示，并计算概念间的向量相似度以完成精确匹配。

4.3.3.1 双语双词主题模型

传统对文本进行建模的方法为 TF-IDF 模型[93]。然而，由于概念对应的文本上下文抽取自万维网网页片段，所以不同措辞风格的描述均可能出现，如一条随意的推特或一则使用正式场合措辞的新闻，这意味着不同片段中抽取得到的词可能完全不同，即拥有较低的词频，因此 TF-IDF 模型在该场景下可能并不适用。为解决此问题，选择双语主题模型挖掘每个概念对应的双语文本上下文中的主题，故理论上可直接使用双语隐含狄利克雷分布[111]（Bilingual Latent Dirichlet Allocation, BiLDA），但由于该模型在短文档中面临数据稀疏问题[112]时的主题建模效果不佳，所以 BiLDA 在本章场景依旧适用程度不高。基于此，本节提出一种新的适用于上述所得双语文本上下文的数据特点的双语主题模型，即双语双词主题模型 BiBTM，该模型直接对每对双语短文档中的词共现进行显式建模以克服数据稀疏问题。

BiBTM 是用于对双词生成建模的双词主题模型[117, 118]（Biterm Topic Model, BTM）的扩展。BTM 的核心思想是如果两个词共现越频繁，则它们属于同一主题的可能性越高。与 BTM 不同，BiBTM 中的双词（Biterm）指在一对双语文档中共现的无序词对。一对双语文档中任意两个词均可构成一个双词。比如，给定一对双语文档 (d^s, d^t)，其中 d^s 由 n 个语言为 s 的不同的词构成，而 d^t 由 m 个语言为 t 的不同的词构成，因此将生成共 $C_n^2 + C_m^2 + m \times n$ 个双词，其中 C_n^2 与 C_m^2 均为组合数。图 4.3给出了一个关于 BiBTM 中双词生成的示例。BiBTM 假设从整个语料库中抽取得到的双词共享一个主题分布，而每个主题由两个不同语言的词上的离散分布分别构成。

模型描述. 给定一个由多对双语文档构成的语料库，假设其包含 $|\mathbf{B}|$ 个双词，$\mathbf{B} = \mathbf{B}^s \cup \mathbf{B}^{st} \cup \mathbf{B}^t = \{b_i^s\}_{i=1}^{|\mathbf{B}^s|} \cup \{b_i^{st}\}_{i=1}^{|\mathbf{B}^{st}|} \cup \{b_i^t\}_{i=1}^{|\mathbf{B}^t|}$，其中双词 $b_i^s = (w_{i,1}^s, w_{i,2}^s)$ 且 b_i^s 中每个词的语言均为 s，双词 $b_i^{st} = (w_{i,1}^s, w_{i,2}^t)$ 且 b_i^{st} 中包含两个不同语言的词，双词 $b_i^t = (w_{i,1}^t, w_{i,2}^t)$ 且 b_i^t 中每个词的语言均为 t。此外，还假设该语料库包含 K 个主题，这些主题可由 W^s 个不同的语言为 s 的词表达或 W^t 个不同的语言为 t 的词表达。针对

上述三种类型的双词，主题指示变量 $z \in [1, K]$ 可分别表示成 z^s、z^{st}、z^t。语料库中的主题可由一个 K 维的多项式分布（Multinomial Distribution）$\boldsymbol{\theta} = \{\theta_k\}_{k=1}^K$ 表示，其中 $\theta_k = P(z = k)$。语言为 s 的词分布由一 $K \times W^s$ 的矩阵 $\boldsymbol{\varphi}^s$ 表示，其中第 k 行 $\boldsymbol{\varphi}_k^s$ 指一 W^s 维的多项式分布且该行中每一项可表示为 $\varphi_{k,w^s}^s = P(w^s|z = k)$。类似地，语言为 t 的词分布则由一 $K \times W^t$ 的矩阵 $\boldsymbol{\varphi}^t$ 表示，其中第 k 行 $\boldsymbol{\varphi}_k^t$ 指一 W^t 维的多项式分布且该行中每一项可表示为 $\varphi_{k,w^t}^t = P(w^t|z = k)$。

与 BTM 类似，BiBTM 也将超参数 α 与 β 设置为对称的狄利克雷先验（Dirichlet Priors）。图 4.4 给出了 BiBTM 的图表示，而算法 1相应地给出了 BiBTM 的生成过程。在算法 1中，首先初始化主题数量 K 与狄利克雷先验 α 及 β 的值（*第 1-2 行*）。然后针对每个主题 k，依据狄利克雷分布（Dirichlet Distribution）分别采样语言为 s 与 t 的主题 - 词分布 $\boldsymbol{\varphi}_k^s$ 与 $\boldsymbol{\varphi}_k^t$，它们同时也是多项式分布（*第 3-4 行*）。之后同样依据狄利克雷分布，针对语料库中所有双词采样一个全局主题分布 $\boldsymbol{\theta}$，其同时也是一个多项式分布（*第 5 行*）。最后针对不同类型的双词，依次先依据全局主题分布 $\boldsymbol{\theta}$ 采样生成一个主题，再依据生成的主题采样生成各类双词（*第 6-14 行*）。当使用 BiBTM 时，在给定超参数 α 与 β 的情况下，生成整个语料的概率计算方式如下：

$$
\begin{aligned}
P(\mathbf{B}|\alpha, \beta) = & \prod_{i=1}^{|\mathbf{B}^s|} \int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^s}^s \varphi_{k,w_{i,2}^s}^s \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^s \\
& \times \prod_{i=1}^{|\mathbf{B}^{st}|} \int\int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^s}^s \varphi_{k,w_{i,2}^t}^t \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^s\mathrm{d}\boldsymbol{\varphi}^t \\
& \times \prod_{i=1}^{|\mathbf{B}^t|} \int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^t}^t \varphi_{k,w_{i,2}^t}^t \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^t
\end{aligned}
\tag{4.2}
$$

参数估计. 在公式 4.2中，由于 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}^s$、$\boldsymbol{\varphi}^t$ 存在耦合关系，所以无法直接使用最大似然估计精确求解这三个参数。此处选择与 BTM 相同的解决方案，即使用 collapsed Gibbs 采样算法 [119] 近似求解 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}^s$、$\boldsymbol{\varphi}^t$。由于狄利克雷分布与多项式分布是共轭分布，所以参数 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}^s$、$\boldsymbol{\varphi}^t$ 可通过积分积掉。基于此，仅需要对每个双词的主题进行采样即可。对于 $b_i^s \in \mathbf{B}^s$、$b_i^{st} \in \mathbf{B}^{st}$、$b_i^t \in \mathbf{B}^t$ 而言，其各自对应的 Gibbs 采样公式（具体推导见附录 A.1）如下：

$$
P(z_i^s = k|z_{\neg b_i^s}, \mathbf{B}) \propto (n_{\neg b_i^s,k} + \alpha) \cdot \frac{(n_{\neg b_i^s,w_{i,1}^s|k} + \beta)(n_{\neg b_i^s,w_{i,2}^s|k} + \beta)}{(n_{\neg b_i^s,\cdot^s|k} + 1 + W^s\beta)(n_{\neg b_i^s,\cdot^s|k} + W^s\beta)} \tag{4.3}
$$

$$
P(z_i^{st} = k|z_{\neg b_i^{st}}, \mathbf{B}) \propto (n_{\neg b_i^{st},k} + \alpha) \cdot \frac{(n_{\neg b_i^{st},w_{i,1}^s|k} + \beta)(n_{\neg b_i^{st},w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^{st},\cdot^s|k} + W^s\beta)(n_{\neg b_i^{st},\cdot^t|k} + W^t\beta)} \tag{4.4}
$$

$$
P(z_i^t = k|z_{\neg b_i^t}, \mathbf{B}) \propto (n_{\neg b_i^t,k} + \alpha) \cdot \frac{(n_{\neg b_i^t,w_{i,1}^t|k} + \beta)(n_{\neg b_i^t,w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^t,\cdot^t|k} + 1 + W^t\beta)(n_{\neg b_i^t,\cdot^t|k} + W^t\beta)} \tag{4.5}
$$

其中，$z_{\neg b}$ 表示除双词 b 以外其他所有双词的主题赋值，$n_{\neg b,k}$ 表示除双词 b 以外属于主题 k 的双词个数，$n_{\neg b,w^s|k}$ 表示除双词 b 以外语言为 s 的词 w^s 被赋予主题 k 的次数，

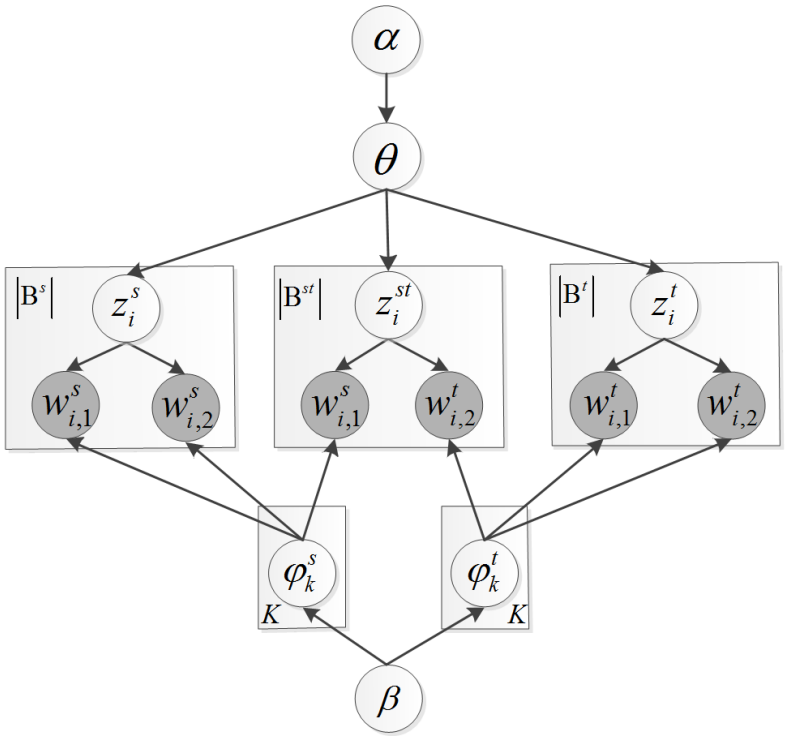


图 4.4 BiBTM 的图表示

算法 1: BiBTM 的生成过程

initialize: (1) 设置主题数量为 K;
(2) 设置狄利克雷先验 α 与 β 的值;
foreach 主题 $k \in [1, K]$ **do**
  sample: $\boldsymbol{\varphi}_k^s$, $\boldsymbol{\varphi}_k^t \sim Dirichlet(\beta)$;
sample: $\boldsymbol{\theta} \sim Dirichlet(\alpha)$;
foreach 双词 $b_i^s \in \mathbf{B}^s$ **do**
  sample: $z_i^s \sim Multinomial(\boldsymbol{\theta})$,
  sample: $w_{i,1}^s, w_{i,2}^s \sim Multinomial(\boldsymbol{\varphi}_{z_i^s}^s)$;
foreach 双词 $b_i^{st} \in \mathbf{B}^{st}$ **do**
  sample: $z_i^{st} \sim Multinomial(\boldsymbol{\theta})$,
  sample: $w_{i,1}^s \sim Multinomial(\boldsymbol{\varphi}_{z_i^{st}}^s)$, $w_{i,2}^t \sim Multinomial(\boldsymbol{\varphi}_{z_i^{st}}^t)$;
foreach 双词 $b_i^t \in \mathbf{B}^t$ **do**
  sample: $z_i^t \sim Multinomial(\boldsymbol{\theta})$,
  sample: $w_{i,1}^t, w_{i,2}^t \sim Multinomial(\boldsymbol{\varphi}_{z_i^t}^t)$;

$n_{\neg b,\cdot^s|k} = \sum_{w^s} n_{\neg b,w^s|k}$，而 $n_{\neg b,w^t|k}$ 表示除双词 b 以外语言为 t 的词 w^t 被赋予主题 k 的次数，$n_{\neg b,\cdot^t|k} = \sum_{w^t} n_{\neg b,w^t|k}$。

经过足够次数的迭代后，即可对全局主题分布 $\boldsymbol{\theta}$ 与主题 -词分布 $\boldsymbol{\varphi}_k^s$、$\boldsymbol{\varphi}_k^t$ 进行估计

（具体推导见附录 A.2）：

$$\theta_k = \frac{\alpha + n_k}{K\alpha + |\mathbf{B}|} \tag{4.6}$$

$$\varphi^s_{k,w^s} = \frac{\beta + n_{w^s|k}}{W^s\beta + n_{\cdot^s|k}} \tag{4.7}$$

$$\varphi^t_{k,w^t} = \frac{\beta + n_{w^t|k}}{W^t\beta + n_{\cdot^t|k}} \tag{4.8}$$

其中，n_k 指属于主题 k 的双词个数，$n_{w^s|k}$ 表示语言为 s 的词 w^s 被赋予主题 k 的次数，$n_{\cdot^s|k} = \sum_{w^s} n_{w^s|k}$，$n_{w^t|k}$ 表示语言为 t 的词 w^t 被赋予主题 k 的次数，$n_{\cdot^t|k} = \sum_{w^t} n_{w^t|k}$。

上下文主题推断. 为了完成精确匹配的任务，需计算每个概念对应的双语文本上下文的主题分布。给定一个概念 c，假设其包含 N_c 个双词 $\{b_j\}_{j=1}^{N_c}$，这些双词均是从 c 对应的双语文本上下文中抽取得到，此处使用如下公式推断 c 对应的双语文本上下文的主题分布：

$$P(z|c) = \sum_{j=1}^{N_c} P(z = k|b_j)P(b_j|c) \tag{4.9}$$

在公式 4.9中，$P(b_j|c)$ 可通过经验分布估计：

$$P(b_j|c) = \frac{n(b_j)}{\sum\limits_{j=1}^{N_c} n(b_j)} \tag{4.10}$$

其中 $n(b_j)$ 表示双词 b_j 在 c 对应的双语文本上下文中的出现次数。与此同时，$P(z = k|b_j)$ 可基于 BiBTM 中学习得到的参数并通过贝叶斯公式进行计算：

$$P(z = k|b_j) = \begin{cases} \dfrac{\theta_k \cdot \varphi^s_{k,w^s_{j,1}} \cdot \varphi^s_{k,w^s_{j,2}}}{\sum\limits_{k'=1}^{K} \theta_{k'} \cdot \varphi^s_{k',w^s_{j,1}} \cdot \varphi^s_{k',w^s_{j,2}}}, & \text{if } b_j \in \mathbf{B}^s \\ \dfrac{\theta_k \cdot \varphi^s_{k,w^s_{j,1}} \cdot \varphi^t_{k,w^t_{j,2}}}{\sum\limits_{k'=1}^{K} \theta_{k'} \cdot \varphi^s_{k',w^s_{j,1}} \cdot \varphi^t_{k',w^t_{j,2}}}, & \text{if } b_j \in \mathbf{B}^{st} \\ \dfrac{\theta_k \cdot \varphi^t_{k,w^t_{j,1}} \cdot \varphi^t_{k,w^t_{j,2}}}{\sum\limits_{k'=1}^{K} \theta_{k'} \cdot \varphi^t_{k',w^t_{j,1}} \cdot \varphi^t_{k',w^t_{j,2}}}, & \text{if } b_j \in \mathbf{B}^t \end{cases} \tag{4.11}$$

在获得每个概念对应的双语文本上下文的主题分布后，即可将不同语言的概念在同一主题向量空间进行表示。由此，可利用向量之间的余弦相似度度量给定的一种语言的层次分类体系中的每个概念与其对应的另一种语言的层次分类体系中的候选匹配概念之间的相似性。

4.3.3.2 基于概念关联关系的双语双词主题模型

虽然利用 BiBTM 即可学习得到每个概念的向量表示，但是 BiBTM 忽略了显式的概念关联关系。第一种概念关联关系是共现关联关系（Co-occurrence Correlation），它存在于概念与其共现的文本中的词之间。一些研究 [113, 115] 已经表明同时对文本与共现的元数据（如作者、标签等）建模可以学习得到对许多应用有价值的高质量的主题向量。另外一种重要的概念关联关系是结构化关联关系（Structural Correlation），它表示层次分类体系中具有祖先 -子孙关系的概念间的联系。在不同语言层次分类体系中的跨语言概念匹配任务中，结构化关联关系的作用是显而易见的，即如果两个源自不同层次分类体系中的概念拥有相似的祖先概念或子孙概念，那么这两个给定概念之间可能具有很高的相关性。因此，如果直接忽略上述两种概念关联关系，而仅仅基于 BiBTM 学习得到的概念的主题向量计算概念间的相似度并不足以很好地完成跨语言概念匹配的任务。

章节 4.3.2中介绍了关于每个概念的双语文本上下文抽取方法，它是首先通过将给定概念作为查询提交到 Google 搜索引擎以获得相应语言的文本上下文（多个网页片段（即短文档）），然后基于 Google 翻译构建每个概念对应的多对双语短文档。换句话说，每对双语短文档至少包含给定的概念（可能更多概念），即每个从双语短文档抽取得到的双词至少对应一个概念。为了获取共现关联关系，将每个双词表示成概念的混合体，即每个双词应对应一个先验概念分布。考虑到概念间的结构化关联关系，则选择利用层次分类体系中的信息内容 [120]（Information Content）与路径长度进一步改善每个双词对应的先验概念分布。在将共现关联关系与结构化关联关系编码进每个双词对应的先验概念分布中之后，则可将这些概念关联关系融合到双语双词主题建模的过程中。由于需要利用每个概念的低维主题向量表示辅助跨语言概念匹配，所以通过假设每个概念对应一个主题概率分布以建立显式的概念与隐式的主题之间联系，即每个概念可以表示为主题的混合体。此外，与 BiBTM 相同的是每个主题可表示成给定语言的词的混合体。综上，即可获得一个新的基于概念关联关系的双语双词主题模型（Concept Corrleation based Bilingual Biterm Topic Model，CC-BiBTM），该模型的详细细节将在下文介绍。

模型描述. 给定一个语料库 $\mathbb{O}$，假设其拥有 $|\mathbf{B}|$ 个双词与 C 个显式的源自两个不同语言层次分类体系的概念。$\mathbf{B} = \mathbf{B}^s \cup \mathbf{B}^{st} \cup \mathbf{B}^t = \{b_i^s\}_{i=1}^{|\mathbf{B}^s|} \cup \{b_i^{st}\}_{i=1}^{|\mathbf{B}^{st}|} \cup \{b_i^t\}_{i=1}^{|\mathbf{B}^t|}$，其中双词 $b_i^s = (w_{i,1}^s, w_{i,2}^s)$ 且 b_i^s 中每个词的语言均为 s，双词 $b_i^{st} = (w_{i,1}^s, w_{i,2}^t)$ 且 b_i^{st} 中包含两个不同语言的词，双词 $b_i^t = (w_{i,1}^t, w_{i,2}^t)$ 且 b_i^t 中每个词的语言均为 t。由于已定义每个双词是概念的混合体，所以一个双词 b_i 可以表示成一个 C 维的多项式分布 $\boldsymbol{\pi}_i = \{\pi_{i,c}\}_{c=1}^{C}$，而 $\boldsymbol{\pi}_i$ 可表示成 $\boldsymbol{\pi}_i^s$ 或 $\boldsymbol{\pi}_i^{st}$ 或 $\boldsymbol{\pi}_i^t$ 以区分三种不同类型的双词。与此同时，$\boldsymbol{\pi}_i$ 即为 CC-BiBTM 中每个双词 b_i 的先验概念分布。令 $x \in [1, C]$ 为概念指示变量，针对仅由语言为 s 的词构成的双词，由语言分别为 s 与 t 的两个词构成的双词，以及仅由语言为 t 的词构成的双词，x 可分别表示成 x^s、x^{st}、x^t。类似地，主题指示变量 $z \in [1, K]$ 可以表示为 z^s 或 z^{st} 或 z^t。然后，定义每个概念由 K 隐式的主题表达，而每个主题可由 W^s 个不同的语言为 s 的词表达或 W^t 个不同的语言为 t 的词表达。此处使用一个 K 维的多项式分布 $\boldsymbol{\theta}_c = \{\theta_{c,k}\}_{k=1}^{K}$ 表示给定概念 c 的主题分布。主题 k 对应的语言为 s 的词分布则由一

个 W^s 维的多项式分布 $\boldsymbol{\varphi}_k^s$ 表示，其中每一项可表示为 $\varphi_{k,w^s}^s = P(w^s|z=k)$；而主题 k 对应的语言为 t 的词分布则由一个 W^t 维的多项式分布 $\boldsymbol{\varphi}_k^t$ 表示，其中每一项可表示为 $\varphi_{k,w^t}^t = P(w^t|z=k)$。与 BiBTM 相同，超参数 α 与 β 均设置为对称的狄利克雷先验。

图 4.5给出了 CC-BiBTM 的图表示，而算法 2则给出了相应的 CC-BiBTM 的生成过程。在算法 2中，首先初始化主题数量 K 与狄利克雷先验 α 及 β 的值（*第 1-2 行*）。然后针对每个主题 k，依据狄利克雷先验分布分别采样语言为 s 与 t 的主题 -词分布 $\boldsymbol{\varphi}_k^s$ 与 $\boldsymbol{\varphi}_k^t$（*第 3-4 行*），它们同时也是多项式分布。之后针对每个概念 c，同样依据狄利克雷先验分布采样一个概念 -主题分布 $\boldsymbol{\theta}_c$（*第 5-6 行*），其同时也是一个多项式分布。最后针对不同类型的双词，依次先计算给定双词的先验概念分布，再依据此分布采样一个概念，然后依据概念 -主题分布 $\boldsymbol{\theta}_c$ 采样生成一个主题，最后依据此主题采样生成各类双词（*第 7-18 行*）。

先验概念分布计算. 本节主要介绍如何利用不同类型的概念关联关系计算每个双词的先验概念分布。首先定义每个双词对应的概念为共现概念（Co-Occurring Concepts），而每个双词的先验概念分布则反映了双词的共现概念与双词中的每个词的共现关系。然后假设给定双词的每个共现概念被采样的概率相同。最后在给定双词 b_i 与其对应的共现概念集合 $CC(b_i)$ 的情况下，针对每个概念 $c \in [1, C]$ 给出关于 b_i 的基于共现关联关系的先验概念概率 $\pi_{i,c}^{CC}$ 的计算方式如下：

$$\pi_{i,c}^{CC} = \begin{cases} \dfrac{1}{|CC(b_i)|}, & \text{if } c \in CC(b_i) \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

其中 $|CC(b_i)|$ 是 $CC(b_i)$ 中概念的数量。

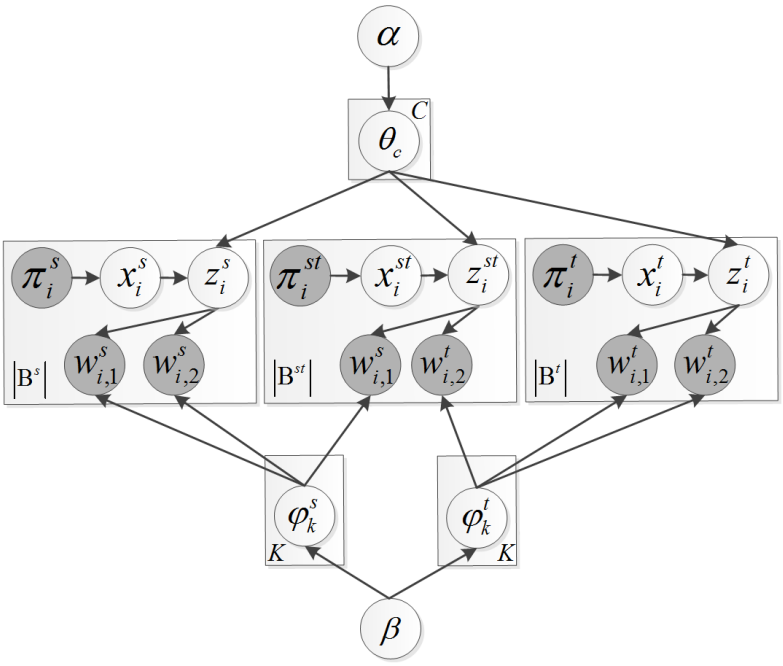


图 4.5 CC-BiBTM 的图表示

算法 2: CC-BiBTM 的生成过程

initialize: (1) 设置主题数量为 K;
(2) 设置狄利克雷先验 α 与 β 的值;
foreach 主题 $k \in [1, K]$ **do**
  sample: $\boldsymbol{\varphi}_k^s$, $\boldsymbol{\varphi}_k^t \sim Dirichlet(\beta)$;
foreach 概念 $c \in [1, C]$ **do**
  sample: $\boldsymbol{\theta}_c \sim Dirichlet(\alpha)$;
foreach 双词 $b_i^s \in \mathbf{B}^s$ **do**
  给定先验概念分布 $\boldsymbol{\pi}_i^s$,
  sample: $x_i^s \sim Multinomial(\boldsymbol{\pi}_i^s)$, $z_i^s \sim Multinomial(\boldsymbol{\theta}_{x_i^s})$,
  sample: $w_{i,1}^s$, $w_{i,2}^s \sim Multinomial(\boldsymbol{\varphi}_{z_i^s}^s)$;
foreach 双词 $b_i^{st} \in \mathbf{B}^{st}$ **do**
  给定先验概念分布 $\boldsymbol{\pi}_i^{st}$,
  sample: $x_i^{st} \sim Multinomial(\boldsymbol{\pi}_i^{st})$, $z_i^{st} \sim Multinomial(\boldsymbol{\theta}_{x_i^{st}})$,
  sample: $w_{i,1}^s \sim Multinomial(\boldsymbol{\varphi}_{z_i^{st}}^s)$, $w_{i,2}^t \sim Multinomial(\boldsymbol{\varphi}_{z_i^{st}}^t)$;
foreach 双词 $b_i^t \in \mathbf{B}^t$ **do**
  给定先验概念分布 $\boldsymbol{\pi}_i^t$,
  sample: $x_i^t \sim Multinomial(\boldsymbol{\pi}_i^t)$, $z_i^t \sim Multinomial(\boldsymbol{\theta}_{x_i^t})$,
  sample: $w_{i,1}^t$, $w_{i,2}^t \sim Multinomial(\boldsymbol{\varphi}_{z_i^t}^t)$;

除了概念与词之间的共现关联关系，以下还将介绍在层次分类体系中具有祖先 -子孙关系的概念之间的两种不同的结构化关联关系。第一种是基于信息内容 [120] 的结构化关联关系。直观而言，由于每个双词的所有共现概念在一个层次分类体系中的不同位置上传递了不同的信息量，所以这些共现概念应该拥有不同的权重。与文献 [120, 121] 中的观点类似，此处认为一个概念越抽象（即在一个层次分类体系中越靠近根节点），其携带的信息量越少，否则无需通过子孙概念对其进行进一步划分。因此，越具体的概念的信息量则越多，故其对于给定双词越重要。给定概念 c，基于 c 在层次分类体系 T 中的子孙概念的集合 $DES(c)$ 可计算 c 的内在信息内容 [120]（Intrinsic Information Content, IIC）如下：

$$IIC(c) = 1 - \frac{\log\left(|DES(c)| + 1\right)}{\log N_T} \tag{4.13}$$

其中 $|DES(c)|$ 是 $DES(c)$ 中概念的数量，N_T 是层次分类体系 T 中所有概念的数量。此外，对于给定的层次分类体系需创建一个虚拟的根节点以避免 $IIC(c) = 0$。基于概念 c 的内在信息内容（公式 4.13），定义双词 b_i 的先验概念概率 $\pi_{i,c}$ 如下：

$$\pi_{i,c} = IIC(c) \cdot \pi_{i,c}^{CC} \tag{4.14}$$

其中 $\pi_{i,c}^{CC}$ 由公式 4.12计算得到，注意需针对 $\pi_{i,c}$ 做归一化处理，即 $\sum_c \pi_{i,c} = 1$。

第二种结构化关联关系是基于路径长度的结构化关联关系。不难发现，一个双词的共现概念的祖先概念（在层次分类体系中）可能与该给定双词十分相关。比如，一个

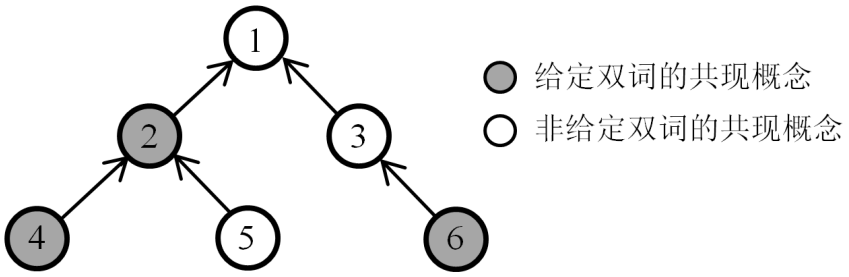


图 4.6 示例：一个层次分类体系中概念的位置

双词拥有一个共现概念“*Computer Vision*”，其祖先概念如“*Artificial Intelligence*”也与该双词相关。因此，将对双词的共现概念的祖先概念赋予先验概率。给定双词 b_i，如果其对应的一个共现概念 c_c 与其祖先概念 c_a 在层次分类体系中的距离越远，那么 c_a 与 b_i 相关的概率越低。基于此，选择 c_c 与 c_a 之间的最短路径长度 $SPL(c_c, c_a, T)$（路径长度即在层次分类体系 T 中两个概念间边的数量）以度量 c_c 到 c_a 的传播概率（Propagation Probability, PP）如下：

$$PP(c_c, c_a) = \pi_{i,c_c} \cdot \frac{1}{SPL(c_c, c_a, T) + 1} \tag{4.15}$$

其中 π_{i,c_c} 是双词 b_i 的关于概念 c_c 的先验概念概率。如图 4.6所示，由于一个祖先概念 *1*可能获取多个不同的传播概率（可分别从概念 *2*、*4*、*6*传播而来），所以只取最高的传播概率与概念 *1*当前的先验概念概率进行比较，若传播概率较高，则用该传播概率替换概念 *1*的先验概念概率。然而，由于一个共现概念 *2*同时也会得到一个由共现概念 *4*传播而来的传播概率，且该传播概率可能被用于替换概念 *2*的先验概念概率，所以可能导致概念 *1*获取的最高传播概率的变化以及概念 *1*自身的先验概念概率的变化。为了确保每个祖先概念可以获取最高的先验概念概率，提出一种先验概念分布更新算法（即算法 3）。在该算法中，首先对给定双词的所有共现概念按其对应的先验概念概率进行降序排列（*第 1 行*），然后按序为每个共现概念的祖先概念计算所有的传播概率并更新先验概念概率（*第 2-5 行*），最后对给定双词对应的所有先验概念概率做归一化处理（*第 6 行*）。

算法 3: 先验概念分布更新

Input: 一个双词 b_i, 其对应的概念分布 $\boldsymbol{\pi}_i$ 与共现概念集合 $CC(b_i)$;

Output: 更新后的 $\boldsymbol{\pi}_i$;

1 依据 $\boldsymbol{\pi}_i$ 对 $CC(b_i)$ 中的所有概念 $c_1, \cdots, c_{|CC(R)|}$ 进行降序排列，得 $c'_1, \cdots, c'_{|CC(R)|}$;

2 **for** $j = 1, \cdots, |CC(R)|$ **do**

3 　**foreach** c'_j 的祖先概念 c_a **do**

4 　　**if** $PP(c'_j, c_a) > \pi_{i,c_a}$ **then**

5 　　　$\pi_{i,c_a} = PP(c'_j, c_a)$

6 归一化处理：$\sum_c \pi_{i,c} = 1$

参数估计. 由于无法精确求解 CC-BiBTM 中的耦合参数 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}^s$、$\boldsymbol{\varphi}^t$，所以采用与

BiBTM 相同的解决方案，即 collapsed Gibbs 采样算法进行近似求解。此处选择对隐变量 x 与 z 进行联合采样，针对双词 $b_i^s \in \mathbf{B}^s$、$b_i^{st} \in \mathbf{B}^{st}$、$b_i^t \in \mathbf{B}^t$ 各自对应的 Gibbs 采样公式（具体推导见附录 A.3）如下：

$$P(x_i^s = c, z_i^s = k | x_{\neg b_i^s}, z_{\neg b_i^s}, \mathbb{O}) \propto \pi_{i,c}^s \cdot \frac{(n_{\neg b_i^s, k|c} + \alpha)}{(n_{\neg b_i^s, \cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^s, w_{i,1}^s|k} + \beta)(n_{\neg b_i^s, w_{i,2}^s|k} + \beta)}{(n_{\neg b_i^s, \cdot^s|k} + W^s\beta)(n_{\neg b_i^s, \cdot^s|k} + 1 + W^s\beta)} \tag{4.16}$$

$$P(x_i^{st} = c, z_i^{st} = k | x_{\neg b_i^{st}}, z_{\neg b_i^{st}}, \mathbb{O}) \propto \pi_{i,c}^{st} \cdot \frac{(n_{\neg b_i^{st}, k|c} + \alpha)}{(n_{\neg b_i^{st}, \cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^{st}, w_{i,1}^s|k} + \beta)(n_{\neg b_i^{st}, w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^{st}, \cdot^s|k} + W^s\beta)(n_{\neg b_i^{st}, \cdot^t|k} + W^t\beta)} \tag{4.17}$$

$$P(x_i^t = c, z_i^t = k | x_{\neg b_i^t}, z_{\neg b_i^t}, \mathbb{O}) \propto \pi_{i,c}^t \cdot \frac{(n_{\neg b_i^t, k|c} + \alpha)}{(n_{\neg b_i^t, \cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^t, w_{i,1}^t|k} + \beta)(n_{\neg b_i^t, w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^t, \cdot^t|k} + W^t\beta)(n_{\neg b_i^t, \cdot^t|k} + 1 + W^t\beta)} \tag{4.18}$$

其中 x 与 z 分别是给定双词当前的概念赋值与主题赋值。对于除双词 b 的以外的所有双词而言，$x_{\neg b}$ 指它们各自的概念赋值，而 $z_{\neg b}$ 则是它们各自的主题赋值。$\pi_{i,c}$ 表示第 i 个双词 $b_i^s \in \mathbf{B}^s$ 或 $b_i^{st} \in \mathbf{B}^{st}$ or $b_i^t \in \mathbf{B}^t$ 对应的关于概念 c 的先验概念概率。在排除双词 b 的情况下，$n_{\neg b,k|c}$ 指同时被赋予概念 c 与主题 k 的双词的数量，$n_{\neg b,\cdot|c} = \sum_k n_{\neg b,k|c}$，$n_{\neg b,w^s|k}$ 表示语言为 s 的词 w_s 被赋予主题 k 的次数，$n_{\neg b,\cdot^s|k} = \sum_{w^s} n_{\neg b,w^s|k}$，而 $n_{\neg b,w^t|k}$ 则是语言为 t 的词 w_t 被赋予主题 k 的次数，$n_{\neg b,\cdot^t|k} = \sum_{w^t} n_{\neg b,w^t|k}$。

经过足够次数的迭代后，即可对 CC-BiBTM 中的所有参数进行估计。由于使用双语主题模型的目的是学习概念的向量表示，所以在跨语言概念匹配的任务中只需要估计概念 -主题分布 $\boldsymbol{\theta}_c$（具体推导见附录 A.4）如下：

$$\theta_{c,k} = \frac{\alpha + n_{k|c}}{K\alpha + n_c} \tag{4.19}$$

其中 n_c 是被赋予概念 c 的双词的数量，而 $n_{k|c}$ 是同时被赋予概念 c 与主题 k 的双词的数量。在获得 CC-BiBTM 中的概念 -主题分布 $\boldsymbol{\theta}_c$ 后，即可在同一主题向量空间表示两个不同语言的层次分类体系中的概念。给定的一种语言的层次分类体系中的每个概念与其对应的另一种语言的层次分类体系中的候选匹配概念之间的相似性可由这两个概念各自对应的主题向量间的余弦相似度度量。

4.4 实验分析

4.4.1 已标注数据集上的评测

本节在两种不同类型的已标注数据集上对所提出的方法进行评测，并给出与其他方法在不同评测指标上的对比结果。

表 4.1 已标注数据集中每个层次分类体系的细节介绍

层次分类体系	京东商品目录	eBay 商品目录	中文 Dmoz 导航目录	Yahoo 导航目录
概念数量	7,741	7,782	2,084	2,353
双语文档对数量	67,594	72,979	19,277	21,467
中文词数量	24,483	18,190	11,064	8,581
英文词数量	15,489	14,729	8,806	8,100

4.4.1.1 数据集

该实验使用两种跨语言、跨领域的层次分类体系以验证所提出方法的有效性，这两种数据集的细节介绍如下：

- **商品目录**：此处选择中文的京东商品目录与英文的 eBay 商品目录，期望为京东商品目录中的每个概念找到其在 eBay 商品目录中的最相关概念。共在京东商品目录中抽取得到 7,741 个中文概念，在 eBay 商品目录中抽取得到 7,782 个英文概念。

- **导航站点目录**：此处选择中文 Dmoz 导航目录与英文的雅虎导航目录，期望为中文 Dmoz 导航目录中的每个概念找到其在雅虎导航目录中的最相关概念。共在中文 Dmoz 导航目录中抽取得到 2,084 个中文概念，在雅虎导航目录抽取得到 2,353 个英文概念。

给定一对不同语言的层次分类体系（即京东商品目录与 eBay 商品目录，或中文 Dmoz 导航目录与雅虎导航目录），为了生成对应的已标注数据，首先在中文层次分类体系中随机选择 100 个概念，然后针对每个随机选择的中文概念，在英文层次分类体系中标注与其最相关的英文概念。该标注过程共有五位从事知识图谱研究的标注人员参与，每位均对所有随机选择的概念进行标注，标注结果基于多数投票机制。

针对每个概念，抽取由 Google 返回的前二十个结果中的网页片段，并在每个片段中仅保留与给定概念及其父概念在相同句子中共现的词（细节于章节 4.3.2介绍）。将所有抽取得到的中文网页片段翻译成英文，而英文网页片段则翻译成中文，以此构建每个概念对应的双语文本上下文，即二十对双语文档构成的集合。针对这些文档还需进行噪音处理，具体方法如下：

- **处理中文文档**：1）使用 FudanNLP [102] 进行分词并去除停用词；2）删除在所有文档中词频低于 10 的词；3）删除文档长度（即文档中词的数量）小于 2 的文档。

- **处理英文文档**：1）删除非拉丁字母构成的词与停用词；2）将所有字母装换为小写并对所有词进行词干化处理；3）删除在所有文档中词频低于 10 的词；4）删除文档长度小于 2 的文档。

经过上述处理，表 4.1给出了每个层次分类体系对应的双语文本上下文的细节，包括双语文档对的数量、中文词的数量、英文词的数量。

4.4.1.2 对比方法

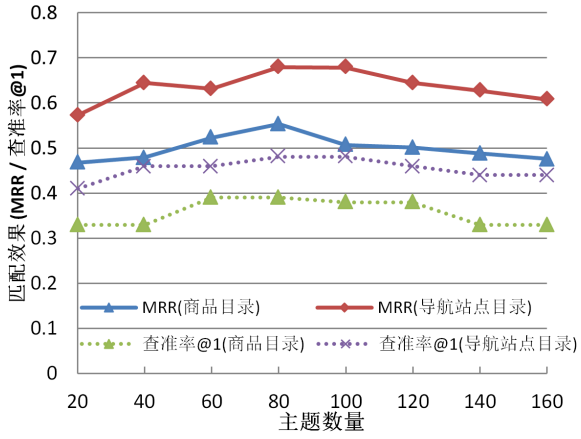
本章所提出方法是基于 BiBTM 的跨语言概念匹配方法（由 BiBTM 表示该方法）与基于 CC-BiBTM 的跨语言概念匹配方法（由 CC-BiBTM 表示该方法），在已标注数据集上进行实验的过程中，BiBTM 中的参数设置为 $\alpha = 50/K$（K 是主题数），$\beta = 0.1$，$K = 120$；在 CC-BiBTM 中的参数设置为 $\alpha = 50/K$，$\beta = 0.1$，$K = 100$。α、β 是依据经验直接设置，主题 K 是依据参数训练的结果确定（细节将在章节 4.4.1.3中介绍）。将本章所提出方法与其他三种不同类型的方法进行对比，包括基于现有模型的跨语言概念匹配的方法、基于 CC-BiBTM 简化版本的跨语言概念匹配方法、现有的跨语言本体匹配系统，具体如下：

- **基于现有模型的跨语言概念匹配的方法**：这一类方法也划分成两个子类别，第一类是基于 Ranking SVM [122] 模型的跨语言概念匹配方法 [31]（由 RSVM 表示该方法），为了在本章的已标注数据集上应用该方法，删除了特定的领域特征，保留了 20 个语言学特征与 8 个结构特征用于训练 Ranking SVM，这些特征均依赖于机器翻译后的字符串相似度。第二类与本章所提出方法的框架相同，唯一不同点在于仅将精确匹配部分的主题模型替换成传统的 TF-IDF 模型 [93] 或 BiLDA 模型 [111]，从而学习每个概念的向量表示。在基于 TF-IDF 的跨语言概念匹配方法（由 TF-IDF 表示该方法）中，将每个概念对应的所有双语文档对合并成一个文档后，再统计词频与逆向文档频率。而在基于 BiLDA 的跨语言概念匹配方法（由 BiLDA 表示该方法）中，设置参数 $\alpha = 50/K$，$\beta = 0.1$，$K = 80$（主题 K 的训练过程将在章节 4.4.1.3中介绍）。

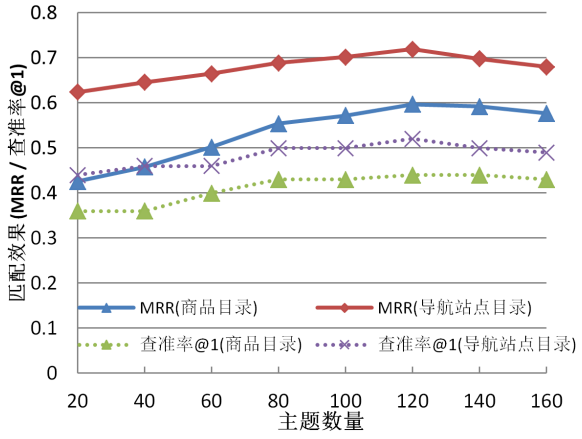
- **基于 CC-BiBTM 简化版本的跨语言概念匹配方法**：这类方法与本章所提出的框架相同，区别在于在精确匹配部分使用 CC-BiBTM 的简化版本学习每个概念的向量表示。CC-BiBTM 的完全版本使用了三种概念关联关系，即共现关联关系、基于信息内容的结构化关联关系、基于路径长度的结构化关联关系。第一种 CC-BiBTM 的简化版本仅使用共现关联关系，基于该版本的跨语言概念匹配方法由 CC-BiBTM(a) 表示。第二种简化版本在使用共现关联关系的基础上再融合了基于信息内容的结构化关联关系，该版本的跨语言概念匹配方法由 CC-BiBTM(b) 表示。CC-BiBTM(a) 与 CC-BiBTM(b) 中的参数均设置为 $\alpha = 50/K$，$\beta = 0.1$，$K = 100$（主题 K 的训练过程同样将在章节 4.4.1.3中介绍）。

- **现有的跨语言本体匹配系统**：虽然不同语言的层次分类体系中的跨语言概念匹配与跨语言本体匹配是不同的任务，但是由于层次分类体系可视为一种特殊的无形式化定义的属性与实例的本体，所以将目前最先进的两种跨语言本体匹配系统 AML [61]、LogMap [60] 作为对比进行评测。

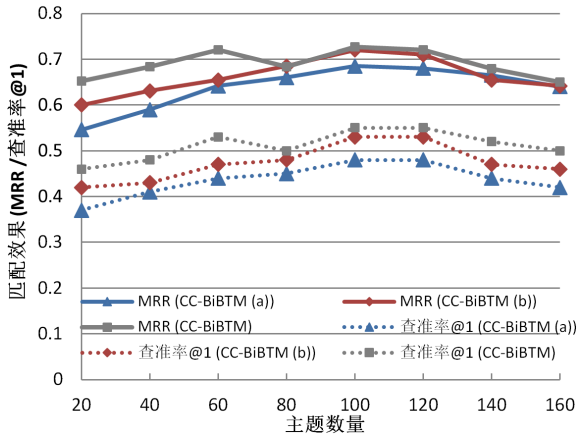
关于评测指标，与工作 [31–33] 类似，本章所提出方法与设计的对比方法均将不同层次分类体系中的跨语言概念匹配视为一个排序任务，即对一种语言的层次分类体系中的每个概念而言，依据其与另一种语言的层次分类体系中的每个概念的相似性对该层次分类体系中的所有概念进行排序。所以此处使用 MRR 与查准率@1 作为评测指标。



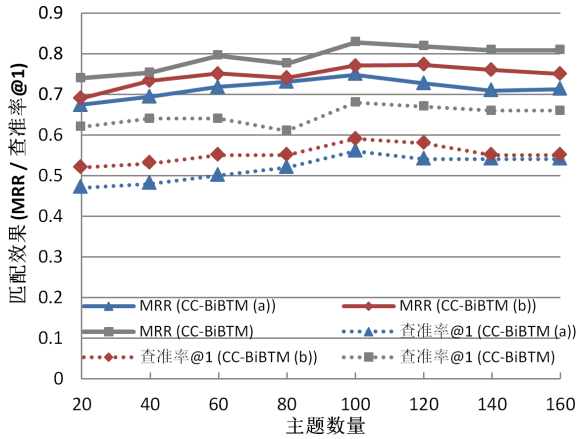
(a) BiLDA 在两种数据集上的效果



(b) BiBTM 在两种数据集上的效果



(c) CC-BiBTM 及其简化版本在商品目录上的效果



(d) CC-BiBTM 及其简化版本在导航站点目录上的效果

图 4.7 在不同主题数量 K 的情况下使用各双语主题模型的评测效果

4.4.1.3 参数训练

在本章所提出方法的框架下，在精确匹配部分可使用不同的双语主题模型，如本章所提出的 BiBTM 与 CC-BiBTM，或是 CC-BiBTM 的简化版本及现有的双语主题模型 BiLDA。由于双语主题模型中的主题数量是影响概念的向量表示学习的关键因素，继而能够影响跨语言概念匹配的效果，所以本节通过变化不同双语主题模型中的主题数量 K，记录在两种数据集上使用各模模型的效果，旨在针对不同双语主题模型选出最优的用于跨语言概念匹配的 K，需要注意的是这些双语主题模型在每个数据集上的 Gibbs 采样均迭代 500 次。图 4.7 给出了在不同主题数量 K 的情况下使用各双语主题模型的评测效果。在图 4.7(a) 中，BiLDA 中的主题数量 $K = 80$ 时，在两个数据集上的 MRR 与查准率@1 均最高。在图 4.7(b) 中，BiBTM 中的主题数量 $K = 120$ 时，在两个数据集上的 MRR 与查准率@1 均最高。在图 4.7(c) 与 (d) 中，可以看出当 CC-BiBTM 及其简化版本中的主题数量 $K \in [100, 120]$ 时，在两种数据集上的 MRR 与查准率@1 的值达到最高，所以考虑到模型训练的效率，则选择 $K = 100$。

4.4.1.4 评测结果分析

表 4.2 各种方法在两种已标注数据集上的评测结果

方法	商品目录		导航站点目录	
	MRR	查准率 @1	MRR	查准率 @1
AML	0.102	0.100	0.314	0.270
LogMap	0.105	0.100	0.265	0.250
RSVM	0.195	0.160	0.261	0.250
TF-IDF	0.423	0.330	0.489	0.400
BiLDA	0.553	0.390	0.679	0.480
BiBTM	0.597	0.440	0.719	0.520
CC-BiBTM (a)	0.685	0.480	0.748	0.560
CC-BiBTM (b)	0.721	0.530	0.771	0.590
CC-BiBTM	**0.727**	**0.550**	**0.828**	**0.680**

表 4.2给出了本章所提出方法与所设计对比方法的评测结果，从中可以看出：

- 本章所提出的基于 BiBTM 的方法与基于 CC-BiBTM 或其简化版本的方法的表现优于所有其他对比方法，特别是相比于在所有对比方法中表现最好的基于 BiLDA 的方法，本章所提出方法中表现最好的基于 CC-BiBTM 的方法在两种已标注数据集上的每个评测指标至少高出 **14.9%**，这充分体现了本章所提出方法的价值。

- 在本章所提出方法的内部进行比较，在 BiBTM 的基础上每增加一种概念关联关系，即在本章所提出方法的框架下将 BiBTM 依次换成 CC-BiBTM (a)、CC-BiBTM (b)、CC-BiBTM，可以发现在两种已标注数据集上的每个评测指标每次均有所提高，这说明关于本章将概念关联关系编码到双语双词主题建模过程中的方案的有效性，以及所提出的三种不同的概念关联关系（即共现关联关系、基于信息内容的结构化关联关系、基于路径长度的结构化关联关系）对于跨语言概念匹配均有积极的作用。

- 跨语言本体匹配系统 AML 与 LogMap 在已标注数据集上的表现均不佳。虽然这两个系统并未对于不同语言层次分类体系中的跨语言概念匹配任务做适应性调谐，但还是反映了在缺少本体通常拥有的内部特征（如实例、属性等）时，现有的跨语言本体匹配系统难以有效地完成不同语言层次分类体系中的跨语言概念匹配任务。

- 基于 RSVM 的方法的效果不如基于 TF-IDF 的方法，而基于 TF-IDF 的方法的各评测指标又低于所有的基于双语主题模型的方法。原因在于 1）基于 RSVM 的方法仅依赖于机器翻译后的字符串相似度，而基于 TF-IDF 的方法与基于双语主题模型的方法不仅依靠字符串相似度还利用了向量相似度，这体现将每个概念转换成向量表示再进行比较对于跨语言概念匹配的显著作用；2）与 TF-IDF 生成的词向量相比，由双语主题模型生成的低维主题向量不单纯依赖于词频信息，而是将双语文本上下文中的共现信息与概念关联关系编码进低维主题向量中，这导致在训练语料为搜索引

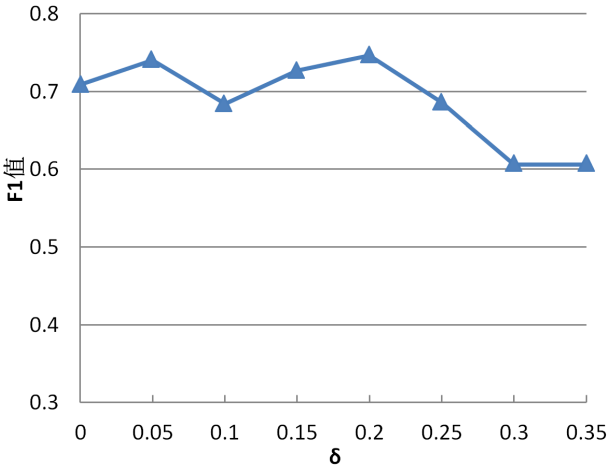


图 4.8 不同语言概念间的等价关系判定的阈值训练结果

擎获得的长度较短的网页片段的情况下，在跨语言概念匹配中使用低维主题向量的效果优于高维词向量。

4.4.2 等价关系发现的评测

本章所提出方法实际是为一种语言的层次分类体系中的每个概念寻找与其最相关的源自另一种语言的层次分类体系的概念，即挖掘不同语言概念间的“*最相关*”关系，而这些“*最相关*”关系中隐含了大量的概念间的等价关系。为了进一步挖掘不同语言概念间的等价关系，首先对章节 4.4.1中的已标注数据进行关于等价关系的标注，即从每个数据集中的 100 个已标注的不同语言概念间的“*最相关*”关系中标注其是否也是等价关系。依据标注结果，商品目录数据集的已标注数据中包含 48 个等价关系，而站点导航目录数据集的已标注数据中包含 62 个等价关系，即目前共包含 110 个已标注的不同语言概念间的等价关系与 90 个非等价关系。此处使用跨语言概念匹配效果最佳的 CC-BiBTM 模型以获得不同语言概念的向量表示，当给定的不同语言概念间的向量相似度大于一定阈值 δ 且二者之间已存在“*最相关*”关系时，则判定这两个不同语言的概念之间存在等价关系。

关于阈值 δ 的训练是在已标注数据上使用五折交叉验证完成。图 4.8给出了当变化 δ 的值（从 0 开始，步长为 0.05）时，关于等价关系判定的 F1 值的变化趋势。在该图中，当 $\delta = 0.2$ 时，F1 值达到峰值 0.746，此时的查准率为 0.920，查全率为 0.627。在整个商品目录与导航站点目录数据集上应用阈值 $\delta = 0.2$ 后，可从商品目录数据集中的 7,741 个“*最相关*”关系中发现 2,752 个等价关系，从导航站点目录数据集中的 2,084 个“*最相关*”关系中发现 1,020 个等价关系。共有五位研究生对上述所得等价关系进行标注。在标注时，每位标注人员共有三个选项：“*正确*”、“*不正确*”、“*不清楚*”。在每位标注人员标注完所有上述所得等价关系后，统计标注结果并计算正确率，标注结果基于多数投票机制。依据标注结果，商品目录数据集中平均有 2,135 个不同语言概念间的等价关系被标注为“*正确*”，正确率为 0.776；而导航站点目录数据集中平均有 825 个不同语言概念间的等价关系被标注为“*正确*”，正确率为 0.809。综上可以看出，经过

阈值训练后，本章所提出的方法能够挖掘出高质量的源自不同语言层次分类体系的概念之间的等价关系。

4.5 本章小结

本章主要介绍了一种基于双语主题模型的不同语言的层次分类体系中的跨语言概念匹配方法。该方法首先利用一种基于 BabelNet 的跨语言字符串相似度为一种语言的层次分类体系中的每个概念识别其在另一种语言的层次分类体系中最相关的概念。然后利用 Google 搜索引擎与 Google 翻译抽取每个概念对应的双语文本上下文。之后提出两种不同的主题模型 BiBTM 与 CC-BiBTM，这两种主题模型均可在给定的两个不同语言层次分类体系中的所有概念对应的双语文本上下文上训练得到每个概念的向量表示。最后利用概念间的余弦相似度完成跨语言概念匹配。实验结果表明，本文所提出的基于 CC-BiBTM 的方法在已标注数据集上的查准率@1 与 MRR 最佳，而基于 BiBTM 的方法的查准率 @1 与 MRR 也优于现有的其他方法或模型；此外，本章所提出的跨语言概念匹配方法还能够有效地挖掘不同语言概念间的等价关系。

本章工作分别发表于：

1) CCF A 类会议 *AAAI*（*AAAI Conference on Artificial Intelligence*）*2016*，论文题目为“Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic Model”（Full Paper）；

2) CCF B 类会议 *ISWC*（*International Semantic Web Conference*）*2017*，论文题目为“Encoding Category Correlations into Bilingual Topic Modeling for Cross-Lingual Taxonomy Alignment”（Full Paper）。

第五章　实例类别推断

实例知识是知识图谱不可或缺的组成部分，由于当前开放链接数据中的大规模多语言知识图谱（如：DBpedia、Yago 等）中的实例知识来源主要是维基百科，所以本章旨在通过实例类别推断将维基百科中的实例与社交站点中的概念链接，从而为面向社交站点的双语知识图谱构建提供一种获取实例知识的途径。又因为多语言的维基百科中的概念是社交站点中概念的重要组成部分，且维基百科中的概念本身存在于实例所在的页面中，即可视为二者存在某种语义关联，所以本章专注于以维基百科中的概念为候选类别进行实例类别推断。本章的具体内容安排如下，首先在章节 5.1中概述现有工作的问题、本章的解决方案、实验结果，然后在章节 5.2中介绍相关工作，之后在章节 5.3 中详细介绍本章所提出的一种新的各语言通用的从维基百科概念中推断实例类别的方法，并在章节 5.4中讨论所提出的方法在中英文维基百科上的实验结果，最后于章节 5.5进行小结。

5.1 概述

类别信息（Type Information）是模式知识与实例知识相关联的重要纽带，可由三元组进行表示，每个三元组以 *TypeOf* 关系为谓词连接概念与实例，比如：“*President of the United States*”*TypeOf* “*Barack Obama*”与“*Country in Europe*”*TypeOf* “*Italy*”，*TypeOf* 关系可视为概念与实例之间的上下位关系。在维基百科中，概念本身存在于实例所在的页面中，但二者之间并不是 *TypeOf* 关系，而仅是一种主题相关关系（由 *TopicOf* 关系表示），比如：“*Obama Family*”*TopicOf* “*Barack Obama*”。目前已有一些基于启发式规则的工作 [34, 35] 试图从这类 *TopicOf* 关系中发现正确的 *TypeOf* 关系，而这些工作所使用的启发式规则非常类似，其中心思想是：“*在维基百科中，当一个实例所在页面中的某个概念的词汇中心词是复数名词时，该概念即为给定实例的类别，即二者之间存在 TypeOf 关系*”。虽然这些规则已被证明十分有效，但依旧存在如下问题：

- 这些规则并不是各语言通用的，它们无法适用于无显式单复数名词的语言，比如：中文、日文、韩文等。
- 这些规则无法获取实例与概念（候选类别）之间的语义关联，易导致类别推断过程中的错误与遗漏。例如，一些概念的词汇中心词虽然是复数名词，但是它们并不是给定实例的正确类别（见图 5.1）。此外，所有词汇中心词为单数名词的概念均被忽略，但是许多这样的概念实际是给定实例的正确类别（见图 5.2）。

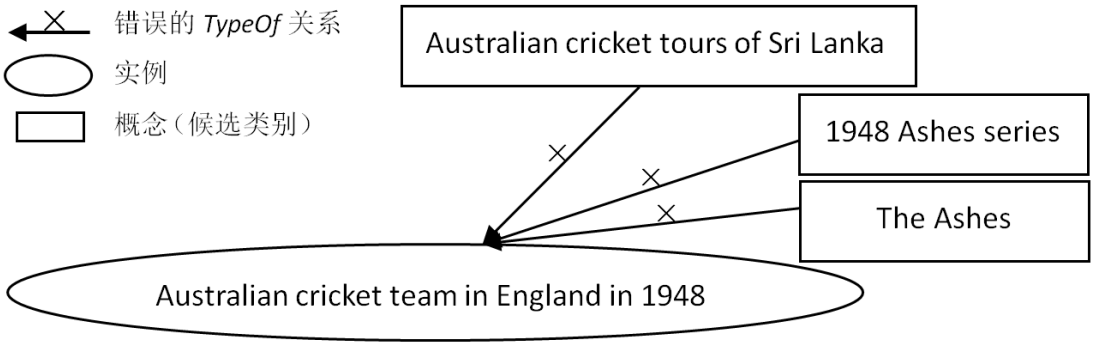


图 5.1 实例“*Australian cricket team in England in 1948*”所在页面中的部分概念（词汇中心词为复数名词）

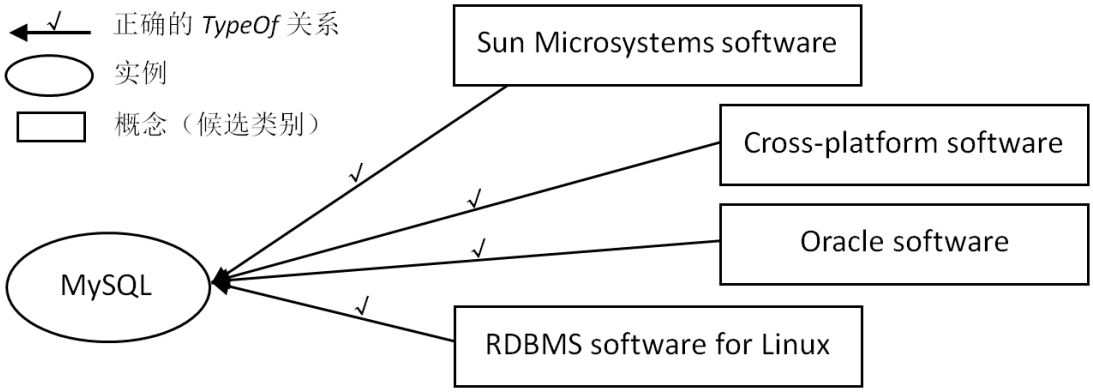


图 5.2 实例“*MySQL*”所在页面中的部分概念（词汇中心词为单数名词）

为了解决上述问题，本章引入了一种语言独立的特征，即属性，以帮助建立实例与概念之间的语义关联并在不同语言的维基百科中推断实例的类别。直观而言，当给定某个实例的属性“*actors, release date, director*”时，人类可能推断该实例的类别是“*Movie*”。然而，当给定属性“*name, foreign name*”时，人类却难以推断给定实例的类别，这是因为存在众多概念拥有属性“*name*”or“*foreign name*”。因此，本章提出了一个**属性驱动的实例类别推断假设**：“*在维基百科中，如果一个实例拥有在其对应页面中的某个概念的具有代表性的属性时，那么有很高的概率在该概念与该实例之间存在 TypeOf 关系*”。

基于上述假设，本章首先从维基百科中的信息框（见图 5.3(a)）抽取实例属性，然后提出了一种包含若干各语言通用的规则的算法抽取概念属性，之后考虑到许多实例的属性依旧缺失或不完全，提出了一种可以获取每个给定实例的拥有属性的最相似实例的方法，最后针对每个实例，构建一张包含给定实例与相应属性、概念、最相似实例的图，并提出一种随机游走模型计算每个概念与给定实例之间 *TypeOf* 关系成立的概率。

实验结果表明本章所提出的方法在测试数据集上的查准率、查全率、F1 值三项评测指标上均优于其他所设计的对比方法。在将本章所提出方法分别应用于整个中英文维基百科后，评测结果表明所提出方法能够获得大规模、高质量的中英文类别信息，与 DBpedia、Yago、LHD [123] 相比，通过本章方法得到的中文类别信息的规模最大，且获得许多并不存在于其他知识图谱中的新的英文类别信息。

(a)

```
{{Infobox school
| name                =
| image               =
| alt                 =
| caption             =
| motto               =
| location            =
| country             =
| coordinates         =
| established         =
| opened              =
| closed              =
| type                =
| district            =
| grades              =
| superintendent      =
| principal           =
| enrollment          =
| faculty             =
| campus_type         =
| campus_size         =
| team_name           =
| newspaper           =
| colors              =
| communities         =
| feeders             =
| website             =
| footnotes           =
}}
```

(b)

图 5.3 示例: (a) 信息框; (b) 信息框模板

5.2 相关工作

本节从四个不同方面介绍本章的相关工作，分别是：知识库构建中的实例类别推断，知识库补全中的实例类别推断，命名实体的类别推断，概念属性抽取。

5.2.1 知识库构建中的实例类别推断

知识库构建中的实例类别推断旨在无任何实例类别信息的情况下进行实例类别推断。Auer 等人 [124] 提出将维基百科中的信息框模板的名称作为相应实例的类别。Gangemi 等人 [125] 提出了一种从维基百科文章摘要中抽取类别信息的方法，核心思想是首先利用一些关于 *TypeOf* 关系的固定的词法、句法模式进行抽取，然后借助 WordNet 进行类别消歧，最后将类别与已有本体进行对齐。类似地，Kliegr [123] 也使用 Hearst 模式 [85] 从维基百科文章的第一句话中抽取实例类别，并将识别到的类别与 DBpedia 中的概念建立映射。本章工作与上述研究不同，其原因在于本章从实例对应的维基百科页

面中的概念中推断实例类别，而不是从信息框模板或文章摘要中抽取类别，即抽取来源不同。

与本章工作最为相关的研究 [34, 35]（抽取来源相同）均使用依赖于特定语言的启发式规则进行实例类别推断，而本章专注于从不同语言的维基百科中进行语言独立的实例类别推断，即提出一种各语言通用的方法。而在实验部分 5.4.1.4已仔细比较本文方法与基于规则的方法，结果显示本文方法在多个评测指标中均优于基于规则的方法。

5.2.2 知识库补全中的实例类别推断

知识库补全中的实例类别推断指使用知识图谱中已存在的实例类别补全缺失的实例类别信息。Nuzzolese 等人 [126] 提出两种分别基于归纳与演绎的技术，利用维基百科中的链接结构对 DBpedia 中的类别信息进行补全。Paulheim 与 Bizer [127, 128] 提出一种启发式的基于知识图谱中节点间链接的实例类别推断机制 SDType，它可用于任意 RDF 知识图谱的类别信息补全。Kliegr 与 Zamazal [129] 引入了一种文本挖掘的方法对知识图谱的类别信息进行补全，该方法首先使用一种统计类别推断（Statistic Type Inference, STI）算法对不同知识图谱中的概念建立映射，其中 STI 算法是一种通用的基于实例共现的计算概念相似度的算法，然后利用一组具有层次结构的支持向量机分类器针对给定的知识图谱中进行实例类别推断。Melo 等人 [130] 使用层次化多标签分类器对 RDF 知识图谱进行实例类别补全。以上研究与本章工作最大的不同点在于，本章所提出的方法并不利用任何现有知识图谱中的实例类别进行实例类别推断。

5.2.3 命名实体的类别推断

命名实体的类别推断是将文本中的识别到的实体分类到预先定义的一组类别中。不同研究领域均提出了大量的解决方案。著名的 FIGER 系统 [131] 与 PEARL 系统 [132] 均依赖于关系模式并利用给定知识图谱中的类别以推断命名实体的类别。另一类方法 [133–137] 设计了不同的特征与分类器，旨在将预定义的类别分配给命名实体。近期的 FINET 系统 [138] 使用多个不同的抽取器与词语含义消歧算法为命名实体选择最合适的类别。与之前的工作相比，FINET 系统不再依赖于某个特定的知识库或训练数据。目前最为流行的是基于词嵌入的方法 [139–141]，他们将实体与类别表示成位于同一向量空间的向量，并且设计不同的得分函数以预测给定类别与实体之间是否存在 *TypeOf* 关系。

上述工作中的各个系统的输入是一段可能包含命名实体的文本，这些系统将从文本中识别实体并推断其类别。与这些工作不同，本章工作的输入是一个实例与其对应的维基百科页面中的概念（如图 5.1与图 5.2所示），而输出即为从这些概念中选取得到的给定实例的正确的类别。

5.2.4 概念属性抽取

在不依赖于特定语言特征的条件下抽取概念属性是本章工作的主要挑战之一。Pasca 等人 [142–144] 提出了多个不同的设计与挖掘抽取模式的方法以从万维网文本与查询

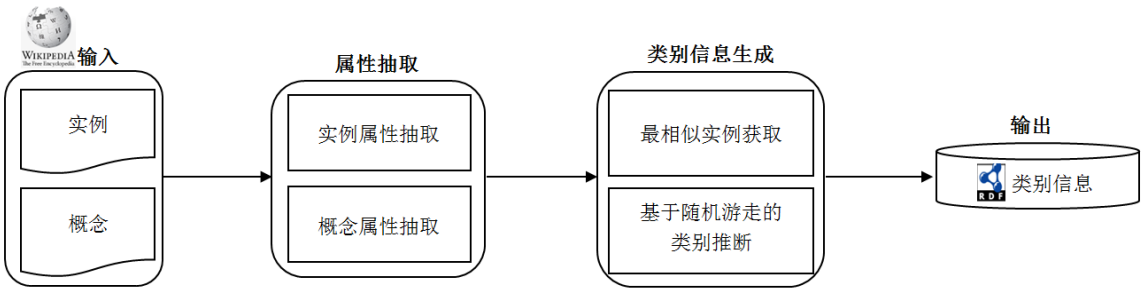


图 5.4 实例类别推断方法的流程示意图

记录中抽取概念属性。Lee 等人 [145] 首先提出了多种不同的方法以从文本、查询记录、知识图谱中抽取实例属性与概念属性，然后将实例属性映射到概念空间。由于无法轻易获得大规模的万维网文本与查询记录，且几乎所有抽取的模式均依赖于特定语言的特征，所以本章无法直接将上述研究成果应用于维基百科中的概念属性抽取。

5.3 方法设计

本小节将详细介绍所提出的在各语言维基百科中通用的实例类别推断方法。图 5.4给出了所提出方法的整个流程。方法的输入为给定语言的维基百科中的所有实例及其对应各自页面中的概念。属性抽取部分旨在从信息框直接抽取实例属性，并利用包含若干各语言通用规则的算法抽取概念属性。之后，在类别信息生成部分中，首先运用一个整合的相似实例得分获取每个给定实例的拥有属性的最相似实例，再使用一种随机游走模型进行实例类别推断。最后，方法的输出为给定语言的大规模类别信息。

5.3.1 属性抽取

由于维基百科中的实例属性显式定义于信息框中，所以关于实例属性的获取可直接从每个实例对应的页面中的信息框中进行抽取，故本节将主要讨论如何使用各语言通用的规则从不同语言的维基百科中抽取概念属性。

在维基百科中，信息框模板（图 5.3(b)给出了概念“*school*”的信息框模板）用于描述概念的特性，因此本节将信息框模板作为概念属性抽取的来源之一。本节首先定义一个基于信息框模板的抽取规则（Infobox Template based Extraction Rule, IT-ER）如下:

定义 5.1 ***基于信息框模板的抽取规则（Infobox Template based Extraction Rule, IT-ER）:*** *给定一个信息框模板 it 与一个概念 c，it 与 c 的字符串名称分别表示为 n(it) 与 n(c)，所有 it 中的属性可以传播给 c 当且仅当*

- *n(it) 与 n(c) 完全相同（在某些语言中忽略大小写与单复数形式的差异），*
- *或者将“Category:n(it)”作为查询提交给维基百科后，可重定向到 c，*
- *或者将 n(it) 作为查询提交给维基百科后，可重定向到 n(c)，*
- *或者将 n(c) 作为查询提交给维基百科后，可重定向到 n(it)。*

针对 **IT-ER**举例说明如下，信息框模板“*Template:Infobox islands*”与概念“*Category:Islands*”拥有相同的字符串名称“*islands*”。“*n(Category:Countries)*”的单数形式 (即“*Country*”) 与“*n(Template:Infobox country)*”(即“*country*”) 在忽略大小写的情况下完全相同。如果在维基百科中提交查询“*Category:n(Template:Infobox university)*”，则直接重定向到“*Category:Universities and colleges*”。而“*n(Category:States of the United States)*”也可以直接重定向到“*n(Template:Infobox U.S. state)*”。由于信息框模板是群体智慧的总结，所以本章假设任意信息框模板中的属性均是正确且完全的，因此不会对由 **IT-ER**抽取得到的概念属性做任何添加。

受面向对象编程 [146] 中的继承性质的启发，一个概念可以继承其父概念的所有属性（比如：概念“*Manager*”可以继承其父概念“*Person*”的属性“*gender*”），所以本章期望在维基百科的层次分类体系中应用该思想以抽取概念属性。当概念之间存在严格的上下位关系时，上述的继承性质是合理的。基于此，本节首先从众多开放知识图谱（包括 DBpedia、Yago、BabelNet、WikiTaxonomy、Zhishi.schema）中收集多个概念间上下位关系的集合。每个集合对应一种语言。然后，使用收集到的概念间上下位关系的集合对不同语言的维基百科的层次分类体系进行提纯，即当两个概念于指定的集合中不存在上下位关系时，便去除这两个概念在层次分类体系中的边。最终，每种语言各自对应一个提纯后的层次分类体系。基于这些提纯后的层次分类体系，本节定义自顶向下的基于层次结构的抽取规则（Top-Down Hierarchy-based Extraction Rule, TDH-ER）如下：

定义 5.2 ***自顶向下的基于层次结构的抽取规则（Top-Down Hierarchy-based Extraction Rule, TDH-ER)****：如果一个概念 c 不拥有从信息框模板中抽取得到的属性，且 c 的父概念拥有属性，那么所有其父概念的属性可由 c 继承。*

由于文献 [147] 提出一个概念传统上是一组拥有相似属性的实例的集合的占位符，所以若干后续工作 [143, 145, 148] 利用与给定概念具有严格上下位关系的子概念或实例的属性对给定概念补充属性信息。此处，同样利用该思想在提纯后的维基百科层次分类体系中抽取概念属性，定义一个基于多数投票的自底向上的基于层次结构的抽取规则（Bottom-Up Hierarchy-based Extraction Rule, BUH-ER）如下：

定义 5.3 ***自底向上的基于层次结构的抽取规则（Bottom-Up Hierarchy-based Extraction Rule, BUH-ER)****：如果一个概念 c 不拥有从信息框模板中抽取得到的属性，且其拥有具有属性的下位词（包括子概念与实例），那么当其中某个属性被超过一半的下位词所拥有时，该属性可以被传播给 c。*

为了共同使用上述提出的各语言通用的规则，本节提出一个通用的迭代算法（如算法 4所示），旨在从不同语言的维基百科中抽取概念属性。在算法 4中，给定一种语言 L，首先输入 L 对应的提纯后的层次分类体系 $CH^L = (C^L, R^L)$ 与一个二元组 $TUP^L = (C^L, A^L)$，其中 C^L 表示提纯后的层次分类体系中所有概念的集合，A^L 表示多个属性集合构成的集合，C^L 中每个概念与 A^L 中每个属性集合一一对应，且每个属性集合均初始化为 $\emptyset$。然后，对 C^L 中每个概念使用 **IT-ER**（*第 1 行*）。为了使用 **TDH-ER**，需要在 CH^L 中按序从根节点到叶子节点处理每个概念。类似地，为了使用

BUH-ER，需要在 CH^L 中按序从叶子节点到根节点处理每个概念。因此，先初始化三个队列 $Queue_{td}$、$Queue_{bu}$、$Queue$ 为 $\emptyset$（*第 2 行*），之后对 CH^L 中每个概念 c^L 依据其最大深度进行升序或降序排列（*第 3 行*），并按升序安排每个 c^L 进队列 $Queue_{td}$ 与按降序安排每个 c^L 进队列 $Queue_{bu}$（*第 4-5 行*）。在使用 **TDH-ER**进行自顶向下的概念属性抽取的过程中，仅需要将 $Queue_{td}$ 赋值于变量 $Queue$，随后按序处理 $Queue$ 中的概念即可（*第 7-10 行*）。而在利用 **BUH-ER**进行自底向上的概念属性抽取的过程中，同样也仅需要将 $Queue_{bu}$ 赋值于 $Queue$，随后按序处理 $Queue$ 中的概念即可（*第 13-16 行*）。最终，在 CH^L 中按序迭代执行自顶向下的概念属性抽取与自底向上的概念属性抽取，直至收敛，即 TUP^L 不再变化（*第 11-12 行*与*第 17-18 行*）。需要注意的是虽然存在一些概念与实例之间的噪音关系（即 *TopicOf* 关系），但由于 **BUH-ER**在 **IT-ER**与 **TDH-ER**之后使用，故抽取的质量可有所保障。另一方面，基于多数投票的思想，大量正确的上下位关系可以降低噪音关系在算法中的消极影响。

算法 4: 概念属性抽取

Input: 一个提纯后的层次分类体系 $CH^L = (C^L, R^L)$,
二元组 $TUP^L = (C^L, A^L)$,
L 指某种语言;

Output: 补充概念属性后的二元组 $TUP^L = (C^L, A^L_*)$

针对每个 $c^L \in C^L$ 使用 ***IT-ER***并更新 TUP^L;
$Queue_{td} \leftarrow \emptyset$, $Queue_{bu} \leftarrow \emptyset$, $Queue \leftarrow \emptyset$;
对 CH^L 中每个概念 c^L 依据其最大深度进行升序或降序排列;
按升序安排每个 c^L 进队列 $Queue_{td}$;
按降序安排每个 c^L 进队列 $Queue_{bu}$;
while true **do**
    $Queue = Queue_{td}$;
    while $Queue \neq \emptyset$ **do**
        $c_i^L = DeQueue(Queue)$;
        针对 c_i^L 使用 ***TDH-ER***并更新 TUP^L;
    if TUP^L 没有变化 **then**
        return TUP^L;
    $Queue = Queue_{bu}$;
    while $Queue \neq \emptyset$ **do**
        $c_j^L = DeQueue(Queue)$;
        针对 c_j^L 使用 ***BUH-ER***并更新 TUP^L;
    if TUP^L 没有变化 **then**
        return TUP^L;

5.3.2 类别信息生成

考虑到在维基百科中并不是每个实例对应的页面均包含信息框且信息框中的信息也不一定完全，即许多实例的属性依旧缺失或不完全，本小节首先提出一种获取每个实例的拥有属性的最相似实例的方法，期望利用最相似实例的属性对给定实例的属性信息进行有效的补充。然后，针对每个实例，构建一张包含给定实例与相应属性、概念、最相似实例的图，其中给定实例与概念之间通过属性连接。为了利用图结构计算某个概念是给定实例的类别的概率，本节提出了一种新的用于实例类别推断的随机游走模型。

5.3.2.1 最相似实例获取

本小节将详细介绍获取每个实例的拥有属性的最相似实例的方法，其中主要利用一种上下文相似度与一种已有概念集合相似度度量实例之间的相似程度。

上下文相似度. 为了度量实例之间的上下文相似程度，首先在相应语言的整个维基百科语料库上训练得到给定实例各自的向量表示，然后直接计算向量之间相似度。每个实例对应的向量表示可视为其上下文表示。如果两个实例出现在相似的上下文环境中，那么它们在同一向量空间中的距离则十分接近。在本章的工作中，使用 word2vec [149] 生成每个实例的向量表示。在训练的过程中，每个实例作为一个词单元进行训练。给定两个实例 i_1 与 i_2，它们之间的上下文相似度（Context Similarity, CSim）的定义如下:

$$CSim(i_1, i_2) = \frac{v(i_1) \cdot v(i_2)}{|v(i_1)| \times |v(i_2)|} \tag{5.1}$$

其中 $v(i_1)$ 与 $v(i_2)$ 表示 i_1 与 i_2 各自对应的向量，而上下文相似度的计算方式的实质为两个向量之间的余弦相似度。

已有概念集合相似度. 相似的实例倾向于共享相似的主题（即每个实例对应的维基百科页面中的概念）。给定两个实例 i_1 与 i_2，定义其各自对应的维基百科页面中的概念集合为已有概念集合 $ECset(i_1)$ 与 $ECset(i_2)$，而已有概念集合相似度（Existing Class Set Similarity, ECSSim）的定义如下：

$$ECSSim(i_1, i_2) = \frac{|ECset(i_1) \cap ECset(i_2)|}{|ECset(i_1) \cup ECset(i_2)|} \tag{5.2}$$

已有概念集合相似度的计算方式的实质为两个集合之间的 Jaccard 相似度。

为了在度量给定实例之间的相似性的过程中平衡上述两种相似度，则选择将两种相似度相乘，并将乘积结果排序以获取给定实例的拥有属性的最相似实例。此处，定义一种给定实例 i_1 与 i_2 之间的整合的相似实例得分（Integrated Instance Similar Score, IISS）如下：

$$IISS(i_1, i_2) = (CSim(i_1, i_2) + 1) \times (ECSSim(i_1, i_2) + 1) \tag{5.3}$$

由于可能存在某个相似度的值为 0，所以若是将相似度直接相乘，则导致整合的相似实例得分为 0，从而忽略了值不为 0 的相似度的作用。基于此，在公式 5.3中，在将两个相似度的值相乘之前对其各自加 1。如果不存在一个拥有属性的实例 i_2 满足 $IISS(i_1, i_2) > 1$，则表示无法获取实例 i_1 的拥有属性的最相似实例。

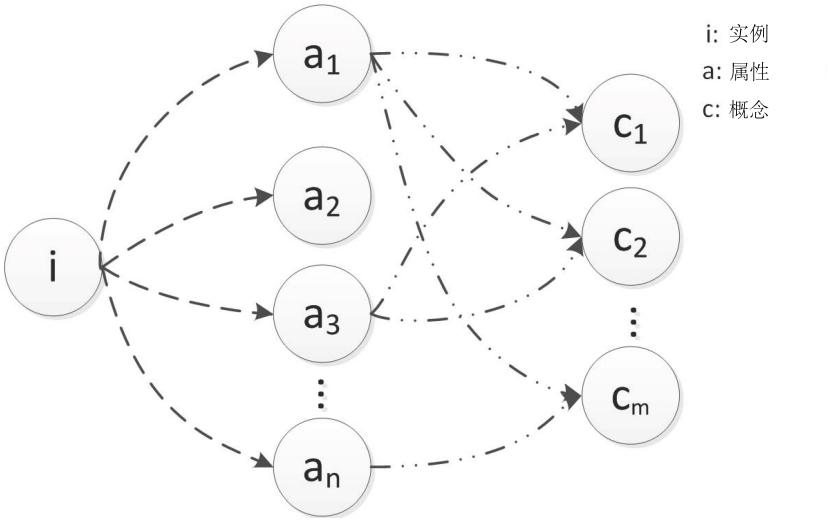


图 5.5 实例 i 对应的图的构建完成时的示例（不考虑给定实例的拥有属性的最相似实例）

5.3.2.2 随机游走模型

图构建 -不考虑给定实例的拥有属性的最相似实例. 给定一个实例 i，它可能拥有一组属性 $\{a_j^i\}_{j=1}^n$，并在其对应的维基百科页面中存在一组概念 $\{c_k^i\}_{k=1}^m$。每个概念 c_k^i 可能也对应一个属性集合 $A^{c_k^i}$。此处，从 i 到其对应的每个属性 a_j^i 之间建立一条有向边。如果属性 a_j^i 同时也是概念 c_k^i 对应的属性集合 $A^{c_k^i}$ 中的元素，那么将建立一条从 a_j^i 指向 c_k^i 的有向边。由于所有的属性抽取自维基百科，所以用户可以使用不同标签表示具有相同含义的属性（如：“*author*”与“*writer*”），这可能导致本应从属性指向概念的边的缺失。因此利用 BabelNet 检测具有相同含义的属性，如果某个概念 c_k^i 并不拥有属性 a_j^i，但是拥有一个 a_j^i 的同义属性，那么将建立一条从 a_j^i 指向 c_k^i 的有向边。在不考虑给定实例的拥有属性的最相似实例的情况下，图 5.5给出一个关于实例 i 对应的图的构建完成时的示例。

图构建 -考虑给定实例的拥有属性的最相似实例. 在构建实例 i 对应的图的过程中，考虑添加与 i 最相似的多个实例 $\{s_r^i\}_{r=1}^t$ 与其各自对应的属性。首先建立从 i 分别指向 $\{s_r^i\}_{r=1}^t$ 中每个元素的 t 条有向边，以及从每个最相似实例 s_r^i 分别指向其对应的每个属性的有向边。然后将最相似实例对应的属性加入到 i 对应的属性集合 $\{a_j^i\}_{j=1}^n$ 中，可获得一个更新后的属性集合 $\{a_j^i\}_{j=1}^{n^+}$ （$n^+ \geq n$），在此过程中，若一待加入的属性与某已有属性为同义属性，则去除此待加入属性，所有关联到此待加入属性的边与相应已有属性建立关联，可视为继承此待加入属性所关联的边。之后，如果一个新加入 i 的属性集合的属性也是概念 c_k^i 对应的属性集合 $A^{c_k^i}$ 中的元素，或与 $A^{c_k^i}$ 中某个元素为同义属性，则建立一条从该新加入的属性指向 c_k^i 的有向边。至此，实例 i 对应的图的构建全部完成。在考虑给定实例的拥有属性的最相似实例的情况下，图 5.6给出一个关于实例 i 对应的图的构建全部完成时的示例。

随机游走模型. 依据章节 5.1中提出的**属性驱动的实例类别推断假设**，如果一个实例包含越多其对应的概念的具有代表性的属性时，这个概念则很有可能是给定实例的类别。基于此思想，在每个实例对应的构建完成的图中，若某个概念可以通过越多的其

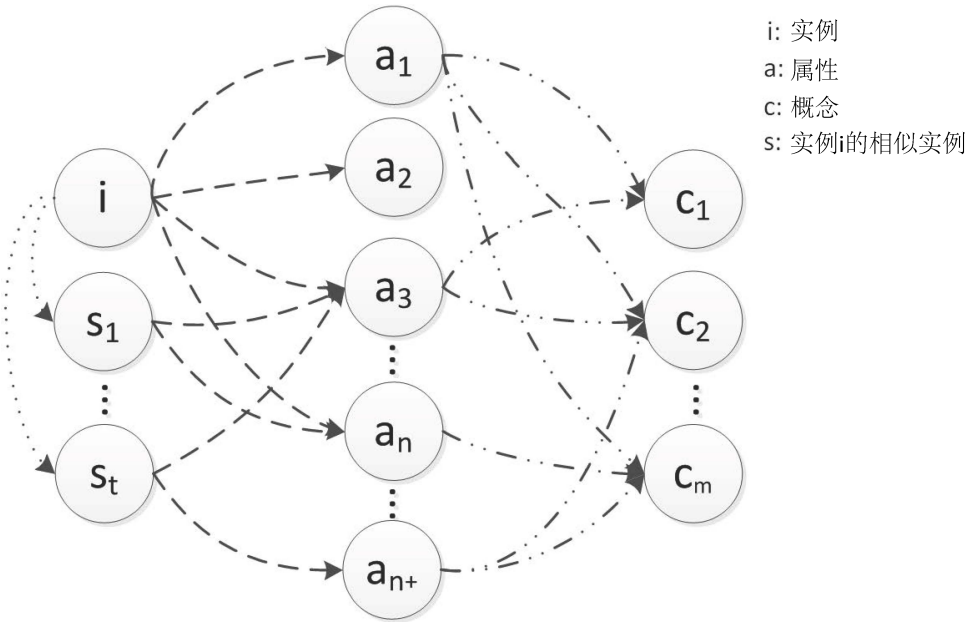


图 5.6 实例 i 对应的图的构建全部完成时的示例（考虑给定实例的拥有属性的最相似实例）

对应的有代表性的属性到达，那么该概念是给定实例类别的概率则更高。基于此分析，本节设计了一种用于实例类别推断的随机游走模型。在该模型中，首先计算两种给定实例对应的图中的不同节点之间的转移概率，然后通过整个图上的随机游走获取某个概念是给定实例类别的概率。

从给定实例出发的转移概率. 给定一个实例 i，当随机游走从 i 出发，下一步可能到达属性 a_j^i，或最相似实例 s_r^i。因此定义一个从 i 到 a_j^i 的转移概率 $P_{IA}(i, a_j^i)$ 与一个从 i 到 s_r^i 的转移概率 $P_{IS}(i, s_r^i)$ 如下：

$$P_{IA}(i, a_j^i) = \alpha \cdot \frac{Weight(a_j^i)}{\sum_{j=1}^{n} Weight(a_j^i)} \tag{5.4}$$

$$P_{IS}(i, s_r^i) = (1-\alpha) \cdot \frac{IISS(i, s_r^i)}{IISS(i, *)} \tag{5.5}$$

以上两个公式中的 $\alpha \in [0, 1]$ 是一个常数。在公式 5.4中，$Weight(a_j^i)$ 是 a_j^i 在维基百科中所有概念各自对应的属性集合中的出现次数的倒数。本章认为若一个属性被越少的概念所共享，那么该属性则越具有代表性，所以如果 a_j^i 出现的次数较低，则 $Weight(a_j^i)$ 越高，由于 a_j^i 被较少的概念共享，故 a_j^i 可被视为一个具有代表性的属性。当从 i 走一步至 a_j^i 时，随机游走模型更倾向于选择最具代表性的属性，从而能以更高的概率最终到达可作为 i 的类别的概念。在公式 5.5中，$IISS(i, s_r^i)$ 是 i 与 s_r^i 之间的整合的相似实例得分（见公式 5.3），而 $IISS(i, *)$ 是 i 与其对应的每个最相似实例之间的整合的相似实例得分的总和。关于公式 5.4与公式 5.5中的常数 α 的取值存在两种极端情况：1）如果 i 在属性抽取阶段已获取其所对应的属性，但并不存在拥有属性的最相似实例，则 $\alpha = 1$；2）如果 i 在属性抽取阶段无法获取属性，但是存在拥有属性的最相似实例，则 $\alpha = 0$。

从属性出发的转移概率. 给定实例 i，当随机游走从 i 的属性 a_j^i 出发，下一步即可到达某个概念 c_k^i，此处定义到达与 a_j^i 相连的所有概念的概率相同，关于从 a_j^i 到 c_k^i 的转移概率 $P_{AC}(a_j^i, c_k^i)$ 的公式如下：

$$P_{AC}(a_j^i, c_k^i) = \frac{1}{|N_C|} \tag{5.6}$$

其中 $|N_C|$ 表示在 i 对应的图中与 a_j^i 之间存在相连的边的概念的数量。以图 5.6为例，存在两条从属性 a_3 分别指向概念 c_1 与 c_2 的有向边，则转移概率 $P_{AC}(a_3, c_1)$ 与 $P_{AC}(a_3, c_2)$ 均为 0.5。

根据上述三种转移概率，定义适用于每个实例对应的图上的随机游走过程如下：

1) 从实例 i 出发，可通过两种方式抵达某个属性 a_j^i。一是直接通过 i 指向 a_j^i 的有向边，转移概率为 $P_{IA}(i, a_j^i)$（公式 5.4）；二是先到达 i 对应的某个最相似实例 s_r^i，转移概率为 $P_{IS}(i, s_r^i)$（公式 5.5），然后通过 s_r^i 指向 a_j^i 的有向边抵达 a_j^i，转移概率为 $P_{IA}(s_r^i, a_j^i)$（公式 5.4）。

2) 从属性 a_j^i 出发，通过 a_j^i 指向某概念 c_k^i 的有向边抵达 c_k^i，转移概率为 $P_{AC}(a_j^i, c_k^i)$（公式 5.6）。

基于此随机游走过程，随机游走模型可计算从实例 i 出发抵达概念 c_k^i 的概率 $P_{rgw}(i, c_k^i)$ 如下：

$$\begin{aligned} P_{rgw}(i, c_k^i) = &\sum_{j=1}^{n^+} P_{IA}(i, a_j^i) \cdot P_{AC}(a_j^i, c_k^i) + \\ &\sum_{r=1}^{t} P_{IS}(i, s_r^i) \cdot \sum_{j=1}^{n^+} P_{IA}(s_r^i, a_j^i) \cdot P_{AC}(a_j^i, c_k^i) \end{aligned} \tag{5.7}$$

将 $P_{rgw}(i, c_k^i)$ 归一化后，即为实例 i 与概念 c_k^i 之间存在 *TypeOf* 关系的概率。若该概率大于一固定阈值，则可判定 c_k^i 是 i 的类别。

5.4 实验分析

5.4.1 已标注数据集上的评测

本节在中英文已标注数据集上对所提出的方法进行评测，并给出与其他方法在不同评测指标上的对比结果。

5.4.1.1 数据集与评测指标

数据集. 在中英文维基百科中随机选择 1,000 个中文实例与 1,000 个英文实例，每个实例在其对应的维基百科页面中至少拥有 1 个概念。共五位从事知识图谱研究的硕士研究生参与标注，标注内容为给定概念与相应实例之间是否存在 *TypeOf* 关系。标注结果基于多数投票机制。表 5.1已给出标注为“*正确*”（或“*不正确*”）的表示 *TypeOf* 关系的三元组的细节。

表 5.1 中英文数据集的详细标注结果

数据集	正确三元组数量（百分比）	错误三元组数量（百分比）
英文	4,484 (76.2%)	1,402 (23.8%)
中文	4,972 (85.8%)	821 (14.2%)

评测指标. 针对本章所提出的方法与其他对比方法（将在章节 5.4.1.2中介绍）在已标注数据集中进行参数训练（将在章节 5.4.1.3中介绍）后，即可得到实例类别推断结果，再选择使用三种不同的指标：查准率、查全率、F1 值对结果进行评测。

5.4.1.2 对比方法

本章所提出的方法是基于属性驱动的随机游走（Attribute-Driven Random Walk, ARW）模型，将其（由 ARW 表示）与下列方法进行对比：

- **启发式规则（Heuristic Rules, HR）**：文献 [34, 35] 中使用启发式规则对维基百科中的实例进行类别推断。这些规则通过检测给定实例对应的概念是否是复数名词或可数名词以决定该概念是否为给定实例的类别。这是目前最先进的从维基百科已有概念中进行实例类别推断的方法。需注意的是，当在中文数据集中应用该方法时，需首先通过维基百科中的跨语言链接将中文实例与概念转换成英文，再进行推断。

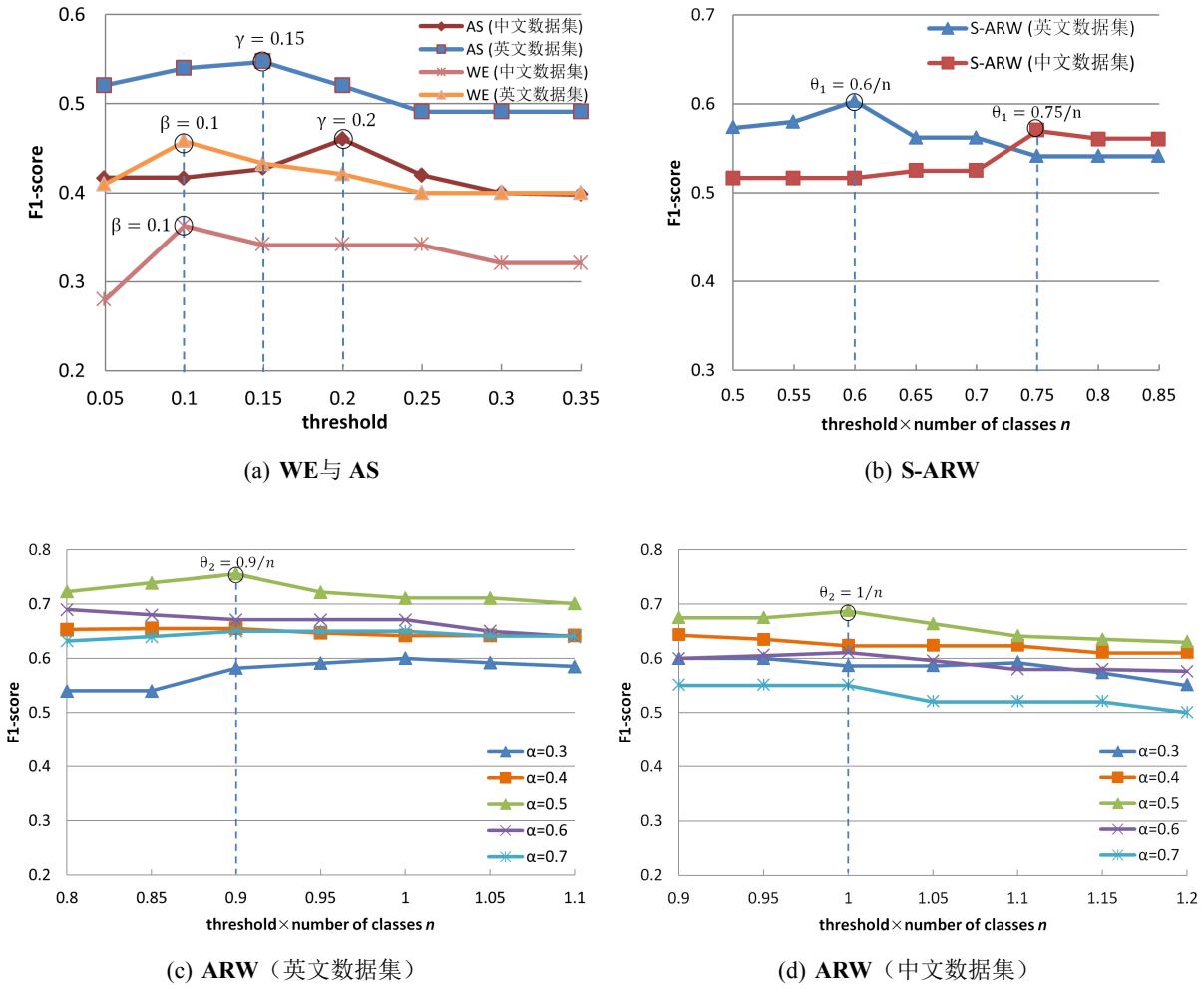
- **词嵌入（Word Embedding, WE）**：针对任意特定语言，本章所提出方法使用 word2vec 在给定语言的维基百科中的全部文本中训练得到每个实例的向量表示（于章节 5.3.2.1中介绍）。并且在训练之后可得到每个词的向量表示，所以每个概念可由构成该概念的所有词的平均向量表示。给定一个实例及其对应的维基百科页面中的一个概念，此方法计算实例与概念各自向量之间的余弦相似度，若相似度高于给定阈值，则判定该概念是给定实例的类别。

- **属性相似度（Attribute Similarity, AS）**：给定一个实例与其对应的维基百科页面中的一个概念，直接计算各自对应的属性集合的 Jaccard 相似度，如果相似度高于给定阈值，则判定该概念是给定实例的类别。

- **本章方法的简单版本（Simplified Version of Our Approach, S-ARW）**：在该方法中，在完成了本章所提出的实例属性抽取与概念属性抽取后，不再考虑每个实例的最相似实例及其对应属性，即将随机游走模型退化到如图 5.5所示。从公式的角度出发，直接将公式 5.4与公式 5.5中的常数 α 设置为 1 即可。

5.4.1.3 参数训练

在实验中，本章所提出方法与其他对比方法均应用五折交叉验证训练参数。一类参数是用于推断给定概念与实例之间是否存在 *TypeOf* 关系的方法中固定阈值，即 **WE**中的 β、**AS**中的 γ、**S-ARW**中的 θ_1、**ARW**中的 θ_2。另一类是 **ARW**中使用的公式 5.4与公式 5.5中的常数 α。此处，当变化方法 **WE**中的阈值 β 或 **AS**中的阈值 γ 时，记录 F1 值的变化。β 与 γ 的变化从 0 开始，步长为 0.05。针对 **S-ARW**中阈值的 θ_1 与 **ARW**中阈



(a) **WE**与 **AS**　　(b) **S-ARW**

(c) **ARW**（英文数据集）　　(d) **ARW**（中文数据集）

图 5.7 **WE**、**AS**、**S-ARW**、**ARW**中的参数训练结果

值的 θ_2，观察得到它们与给定实例对应的维基百科页面中的概念数量 n 呈负相关关系。换言之，越多概念出现于给定实例对应的维基百科页面中，有越高的可能性在这些概念中存在更多的类别，所以更多的类别会作为 **S-ARW**与 **ARW**中随机游走模型的终点，那么通过随机游走最终抵达每个概念的归一化概率越低。因此，选择在变化 $\theta_1 \times n$（或 $\theta_2 \times n$）与 α 的同时，记录 F1 值的变化。关于 $\theta_1 \times n$ 与 $\theta_2 \times n$ 的变化，均从 0 开始，步长为 0.05，而 α 变化的区间为 $[0.1, 0.9]$，变化的步长为 0.1。

图 5.7给出了当 F1 值达到峰值后的各参数的取值。在 **WE**中，阈值 β 在中英文数据集中均设置为 0.1。在 **AS**中，阈值 γ 在英文数据集中设置为 0.15，在中文数据集中设置为 0.2。对于 **S-ARW**而言，阈值 θ_1 在英文数据集中设置为 $0.6/n$，在中文数据集中设置为 $0.75/n$。在 **ARW**中，针对英文数据集设置阈值 $\theta_2 = 0.9/n$，常数 $\alpha = 0.5$，而在中文数据集中设置 $\theta_2 = 1/n$，$\alpha = 0.5$。

5.4.1.4 效果与分析

在本章所提出方法（即 **ARW**）与对比方法 **WE**、**AS**、**S-ARW**中，均需要抽取实例属性与概念属性，此处使用章节5.3.1中所提出属性抽取策略以完成此任务。当使用

表 5.2 关于各实例类别推断方法的评测结果

方法	英文数据集			中文数据集		
	查准率	查全率	F1 值	查准率	查全率	F1 值
HR	0.839	0.531	0.650	0.894	0.481	0.625
WE	0.771	0.326	0.458	0.805	0.234	0.363
AS	0.852	0.403	0.547	0.918	0.308	0.461
S-ARW	0.903	0.453	0.603	0.954	0.406	0.570
ARW	**0.917**	**0.643**	**0.756**	**0.981**	**0.528**	**0.687**

ARW时，为了确保类别推断的质量，针对每个实例均只使用其排名前五的拥有属性的最相似实例。这些最相似实例由章节5.3.2.1中所提出的方法获取。在另外一个对比方法**HR**中，应用文献 [98] 中的词汇中心词抽取方法以获取每个概念的词汇中心词。表 5.2给出了本章所提出方法与其他对比放在已标注数据集上的全部评测结果，可以看出：

- **ARW**在所有评测指标上均超出其他方法，这充分说明本章所提出方法能够更好地在维基百科中进行语言独立的实例类别推断。所有基于属性的方法（即 **AS**、**S-ARW**、**ARW**）在所有评测指标上均超过基于词嵌入的方法（即 **WE**），且在查准率上优于基于规则的方法 **HR**，这恰恰显示本章引入的语言独立的特征（即属性）的积极作用。由于一些实例与概念均存在属性缺失的情况，所以一些基于属性的方法在查全率与 F1 值上的表现不如基于规则的方法 **HR**，然而本章所提出方法 **ARW**在这两个评测指标上优于 **HR**，这反映了拥有属性的最相似实例对于实例类别推断的价值。

- **ARW**在查准率与查全率上均优于 **S-ARW**，这是因为所获取的最相似实例的属性能够对于两类实例进行补充。一类是拥有属性的实例，但并不完全，所以源自最相似实例的属性可以帮助这类实例更加准确地进行类别推断，从而导致了查准率的提升。另一类是没有属性的实例，**ARW**可以帮助这类实例完成类别推断，从而提升了查全率。此外，**ARW**与 **S-ARW**一致在查准率、查全率、F1 值上的表现优于 **AS**，这意味着应用本章所提出的**属性驱动的实例类别推断假设**是合理且有效的。

- 另一个有趣的现象是相对于英文数据集，本章所提出的方法 **ARW**在中文数据集上的查准率更高，但查全率较低。依据数据统计，中英文数据集中的每个实例均至少获取了一个拥有属性的最相似实例，但仅有一部分概念（英文数据集中 86% 的概念与中文数据集中 72% 的概念）能够通过章节5.3.1中的属性抽取策略获得属性。因此，较低比例的拥有属性的中文概念是导致较低查全率的主要原因。此外，在英文维基百科中，每个属性平均被 718 个概念共享；而在中文维基百科中，每个属性平均被 77 个概念共享。本章所提出的**属性驱动的实例类别推断假设**依赖于每个概念的具有代表性的属性，且一个属性被越少的概念共享，其代表性越强（由公式 5.4体现）。因此，英文维基百科中较高的共享每个属性的平均概念数量导致较低的属性代表性，从而影响 **ARW**在英文数据集上的查全率。这也同样说明了本章所提出的方法在计算每个概念的属性代表性方面存在进一步的改进空间。

5.4.2 整个中英文维基百科上的评测

利用在已标注数据集上训练得到的参数，将本章所提出的方法应用于整个英文维基百科与中文维基百科后，共得 7,571,009 个不同的英文 *TypeOf* 关系与 400,349 不同的中文 *TypeOf* 关系。本节评测所有得到的中英文类别信息（即表示 *TypeOf* 关系的三元组）的正确率，并分析本章所提出方法中的各个模块的运行时间（用于生成上述中英文类别信息）。

5.4.2.1 类别信息正确率评测

由于并不存在关于中英文维基百科中所有 *TypeOf* 关系的已标注数据，所以不可能对所有得到的中英文类别信息进行自动化评测。与此同时，由于类别信息的规模庞大，所以也不可能手工对所有类别信息进行评测。因此，本节设计了一个抽样及标注的策略以近似估计所得所有中英文类别信息的正确率。

抽样. 给定某种语言 L 的所有类别信息，抽样旨在抽取出能够反映整个数据分布的一组小规模的类别信息。本节首先提取出所有类别信息中的实例，从中随机选择 100 个实例。随后，关于语言 L 的样本由表示随机选择的实例及其对应的概念之间的 *TypeOf* 关系的三元组构成。该抽样策略与 Yago 类似。

标注. 整个标注过程与 Yago 及 Zhishi.schema 类似。同样五位从事知识图谱研究的硕士研究生参与标注，在标注样本中的每个三元组时，标注人员有三个选项：“*正确*”、“*不正确*”、“*不清楚*”。在每位学生标注完所有样本后，在显著性水平 $\alpha = 5\%$ 时利用威尔逊区间 [103] 将样本的平均正确率泛化至语言为 L 的所有类别信息上。

针对使用本章方法获得的所有中英文类别信息，均采用上述抽样与标注策略进行正确率评测。下列结果显示了所得中英文类别信息的高正确率：

- **英文类别信息的正确率**：给定 100 个随机选择的英文实例，可获得 582 个表示 *TypeOf* 关系的三元组。在标注后，平均“正确”的票数为 534，正确率为 91.42% ± 2.25%。
- **中文类别信息的正确率**：给定 100 个随机选择的中文实例，共获得 641 个表示 *TypeOf* 关系的三元组。在标注后，平均“正确”的票数为 622，正确率为 96.74% ± 1.33%。

5.4.2.2 生成类别信息的时间分析

按序在中英文维基百科中运行本章所提出方法中的各个模块，计算机配置为 Intel Xeon E5-2630 v4 2.20 GHz CPU，256GB 内存，以及 Linux 服务器。表 5.3给出生成中英文类别信息的各模块运行时间，从中可以看出：

- 最消耗时间的模块是最相似实例获取，其原因在于它使用了整个英文（或中文）维基百科中的文本作为 word2vec 的训练语料以获得每个实例的向量表示。
- 生成所有英文类别信息的时间远高于生成所有中文类别信息，主要原因是英文维基百科中实例、概念、属性的数量均远远少于其在中文维基百科中的数量。

表 5.3 所提出方法各模块在中英文维基百科上的运行时间

模块	各模块运行时间（小时）	
	英文类别信息	中文类别信息
实例属性抽取	1.93	0.24
概念属性抽取	0.72	0.21
最相似实例获取	8.23	2.10
类别推断	1.79	0.37
总计	12.67	2.92

- 生成所有的中英文类别信息共只需 $12.67 + 2.62 = 15.29$ 个小时。该时间是在中英文维基百科中顺序执行所提出方法中的各模块所得总计时间，一方面该时间长度是可接受的，另一方面还是存在很大的提高空间，比如可以尝试并行化各模块中的一些任务以降低时间消耗。

5.4.3 与其他知识图谱的对比

通过本章所提出方法获得的中文类别信息共包含 400,349 不同的中文 *TypeOf* 关系，216,727 个不同的拥有类别的中文实例，25,406 个不同的中文类别。由于现有的开放知识图谱并不包含中文的类别信息，故本节仅将所得英文类别信息与其他知名知识图谱 DBpedia、Yago、LHD 中的英文类别信息进行比较，这三个知识图谱中的英文类别信息同样挖掘自英文维基百科。在比较的过程中，不仅记录了 *TypeOf* 关系的数量、拥有类别的实例的数量、类别的数量，还通过完全字符串匹配的方式计算了本章所得英文类别信息与其他知识图谱中的英文类别信息之间的关于 *TypeOf* 关系的交集、拥有类别的实例的交集、类别的交集，以及统计了平均每个实例拥有的类别数量。具体结果见表 5.4.

在本章所得的英文类别信息中，共有 7,571,009 个不同的英文 *TypeOf* 关系，3,207,668 个不同的拥有类别的英文实例，475,148 个不同的英文类别。与 DBpedia、Yago、LHD 相比，通过本章方法所得的英文类别信息包含第二多的 *TypeOf* 关系，而重合的 *TypeOf* 关系的数量并不多，这恰恰说明本章所得的英文类别信息是现有知识图谱的良好补充。在本章所得英文类别信息中去除 DBpedia、Yago、LHD 中已有的英文 *TypeOf* 关系后，发现本章工作共贡献 3,927,727 个新的英文 *TypeOf* 关系。

本章所得英文类别信息中的拥有类别的实例的数量（排名第三）与其他知识图谱中的拥有类别的英文实例的数量属于一个量级，即 300 多万个。基于重合的拥有类别的实例数量可发现本章所得英文类别信息对于其他每个知识图谱至少贡献超过 27 万的新的拥有类别的实例。此外，本章所得英文类别信息拥有最大数量的英文类别，其中包含 148,395 新的词汇中心词为单数名词的类别，由此产生了大量新的 Yago 所无法拥有的英文 *TypeOf* 关系（Yago 只包含词汇中心词为复数名词的类别），这也体现了本章所提出方法的价值。由于许多实例在真实世界中扮演了不同的角色，所以每个实例可能拥有不止一个类别，因此计算了平均每个英文实例拥有的类别信息，而在这一点上本章所得英文类别信息也优于 DBpeida 与 LHD 中的英文类别信息。

表 5.4 本章所得英文类别信息与其他知识图谱中的类别信息的比较

	本章的英文类别信息	DBpedia	Yago	LHD
TypeOf 关系数量	7,571,009	3,121,552	14,164,808	6,505,492
重合的 *TypeOf* 关系数量	/	80,693	3,521,747	123,531
拥有类别的实例数量	3,207,668	3,121,054	3,632,793	3,618,094
重合的拥有类别的实例数量	/	2,445,095	2,928,265	2,749,709
类别数量	475,148	401	445,337	47,626
重合的类别数量	/	29	326,753	3,648
平均每个实例拥有的类别数量	2.36	1.00	3.90	1.80

5.5 本章小结

本章主要介绍了一种新的各语言通用的面向不同语言维基百科的实例类别推断方法，该方法包含两部分。第一部分是属性抽取模块，旨在从维基百科信息框中直接生成实例属性，且使用一种通用的融合了多条语言独立的规则的迭代算法抽取概念属性。第二部分是类别信息生成模块，首先基于一种整合的相似实例得分获取每个实例的拥有属性的最相似实例，然后将每个给定实例、属性、概念、最相似实例组织成一张图，最后在图上进行随机游走以计算每个概念与给定实例之间存在 *TypeOf* 关系的概率。在实验中进行了两类评测。第一类是在中英文已标注数据集上的评测，实验结果表明本章所提出方法在多个评测指标中均优于现有的其他方法。第二类是在整个中英文维基百科上的评测，实验结果表明所得大规模中英文类别信息[1]均拥有较高的正确率，且包含大量现有知识图谱中不存在的中英文类别信息，它们是现有知识图谱中的类别信息的有效补充。

本章工作已被 CCF C 类期刊 *International Journal on Semantic Web and Information Systems*接收（待正式发表），论文题目为“Language-Independent Type Inference of the Instances from Multilingual Wikipedia”。

[1] http://www.multype.org/

第六章　总结与展望

6.1 论文总结

近年来，知识图谱作为语义网与人工智能发展的重要助力得到了学术界与工业界的大量关注。目前，知识图谱已经成为众多智能应用（如语义搜索、问答系统、情报分析等）的重要基础资源，因此如何自动化构建高质量的知识图谱是具有极大意义与价值的研究课题。然而，当前的相关研究并未全面关注万维网中的一种十分重要的知识挖掘来源，即不同类型的社交站点，包括电子商务、百科、问答、博客、游戏、旅行等站点，其中存在的大量表示概念的层次分类体系中的分类与分众分类系统中的标签等信息可视为知识图谱构建的重要资源。此外，随着信息全球化的发展，跨语言知识分享的需求也日益迫切，特别是在我国一带一路战略的背景下，如何将中文知识与其他语言的知识进行对齐，以促进中文与其他语言的知识分享，是扩大我国科技影响力的重要手段。在现有的多语言知识图谱中，英文知识占绝对主导地位，而其他语言知识较少是跨语言知识分享的主要障碍之一。虽然已经存在关于双语知识图谱构建（即针对两种任意给定的语言分别构建一个知识图谱并进行跨语言知识对齐）的研究工作，但其知识挖掘来源仅局限于百科站点。

基于上述背景，本文对从社交站点中构建双语知识图谱进行了深入研究，旨在提出各领域、各语言通用的构建方法。具体而言，本文选择研究三项关于双语知识图谱构建的子任务，分别是：模式知识挖掘、跨语言概念匹配、实例类别推断，研究内容如下：

1) 提出了一种新的结合机器学习与规则的模式知识挖掘方法，旨在挖掘概念间的不同关系。该方法首先使用一种分块机制减少待判定的概念对的数量，以尽可能保证方法能够在大规模的场景下进行应用。然后，利用一种自动化的策略从给定的概念对中生成标注数据。最后，提出一种半监督学习方法检测给定的两个概念之间是否存在 *equal*、*subClassOf*、*relate*关系，其中包含一个基于各语言通用的规则的后处理步骤，该步骤在迭代学习的过程中用于修正错误分类的结果。该方法克服了现有相关工作依赖特定语言的特征与规则的缺陷，整个方法中提出的特征与规则均是各语言通用的，因此这是一个各语言通用的面向社交站点的模式知识挖掘方法。在实验中，所提出方法不仅在测试数据集上的查准率、查全率、F1 值均优于其他基准对比方法，而且在大量中英文社交站点中应用后可生成大规模、高质量的概念间的不同关系。

2) 提出了一种新的基于双语主题模型的跨语言概念匹配方法，旨在为一种语言的层次分类体系中的每个概念寻找与其最相关的源自另一种语言的层次分类体系中的概

念。该方法首先提出一种基于 BabelNet 的跨语言字符串相似度以识别候选匹配概念对。然后利用 Google 搜索引擎与 Google 翻译抽取每个概念的双语文本上下文。之后提出两种新的主题模型：双语双词主题模型与基于概念关联关系的双语双词主题模型，二者均可在给定不同语言的两个层次分类体系中所有概念对应的双语文本上下文中训练得到每个概念的向量表示。最后通过不同语言概念之间的向量余弦相似度得到匹配结果。该方法克服了现有相关工作依赖于特定的领域信息的缺陷，整个方法不涉及任何特定的领域信息，因此这是一个综合考虑字符串相似度与向量相似度的各领域通用的面向社交站点的跨语言概念匹配方法。在实验中，当所提出方法使用基于概念关联关系的双语双词主题模型时，其在测试数据集上的查准率@1 与 MRR 均远远超过其他基准对比方法。此外，通过控制最终的相似度阈值，可生成高质量的源自不同层次分类体系的不同语言的概念间的等价关系。

3) 提出了一种新的基于随机游走模型的实例类别推断方法，旨在挖掘概念与实例之间的上下位关系，从而为双语知识图谱引入维基百科中的实例知识。该方法首先从维基百科信息框中抽取实例属性，然后提出一种包含若干各语言通用的规则的算法以抽取概念属性，之后考虑到许多实例的属性依旧缺失或不完全，提出一种可以获取每个给定实例的拥有属性的最相似实例的方法，最后针对每个实例，构建一张包含给定实例与相应属性、概念、最相似实例的图，并利用一种随机游走模型计算每个概念是给定实例类别的概率。该方法克服了现有工作依赖于特定语言的规则的缺陷，不涉及任何特定语言的信息，达成了各语言通用的目标。在实验中，所提出方法不仅在测试数据集上的查准率、查全率、F1 值均优于其他基准对比方法，而且在整个中英文维基百科中应用后可生成大规模、高质量的实例类别信息。

6.2 工作展望

通过本文提出的面向社交站点的模式知识挖掘、跨语言概念匹配、实例类别推断的方法，可初步构建任意给定的两种语言对应的双语知识图谱，但这些方法依旧存在一些问题，可以在将来的工作中进行改进，具体如下：

1) 关于模式知识挖掘，目前的方法仅能够挖掘概念间的 *equal*、*subClassOf*、*relate*关系。在后续工作中需要考虑更多概念间的关系，如 *partOf*、*madeOf*、*cause*等关系。由此需设计更多的能够捕获不同关系特性的特征与规则，将三分类的机器学习任务扩展到 N（$N > 3$）分类，但是这无形中大大增加了特征选取的难度，因此可以考虑使用深度学习模型，并将所设计的规则融入到深度学习的过程中。此外，模式知识不仅包含概念间的关系，还有概念属性、属性定义域与值域等，所以如何面向社交站点挖掘其他种类的模式知识也是后续工作必须研究的问题。

2) 关于跨语言概念匹配，由于在跨语言概念匹配任务中表现最佳的基于概念关联关系的双语双词主题模型的复杂度较高，所以如何有效地降低该模型的训练时间是后续研究的主要问题。计划的解决方案分为三种方向。第一种是考虑将现有的 Gibbs

采样加速算法（如 FastLDA [150] 等）应用于基于概念关联关系的双语双词主题模型中。第二种是将 Gibbs 采样算法并行化，在这个过程中需要做近似处理，考虑采用 PLDA+ [151] 等算法中的解决方案。第三种是在不影响主题模型训练质量的前提下减少用于训练的双词的数量，此处需要对双词进行分类，并定义什么样的双词是有价值的，而什么样的双词是可以删除的。

3) 关于实例类别推断，目前的方法的主要问题在于相当数量的实例与概念的属性是缺失的。针对这一问题，计划采用文本挖掘的方法从维基百科中的文本以及万维网中的文本中抽取实例属性与概念属性，或众包的方法补全实例与概念的属性。此外，在实验中（于章节 5.4.1.4中介绍）发现，完全由拥有给定属性的概念的数量决定该属性的代表性并不合理，因为这样计算得到的属性的代表性易受给定语言的知识稀疏程度影响，所以设计一种更加合理的计算属性代表性的方法也是后续需要解决的问题。

除了方法的改进外，后续考虑将方法工具化，使得其他研究人员能够更加便利地构建双语知识图谱。此外，对于利用所提出方法构建的双语知识图谱，计划将其以链接数据的形式发布在万维网中，以进一步促进整个世界的知识分享。

参考文献

[1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web[J]. Scientific American Magazine, 2001, 284(October):34–43.

[2] Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far[J]. International Journal on Semantic Web and Information Systems, 2009, 5(3):1–22.

[3] Fernández J D, Beek W, Martínez-Prieto M A, et al. LOD-a-lot - A Queryable Dump of the LOD Cloud[C]. In: Proceedings of International Semantic Web Conference, Part II. 2017. 75–83.

[4] 漆桂林, 高桓, 吴天星. 知识图谱研究进展 [J]. 情报工程, 2017, 3(1):4–25.

[5] Hua W, Wang Z, Wang H, et al. Understand Short Texts by Harvesting and Analyzing Semantic Knowledge[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(3):499–512.

[6] Patel-Schneider P F. Analyzing Schema. org[C]. In: Proceedings of International Semantic Web Conference, Part I. 2014. 261–276.

[7] Ferrucci D A, Brown E W, Chu-Carroll J, et al. Building Watson: An Overview of the DeepQA Project[J]. AI Magazine, 2010, 31(3):59–79.

[8] Ma Y, Crook P A, Sarikaya R, et al. Knowledge Graph Inference for spoken dialog systems[C]. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. 2015. 5346–5350.

[9] Nickel M, Murphy K, Tresp V, et al. A Review of Relational Machine Learning for Knowledge Graphs[J]. Proceedings of the IEEE, 2016, 104(1):11–33.

[10] Fellbaum C. WordNet: An Electronic Lexical Database[M].[S.l.]: MIT Press, 1998.

[11] Lenat D B. CYC: A Large-Scale Investment in Knowledge Infrastructure[J]. Communications of the ACM, 1995, 38(11):32–38.

[12] Bollacker K D, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. In: Proceedings of ACM SIGMOD International Conference on Management of Data. 2008. 1247–1250.

[13] Dong Z, Dong Q, Hao C. HowNet and Its Computation of Meaning[C]. In: Proceedings of International Conference on Computational Linguistics. 2010. 53–56.

[14] 梅家驹. 同义词词林 [M].[S.l.]: 上海辞书出版社, 1983.

[15] Lehmann J, Isele R, Jakob M, et al. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2):167–195.

[16] Zaveri A, Kontokostas D, Sherif M A, et al. User-driven quality evaluation of DBpedia[C]. In: Proceedings of International Conference on Semantic Systems. 2013. 97–104.

[17] Mahdisoltani F, Biega J, Suchanek F M. YAGO3: A Knowledge Base from Multilingual Wikipedias[C]. In: Proceedings of Biennial Conference on Innovative Data Systems Research. 2015.

[18] Vrandecic D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10):78–85.

[19] Navigli R, Ponzetto S P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial Intelligence, 2012, 193:217–250.

[20] Niu X, Sun X, Wang H, et al. Zhishi.me - Weaving Chinese Linking Open Data[C]. In: Proceedings of International Semantic Web Conference, Part II. 2011. 205–220.

[21] Wang Z, Li J, Wang Z, et al. XLore: A Large-scale English-Chinese Bilingual Knowledge Graph[C]. In: Proceedings of International Semantic Web Conference, Posters & Demonstrations Track. 2013. 121–124.

[22] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[C]. In: Proceedings of International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, Part II. 2017. 428–438.

[23] Mitchell T M, Cohen W W, Jr. E R H, et al. Never-Ending Learning[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2015. 2302–2310.

[24] Wu W, Li H, Wang H, et al. Probase: a probabilistic taxonomy for text understanding[C]. In: Proceedings of ACM SIGMOD International Conference on Management of Data. 2012. 481–492.

[25] Speer R, Chin J, Havasi C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2017. 4444–4451.

[26] Tang J, Leung H, Luo Q, et al. Towards Ontology Learning from Folksonomies[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2009. 2089–2094.

[27] García-Silva A, Corcho Ó, Alani H, et al. Review of the state of the art: discovering and associating semantics to tags in folksonomies[J]. The Knowledge Engineering Review, 2012, 27(1):57–85.

[28] Wang H, Wu T, Qi G, et al. On Publishing Chinese Linked Open Schema[C]. In: Proceedings of International Semantic Web Conference, Part I. 2014. 293–308.

[29] Zhang L, Färber M, Rettinger A. XKnowSearch!: Exploiting Knowledge Bases for Entity-based Cross-lingual Information Retrieval[C]. In: Proceedings of ACM International Conference on Information and Knowledge Management. 2016. 2425–2428.

[30] Zhang L, Rettinger A. X-LiSA: Cross-lingual Semantic Annotation[J]. Proceedings of the VLDB Endowment, 2014, 7(13):1693–1696.

[31] Spohr D, Hollink L, Cimiano P. A Machine Learning Approach to Multilingual and Cross-Lingual Ontology Matching[C]. In: Proceedings of International Semantic Web Conference, Part I. 665–680.

[32] Boldyrev N, Spaniol M, Weikum G. ACROSS: A framework for multi-cultural interlinking of web taxonomies[C]. In: Proceedings of ACM Conference on Web Science. 2016. 127–136.

[33] Prytkova N, Weikum G, Spaniol M. Aligning Multi-Cultural Knowledge Taxonomies by Combinatorial Optimization[C]. In: Proceedings of International Conference on World Wide Web, Companion Volume. 2015. 93–94.

[34] de Melo G, Weikum G. MENTA: inducing multilingual taxonomies from wikipedia[C]. In: Proceedings of ACM Conference on Information and Knowledge Management. 2010. 1099–1108.

[35] Suchanek F M, Kasneci G, Weikum G. YAGO: A Large Ontology from Wikipedia and WordNet[J]. Journal of Web Semantics, 2008, 6(3):203–217.

[36] Sowa J F. Principles of semantic networks: Explorations in the representation of knowledge[M].[S.l.]: Morgan Kaufmann, 2014.

[37] Simmons R F, Bruce B C. Some Relations Between Predicate Calculus and Semantic Net Representations of Discourse.[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 1971. 2:524–529.

[38] Schubert L K. Extending the expressive power of semantic networks[J]. Artificial intelligence, 1976, 7(2):163–198.

[39] Schild K. A correspondence theory for terminological logics: Preliminary report[M].[S.l.]: Techn. Univ., 1991.

[40] Baader F. The description logic handbook: Theory, implementation and applications[M].[S.l.]: Cambridge university press, 2003.

[41] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic Web, 2017, 8(3):489–508.

[42] Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2):443–460.

[43] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1):3–26.

[44] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of machine learning research, 2003, 3(Feb):1083–1106.

[45] Zhou G, Su J, Zhang J, et al. Exploring Various Knowledge in Relation Extraction[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2005. 427–434.

[46] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2014. 1:402–412.

[47] Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2016. 1:1105–1116.

[48] Kushmerick N. Wrapper induction: Efficiency and expressiveness[J]. Artificial Intelligence, 2000, 118(1-2):15–68.

[49] Liu B. Web data mining: exploring hyperlinks, contents, and usage data[M].[S.l.]: Springer Science & Business Media, 2007.

[50] Muñoz E, Hogan A, Mileo A. Using linked data to mine RDF from wikipedia's tables[C]. In: Proceedings of ACM International Conference on Web Search and Data Mining. 2014. 533–542.

[51] Zhang Z. Effective and efficient Semantic Table Interpretation using TableMiner^{+}[J]. Semantic Web, 2017, 8(6):921–957.

[52] Bizer C, Seaborne A. D2RQ-treating non-RDF databases as virtual RDF graphs[C]. In: Proceedings of International Semantic Web Conference. 2004. 2004.

[53] Li M, Du X Y, Wang S. Learning ontology from relational database[C]. In: Proceedings of International Conference on Machine Learning and Cybernetics. 2005. 6:3410–3415.

[54] Astrova I. Rules for mapping SQL relational databases to OWL ontologies[M]//. In: Metadata and semantics.[S.l.]: Springer, 2009. 415–424.

[55] Santoso H A, Haw S C, Abdul-Mehdi Z T. Ontology extraction from relational database: Concept hierarchy as background knowledge[J]. Knowledge-Based Systems, 2011, 24(3):457–464.

[56] Euzenat J, Shvaiko P, et al. Ontology matching[M]. Vol. 18.[S.l.]: Springer, 2007.

[57] Jian N, Hu W, Cheng G, et al. Falcon-ao: Aligning ontologies with falcon[C]. In: Workshop on Integrating Ontologies (K-CAP). 2005. 85–91.

[58] Jain P, Hitzler P, Sheth A P, et al. Ontology alignment for linked open data[C]. In: Proceedings of International Semantic Web Conference. 2010. 402–417.

[59] Suchanek F M, Abiteboul S, Senellart P. Paris: Probabilistic alignment of relations, instances, and schema[J]. Proceedings of the VLDB Endowment, 2011, 5(3):157–168.

[60] Jiménez-Ruiz E, Grau B C. Logmap: Logic-based and scalable ontology matching[C]. In: Proceedings of International Semantic Web Conference. 2011. 273–288.

[61] Faria D, Pesquita C, Santos E, et al. The AgreementMakerLight Ontology Matching System[C]. In: Proceedings of OTM Conferences - Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE. 2013. 527–541.

[62] Niu X, Rong S, Wang H, et al. An effective rule miner for instance matching in a web of data[C]. In: Proceedings of ACM International Conference on Information and Knowledge Management. 2012. 1085–1094.

[63] Rong S, Niu X, Xiang E W, et al. A machine learning approach for instance matching based on similarity metrics[C]. In: Proceedings of International Semantic Web Conference. 2012. 460–475.

[64] Kejriwal M, Miranker D P. An unsupervised instance matcher for schema-free RDF data[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2015, 35:102–123.

[65] Nguyen K, Ichise R. ScLink: supervised instance matching system for heterogeneous repositories[J]. Journal of Intelligent Information Systems, 2017, 48(3):519–551.

[66] Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2017. 1511–1517.

[67] Sun Z, Hu W, Li C. Cross-lingual Entity Alignment via Joint Attribute-Preserving Embedding[C]. In: Proceedings of International Semantic Web Conference. 2017. 628–644.

[68] Wang Z, Li J, Wang Z, et al. Cross-lingual knowledge linking across wiki knowledge bases[C]. In: Proceedings of International Conference on World Wide Web. 2012. 459–468.

[69] Wang Z, Li J, Tang J. Boosting Cross-Lingual Knowledge Linking via Concept Annotation.[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2013. 2733–2739.

[70] Zhang Y, Paradis T, Hou L, et al. Cross-Lingual Infobox Alignment in Wikipedia Using Entity-Attribute Factor Graph[C]. In: Proceedings of International Semantic Web Conference. 2017. 745–760.

[71] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion[J]. Proceedings of the VLDB Endowment, 2014, 7(10):881–892.

[72] Dong X L, Gabrilovich E, Murphy K, et al. Knowledge-based trust: Estimating the trustworthiness of web sources[J]. Proceedings of the VLDB Endowment, 2015, 8(9):938–949.

[73] Wang H, Fang Z, Zhang L, et al. Effective online knowledge graph fusion[C]. In: Proceedings of International Semantic Web Conference. 2015. 286–302.

[74] Horrocks I. The FaCT System[C]. In: Proceedings of International Conference of Automated Reasoning with Analytic Tableaux and Related Methods. 1998. 307–312.

[75] Goodman E L, Jimenez E, Mizell D, et al. High-performance computing applied to semantic databases[C]. In: Proceedings of Extended Semantic Web Conference. 2011. 31–45.

[76] Urbani J, Kotoulas S, Maassen J, et al. OWL reasoning with WebPIE: calculating the closure of 100 billion triples[C]. In: Proceedings of Extended Semantic Web Conference. 2010. 213–227.

[77] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. In: Proceedings of Annual Conference on Neural Information Processing Systems. 2013. 2787–2795.

[78] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes.[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2014. 14:1112–1119.

[79] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion.[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2015. 15:2181–2187.

[80] Quinlan J R. Learning logical definitions from relations[J]. Machine learning, 1990, 5(3):239–266.

[81] Galárraga L, Teflioudi C, Hose K, et al. Fast rule mining in ontological knowledge bases with AMIE + +[J]. The VLDB Journal, 2015, 24(6):707–730.

[82] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base[C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011. 529–539.

[83] Ponzetto S P, Strube M. WikiTaxonomy: A Large Scale Knowledge Resource[C]. In: Proceedings of European Conference on Artificial Intelligence. 2008. 751–752.

[84] Wu F, Weld D S. Automatically refining the wikipedia infobox ontology[C]. In: Proceedings of International Conference on World Wide Web. 2008. 635–644.

[85] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora[C]. In: Proceedings of International Conference on Computational Linguistics. 1992. 539–545.

[86] Zhou M, Bao S, Wu X, et al. An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations[C]. In: Proceedings of International Semantic Web Conference and Asian Semantic Web Conference. 2007. 680–693.

[87] Lin H, Davis J G, Zhou Y. An Integrated Approach to Extracting Ontological Structures from Folksonomies[C]. In: Proceedings of European Semantic Web Conference. 2009. 654–668.

[88] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia - A crystallization point for the Web of Data[J]. Journal of Web Semantics, 2009, 7(3):154–165.

[89] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194:28–61.

[90] Smeros P, Gupta A, Catasta M, et al. deepschema.org: An Ontology for Typing Entities in the Web of Data[C]. In: Workshop on Linked Data on the Web. 2017.

[91] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis.[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2007. 7:1606–1611.

[92] Wu Z, Palmer M. Verbs semantics and lexical selection[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 1994. 133–138.

[93] Baeza-Yates R, Ribeiro-Neto B, et al. Modern information retrieval[M]. Vol. 463.[S.l.]: ACM press New York, 1999.

[94] Shen D, Qin M, Chen W, et al. Mining web query hierarchies from clickthrough data[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2007. 7:341–346.

[95] Nigam K, Ghani R. Analyzing the Effectiveness and Applicability of Co-training[C]. In: Proceedings of ACM Conference on Information and Knowledge Management. 2000. 86–93.

[96] Vapnik V. The nature of statistical learning theory[M].[S.l.]: Springer science & business media, 2013.

[97] Delgado M F, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems?[J]. Journal of Machine Learning Research, 2014, 15(1):3133–3181.

[98] Ponzetto S P, Strube M. Deriving a large scale taxonomy from Wikipedia[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2007. 7:1440–1445.

[99] Klein D, Manning C D. Fast exact inference with a factored model for natural language parsing[C]. In: Proceedings of Annual Conference on Neural Information Processing Systems. 2003. 3–10.

[100] Collins M. Head-driven statistical models for natural language parsing[J]. Computational linguistics, 2003, 29(4):589–637.

[101] Wu T, Ling S, Qi G, et al. Mining type information from chinese online encyclopedias[C]. In: Proceedings of Joint International Semantic Technology Conference. 2014. 213–229.

[102] Qiu X, Zhang Q, Huang X. Fudannlp: A toolkit for chinese natural language processing[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2013. 49–54.

[103] Brown L D, Cai T T, DasGupta A. Interval estimation for a binomial proportion[J]. Statistical science, 2001:101–117.

[104] Do H H, Rahm E. Matching large schemas: Approaches and evaluation[J]. Information Systems, 2007, 32(6):857–885.

[105] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching[J]. The VLDB Journal, 2001, 10(4):334–350.

[106] Berlin J, Motro A. Database schema matching using machine learning with feature selection[C]. In: Proceedings of International Conference on Advanced Information Systems Engineering. 2002. 452–466.

[107] Do H H, Melnik S, Rahm E. Comparison of schema matching evaluations[C]. In: Proceedings of International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World. 2002. 221–237.

[108] Trojahn C, Fu B, Zamazal O, et al. State-of-the-art in multilingual and cross-lingual ontology matching[M]//. In: Towards the Multilingual Semantic Web.[S.l.]: Springer, 2014. 119–135.

[109] Hofmann T. Probabilistic latent semantic analysis[C]. In: Proceedings of Conference on Uncertainty in Artificial Intelligence. 1999. 289–296.

[110] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan):993–1022.

[111] Vulić I, De Smet W, Tang J, et al. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications[J]. Information Processing & Management, 2015, 51(1):111–147.

[112] Hong L, Davison B D. Empirical study of topic modeling in twitter[C]. In: Proceedings of Workshop on Social Media Analytics. 2010. 80–88.

[113] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]. In: Proceedings of Conference on Uncertainty in Artificial Intelligence. 2004. 487–494.

[114] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. 2009. 248–256.

[115] Li S, Li J, Pan R. Tag-Weighted Topic Model for Mining Semi-Structured Documents[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2013. 2855–2861.

[116] Li S, Huang G, Tan R, et al. Tag-weighted dirichlet allocation[C]. In: Proceedings of International Conference on Data Mining. 2013. 438–447.

[117] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]. In: Proceedings of International Conference on World Wide Web. 2013. 1445–1456.

[118] Cheng X, Yan X, Lan Y, et al. BTM: Topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12):2928–2941.

[119] Liu J S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem[J]. Journal of the American Statistical Association, 1994, 89(427):958–966.

[120] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet[C]. In: Proceedings of European Conference on Artificial Intelligence. 2004. 16:1089.

[121] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 1995. 448–453.

[122] Joachims T. Optimizing search engines using clickthrough data[C]. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002. 133–142.

[123] Kliegr T. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery[J]. Journal of Web Semantics, 2015, 31:59–69.

[124] Auer S, Lehmann J. What have innsbruck and leipzig in common? extracting semantics from wiki content[C]. In: Proceedings of European Semantic Web Conference. 2007. 503–517.

[125] Gangemi A, Nuzzolese A G, Presutti V, et al. Automatic typing of DBpedia entities[C]. In: Proceedings of International Semantic Web Conference. 2012. 65–81.

[126] Nuzzolese A G, Gangemi A, Presutti V, et al. Type inference through the analysis of Wikipedia links[C]. In: Workshop on Linked Data on the Web. 2012.

[127] Paulheim H, Bizer C. Type inference on noisy rdf data[C]. In: Proceedings of International Semantic Web Conference. 2013. 510–525.

[128] Paulheim H, Bizer C. Improving the quality of linked data using statistical distributions[J]. International Journal on Semantic Web and Information Systems, 2014, 10(2):63–86.

[129] Kliegr T, Zamazal O. LHD 2.0: A text mining approach to typing entities in knowledge graphs[J]. Journal of Web Semantics, 2016, 39:47–61.

[130] Melo A, Völker J, Paulheim H. Type prediction in noisy RDF knowledge bases using hierarchical multilabel classification with graph and latent features[J]. International Journal on Artificial Intelligence Tools, 2017, 26(02):1760011.

[131] Lin T, Etzioni O, et al. No noun phrase left behind: detecting and typing unlinkable entities[C]. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 893–903.

[132] Nakashole N, Tylenda T, Weikum G. Fine-grained semantic typing of emerging entities[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2013. 1:1488–1497.

[133] Fleischman M, Hovy E. Fine grained classification of named entities[C]. In: Proceedings of International Conference on Computational Linguistics. 2002. 1–7.

[134] Rahman A, Ng V. Inducing fine-grained semantic classes via hierarchical and collective classification[C]. In: Proceedings of International Conference on Computational Linguistics. 2010. 931–939.

[135] Ling X, Weld D S. Fine-Grained Entity Recognition.[C]. In: Proceedings of AAAI Conference on Artificial Intelligence. 2012. 94–100.

[136] Yosef M A, Bauer S, Hoffart J, et al. Hyena: Hierarchical type classification for entity names[J]. Proceedings of International Conference on Computational Linguistics, 2012:1361–1370.

[137] Yosef M A, Bauer S, Hoffart J, et al. Hyena-live: Fine-grained online entity type classification from natural-language text[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2013. 133–138.

[138] Corro L D, Abujabal A, Gemulla R, et al. FINET: Context-Aware Fine-Grained Named Entity Typing[C]. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. 2015. 868–878.

[139] Ren X, He W, Qu M, et al. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding[C]. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. 2016. 1369–1378.

[140] Ma Y, Cambria E, Gao S. Label embedding for zero-shot fine-grained named entity typing[C]. In: Proceedings of International Conference on Computational Linguistics. 2016. 171–180.

[141] Abhishek A, Anand A, Awekar A. Fine-grained entity type classification by jointly learning representations and label embeddings[C]. In: Proceedings of Conference of the European Chapter of the Association for Computational Linguistics. 2017. 1:797–807.

[142] Paşca M. Organizing and searching the world wide web of facts–step two: harnessing the wisdom of the crowds[C]. In: Proceedings of International Conference on World Wide Web. 2007. 101–110.

[143] Pasca M, Van Durme B. What You Seek Is What You Get: Extraction of Class Attributes from Query Logs[C]. In: Proceedings of International Joint Conference on Artificial Intelligence. 2007. 7:2832–2837.

[144] Pasca M, Durme B V. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs[C]. In: Proceedings of Annual Meeting of the Association for Computational Linguistics. 2008. 19–27.

[145] Lee T, Wang Z, Wang H, et al. Attribute extraction and scoring: A probabilistic approach[C]. In: Proceedings of IEEE International Conference on Data Engineering. 2013. 194–205.

[146] Meyer B. Object-oriented software construction[M]. Vol. 2.[S.l.]: Prentice hall New York, 1988.

[147] Dowty D R, Wall R, Peters S. Introduction to Montague semantics[M]. Vol. 11.[S.l.]: Springer Science & Business Media, 2012.

[148] Stojanovic L. Methods and tools for ontology evolution[D]:[PhD Thesis]. Karlsruhe, Germany: Karlsruhe Institute of Technology, 2004.

[149] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]. In: Workshop of International Conference on Learning Representations. 2013.

[150] Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008. 569–577.

[151] Liu Z, Zhang Y, Chang E Y, et al. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):26.

附录 A　公式推导

为了保证前述章节的可读性，将部分公式的推导细节移到附录部分，详见下文。

A.1 BiBTM 中 Gibbs 采样公式的推导

给定一个双词 b，在 Gibbs 采样中需计算 $P(z_b|z_{\neg b}, \mathbf{B})$，该条件概率的推导过程如下：

$$P(z_b|z_{\neg b}, \mathbf{B}) = \frac{P(z, \mathbf{B})}{P(z_{\neg b}, \mathbf{B})} = \frac{P(\mathbf{B}|z)P(z)}{P(\mathbf{B}_{\neg b}|z_{\neg b})P(z_{\neg b})P(b)} \propto \frac{P(\mathbf{B}|z)P(z)}{P(\mathbf{B}_{\neg b}|z_{\neg b})P(z_{\neg b})} \tag{A.1}$$

依据公式 A.1，需分别计算 $P(\mathbf{B}|z)$、$P(\mathbf{B}_{\neg b}|z_{\neg b})$、$P(z)$、$P(z_{\neg b})$。此处先令 n_k 表示被赋予主题 k 的双词的数量，$n_{\neg b,k}$ 表示除双词 b 外被赋予主题 k 的双词的数量，再对 $\boldsymbol{\theta}$ 积分可得 $P(z)$ 如下：

$$\begin{aligned} P(z) = \int P(z|\boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &= \int (\prod_{\mathbf{B}} P(z_b|\boldsymbol{\theta}))P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \int \prod_{k=1}^{K} \theta_k^{n_k} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_k^{\alpha-1}\mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_k^{n_k+\alpha-1}\mathrm{d}\boldsymbol{\theta} \\ &= \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \cdot \frac{\prod_{k=1}^{K} \Gamma(n_k+\alpha)}{\Gamma(|\mathbf{B}|+K\alpha)} \end{aligned} \tag{A.2}$$

其中 $\Gamma(\cdot)$ 为 Gamma 函数[1]。与公式 A.2类似，可计算 $P(z_{\neg b})$ 如下：

$$P(z_{\neg b}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \cdot \frac{\prod_{k=1}^{K} \Gamma(n_{\neg b,k}+\alpha)}{\Gamma(|\mathbf{B}|-1+K\alpha)} \tag{A.3}$$

此处令 $n_{w^s|k}$ 表示语言为 s 的词 w^s 被赋予主题 k 的次数，$n_{\cdot^s|k} = \sum_{w^s} n_{w^s|k}$，$n_{\neg b,w^s|k}$ 表示除双词 b 以外语言为 s 的词 w^s 被赋予主题 k 的次数，$n_{\neg b,\cdot^s|k} = \sum_{w^s} n_{\neg b,w^s|k}$，$n_{w^t|k}$ 表示语言为 t 的词 w^t 被赋予主题 k 的次数，$n_{\cdot^t|k} = \sum_{w^t} n_{w^t|k}$，$n_{\neg b,w^t|k}$ 表示除双词 b 以外语言为 s 的词 w^t 被赋予

[1] http://en.wikipedia.org/wiki/Gamma_function

主题 k 的次数，$n_{\neg b,\cdot^t|k}=\sum_{w^t} n_{\neg b,w^t|k}$。基于此，通过对 $\boldsymbol{\varphi}^s$ 与 $\boldsymbol{\varphi}^t$ 积分计算 $P(\mathbf{B}|z)$ 如下：

$$\begin{aligned}
P(\mathbf{B}|z) &= \iint P(\mathbf{B}^s|z^s,\boldsymbol{\varphi}^s)P(\mathbf{B}^{st}|z^{st},\boldsymbol{\varphi}^s,\boldsymbol{\varphi}^t)P(\mathbf{B}^t|z^t,\boldsymbol{\varphi}^t)P(\boldsymbol{\varphi}^s)P(\boldsymbol{\varphi}^t)d\boldsymbol{\varphi}^s d\boldsymbol{\varphi}^t \\
&= \prod_{k=1}^{K}\iint \prod_{i=1}^{|\mathbf{B}|} P(b_i|z_i=k,\varphi^s_{i,k},\varphi^t_{i,k})P(\boldsymbol{\varphi}^s)P(\boldsymbol{\varphi}^t)d\boldsymbol{\varphi}^s d\boldsymbol{\varphi}^t \\
&= \prod_{k=1}^{K}\iint \prod_{w^s}{\varphi^s_{w^s,k}}^{n_{w^s|k}} \prod_{w^t}{\varphi^t_{w^t,k}}^{n_{w^t|k}} \frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}} \prod_{w^s}{\varphi^s_{w^s,k}}^{\beta-1} \frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}} \prod_{w^t}{\varphi^t_{w^t,k}}^{\beta-1} d\boldsymbol{\varphi}^s d\boldsymbol{\varphi}^t \\
&= \left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K \cdot \prod_{k=1}^{K}\iint \prod_{w^s}{\varphi^s_{w^s,k}}^{n_{w^s|k}+\beta-1} \prod_{w^t}{\varphi^t_{w^t,k}}^{n_{w^t|k}+\beta-1} d\boldsymbol{\varphi}^s d\boldsymbol{\varphi}^t \\
&= \left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K \cdot \prod_{k=1}^{K}\int \prod_{w^s}{\varphi^s_{w^s,k}}^{n_{w^s|k}+\beta-1} d\boldsymbol{\varphi}^s \prod_{k=1}^{K}\int \prod_{w^t}{\varphi^t_{w^t,k}}^{n_{w^t|k}+\beta-1} d\boldsymbol{\varphi}^t \\
&= \left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K \cdot \prod_{k=1}^{K}\frac{\prod_{w^s}\Gamma\left(n_{w^s|k}+\beta\right)}{\Gamma\left(n_{\cdot^s|k}+W^s\beta\right)}\cdot\frac{\prod_{w^t}\Gamma\left(n_{w^t|k}+\beta\right)}{\Gamma\left(n_{\cdot^t|k}+W^t\beta\right)}
\end{aligned} \tag{A.4}$$

与公式 A.4类似，可计算 $P(\mathbf{B}_{\neg b}|z_{\neg b})$ 如下：

$$P(\mathbf{B}_{\neg b}|z_{\neg b}) = \left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K \cdot \prod_{k=1}^{K}\frac{\prod_{w^s}\Gamma\left(n_{\neg b,w^s|k}+\beta\right)}{\Gamma\left(n_{\neg b,\cdot^s|k}+W^s\beta\right)}\cdot\frac{\prod_{w^t}\Gamma\left(n_{\neg b,w^t|k}+\beta\right)}{\Gamma\left(n_{\neg b,\cdot^t|k}+W^t\beta\right)} \tag{A.5}$$

针对三种不同类型的双词 $b^s_i\in\mathbf{B}^s$、$b^{st}_i\in\mathbf{B}^{st}$、$b^t_i\in\mathbf{B}^t$，将公式 A.2、A.3、A.4、A.5带入到公式 A.1中，可得：

$$P(z^s_i=k|z_{\neg b^s_i},\mathbf{B}) \propto (n_{\neg b^s_i,k}+\alpha)\frac{(n_{\neg b^s_i,w^s_{i,1}|k}+\beta)(n_{\neg b^s_i,w^s_{i,2}|k}+\beta)}{(n_{\neg b^s_i,\cdot^s|k}+1+W^s\beta)(n_{\neg b^s_i,\cdot^s|k}+W^s\beta)} \tag{A.6}$$

$$P(z^{st}_i=k|z_{\neg b^{st}_i},\mathbf{B}) \propto (n_{\neg b^{st}_i,k}+\alpha)\frac{(n_{\neg b^{st}_i,w^s_{i,1}|k}+\beta)(n_{\neg b^{st}_i,w^t_{i,2}|k}+\beta)}{(n_{\neg b^{st}_i,\cdot^s|k}+W^s\beta)(n_{\neg b^{st}_i,\cdot^t|k}+W^t\beta)} \tag{A.7}$$

$$P(z^t_i=k|z_{\neg b^t_i},\mathbf{B}) \propto (n_{\neg b^t_i,k}+\alpha)\frac{(n_{\neg b^t_i,w^t_{i,1}|k}+\beta)(n_{\neg b^t_i,w^t_{i,2}|k}+\beta)}{(n_{\neg b^t_i,\cdot^t|k}+1+W^t\beta)(n_{\neg b^t_i,\cdot^t|k}+W^t\beta)} \tag{A.8}$$

A.2 BiBTM 中参数 θ_k、φ^s_{k,w^s}、φ^t_{k,w^t} 的估计

给定超参数 α 与 β，所有双词所在集合 $\mathbf{B}$，以及经过多次迭代后所有双词的主题赋值 z，依据狄利克雷分布（Dirichlet Distribution）与多项式分布（Multinomial Distribution）的共轭关系及贝叶斯法则可推导 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}^s_k$、$\boldsymbol{\varphi}^t_k$ 的概率分布如下：

$$P(\boldsymbol{\theta}|z,\alpha) = \frac{1}{Z_{\boldsymbol{\theta}}}\prod_{i=1}^{|\mathbf{B}|}P(z_i|\boldsymbol{\theta})P(\boldsymbol{\theta}|\alpha) = Dirichlet(\boldsymbol{\theta}+\boldsymbol{n}) \tag{A.9}$$

$$P(\boldsymbol{\varphi}^s_k|z,\mathbf{B},\beta) = \frac{1}{Z_{\boldsymbol{\varphi}^s_k}}\prod_{\{i:z=k\}}P(w^s_i|\boldsymbol{\varphi}^s_k)P(\boldsymbol{\varphi}^s_k|\beta) = Dirichlet(\beta+\boldsymbol{n}^{\boldsymbol{s}}_{\boldsymbol{k}}) \tag{A.10}$$

$$P(\boldsymbol{\varphi}^t_k|z,\mathbf{B},\beta) = \frac{1}{Z_{\boldsymbol{\varphi}^t_k}}\prod_{\{i:z=k\}}P(w^t_i|\boldsymbol{\varphi}^t_k)P(\boldsymbol{\varphi}^t_k|\beta) = Dirichlet(\beta+\boldsymbol{n}^{\boldsymbol{t}}_{\boldsymbol{k}}) \tag{A.11}$$

其中 $Z_{\boldsymbol{\theta}}$、$Z_{\boldsymbol{\varphi}_k^s}$、$Z_{\boldsymbol{\varphi}_k^t}$ 为归一化因子，$Dirichlet(\cdot)$ 表示狄利克雷分布。$\boldsymbol{n}$ 是一个 K 维向量，其中第 k 项为 n_k；$\boldsymbol{n_k^s}$ 是一个 W^s 维的向量，其中每一项可表示为 $n_{w^s|k}$；$\boldsymbol{n_k^t}$ 是一个 W^t 维的向量，其中每一项可表示为 $n_{w^t|k}$。此处可依据狄利克雷分布的期望估计 θ_k、φ^s_{k,w^s}、φ^t_{k,w^t} 如下：

$$\theta_k = \frac{\alpha + n_k}{K\alpha + |\mathbf{B}|} \tag{A.12}$$

$$\varphi^s_{k,w^s} = \frac{\beta + n_{w^s|k}}{W^s\beta + n_{.^s|k}} \tag{A.13}$$

$$\varphi^t_{k,w^t} = \frac{\beta + n_{w^t|k}}{W^t\beta + n_{.^t|k}} \tag{A.14}$$

A.3 CC-BiBTM 中 Gibbs 采样公式的推导

给定一个双词 b，在 Gibbs 采样中需计算 $P(x_b, z_b|z_{\neg b}, x_{\neg b}, \mathbb{O})$，该条件概率的推导过程如下：

$$\begin{aligned} P(x_b, z_b|z_{\neg b}, x_{\neg b}, \mathbb{O}) &= P(x_b, z_b|z_{\neg b}, x_{\neg b}, \mathbf{B}) \\ &= \frac{P(z, x, \mathbf{B})}{P(z_{\neg b}, x_{\neg b}, \mathbf{B})} \\ &= \frac{P(\mathbf{B}|z)P(z|x)P(x)}{P(\mathbf{B}_{\neg b}|z_{\neg b})P(z_{\neg b}|x_{\neg b})P(x_{\neg b})P(b)} \\ &\propto \frac{P(\mathbf{B}|z)P(z|x)P(x)}{P(\mathbf{B}_{\neg b}|z_{\neg b})P(z_{\neg b}|x_{\neg b})P(x_{\neg b})} \end{aligned} \tag{A.15}$$

令 $\pi_{b,c}$ 为双词 b 对应的关于概念 c 的先验概念概率，计算 $P(x)/p(x_{\neg b})$ 如下：

$$\frac{P(x)}{p(x_{\neg b})} = \pi_{b,c} \tag{A.16}$$

此处令 n_c 为被赋予概念 c 的双词的数量，$n_{\neg b,c}$ 表示除双词 b 以外被赋予概念 c 的双词的数量，$n_{k|c}$ 为被同时赋予概念 c 与主题 k 的双词的数量，$n_{\neg b,k|c}$ 表示除双词 b 以外被同时赋予概念 c 与主题 k 的双词的数量，$\boldsymbol{\theta}$ 表示一个 $C \times K$ 的矩阵，第 c 行即为概念 -主题分布 $\boldsymbol{\theta}_c$，之后对 $\boldsymbol{\theta}$ 积分可得 $P(z|x)$ 如下：

$$\begin{aligned} P(z|x) &= \int P(z|x, \boldsymbol{\theta})P(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta} \\ &= \int (\prod_{b\in\mathbf{B}} P(z_b|x_b, \boldsymbol{\theta}_{x_b}))P(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int \prod_{c=1}^{C}\prod_{k=1}^{K} \theta_{c,k}^{n_{k|c}} \cdot \prod_{c=1}^{C} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \cdot \prod_{k=1}^{K} \theta_{c,k}^{\alpha-1} d\boldsymbol{\theta}_c \\ &= \prod_{c=1}^{C} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \int \prod_{k=1}^{K} \theta_{c,k}^{n_{k|c}+\alpha-1} d\boldsymbol{\theta}_c \\ &= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^C \prod_{c=1}^{C} \frac{\prod_{k=1}^{K}\Gamma(n_{k|c}+\alpha)}{\Gamma(n_c + K\alpha)} \end{aligned} \tag{A.17}$$

与公式 A.17类似，可计算 $P(z_{\neg b}|x_{\neg b})$ 如下：

$$P(z_{\neg b}|x_{\neg b}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^C \prod_{c=1}^{C} \frac{\prod_{k=1}^{K}\Gamma(n_{\neg b,k|c}+\alpha)}{\Gamma(n_{\neg b,c} + K\alpha)} \tag{A.18}$$

关于 $P(\mathbf{B}|z)$ 与 $P(\mathbf{B}_{\neg b}|z_{\neg b})$ 的计算方式参见公式 A.4与公式 A.5，具体如下：

$$P(\mathbf{B}|z)=\left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K\cdot\prod_{k=1}^{K}\frac{\prod_{w^s}\Gamma\left(n_{w^s|k}+\beta\right)}{\Gamma\left(n_{\cdot^s|k}+W^s\beta\right)}\cdot\frac{\prod_{w^t}\Gamma\left(n_{w^t|k}+\beta\right)}{\Gamma\left(n_{\cdot^t|k}+W^t\beta\right)} \tag{A.19}$$

$$P(\mathbf{B}_{\neg b}|z_{\neg b})=\left(\frac{\Gamma(W^s\beta)}{\Gamma(\beta)^{W^s}}\cdot\frac{\Gamma(W^t\beta)}{\Gamma(\beta)^{W^t}}\right)^K\cdot\prod_{k=1}^{K}\frac{\prod_{w^s}\Gamma\left(n_{\neg b,w^s|k}+\beta\right)}{\Gamma\left(n_{\neg b,\cdot^s|k}+W^s\beta\right)}\cdot\frac{\prod_{w^t}\Gamma\left(n_{\neg b,w^t|k}+\beta\right)}{\Gamma\left(n_{\neg b,\cdot^t|k}+W^t\beta\right)} \tag{A.20}$$

针对三种不同类型的双词 $b_i^s\in\mathbf{B}^s$、$b_i^{st}\in\mathbf{B}^{st}$、$b_i^t\in\mathbf{B}^t$，将公式 A.16、A.17、A.18、A.19、A.20带入到公式 A.15中，可得：

$$\begin{aligned}P(x_i^s=c,z_i^s=k|x_{\neg b_i^s},z_{\neg b_i^s},\mathbb{O})\propto\pi_{i,c}^s\cdot\frac{(n_{\neg b_i^s,k|c}+\alpha)}{(n_{\neg b_i^s,\cdot|c}+K\alpha)}\cdot\\\frac{(n_{\neg b_i^s,w_{i,1}^s|k}+\beta)(n_{\neg b_i^s,w_{i,2}^s|k}+\beta)}{(n_{\neg b_i^s,\cdot^s|k}+W^s\beta)(n_{\neg b_i^s,\cdot^s|k}+1+W^s\beta)}\end{aligned} \tag{A.21}$$

$$\begin{aligned}P(x_i^{st}=c,z_i^{st}=k|x_{\neg b_i^{st}},z_{\neg b_i^{st}},\mathbb{O})\propto\pi_{i,c}^{st}\cdot\frac{(n_{\neg b_i^{st},k|c}+\alpha)}{(n_{\neg b_i^{st},\cdot|c}+K\alpha)}\cdot\\\frac{(n_{\neg b_i^{st},w_{i,1}^s|k}+\beta)(n_{\neg b_i^{st},w_{i,2}^t|k}+\beta)}{(n_{\neg b_i^{st},\cdot^s|k}+W^s\beta)(n_{\neg b_i^{st},\cdot^t|k}+W^t\beta)}\end{aligned} \tag{A.22}$$

$$\begin{aligned}P(x_i^t=c,z_i^t=k|x_{\neg b_i^t},z_{\neg b_i^t},\mathbb{O})\propto\pi_{i,c}^t\cdot\frac{(n_{\neg b_i^t,k|c}+\alpha)}{(n_{\neg b_i^t,\cdot|c}+K\alpha)}\cdot\\\frac{(n_{\neg b_i^t,w_{i,1}^t|k}+\beta)(n_{\neg b_i^t,w_{i,2}^t|k}+\beta)}{(n_{\neg b_i^t,\cdot^t|k}+W^t\beta)(n_{\neg b_i^t,\cdot^t|k}+1+W^t\beta)}\end{aligned} \tag{A.23}$$

A.4 CC-BiBTM 中参数 $\theta_{c,k}$ 的估计

给定超参数 α 与 β，语料 $\mathbb{O}$ 中所有双词所在集合 $\mathbf{B}$ 与 C 个源自两个不同语言层次分类体系的所有概念，以及经过多次迭代后所有双词的概念赋值 x 与主题赋值 z。依据狄利克雷分布与多项式分布的共轭关系及贝叶斯法则可推导 $\boldsymbol{\theta}_c$ 的概率分布如下：

$$P(\boldsymbol{\theta}_c|z,x,\alpha)=\frac{1}{Z_{\boldsymbol{\theta}_c}}\prod_{i=1}^{|\mathbf{B}|}P(z_i|\boldsymbol{\theta}_c)P(\boldsymbol{\theta}_c|\alpha)=Dirichlet(\boldsymbol{\theta}_c+\boldsymbol{n_c}) \tag{A.24}$$

其中 $Z_{\boldsymbol{\theta}_c}$ 为归一化因子，$\boldsymbol{n_c}$ 是一个 K 维向量，其中第 k 项为 $n_{k|c}$。然后依据狄利克雷分布的期望估计 $\theta_{c,k}$ 如下：

$$\theta_{c,k}=\frac{\alpha+n_{k|c}}{K\alpha+n_c} \tag{A.25}$$

作者简介 (包括论文和成果清单)

博士期间学术论文成果

1) **Tianxing Wu**, Guilin Qi, Haofen Wang, Kang Xu and Xuan Cui. Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic Model[C]. Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2016, 287-293. (CCF A 类会议，EI 索引)

2) **Tianxing Wu**, Haofen Wang, Guilin Qi, Jiangang Zhu and Tong Ruan. On Building and Publishing Linked Open Schema from Social Web Sites[J]. Journal of Web Semantics (JWS), 2018, 51: 39-50. (CCF B 类期刊，SCI 索引)

3) **Tianxing Wu**, Lei Zhang, Guilin Qi, Xuan Cui and Kang Xu. Encoding Category Correlations into Bilingual Topic Modeling for Cross-Lingual Taxonomy Alignment[C]. Proceedings of International Semantic Web Conference (ISWC), Part I, 2017, 738-744. (CCF B 类会议，EI 索引)

4) **Tianxing Wu**, Guilin Qi, Bin Luo, Lei Zhang and Haofen Wang. Language-Independent Type Inference of the Instances from Multilingual Wikipedia[J]. To Appear in International Journal on Semantic Web and Information Systems (IJSWIS), 2019. (CCF C 类期刊，SCI 索引)

5) **Tianxing Wu**, Du Zhang, Lei Zhang and Guilin Qi. Cross-Lingual Taxonomy Alignment with Bilingual Knowledge Graph Embeddings[C]. Proceedings of Joint International Semantic Technology Conference (JIST), 2017, 251-258. (EI 索引)

6) **Tianxing Wu**, Cong Gao, Guilin Qi, Lei Zhang, Chuanqi Dong, He Liu and Du Zhang. KG-Buddhism: The Chinese Knowledge Graph on Buddhism[C]. Proceedings of Joint International Semantic Technology Conference (JIST), 2017, 259-267. (EI 索引)

7) **Tianxing Wu**, Shengjia Yan, Zhixin Piao, Liang Xu, Ruiming Wang and Guilin Qi. Entity Linking in Web Tables with Multiple Linked Knowledge Bases[C]. Proceedings of Joint International Semantic Technology Conference (JIST), 2016, 239-253. (EI 索引)

8) Kang Xu, Feng Liu, **Tianxing Wu**, Sheng Bi and Guilin Qi. A Fast and Effective Framework for Lifelong Topic Model with Self-learning Knowledge[C]. Proceedings of China National Conference on Computational Linguistics (CCL), 2017, 147-158. (EI 索引)

9) 漆桂林, 高桓, **吴天星**. 知识图谱研究进展 [J]. 情报工程, 2017, 3(1): 4-25.

10) Kang Xu, Junheng Huang and **Tianxing Wu**. A Sentiment and Topic Model with Timeslice, User and Hashtag for Posts on Social Media[C]. Proceedings of China Conference on Knowledge Graph and Semantic Computing (CCKS), 2017, 59-65. (EI 索引)

11) Kang Xu, Guilin Qi, Junheng Huang and **Tianxing Wu**. A Joint Model for Sentiment-Aware Topic Detection on Social Media[C]. Proceedings of European Conference on Artificial Intelligence(ECAI), 2016, 338-346. (CCF B 类会议，EI 索引)

12) Kang Xu, Guilin Qi, Junheng Huang, **Tianxing Wu** and Xuefeng Fu. Detecting bursts in sentiment-aware topics from social media[J]. Knowledge Based Systems (KBS), 2017, 141: 44-54. (CCF C 类期刊，SCI 索引)

13) Kang Xu, Guilin Qi, Junheng Huang and **Tianxing Wu**. Incorporating Wikipedia concepts and categories as prior knowledge into topic models[J]. Intelligent Data Analysis (IDA), 2017, 21(2): 443-461. (CCF C 类期刊，SCI 索引)

博士期间申请国家发明专利

1) **吴天星**，漆桂林，罗斌，陆彬. 一种基于机器学习的图书本体匹配方法. 申请号：201410799922.3（已授权）.

2) **吴天星**，李丞，漆桂林. 一种基于机器学习的社交网络本体构建方法. 申请号：201610115254.7（待授权）.

3) **吴天星**，漆桂林，刘太云，严晟嘉，朴智新，许亮，王瑞明. 一种基于多知识库的表格实体链接方法. 申请号：201610920031.8（待授权）.

4) 方一曙，漆桂林，**吴天星**. 一种基于机器学习的跨语言分类结构匹配方法. 申请号：201510105414.5（已授权）.

5) 花云程，漆桂林，**吴天星**，高桓. 基于机器学习的本体匹配方法和系统. 申请号：201610595524.9（待授权）.

6) 漆桂林，崔轩，**吴天星**. 一种基于主题模型的跨语言层次分类体系匹配方法. 申请号：201710441927.2（待授权）.

7) 漆桂林，李丞，李林，**吴天星**. 一种知识图谱动态更新方法. 申请号：201810627957.7（待授权）.

博士期间参与的研究项目

1) 2015 年 1 月 - 2017 年 12 月，开放域知识集成、推理与检索关键技术及系统，国家 863 项目。

2) 2013 年 1 月 - 2016 年 12 月，描述逻辑中的本体融合方法研究，国家自然科学基金面上项目。

3) 2015 年 1 月 - 2015 年 9 月，互联网语义分析，企业横向项目。

致　谢

现在的时间是 2018 年 8 月 11 日下午，我在新加坡南洋理工大学计算机学院 DMAL 实验室撰写这篇致谢，心中万分感慨时光的力量，仿佛一切都在昨天，博士生涯 3 年有余，恍然便要划下句点。虽从不曾想过会将学生生涯延续至此，却从未后悔走上科研这条道路，一路走来，十分清楚自我性格中的缺陷，强迫、敏感、严苛，但又有幸有这一群支持、理解、包容我的同伴。对，是同伴，些许感谢之言虽不能完全抒发此时的情感，但我始终心怀感恩。

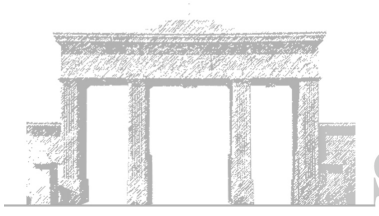
导师漆桂林教授，在我心中始终亦师亦友，回想当初选择攻读博士的原因，老师这个人应是最为主要的原因之一，老师的人品与自律令我敬重，老师学术水平令我向往。这些年来，老师是我科研路上的指明灯，从如何做研究，到如何写论文，再到如何作报告，老师毫无保留，倾囊相授，才使我慢慢能够独立进行研究，进而取得一些成绩。老师不仅在学术研究上给予我莫大的帮助，还提供大量的出外交流访问的机会，美国、德国、荷兰、奥地利，在这些地方的学术交流都极大地开阔了我的眼界，令我迅速成长。更为重要的是，当我面临生活与家庭的重要事件时，老师的充分理解与支持，更是温暖了我的内心。感恩、感谢！

实验室里的师兄弟姐妹，能与优秀的你们并肩前行，我无比骄傲。自 13 年回到南京，正式进入实验室，送走了一批又一批的兄弟姐妹，如今也终于要轮到自己，我们永远都是一个大家庭。感谢付雪峰学长、高桓学长、徐康学长、周张泉学长、马艳芳学姐、陆彬同学、张勇同学、任建欢同学、凌绍伟同学、方一曙同学、吴自勉、刘太云、孙松、罗斌、张晓、崔轩、沈飞、欧阳春、黄超、姚胜、石珺、张梦易、张良、刘兵、花云程、毕胜、高丛、李丞、许亮等等，与你们互相学习、交流、帮助，是我人生成长中不可或缺的部分。此处特别感谢几位我指导的同学，刘太云、罗斌、崔轩、张梦易、李丞、许亮，是你们的包容与大度，使得急躁的我能够平静、平稳，没有你们就没有研究的顺利推进，没有你们就没有我所取得的成绩，感谢你们的真诚，感谢你们对我的肺腑之言，希望一生我们都是这样的好友。

一路走来，最对不住的是父母妻子，现在还有我的儿子。在为了梦想前行时，我为人子，为人夫，为人父，却从未给你们我理想中的殷实生活，今年 28 岁，却还要远赴他国进行为期一年七月的博后工作，让我心中有愧。爸爸妈妈，你们的一直支持，你们最大的帮助与爱护，让我觉得有你们这样的父母，今生无憾；老婆，为了嫁给我，你来到陌生的城市工作，没有朋友，没有亲人，而我却为了心底所谓的梦想经常离开，我很歉疚，此生有你陪伴，足矣；儿子，今天是你出生第 310 天，爸爸却在新加坡，无法照看你，没有尽到一个父亲的责任与义务，爸爸很难过，可爸爸希望你知道，爸爸爱你，非常爱你，要健康成长。你们是我的至亲，我爱你们，没有你们就没有今天的我，是你们允许我的自私与任性，你们始终是我科研道路与人生道路上的最大动力，我一定会更加努力，此生不负！

写到此处，已抑制不住情绪，太多人需要感谢、感恩，如张磊博士、王昊奋博士、好友卢呈远、王昆峰等等，无法一一言尽，无法一一回报，仅以此文向所有关心与帮助过我的人表达诚挚的谢意！我定不忘初心，继续前行！

心求至善



SOUTHEAST UNIVERSITY