· 计算机软件理论、技术与应用 ·

微博知识图谱构建方法研究

杜亚军,吴 越

(西华大学数学与计算机学院,四川 成都 610039)

摘 要:传统搜索引擎需要用户从返回网页中提炼有用知识;社交网络搜索根据人物的社会关系、共同爱好,提供人物和兴趣间的关系等方面的搜索结果。当前,社交网络搜索主要存在 2 个问题:不能从语义上理解用户查询词;仅局限于人物、兴趣搜索,限制了查询范围。为解决微博搜索中存在的一些问题,并主动返回更多知识,基于微博这一社交网络的重要平台,研究微博社区知识图谱构建方法,重点提出 5 方面的研究:微博社区中概念提取,其概念包括人物、事物、地点、事件和话题等 5 种类型;微博社区概念间的关系提取,其关系包括上述 5 种概念间的组合关系;知识图谱是带有语义的网络图谱,将概念作为顶点并将概念间关系作为边,研究知识图谱的构建方法;分析微博社区知识图谱,包括构建效果、演化特征、应用效果分析;研发基于微博知识图谱的应用系统等内容。

关键词:微博;知识图谱;图谱构建;概念提取;关系提取

中图分类号:TP393.09 文献标志码:A 文章编号:1673-159X(2015)01-0027-09

doi:10.3969/j.issn.1673-159X.2015.01.005

Research on Constructing the Knowledge Graph Based on Microblog

DU Ya-jun, WU Yue

(*School of Mathematics and Computer Engineering, Xihua University, Chengdu* 610039 *China*)

Abstract: Search engine only returns the Web page set for the user queries, it needs the user refine useful knowledge from it; Social Network Search (SNS) directly provides people and their interest to users by using characters' social relations and common hobbies. However, the SNS mainly exists two unresolved problems. On the one hand, the SNS can't semantically understand user queries submitted by users. On the other hand, the SNS only provides people search and interest search, and confines query domains for users. Microblog has become an important platform for social network. To address these problems of information retrieval about microblog and provide more knowledge for user queries, this project researches knowledge graph construction and analysis based on the microblog community. The project focuses on five contents. (1) It researches concept extractions for the microblog community, and concepts have five types including people, things, locations, events and topics. (2) It researches relationships extractions for the microblog community. The relationships among concepts include collection types formed by combining two arbitrary types above concepts. (3) It researches knowledge graph construction, and the knowledge graph is a semantic network graph which takes concepts and relationships respectively as vertices and edges. (4) It researches knowledge graph analysis. It includes construction effect analysis, evolution characteristics and rules analysis and application effect analysis. (5) It researches the application interface and system based the knowledge graph.

Keywords: Microblog; knowledge graph; knowledge graph construction; concept extraction; relationships extraction

1 研究意义

从 20 世纪后期,借助关键词匹配的信息检索技术,Google、中文 Baidu 等世界著名的搜索引擎给人们生活、学习、工作带来巨大变化。基于关键词的搜索引擎已成为互联网中信息获取的主要工具。它们主要使用向量空间模型(VSM),因此无法摆脱不能满足人们搜索精确的事实。该问题主要表现

收稿日期:2014-11-12

基金项目:国家自然科学基金(61271413;61472329)

第一作者:杜亚军(1967—),男,教授,博士,硕士生导师,主要研究方向为网上信息挖掘与搜索引擎、计算机软件开发技术。

在 2 方面。1) 由于自然语言的模糊性,相同用户提交不同查询词,虽然查询词语义上是相同的,但是搜索引擎返回的结果是不同的; 不同用户提交相同查询词,虽然查询词对不同用户来说语义是不相同的,但是搜索引擎返回的结果是相同的。2) 由于搜索引擎缺乏网页中人、事、物之间的语义关系,查询结果仅仅从字符串相同与否来匹配用户查询与网页之间的关系,使得一些不包含查询词,但与查询词内容特别相关的信息流失; 一些本身与查询词在语义上关系不大的信息,仅仅包含了查询字符串,却被检索出来。这些问题产生的根本原因在于传统搜索引擎的主要工作流程(图 1) 是网页→网页数据→用户查询→字符匹配→返回结果网页,这种流程模式需要用户把结果网页消化成知识,这个过程对用户而言效率很低(需要花费时间理解),同时精确度也低(很多时候找到的网页了,回答不了我们的问题)。比如"× × ×的儿子的老婆的情人是谁",这个问题目前的搜索引擎是回答不了的。人们更需要的信息应该是以知识的形式展示在搜索结果中,而不是直接输出网页。让搜索能提炼用户查询和网页中的知识,从自然语言语义的角度,理解用户的真正需求,让搜索引擎更能模拟人与人之间自然的交流方式,实现人与 Web 之间的问答、会话,以至让 Web 能预测和理解到人搜索的目的,给出用户真正需要的知识,是未来搜索技术的又一大革新。

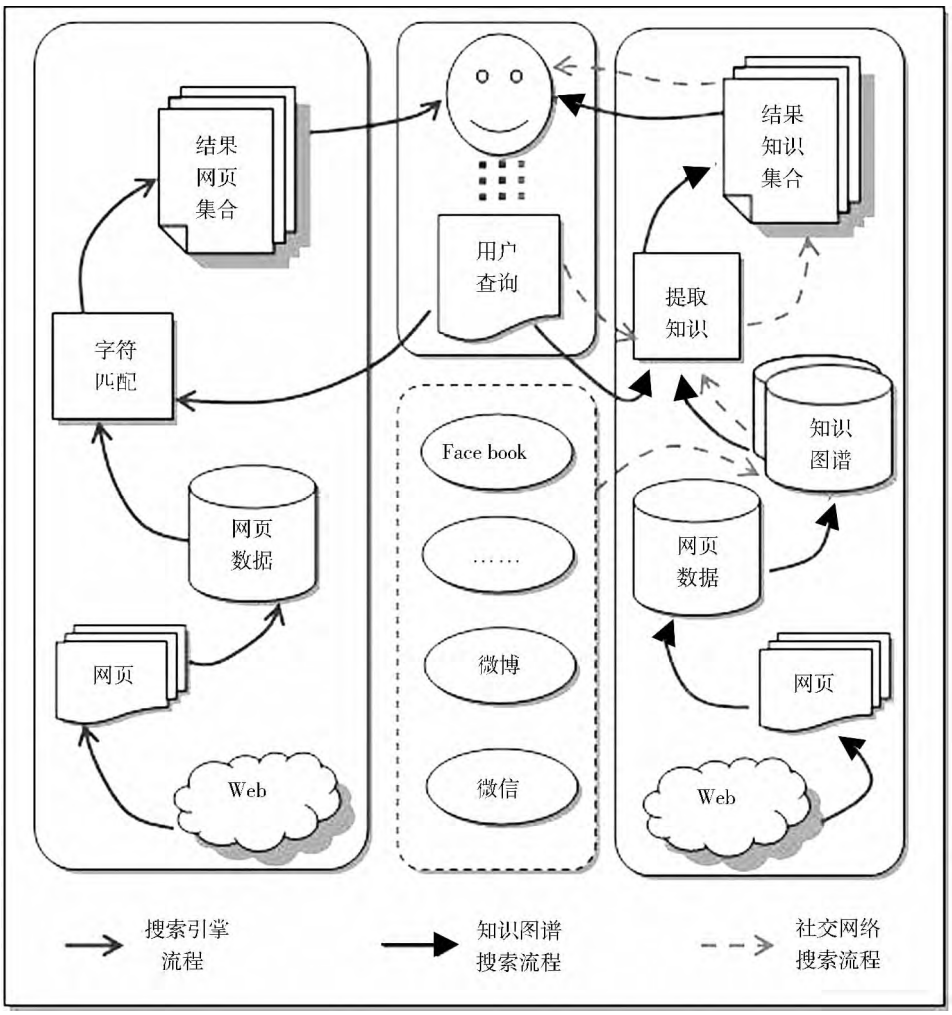

图 1 搜索引擎、知识图谱搜索引擎、社交网络搜索流程对比

为解决查询词的歧义和同义问题,下一代智能搜索引擎应运而生。它能够将用户所提交的查询词理解成实体或者概念,通过实体或者概念匹配为用户返回其所可能关心的全部网页内容。Google 在 2012 年 5 月提出 Knowledge Graph(知识图谱) 搜索[1],将人物、事物或者地点当作实体并构建它们之间的关系,以满足用户语义高级检索需求。Google 知识图谱在搜索中的主要功能表现在 3 方面: 增强语言模糊性处理,以便找出更正确的搜索结果; 对查询主题有关的内容进行语义摘要; 在对

话搜索中,寻找更深更广的信息。Google 借助知识图谱重点解决 2 个问题[2]: 与 Google 会话搜索; Google 预测你的需求。微软提出 Satori[3] 搜索,从互联网非结构化文档中提取结构化的实体所包含的属性信息,这些属性都是与动作相关的,以实现用户的语义检索。百度实体搜索[4] 先根据用户提交的查询词找到实体类型,再利用查询词与这类实体的属性匹配找到相关的实体,从而达到理解查询词语义的目的。实体搜索构建了实体数据库,该数据库包含实体信息和实体属性,通过对实体信息和关键属性进行分类,提高了实体信息的搜索广度和实体属性的搜索精度。搜狗知立方搜索[5] 从网页集中抽取实体及其关系信息,同时利用语义推理补充实体信息,从而实现对用户查询意图的理解。知立方通过数据挖掘、分析等技术准确地构建知识数据库,并通过推理计算分析用户查询语义,不仅向用户提供核心的内容信息,而且展现了较完整的知识体系。这些搜索引擎中知识图谱的搜索过程(图 1) 可以总结概括为: 网页→网页数据→知识图谱→用户查询→知识提取→返回知识。现有知识图谱中的知识来源于网页,从 Web 结构化和半结构化网页中提取知识。由于微博等社交网络中,实体形成的知识具有突发性和实时性,据调查和分析,现在很少有文献研究基于微博的知识图谱的构成。

在 Web 网络中,国内外各大知名搜索引擎都利用知识图谱中的实体及其关系理解用户查询意图并已初步实现下一代智能搜索。在社交网络搜索 SNS(social network search) 中,国内外各大社交网站推出利用类似知识图谱的社交图谱、兴趣图谱,实现社交搜索的功能[6]。社交图谱是表明“我认识你”的网络图谱,研究人物与人物之间的社会关系; 兴趣图谱是表明“我喜欢这个”的网络图谱,研究人物与人物之间的共同兴趣。Facebook 在 2013 年 1 月推出 Graph Search[7],其利用人物间关系与他们所关心的事物,构建成网络图谱,直接向用户提供人物、照片、地点和兴趣的搜索结果。Twitter 提出的兴趣图谱[8],是以人物与人物间的兴趣爱好为线索的网络图谱,在庞大的社交网络挖掘出基于兴趣的海量基础数据,满足用户寻找具有相同兴趣爱好的人物需求。与 Facebook 所提出的社交图谱相比,兴趣图谱基于兴趣与人物之间的相互关联,这些人物可能是不认识的,从而很大程度上扩展了社交网络搜索的深度和广度。腾讯推出 QQ 圈子[9],用户可应用其将自身的真实社会关系进行自动分圈,并向同一交际圈内的陌生人拓展人脉。在 QQ 圈子中,用户可以智能地备注好友的真实姓名,此功能间接实现用户实名制,同时用户向其朋友的朋友发起对话,以达到拓展交际人脉的目的。新浪在兴趣图谱方面做了尝试[6],在很大程度上增强了用户活跃度,但是,未在微博最盛行的时候,将用户好友间的兴趣点推向商业化。相反,阿里推出淘宝数据盛典[10],其在没有社交网络的场景中挖掘海量交易数据,分析网购人群的消费兴趣,展现出很多兴趣图谱的功能。

当前,Web 语义搜索通过实体及其关系信息,智能地理解用户的查询意图,并向用户提供更全面、更精准的搜索结果。同时,社交网络搜索利用人物及其社会关系、共同爱好直接向用户提供人物、兴趣等方面的搜索结果; 但是,与 Web 语义搜索相比,由于查询词的歧义性、多样性,社交搜索仍然存在一些问题。这些问题具体表现为 2 方面。1) 在 SNS 网络中,大多数社交网站没有实现从语义上理解用户所提交的查询词的社交搜索,当前只有 Facebook 已实现社交语义搜索,但是其只专注于人物、照片、地点和兴趣 4 方面的语义搜索; 而在 Web 网络中,许多搜索引擎已经实现查询词的语义理解,并提高了搜索结果的覆盖率和准确度。2) 在 SNS 网络中,由于各大社交网站都利用人物与人物间的社会关系、共同兴趣构建社交图谱和兴趣图谱,这些网络图谱将用户的社交搜索局限到人物、兴趣方面,从而限制用户的查询范围。SNS 网络存在事物、事件、话题等海量信息,对这些信息进行数据挖掘并添加到网络图谱中,能够拓宽用户查询需求,延伸社交网络搜索范围。

微博已成为网民广泛传播信息的社交网络平台,同时基于微博的朋友推荐、信息检索、舆情分析已经越来越受到政府、网民的关注。微博,是一个基于用户关系的信息分享、传播以及获取平台,用户可以通过 Web、Wap 等各种客户端组建个人社区,以 140 字左右的文字更新信息,并实现即时分享。微博社区指一群在微博社交网站中彼此沟通的用户,他们分享某种程度的知识和信息,从而所形成的团体。知识图谱是描述概念及其关系的知识库,其可以利用网络图谱表示,图谱的顶点是概念,边是概念关系。概念类型包括人物、事物、地点、事件、话题,而概念关系类型包括上述 5 种类型任意排列所组成的类型。

随着社交网络中微博用户日益巨增及共享信

息迅猛增长,在微博社区中,知识的产生和发展越来越具有规律性。通过挖掘微博社区中的海量信息,用来构建知识图谱,推进数据挖掘、网络图论、Web、社交网络、搜索引擎理论等研究,促进它们朝着更智能化、语义化的方向发展。同时,所构建的知识图谱能够广泛地应用到社交信息检索、社交朋友推荐、社交舆情分析等系统研发,提高这些系统语义理解能力,向用户提供更全面、更准确的信息。

2 研究现状分析

知识图谱是实体和实体之间关系的一个知识库,它的本质是一种揭示实体知识之间的语义网络图。知识图谱源于引文分析理论、复杂网络系统、社会网络分析,以及信息可视化。1963 年 Garfield 成立美国费城科学情报研究所(简称 SCI),SCI 的设计初衷是为人们提供一个检索工具,它是从被引文献中去检索引用文献的索引,揭示科学文献之间、作者之间的引用与被引用关系,从而提供了引文分析所必需的大量数据,并于次年手工完成了 DNA 领域的引文编年图;1965 年普耐斯运用相同的数据完成了他的经典论文——科学论文网络[11]。这 2 个事件就是知识图谱的最早雏形。知识图谱发展历经了 4 个阶段。1)知识的集成和知识结构化表达。20 世纪 70 年代末期,知识图谱理论诞生,Bondy 等[12]在处理医学和社会学文本时,使用图理论来表达文本内容中由 cause 和 by 产生的因果关系,便于用于决策支持系统。Brachman[13]采用图理论来表达医学论文中名词间 Isa、Partof 和 Kindof 语义关系,构建语义网络,并成功用于 Medical 专家系统。本体 Ontology 是这类用图表示知识的典型代表。2)文本知识的获取与表达。20 世纪 90 年代初期,用图理论表达知识广泛用于自然语言处理中。Sowa[14]提出概念图的基本理论,在图中广泛地考虑了文本名词之间的关系,将其分为 Eou(相等关系)、Sub(子关系)、Ali(似乎像关系)、Dis(变异关系)、Ord(先后关系)、Cau(因果关系)、Par(部分属性关系)、Sko(信息相关关系)8 类。与第 1 阶段本体不同之处在于概念图重点加入了由名词组成的概念之间的关系。3)句子之间的语义知识表达。21 世纪初,基于图的知识表达广泛应用于自然语言的处理中。Zhang[15]建立了不同语言单词之间的语义关系图,将每个单词的语义在图中进行标注,构建基于每个词的语义图谱,成功用于 2 种不同语言之间句子的翻译系统。4)知识图谱在 Web 中的应用。其典型代表是 Google 公司的知识图谱、微软公司的 Satori 知识图谱。基于图的知识表示方法在不同的发展历史阶段,有着不同的表现形式和应用,发展了不同的基于 Web 的知识图谱。目前在国际国内主要的知识图谱有:Wolframalpha,一个计算知识引擎;Freebase,6 800 万个实体,10 亿的关系;Google 知识图谱[2],一个实体及其关系的知识数据库,包括 5 亿实体、180 亿实体关系信息,这些信息不仅为用户准确查找网页内容提供知识库,而且给用户提供扩展知识的工具。与 Google 知识图谱相似,Satori 也构建实体及其关系的知识数据库,该库包括 4 亿多实体[3]。DBpedia 是一个在线关联数据知识库项目,它从维基百科的词条中抽取结构化数据,并将这些数据以关联数据的形式发布到互联网上,提供给需要这些关联数据的在线网络应用、社交网站或者其他在线关联数据知识库。此外,还有 Cyc、KnowItAll、ConceptNet、DBpedia、YAGO 等知识库和知识图谱[16]。大数据的处理离不开知识库,现在知识图谱面临 2 个重要的挑战:一个是发展构建知识图谱的软件系统;另一个是面向现代 Web 网页、博客、微博等社交网络产生的大数据,如何去构建它们的知识图谱和怎样利用知识图谱的知识去建立不同的应用系统。这 2 方面既是挑战也是一种新的尝试。微博社区知识图谱构建的相关研究工作除微博社区的发现外,还可以借鉴的相关研究有网页中实体和概念的获取、关系提取、图的建立等方面。

2.1 微博社区发现

近几年来,借助微博平台的热点研究主要有:微博用户行为分析[17],微博语言的分析与识别、单词规范处理[18],信息传播与舆情模型[19],微博社区划分与发现[20]。随着理论、方法和技术上的突破,在应用领域,以微博信息为基础,开展了基于移动互联用户的微博内容浏览工具[21]、信息检索与搜索、主题或任务推荐系统[22]等研究工作。其中,主要工作是微博社区的发现。目前主流的微博社区划分的典型方法有:基于兴趣,基于链接,基于信任度、主题与链接相结合的方法。Fu 等[23]对微博中微贴的观点和情感进行分析,研究了观点一致、情感一致的跟贴人之间的相似关系,对某个和某些主题观点和情感一致的用户进行层次聚类分析,提出了在微博中隐式社区发现和挖掘的方法。微博中的跟贴人,往往会追逐社会网络中的明星和名流,但是明星们往往又不向他们的粉丝跟贴。结合微博这一特点,Qin 等[24]研究了微博跟贴人之间的链

接关系,提出了一种从微博中挖掘真实朋友之间的关系,从而为每一位微博跟贴人提供一个朋友圈,达到社区发现的目的。Wu 等[25]分析微博中跟贴人的行为和信誉,将微博中用户分为 3 种人群: 正常人群,一些受利益和经济驱使、操纵跟贴人的人群,受操纵的人群,并分别给他们不同的信誉等级,采用半监督学习方法,从不同种类人群中提取他们的信息,然后对整个微博空间中用户进行划分和分类。微博用户参与微贴的回复、评论,构成了人与人、信息与信息的网络。将微博参与者之间构成的社会网络和信息网络相结合,闫光辉等[26]定义了用户相关度,它是链接相关和主题相关的加权和,给出了链接和主题相结合的微博社区发现算法。此外,传统的一些用于社会网络的算法,也广泛地用于微博社区挖掘上。如 Newman 提出了一种快速聚类算法,算法优化目标函数 Q 定义为簇内随机连接与簇内期望连接数目之差,通过合并使 ΔQ 最大,形成一个自底向上的聚类过程[27]。Clauset 等提出的启发式的 CNM 算法[28],能够快速找到网络中社区的近似最优解,它通过迭代计算分割最大介数边的方法来划分网络社区。史春永[29]改善传统社交网络社区中结点只能属于一个社区的弊端,提出了重叠社区发现算法。

2.2　实体和概念的获取

在微博社区中,概念的本质就是现实世界中的语义对象,由描述该对象的属性集合构成,而对象的属性集本质上就是能够表示对象的名词集。微博中概念的提取,不仅有效地简化了微博文本的结构形式,还能够在一定程度上解决微博语言中的一词多义、多词同义的概念匹配问题,从而有效地帮助用户从海量的、实时动态更新的微博信息中检索到感兴趣的内容,并为知识图谱的构建奠定良好的基础。目前,关于微博社区中概念提取方面的研究工作尚处于起步阶段。国内外学者对于 Web 网页文本中的概念提取进行了大量地探索,如: Chen 等[30]利用语义相关的对数似然比和 k-means 方法,对文献资料搜索结果组织的概念进行提取和聚类,同时通过聚类和引文耦合实现搜索结果的组织和可视化呈现; 基于每一个词在网页文本中的相关估计,Fresno 等[31]提出了从 Web 网页中获得相关概念集的方法,对 HTML 环境下的概念进行了提取,其中词的相关性同时考虑到 HTML 语言的特征; 类似于词在文档中的频率和位置,章成志等[32]依据网页标引源加权方案进行文本的概念提取,利用语义相似度算法进行文本的自动分类,从而设计了一个中文 Web 概念挖掘系统,为使系统能及时提取新词,系统中还加入了未登录词挖掘的功能。在网页、微博、博客中,实体的名字往往是模糊的,特别是在微博中,由于文本短小,在一个微贴中,很少有确定性的信息内容。据此,Dalvi 等[33]认为概念是 Web 用户感兴趣的实体、事件和主题。Spina 等[34]先过滤掉大量与实体无关的文本或关键词,然后区分剩下关键词的情感词性,再过滤掉负面关键词,从正面关键词中抽取实体(公司)名,并提出了相应的算法。由于微博中用户对一个事件的观点随着时间的推移在发生改变,从微博中提取人们对某件事的具体观点是一个挑战性的工作。Zhao 等[35]充分利用微博中文本、时间信息、社区结构,建立了 Term-Tweet-User 图模型,提出时间感知的随机游走算法,解决了词项随时间变化的相关性计算,然后再分类词项,提取群体对某件事的观点。潘虹等[36]提出了一种基于最大公共子串(longest common substring, LCS)算法的术语抽取方法,针对学前教育领域进行抽取实验,验证了 LCS 算法可以有效地抽取中文领域术语。Nie 等[37]首次提到了基于视觉的网络实体提取,在理解网页结构和文本内容中综合考虑了信息的可视化布局和知识库的特征,并结合统计滚雪球的方法自动发现文本模式,并在 Microsoft Academic Search (aka Libra) 和 EntityCube 中有很好的应用。梁健等[38]研究了文本的本体学习方法以及本体对知识概念表达的层次结构,利用统计和规则 2 种方法抽取与种子概念相关的领域术语,将种子概念方法用于文本中知识和术语概念的提取。Cui 等[39]提出了基于词语共现的领域概念提取方法,在概念选择时只考虑了频率的作用,没有加入概念之间距离因素的影响。Górriz 等[40]将遗传算法和种子概念相结合,用于本体概念提取,利用种子概念的一些固有关系,得到遗传算法的初始种群,利用遗传算法对种子概念进行扩展,得到扩展种子概念,用户只需要给出基本的种子概念,就可以得到更多、更全面的种子概念。刘竞等[41]提出了基于免疫计算的概念提取方法,它在有效降低特征个数的同时,提取各类的中心,以此为实例模式对待识别样本进行分类决策。Zhang 等[42]提出了对数似然比的领域本体概念提取方法,采用对数似然比计算概念与领域、概念与概念之间的相关性,将其应用到概念提取中,能够有效地提高概念提取的准确度。

2.3 关系提取

美国国家标准技术研究院 2005 年自动内容提取(automatic content Ext ract ion,ACE)会议将文档中的实体之间的关系[43]分为局部整体关系(Part-whole)、地理位置关系(Phys)、类属关系(Gen-aff)、转喻关系(Metonymy)、制造使用关系(Art)、组织结构从属关系(Org-Aff)、人物关系(Per-Soc)。信息提取领域的学者对关系提取进行了长期的探索,目前已经有许多关系提取方法被应用到各种实验系统中,主要有基于知识库的方法和基于机器学习的方法。根据对训练语料数据的不同需求又将基于机器学习的方法归纳为有指导的机器学习方法、半指导的机器学习方法和无指导的机器学习方法。有指导的关系提取方法在关系提取领域占有主导地位,主要方法有基于特征的机器学习方法和 Kernel 方法。SVM(support vector machine)[44]是 Cortes 和 Vapnik 于 1995 年最先提出的,是一个能够将不同类别的样本在样本空间用超平面分隔,也就是说,给定一些标记好的训练样本,SVM 算法输出一个最优化的分隔超平面。MaxEnt(maximum entropy)模型[45]是 MaxEnt 分类器[46]的理论基础,模型是由 Jaynes 首次提出,基本思想是为所有已知的因素建立模型,而把所有未知的因素排除在外。也就是说,要找到一个满足所有已知的事实并且不受任何未知因素影响的概率分布。Zhou 等[47]在基于特征的关系提取中使用了 SVM,广泛合并多样的词法、句法和语义知识,融合基础短语的组块信息,大大提高了在句法方面的性能,还证明了将语义信息如词汇网 WordNet 和名单 Name List 用到基于特征的关系提取中能进一步提高性能。Choi 等[48]采用基于核的支撑向量机的方法,从文本句子中提取人名,构建 2 个不同的人名之间的社会关系,并开发了社会网络的社会关系提取系统。Xu 等[49]研究了挖掘文本中实体、关系的语法和语义模型、关系的上下文句子、关系背景知识图、关系出现的背景区域等识别方法,对实体的不同语义关系进行挖掘,从而探讨实体中不同语义关系随时间变化的演化规律。Chaveevan[50]结合贝叶斯网络和最大熵确定文本中有效的基本篇章单位,从相邻基本的篇章单元采用机器学习方法,提出动词对提取因果关系的方法。Mausam 等[51]从 5 亿个网页中,提取实体之间的链接关系,建立一个 Web 中实体关系的知识库。Furlan 等[52]使用自然语言的处理方法,研究出短文本中不同句子之间的语义关系。文献[53-54]探讨了概念之间的语义关系的计算方法。Liang 等[55]分析概念和实体的关系,研究它们之间的语义,提出了概念间隐式关系的发现方法。

2.4 知识图谱的建立

在 Hoede(离散数学家)和 Stokman(数学社会学家)提出的知识图谱中,用图的概念来表达知识,一个图或者一个有向图 $G=(V,E)$ 是由节点集合 V 和边集合 E 组成。其中节点就是实体或概念,边或者弧就是节点之间的关系。到目前为止,知识图谱能表达 7 种类型的关系[43]和 4 种类型的框架。在知识图谱的构建过程中,实体与实体间、概念与概念之间有各种各样的关系,然后对它们有效合理地进行量化。Bondy 等[12]作为医学内容和社会学文本研究者的代表,用这些文本中积累的知识构建越来越大的图,由此形成一个专家系统。Bakker[56]建立了一个知识图谱的知识管理系统 KISS(knowledge integration and structuring system),并给出了其知识图谱的建立过程:1)文本分析,将文本映射到一个图上;2)构造分析,确定子图,形成一个“自然”单位;3)链接集成,使用路径代数从提取的知识中派生新的知识。Wang 等[57]对稀疏大图的结构相似性进行有效地研究,通过分解图成为不同的由邻居树模式组成的图粒,逐步匹配不同图的图粒,使用编辑距离最小下限估计方法解决了不同图的相似匹配问题。在知识图谱应用方面,Steiner 等[58]在浏览器中开发了一个扩展程序,实现了其搜索结果一旦有知识图谱中实体,就从 Wikipedia 中选择一个真实概念和它的链接,从当前主流的社交网络 Twitter 和 Facebook 的前 n 个结果独立地显式在浏览器的某个面板中,最后从检索结果的有用性和相关性角度分析知识图谱的性能。Pujara 等[59]针对现有知识提取系统输出常常会发生丢失节点和边、不准确的节点分类标记等问题,通过组合实体的解析、协同过滤,提出了基于本体约束的链接预测来进行知识图的节点和边的自动识别与标识。在基于本体的知识库应用中,文献[60-61]探索了知识的更新,对给定的一个知识,将知识分解为概念和概念之间的关系,然后将新的知识插入到知识库中,达到系统新的知识动态更新的目的。在科学研究和工程应用中,知识库和本体等已解决许多问题。

此外,随着应用的深入,一些研究者对已有知识图谱、本体库、知识库等的性能测试做了一些尝试性的研究。Santoso 等[62]研究了本体概念提取方法的精确度问题,提出了从文本中提取知识的方

法,并对这些方法进行测评的策略。面向本体及其应用,该方法比较客观地对知识获取算法进行了较好的评价。Ittoo 等[63] 研究了本体中概念之间的关系提取算法的精度和召回率问题。Luis 等[64] 利用形式概念分析方法,综合考虑了形式概念格中概念和概念关系,研究了从文本中获取知识的方法的评价策略。Zhou 等[65] 在分析关系的语义基础上,探讨了基于树的语义关系表示方法在不同应用问题中的适应能力。Khattak 等[66] 探讨本体中,概念和关系等知识随着时间的变化规律和演化趋势。Liu 等[67] 针对本体中概念和关系的变化规律,探讨了随时间变化的演化过程,从而总结性地提出了本体演化模型。

综上,面对微博社会网络平台中海量信息,几乎很少发现有对其实体、概念、关系提取方法的研究,因此,微博社区知识图谱的构建具有广阔的研究前景。

3　微博知识图谱的几个新研究

3.1　微博社区概念提取方法

微博社区的概念本质上就是现实世界中的语义对象和描述该对象的属性集合构成。这些对象具有 5 种类型: 人物、事物、地点、事件和话题。对象的属性集本质上就是能够表示对象的名词集。本项目重点研究以下几方面的内容。1) 从微博社区所分享的海量信息中抽取概念。概念是由微博社区中的频繁属性和信息属性所组成。频繁属性是指在微博社区中出现次数较多的名词集,例如爆炸词、流行词等; 信息属性是指在微博社区中具有丰富语义信息的名词集,例如敏感词、新近词等,需要研究频繁属性和信息属性的提取方法。2) 在微博社区中,概念理解为描述语义对象的属性名词集构成,由于语言中存在一词多义、多词同义现象,需要研究所提取的属性名词同义匹配、多义匹配的方法。同义匹配是指不同属性名词属于相同概念的匹配方法; 多义匹配是指相同属性名词属于不同概念的匹配方法。3) 概念类型包括人物、事物、地点、事件和话题 5 种类型,但是每种概念类型可能包括更多的子概念类型,例如体育话题概念包括田径运动、球类运动、棋牌运动等子概念类型; 因此,需要研究概念层次聚类方法,将不同的概念划分到不同的类别。

3.2　微博社区关系提取方法

微博社区的概念间关系本质上就是现实世界中对象间的语义联系,由描述该联系的链接集构成,这些联系类型是由上述对象的 5 种类型任意排列所组成的关系类型,例如人物与地点的联系、事物与事物的联系等,而联系的链接集本质上就是表示任意 2 对象的不同属性间联系动词集。本项目重点研究以下几方面的内容。1) 从微博社区所分享的海量信息中抽取概念间关系,其是由不同概念的属性集间的频繁链接和信息链接所组成。频繁链接是指能够链接不同概念的属性并且出现次数较多的动词集; 信息链接是指能够链接不同概念的属性并且具有丰富语义信息的动词集。需要研究频繁链接和信息链接的提取方法。2) 在微博社区中,概念间关系是由描述不同概念间联系的链接动词集构成,由于语言中存在一词多义、多词同义现象,需要研究所提取的链接动词同义匹配、多义匹配的方法。同义匹配是指不同链接动词属于相同概念间关系的匹配方法; 多义匹配是指相同链接动词属于不同概念间关系的匹配方法。3) 概念间关系类型是由人物、事物、地点、事件、话题 5 种类型任意排列所组成的,但是每种关系类型可能包括更多子关系类型,例如人物与人物间的关系包括师生、朋友、亲戚、同学等子关系类型; 因此,需要研究概念间关系层次聚类方法,将不同的概念间关系划分到不同的类别中。

3.3　微博社区知识图谱构建

微博社区知识图谱是由实体及关系、概念及其关系所构成的语义性知识库,将该知识库概念作为顶点,实体或概念关系作为边,可以建立带有语义的网络图谱。根据所建立的语义网络图谱,可能将挖掘出新的隐式概念和关系,并不断地扩展知识图谱。同时,随着微博社区的信息不断增长,将会发现新的实体及关系、概念及其关系,需要不断地更新知识图谱。本项目重点研究以下几方面的内容。1) 在微博社区中,将抽取的全部实体及关系、概念及关系存储到知识库中,建立知识图谱,但是该知识图谱可能存在重复的、歧义的概念及其关系(或实体及关系) 。其重复性表现为这些实体属性集、实体关系集、概念属性集、概念关系的链接集中存在多词同义的现象,因此需要合并重复的实体及关系、概念及关系。实体及关系、概念及关系的歧义性表现为这些实体的属性集合、实体关系集合、概念属性集、概念的关系链接集中存在一词多义的现象,因此需要拆分歧义的实体及关系、概念及关系。2) 在微博社区中,知识图谱是利用微博社区中所分

享的信息建立的,这些分享信息不能体现语言的整体特征,需要利用同义扩展、多义扩展方法来完善所建立的知识图谱。同义扩展是通过实体属性集、实体关系集、概念属性集、概念的关系链接集中的同义词来扩展知识图谱;多义扩展是通过实体属性集、实体关系集、概念属性集、概念的关系链接集中的多义词来扩展知识图谱。3)随着微博社区所分享的信息不断变化,此微博社区将会抽取到新的实体属性集、实体关系集、概念属性集、概念的关系链接集,需要利用它们来更新与完善所建立的知识图谱。概念(或实体)更新是通过新抽取到的概念(或实体)集、概念(或实体)的属性集中的同义词和多义词实现知识图谱的更新;关系更新是通过新抽取到的概念(或实体)关系集、关系的链接集中的同义词和多义词实现知识图谱的更新。

3.4 微博社区知识图谱分析

微博社区知识图谱是带有语义性的知识库,实现在应用层面上,从语义方面理解用户的意图。从知识图谱自身的意义角度,需要分析实体及关系、概念及其关系抽取的覆盖度和精准度;从知识图谱的时间特征角度,需要分析实体及关系、概念及其关系抽取方法的适应性和演化性;从知识图谱的应用效果,需要分析知识图谱在社交网络系统中的召回率和准确率。对于此内容,本项目重点研究以下几方面的内容。1)从微博社区海量信息中获得了实体、概念及其关系,需要分析实体、概念及其关系抽取的覆盖度和精准度。2)由于不同微博社区的描述信息和其他社交平台的呈现信息的差异,需要分析实体、概念及其关系抽取方法、知识图谱构建方法的社区适应性和平台适应性。3)由于微博社区的用户规模、话题内容不断发生变化,需要分析实体、概念及其关系抽取方法的用户演化性和话题演化性。用户演化性反映了随着用户规模不断更新,实体、概念及其关系的变化趋势,而话题演化性反映了随着话题内容不断更新,实体、概念及其关系的变化趋势。此2个趋势反映出知识图谱的演化规律。4)知识图谱能够从语义上理解用户寻找朋友、查询信息等各种意图,主要应用到社交朋友推荐、信息检索、舆情监测等系统服务中,需要分析知识图谱在这些应用中的召回率和准确率。召回率反映了用户所感兴趣的所有信息中,利用知识图谱方法所返回相关信息的比率,而准确率反映了利用该知识图谱所返回的所有信息中相关信息所占的比率,这2种比率反映出知识图谱的应用效果。

3.5 微博社区知识图谱的应用

面向微博社区海量信息,有许多值得研究的应用问题,如社交朋友推荐、信息检索、舆情监测、问答系统、推荐系统、广告投放。基于微博的诸如此类的应用,还是一个尝试性的研究领域。本项目重点解决以下几方面的内容。1)选择一个典型的应用问题(拟为基于微博的问答系统),研究知识图谱中实体、概念提取和利用方法、接口;研究实体关系、概念关系、实体与概念之间的关系提取、快速检索、接口的方法。2)建立和开发与微博有关的、面向不同应用的知识(实体、概念、实体关系、概念关系、实体与概念之间的关系)高效访问的原理和方法、开发访问接口 API 或控件。3)面向不同应用问题,研究知识图谱的知识获取原理和方法。4)面向不同应用问题,研究知识图谱的知识更新原理和方法。

参 考 文 献

[1] Amit S. Introducing the Knowledge Graph: Things, Not Strings [EB/OL]. [2014-10-10]. http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.

[2] Lee J. OK Google: 'The End of Search as We Know It' [EB/OL]. [2014-10-10]. http://searchenginewatch.com/article/2268726, 2013.

[3] Sean G. How Google and Microsoft Taught Search to Understand the Web [EB/OL]. [2014-10-10]. http://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles-knowledge-graph-and-microsofts-satori/.

[4] 辜斯缪. 解密百度实体搜索 [EB/OL]. [2014-10-10]. http://tieba.baidu.com/p/2008266622.

[5] 东坡下载. 搜狗"知立方"让搜索更加准确全面 [EB/OL]. [2014-10-10]. http://www.uzzf.com/news/5986.html.

[6] 新浪科技. 下一代搜索引擎:知识图谱的用户体验报告 [EB/OL]. [2014-10-10]. http://tech.sina.com.cn/I/2013-08-27/18208681029.shtml.

[7] 搜狐 IT. FaceBook 发布社交图谱搜索 Graph Search [EB/OL]. [2014-10-10]. http://it.sohu.com/20130116/n363567269.shtml.

[8] Tian J. 为什么兴趣图谱将重塑我们的社交网络及未来的商业 [EB/OL]. [2014-10-10]. http://www.36kr.co.2013m/p/84790.html.

[9] TechWeb. 腾讯 QQ 圈子实现真人社交 [EB/OL]. [2014-10-10]. http://www.techweb.com.cn/internet/2012-03-21/1169220.shtml.

[10] 艾瑞网. 淘宝网数据盛典公布首份 2011 年度趣味数据 [EB/OL]. [2014-10-10]. http://ec.iresearch.cn/shopping/20120228/164451.shtml.

[11] Garfield E. Scientography: Mapping the Tracks of Science

[1]. Current Contents: Social & Behavioral Science, 1994, 7(45): 5-10.

[12] Bondy J A, Murty U S R. Graph Theory with Applications [M]. London and Basingstoke: McMillan Press, 1976.

[13] Brachman R J. What IS-A Is and Isn't: an Analysis of Taxonomic Links in Semantic Networks [J]. IEEE Transactions on Computers, Special Issue on Knowledge Representation, 1983, 16(10): 30-35.

[14] Sowa J F. Conceptual Structures: Information Processing in Mind and Machine [M]. [s. n.]: Addison-Wesley, 1984: 45 - 86.

[15] Zhang L. Knowledge Graph Theory and Structural Parsing [D]. Enschede, The Netherlands: University of Twente, 2002.

[16] Suchanek F, Weikum G. Knowledge Harvesting in the Big-Data Era [C] // The 2013 ACM SIGMOD International Conference on Management of Data. New York, USA: [s. n.], 2013: 933-938.

[17] Yan Q, Wu L R, Zheng L. Social Network Based Microblog User Behavior Analysis [J]. Physica A-Statistical Mechanics and Its Applications, 2013, 392(7): 1712-1723.

[18] Simon C, Wouter W, Manos T. Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text [J]. Language Resources and Evaluation, 2013, 47(1): 195-215.

[19] Yan Q, Wu L R, Liu C, et al. Information Propagation in Online Social Network Based on Human Dynamics. Abstract and Applied Analysis [EB/OL]. [2014-10-10]. http://Dx. Doi. Org/10. 1155/2013/953406.

[20] Yan Q, Yi L L, Wu L R. Human Dynamic Model Co-Driven by Interest and Social Identity in the MicroBlog Community [J]. Physica A-Statistical Mechanics and Its Applications, 2012, 391(4): 1540-1545.

[21] Han J H, Xie X, Woontack W. Context-Based MicroBlog Browsing for Mobile Users [J]. Journal of Ambient Intelligence and Smart Environments, 2013, 5(1): 89-104.

[22] 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法 [J]. 计算机学报, 2013, 36(2): 349-359.

[23] Fu M H, Peng C H, Kuo Y H, et al. Hidden Community Detection Based on MmicroBlog by Opinion-Consistent Analysis [C] //International Conference on Information Society, I-Society. London, UK: [s. n.], 2012: 83-88.

[24] Qin H L, Liu T, Ma Y J. Mining User's Real Social Circle in MicroBlog. [C] //The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul Turkey: [s. n.], 2012: 348-352.

[25] Wu X, Feng Z M, Fan W, et al. Detecting Marionette MicroBlog Users for Improved Information Credibility [J]. Lecture Notes in Computer Science, 2013, 8190(3): 483-498.

[26] 闫光辉, 舒昕, 李祥. 基于主题和链接分析的微博社区发现算法 [J]. 计算机应用研究, 2013, 30(7): 1953-1957.

[27] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks [J]. Physical Review E, 2004, 69(62): 066133/1-066133/5.

[28] Clauset A, Newman M E J. Finding Community Structure in Very Large Networks [J]. Physics Review E, 2004, 70(62): 066111/1-066111/6.

[29] 史春永. 面向新浪微博的数据采集和社区发现算法研究 [D]. 武汉: 华东科技大学, 2012.

[30] Chen S Y, Chang C N, Nien Y H, et al. Concept Extraction and Clustering for Search Result Organization and Virtual Community Construction [J]. Computer Science and Information Systems, 2012, 9(1): 323-354.

[31] Fresno V, Ribeiro A. An Analytical Approach to Concept Extraction in HTML Environments [J]. Journal of Intelligent Information Systems, 2004, 22(3): 215-235.

[32] 章成志, 侯汉清, 丁璇. 中文 Web 概念挖掘系统设计与测评 [J]. 上海交通大学学报: 自然科学版, 2003, 37(sup): 207-211.

[33] Dalvi N, Kumar R, Pang B, et al. A Web of Concepts [C] // The Twenty-Eighth ACM SIGMOD-SIGACT- SIGART Symposium on Principles of Database System. New York, USA: [s. n.], 2009: 1-12.

[34] Spina D, Gonzalo J, Amigó E. Discovering Filter Keywords for Company Name Disambiguation in Twitter [J]. Expert Systems with Applications, 2013, 40(12), 4986-5003.

[35] Zhao B, Zhang Z, Qian W N, et al. Identification of Collective Viewpoints on MicroBlogs. Data & Knowledge Engineering, 2013, 87: 374-393.

[36] 潘虹, 徐朝军. LCS 算法在术语抽取中的应用研究 [J]. 情报学报, 2010, 29(5): 853-857.

[37] Nie Z Q, Wen J R, Ma W Y. Statistical Entity Extraction from the Web [J]. The proceedings of the IEEE, 2012, 100(9): 2675-2687.

[38] 梁健, 吴丹. 种子概念方法及其在基于文本的本体学习中的应用 [J]. 图书情报工作, 2006, 50(9): 17-21.

[39] Cui G Y, Lu Q, Li W J, et al. Mining Concepts from Wikipedia for Ontology Construction [C] //The IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Workshops. Milan, Italy: [s. n.], 2009: 287-290.

[40] Górriz J M, Puntonet C G, Rojas F, et al. Optimizing Blind Source Separation with Guided Genetic Algorithms [J]. Neurocomputing, 2006, 69(13/15): 1442 - 1457.

[41] 刘竞, 赵友刚, 韩仲志. 基于免疫计算的概念提取方法研究 [J]. 微计算机信息, 2009, 25(1/3): 251 - 252.

[42] Zhang Y F, Shu W L, Xiong Z Y. Domain Ontology Concept and Relation Extraction Using Log-Likelihood Ratio [J]. Computer Engineering and Application, 2013, 49(6): 148-151.

[43] 黄鑫, 朱巧明, 钱龙华, 等. 基于特征组合的中文实体关系抽取 [J]. 微电子学与计算机, 2010, 27(4): 198 - 204.

[44] Cortes C, Vapnik V. Support Vector Networks [J]. Machine Learning, 1995, 20(3): 273 - 297.

[45] Berger A L, Pietra S A D, Pietra V J D. A Maximum Entropy Approach to Natural Language Processing [J]. Computational Linguistics, 1996, 22(1): 39 - 71.

[46] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧 [J]. 软件学报, 2010, 21(6): 1287 - 1295.

[47] Zhou G, Su J, Zhang J, et al. Exploring Various Knowledge in

(下转第 89 页)

－1627.

[12] 王桃,孔繁余,何玉洋,等. 离心泵作透平的研究现状[J]. 排灌机械工程学报,2013,31(8):674－680.

[13] 关醒凡. 现代泵理论与设计[M]. 北京: 中国宇航出版社, 2011:303－313.

[14] 王立文, 高殿荣, 吴建伟. 无阀微泵扩散管及收缩管流动特性分析[J]. 机床与液压, 2006 (6):159－162.

[15] 李俊,沈雪明,应济,等. 基于 MEMS 的扩张管/收缩管内流动特性分析[J]. 机床与液压,2008,36(12):61－63.

[16] 杨军虎,龚朝晖,夏书强,等. 导叶对液力透平性能影响的数值分析[J]. 排灌机械工程学报,2014,32(2):113－118.

(编校: 夏书林)

(上接第 35 页)

Relation Extraction[C]//The 43rd Annual Meeting of the Association for Computational Linguistics. Sydney, Australia: [s. n.],2005: 427－434.

[48] Choi M, Kim H. Social Relation Extraction from Texts Using a Support Vector Machine Based Dependency Trigram Kernel[J]. Information Processing & Management, 2013, 49(1):303－311.

[49] Xu Z, Luo X F. Mining Temporal Explicit and Implicit Semantic Relations between Entities Using Web Search Engines[J]. Future Generation Computer Systems, 2014, 37:468－477.

[50] Chaveevan P. Explanation Knowledge Graph Construction through Causality Extraction from Texts[J]. Journal of Computer Science and Technology, 2010, 25(5):1055－1070.

[51] Mausam T L, Etzioni O. Entity Linking at Web Scale[C]//The Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction. Montreal, Canada: [s. n.], 2012:84－88.

[52] Furlan B, Batanovi ć V, Nikoli ć B. Semantic Similarity of Short Texts in Languages with a Deficient Natural Language Processing Support[J]. Decision Support Systems, 2013, 55(3):710－719.

[53] Li W J, Xia Q X. A Method of Concept Similarity Computation Based on Semantic Distance[J]. Procedia Engineering, 2011, 15: 2852－2859.

[54] Li Y F, Qin K Y, He X X. Some New Approaches to Constructing Similarity Measures[J]. Fuzzy Sets and Systems, 2014, 234: 46－60.

[55] Liang T P, Yang Y F. A Semantic-Expansion Approach to Personalized Knowledge Recommendation[J]. Decision Support Systems, 2008, 45(3):401－412.

[56] Bakker R R. Knowledge Graphs: Representation and Structuring of Scientific Knowledge[D]. [S. l.]: University of Twente, Enschede, 1987.

[57] Wang G R, Wang B, Yang X C, et al. Efficiently Indexing Large Sparse Graphs for Similarity Search[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3):440－451.

[58] Steiner T, Verborgh R, Troncy R, et al. Adding Realtime Coverage to the Google Knowledge Graph[C]//The 11th International Semantic Web Conference. Boston, USA: [s. n.], 2012:1－4.

[59] Pujara J, Miao H. Getoor L Ontology-Aware Partitioning for Knowledge Graph Identification[C]//Proceeding of Automated Knowledge Base Construction. Atlanta, Georgia: [s. n.], 2013:19－24,.

[60] Khan M D A, Banerjee M. An Update Logic for Information Systems[J]. International Journal of Approximate Reasoning, 2014, 55(1):436－456.

[61] Du Y J, Peng Q Q, Gao Z Q. A Topic-Specific Crawling Strategy Based on Semantics Similarity[J]. Data Knowledge & Engineering, 2013, 88: 75－93.

[62] Santoso H A, Haw S C. Ontology Extraction from Relational Database: Concept Hierarchy as Background Knowledge[J]. Knowledge－Based Systems, 2011, 24(3):457－464.

[63] Ittoo A, Bouma G. Minimally-Supervised Extraction of Domain-Specific Part-Whole Relations Using Wikipedia as Knowledge-Base[J]. Data & Knowledge Engineering, 2013, 85: 57－79.

[64] Luis E, Zúrate S, Mariano D, et. al. FCANN: A New Approach for Extraction and Representation of Knowledge from ANN Trained via Formal Concept Analysis[J]. Neurocomputing, 2008, 71(13/15): 2670－2684.

[65] Zhou G D, Qian L H, Fan J X. Tree Kernel-Based Semantic Relation Extraction with Rich Syntactic and Semantic Information[J]. Information Sciences, 2010, 180(8):1313－1325.

[66] Khattak A M, Pervez Z. Time Efficient Reconciliation of Mappings in Dynamic Web Ontologies[J]. Knowledge-Based Systems, 2012, 35: 369－374.

[67] Liu L, Zhang P. Modeling Ontology Evolution with SetPi[J]. Information Sciences, 2014, 255(10):155－169.

(编校: 饶莉)