DOI: 10.3785/j.issn.1008-973X.2017.01.010

永恒语言学习研究与发展

丰小月[1], 梁艳春[1,2], 林希珣[1], 管仁初[1,2]
(1. 吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012;
2. 吉林大学珠海学院,符号计算与知识工程教育部重点实验室珠海市分实验室,广州 珠海 519041)

摘 要: 为了构建智能的语言学习模型,Tom M. Mitchell 教授 2010 年在美国人工智能协会(AAAI)上提出永恒语言学习(NELL)的概念. NELL 模型主要运用半监督学习和自然语言处理技术,持续不断地从互联网上获取大量文本,抽取知识,丰富知识库,使永恒语言学习模型变得更加智能. 介绍了永恒语言学习模型及模型的组成;描述了 NELL 的孕育和发展以及面临的 6 个主要问题,包括自省能力开发,每天需要短暂的人工监督,新谓词学习,新类型知识的学习;命名实体建模和更精确的统计学习模型构建;提出拟解决现有问题的新永恒语言学习模型.

关键词: 永恒语言学习(NELL);半监督学习;语义漂移;知识抽取;知识整合
中图分类号: TP 181 **文献标志码**: A **文章编号**: 1008-973X(2017)01-0082-07

Research and development of never-ending language learning

FENG Xiao-yue[1], LIANG Yan-chun[1,2], LIN Xi-xun[1], GUAN Ren-chu[1,2]
(1. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun* 130012, *China*; 2. *Zhuhai Labaratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai* 519041, *China*)

Abstract: Tom M. Mitchell proposed the never-ending language learning (NELL) in 2010 at American Association for Artificial Intelligence (AAAI) in order to develop an intelligent language learning model. Using semi-supervised learning and natural language processing technique, NELL continuously gets large number of texts from Internet, extracts knowledge and enriches its knowledge base, which improves its intelligence. The NELL model and its modules were introduced. The incubation and development of NELL were depicted. Six problems about NELL were described, including: self-reflection to decide what to do next; daily human interaction; discovery of new predicates to learn; learning additional types of knowledge about language; entity-level (rather than sting-level) modeling; more sophisticated probabilistic modeling throughout the implementation. A new NELL model tending to be a potential solution was proposed.

Key words: never-ending language learning (NELL); semi-supervised learning; semantic drift; knowledge extraction; knowledge integration

生物信息计算、物联网、社交网络以及云计算等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长,这一特征标志着大数据时代的来临[1]. 2011 年 2 月《Science》杂志推出专刊《Dealing with Data》[2],该专刊有文章指出,随着互联网的蓬勃发展和计算机计算能力的提升,信息抽取、信息检

收稿日期: 2016-08-06. **浙江大学学报(工学版)网址**: www.zjujournals.com/eng
基金项目: 国家自然科学基金资助项目(61602207, 61572228, 61373050, 61272207, 61472158);吉林省科技发展资助项目(20140520070JH);珠海市优势学科、广东省优势重点学科建设资助项目.
作者简介: 丰小月(1977—),女,博士,从事机器学习研究. ORCID: 0000-0003-3954-1333. E-mail: fengxy@jlu.edu.cn
通信联系人:管仁初,男,副教授. ORCID: 0000-0002-7162-7826. E-mail: guanrenchu@jlu.edu.cn

索、自动文摘以及机器阅读等研究方向已发展成为自然语言处理(natural language processing，NLP)研究领域的核心[3]. 在 2010 年 7 月第 24 届人工智能国际会议(AAAI)上，Tom M. Mitchell 教授首次提出永恒语言学习(never-ending language learning，NELL)的概念[4-5]. 永恒语言学习项目由美国国防部高级研究计划局、谷歌和雅虎公司共同资助，该研究的主要目的是不间断地从互联网上获取大量文本并从中抽取知识，不断丰富背后的知识库，使计算机变得更加智能. 永恒语言学习是利用网络大数据资源构建人工智能的典型应用之一. 2010 年发表在人工智能国际会议上的论文指出，永恒语言学习的准确率已经达到 74%.《纽约时报》分别于 2010 年 10 月和 2013 年 3 月两次报道了永恒语言学习的研究进展，引起了公众的广泛关注[6-7].

为了使国内相关领域研究人员对该研究有一个较全面的了解和认识，为自然语言处理、大数据挖掘等相关领域的研究者提供有益的参考，本文详细介绍了永恒语言学习模型，综述了永恒语言学习的研究现状以及目前面临的主要问题，给出了相关问题的改进方案.

1　永恒语言学习模型

坐落在卡耐基梅隆大学的永恒语言学习系统每周 7×24 小时不停歇地学习. 它每天主要执行 2 个任务：阅读任务，不断从网页信息中抽取知识，丰富基于结构化事实和知识的知识库；学习任务，根据已有的文本和系统具备的知识抽取能力，学习更智能的阅读方式，从而抽取更多更准确的信息. 如图 1 所示，由 Carlson 等提出的经典永恒语言学习模型主要包括 4 个部分：数据源、子系统组件(知识抽取)、知识库和知识整合.

数据源(data resources)：是永恒语言学习模型的基础，数据和知识主要来源于互联网和语料库. 自然语言理解的快速发展和互联网大数据的迅速膨胀，为永恒语言学习提供了源源不断的数据. NELL 的原始数据源包括：1)包含上百种类别和关系信息的初始本体. 2)针对初始类别和关系信息定义的部分训练样本. 3)从 ClueWeb09 中收集的 500 万个网页信息和十万个每日更新的谷歌 API 搜索查询. 4)每天短暂的人工修订知识.

子系统组件(subsystem components)：是永恒语言学习模型的核心. 在学习过程中，分别从耦合模式学习机(coupled pattern learner，CPL)、语言无关的耦合集扩展模型(coupled set expander for any language，CSEAL)、耦合构词法分类器(coupled morphological classifier，CMC)以及规则学习机(rule learner，RL) 4 个模块对关系实例进行学习和生成. 在这 4 个模块中，CPL 和 CSEAL 同是基于名词短语共现的统计方法(co-occurrence statistics)，主要区别在于前者将互斥约束和类型检查约束用于 ClueWeb09 数据，后者利用集扩展的方法从互联网实时数据中获取潜在知识[8-9]. CMC 从构词法特征角度对名词短语分类. RL 利用路径排序算法[10-11]发现概率豪恩子句(probabilistic Horn clauses)，进行知识库完善工作. 子系统组件从数据源中提取信息和潜藏知识，并提交给知识库.

知识库(knowledge base)：是永恒语言学习的成果，知识库中的知识可供人们浏览和评价(http://rtw.ml.cmu.edu/rtw/resources). 它由两层构成，第一层是候选事实集合(candidate facts)，第二层是概念集合(beliefs). 知识库从子系统组件模块中获取潜藏知识，由知识整合模块萃取出的最终知识被称为信念.

知识整合机(knowledge integrator)：是永恒语言学习的过滤器. 它负责从候选事实集合中挑选出知识，从而形成相关的结论(称作信念知识).

由图 1 可知，NELL 系统首先将成千上万网页中的语句抽取出来作为数据源，输送给子系统组件. 由子系统组件的 4 种算法从 4 种不同角度，分别对语句进行分析和知识抽取，新抽取的知识作为候选事实提交到知识库中. 通过知识整合算法从候选事实中整理和筛选出信念知识，信念知识和候选事实作为已知信息，将反馈指导子系统组件的知识抽取过程.

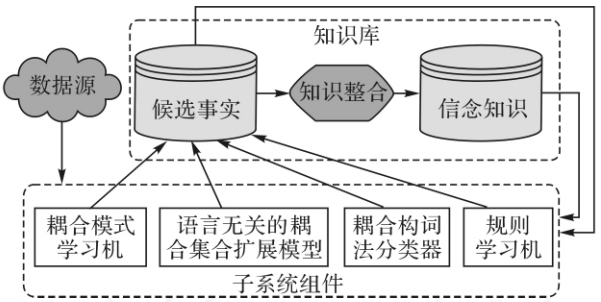

图 1　永恒语言学习架构

Fig. 1　Never-ending language learner (NELL) architecture

2　国内外研究现状

永恒语言学习研究最早可追溯到 1995 年，Tom

M. Mitchell 教授在《Robotics and Autonomous Systems》杂志上提出了终身学习(life-long learning)的概念[12]. 2008 年 5 月,Mitchell 教授等在《Science》杂志上发表文章,提出通过计算语义学和核磁共振图像,将具体名词与大脑活动联系起来的模型[13]. 该模型主要基于两个理论假设:大数据集合上的统计计算能够找到区分不同名词的语义特征;人类思考具体名词时的大脑活动变化可以通过语义特征的线性加权和求得. 2009 年 12 月 Mitchell 教授再次在《Science》上发表文章,指出可以通过互联网上的海量数据挖掘,开展知识发现和热点追踪[14]. 2010 年 7 月,Mitchell 教授在人工智能顶级会议上发表永恒语言学习模型[4],永恒语言学习模型是上述工作的集成和发展. 目前,该领域的研究已得到计算机学术界和社会的广泛关注与认可.《纽约时报》分别于 2010 年 10 月和 2013 年 3 月先后两次报道了永恒语言学习研究,Mitchell 教授及学生先后受邀到普林斯顿大学的图灵百年诞辰庆典[15-16]、机器学习顶级会议 ICML2010[17]、人工智能会议 AAAI2015[5]以及华盛顿大学杰出教授讲课计划[18]作相关专题报告.

自 2010 年以来,为了进一步提高永恒语言学习模型的准确率和学习效率,源源不断的有新研究成果发表. 这些研究工作主要分为两个部分:1)以语义学为基础的噪声去除和知识提取方法研究. 主要包括名词词组词义分析问题[19];名词类别关联分析策略[20];网页命名实体到不完整本体的映射算法[21];基于海量网页的本体比对方法[22];基于隐式句法特征的路径排序推理[10];基于个性化网页排序的概率一阶逻辑[10,19-23]. 2)以提高学习准确率为主的半监督学习算法研究,主要包括对偶贝叶斯集合算法[24];基于约束加权随机游走的知识推理[11];基于无监督信息抽取技术的命名实体识别[25];半结构化网页数据的低维表示[26-27];融合了弱监督信息的混合成员模型[28];能够发现新类别的探索学习[29]. 3)永恒语言学习的推广和应用,例如:Cohen 教授等将生物医学文本引入到 NELL 的学习对象中[30];Gupta 等以 NELL 为模板,提出永恒图像学习(never-ending image learning, NEIL)[31]. 永恒语言学习和永恒图像学习可以统称为永恒学习(never-ending learning). 虽然国外学者已在永恒学习的相关研究上取得了一定的成果,但目前国内专门针对永恒语言学习的研究尚未见文献报道.

永恒语言学习以网络大数据为研究对象,采用半监督学习(semi-supervised learning)策略,从少量标记样本出发,持续不断地在海量网页中抽取和整合知识. 半监督学习是一种介于监督学习和无监督学习之间的方法,主要包括半监督分类、半监督聚类和半监督回归等. 最早的半监督学习可以追溯到上世纪中期出现的自训练、自标记以及定向学习[32-34]. 半监督学习真正的兴起是在上世纪 90 年代,为了有效地解决自然语言处理问题,人们开始将目光转向半监督学习[35-37]. 半监督学习中较流行的算法有半监督支持向量机[38]、约束指导学习[39]、半监督流型学习[40-41]等. 这些算法的共同特点是通过少量已知信息指导大量未知标记样本的学习. 它既不需要监督学习所需的大量标记样本,又能够避免无监督学习对于解空间搜索的盲目性[42-43]. 半监督学习已被广泛地应用到数字图像处理、生物信息学、自然语言处理及信息安全等多个领域[44-46].

3 存在的问题

在永恒语言学习中,知识抽取所用到的关键技术是半监督学习,即:利用少量标记样本集合训练学习模型,然后用该模型去标记更多样本[4],其中每个类别包含少量已知标记的种子实例和模式(NELL 中包括 10～15 个实例和 5 种模式). 该半监督学习过程被称作引导学习(bootstrap learning). 在引导学习过程中,由种子出发,运用多视角学习(multi-view learning)分别从文本背景信息、网页结构信息、构词法特征以及规则学习 4 个角度进行新知识抽取和知识库的扩充. 尽管永恒语言学习采用了二阶学习(two-stage learning)方式且体现出了非凡的机器学习能力,但是网络大数据含有大量噪声且引导学习具有误差累积("语义漂移")现象[47-48],将导致学习准确率的缓慢下降. 永恒语言学习在很多方面仍需改进,主要分为以下 6 个问题:1)自省能力开发;2)每天需要短暂的人工监督;3)学习新谓词;4)学习新类型的知识;5)以命名实体建模代替名词字符串;6)构建更精确的统计学习模型[4,16].

从上述研究现状来看,永恒语言学习研究已经取得了一定进展,并且得到了学术界的认可. 目前,这方面研究处于起步阶段,知识抽取和知识整合的准确率和效率远远不能满足人们的要求. 同时,网络大数据中蕴含着噪声和垃圾网页,为永恒语言学习带来了困难. 面对永恒语言学习存在的问题,可以从下述 4 个方面着手进行研究.

1)减少误差累积,提升永恒语言学习准确率. 由于网络大数据具有复杂性、不确定性和涌现性等特

点[49]，即使采取了多视角学习，也只能达到 74％的准确率[4]. 同时，随着时间推移和知识库的增长，永恒语言学习的准确率逐步下降. 为了保证准确率，只好每天都加入一定的人工监督（约为 10～15 min）. 造成上述现象的主要原因是引导学习算法具有误差累积效应. 目前，永恒语言学习及相关研究的核心算法是引导学习. 设计出能够有效避免或减少误差累积的算法，是该领域研究的难点问题. 文献[24]表明，通过增加约束学习的方法能够改善该问题，例如类别之间的互斥信息和专家知识等[24].

2）利用特征学习降低噪声，提高永恒语言学习效率. 永恒语言学习中的知识抽取部分依靠的是多视角协同学习. 该算法的基本思想是在相互无关的特征集合上构建分类器，然后合并分类结果. 其中特征选择、特征抽取和特征表示是该算法的关键问题. Mitchell 教授指出，在永恒语言学习过程中，若能用特征标记取代具体样本标记，则能够大大减少人工注释的时间[4]. 引入优秀的特征学习算法，能够在一定程度上降低噪声的影响，有效解决大数据量文本挖掘带来的高维稀疏和“维度灾难”等问题，从而提高永恒语言学习的效率.

3）增加新知识和新策略，增强永恒语言学习能力. 目前，永恒语言学习算法主要是从 Google 和原始语料库 ClueWeb09 上获取知识. Wang 等[50]以生物医学知识获取为例，对四大搜索引擎（Google，Yahoo，Bing and Ask.com）进行比较分析，发现它们各有千秋且都有缺陷，建议采用多搜索引擎进行多源知识获取. 永恒语言学习是面向网络大数据的典型应用之一，它是从文本背景、网页结构信息、构词法特征以及规则学习 4 个角度进行知识抽取. 然而，一方面，海量网页数据中蕴含的丰富潜藏信息和知识没有被充分利用和挖掘，新知识或新信息的添加能够增强永恒语言学习的知识抽取能力，例如：Gardner 等[51]尝试构建跨语言模型，从不同的语言中抽取知识，Wijaya 等[52]利用大量未标记网页文本实现了不同语言动词在知识库上的对齐. 另一方面，目前在永恒语言学习的知识库构建过程中，客观事实通过知识整合升级成概念后，将永久存入知识库中[4]. 随着知识库规模的日积月累和引导学习存在的误差，在知识库中难免会存在错误的实例和关系. 研究新的概念降级或回滚操作，将增强永恒语言学习的知识整合能力.

4）提出更加适应大规模知识库的知识表示模型. 在有效学习实体、关系语义信息的基础上，增强永恒语言学习系统的知识表示能力，有效解决知识库长尾分布带来的数据稀疏问题[53]. 目前，永恒语言学习系统主要采用基于随机游走的路径排序算法完成规则学习任务，该算法不能很好地结合全局信息，在处理稀疏实体、关系时面临严重的过拟合问题[54]. 可以在不受逻辑表示图模型限制的基础上[55]，考虑更加合理的知识建模方法，尝试从三阶张量分解[56-57]、非参数贝叶斯聚类[58-59]、神经网络模型[60-61]等多个角度学习知识库中的实体、关系表示，提升 NELL 系统在知识库完善和知识推理领域的表现. 例如：与词和句子级别的语义分析不同，Yang 等[62]在 NAACL2016 上提出文章背景下的事件和本体联合知识抽取.

4　具体改进方案

为了丰富与发展基于永恒语言学习的理论与算法，使新理论和算法能够快速地从海量网页数据中挖掘出更有意义和价值的知识，丰富自身的知识库，提高推理判断能力，针对新知识和新谓词的发现、更准确统计模型构建以及误差累积等问题，充分发掘海量网页数据中的隐知识、构造新学习算法、对影响算法效率的各种因素进行理论分析是解决上述问题的有效手段. 本文建议深入研究如下几个方面.

1）深入挖掘海量网页中的元知识，增加知识抽取视角，解决永恒语言学习框架学习新类型知识的问题.

2）提出增量半监督聚类和中心特征选择算法，为了避免随知识库快速膨胀而计算大规模稀疏矩阵带来的复杂度快速增长和维度灾难问题. 利用筛选的特征与规则学习相结合，发现新谓词.

3）提出分层输入深度神经网络模型实现特征抽取和事实分类. 融合上述新提出的算法及模型，大幅度提高永恒语言学习的准确率和效率. 逐步解决永恒语言学习中的误差累积问题，减少学习过程的人工指导.

4）构建纠错学习算法，使知识整合与知识抽取阶段都能够进行定期评价，使永恒语言学习具有一定自省能力.

如图 2 所示，具体解决方案如下. 考虑将科技文献网页和社交网站等不同类型网页引入到永恒语言学习中；尝试利用复杂网络对网页中蕴含的深层知识进行提取，为知识库提供新的候选知识；结合半监督聚类算法和增量学习，对来自不同数据源的数据进行整合和预处理；研究将特征选择算法、构词法和规则学习相结合，构建谓词发现方法；进一步分析多

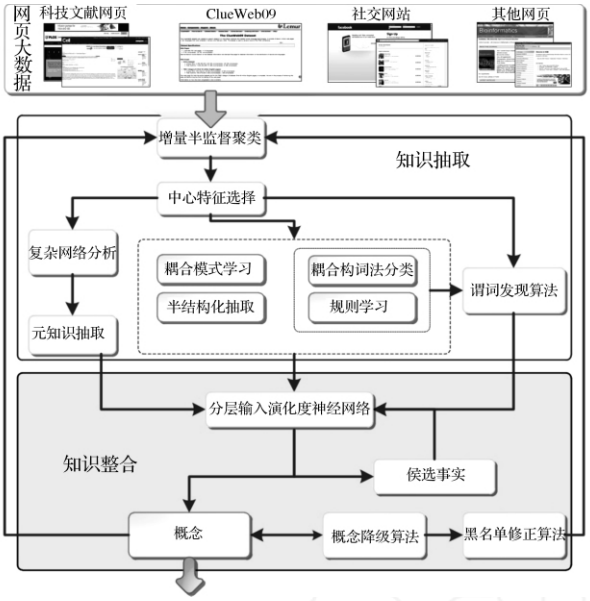

图 2 新永恒语言学习架构

Fig. 2 Updated architecture for NELL

源知识和多种学习策略的非线性融合问题,考虑利用深度神经网络模型进行特征抽取和事实分类;现有 NELL 模型的知识库只能够添加事实,须分析影响永恒语言学习准确率的因素以及它们之间内在的联系,提出纠错学习算法.

5 结 语

永恒语言学习可以被看作是自然语言处理和半监督学习算法相结合的最典型的应用之一.它的研究与发展体现了人工智能领域最前沿的成果.自 2010 年 10 月被提出以来,NELL 已受到包括学术界(如图灵百年诞辰庆典和机器学习顶级会议)和社会(如纽约时报)的广泛关注.从国内外研究现状来看,NELL 经历了研究的孕育和诞生阶段,正处于快速发展时期.从 NELL 存在的问题来看,虽然该研究取得了令人瞩目的成就,但有很多需要完善和改进之处.本文给出一种方案,希望国内外学者能够更多地关注相关研究,完善相关理论,加快永恒语言学习在国内的推广,共同研究构建永恒语言学习的中文模型.

参考文献(References):

[1] HOWE D, COSTANZO M, FEY P, et al. Big data: the future of biocuration [J]. **Nature**, 2008, 455:47-50.

[2] Science. Special online collection: dealing with data [EB/OL]. (2011-02-11). http://www.sciencemag.org/site/special/data/.

[3] EVANS J A, FOSTER J G. Meta knowledge [J]. **Science**, 2011, 331: 721-725.

[4] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning [C] // **Proceedings of AAAI 2010**. Atlanta: AAAI, 2010: 1306-1313.

[5] MITCHELL T M, COHEN W W, TALUKDAR P P, et al. Never-ending learning [C] // **Proceeding of AAAI**. Austin Texas: AAAI, 2015: 2302-2310.

[6] LOHR S. NELL is a computer that reads the web-with a little human help [N/OL]. New York Times, 2013-03-11. http://bits.blogs.nytimes.com/2013/03/11/nell-is-a-computer-that-reads-the-web-with-a-little-human-help/.

[7] LOHR S. Aiming to learn as we do, a machine teaches itself [N/OL]. New York Times, 2010-10-05. http://www.nytimes.com/2010/10/05/science/05compute.html.

[8] CARLSON A, BETTERIDGE J, WANG R C, et al. Coupled semi-supervised learning for information extraction [C] // **Proceeding of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)**. New York: ACM, 2010: 101-110.

[9] WANG R C, COHEN W W. Character-level analysis of semi-structured documents for set expansion [C] // **Proceeding of EMNLP**. Singapore: ACL, 2009: 1503-1512.

[10] GARDNER M, TALUKDAR P P, KISIEL B, et al. Improving learning and inference in a large knowledge-base using latent syntactic cues [C] // **Proceeding of EMNLP**. Seattle: ACL, 2013: 833-838.

[11] LAO N, MITCHELL T M, COHEN W W. Random walk inference and learning in a large scale knowledge base [C] // **Proceeding of EMNLP**. Edinburgh: ACL, 2011: 529-539.

[12] THRUN S, MITCHELL T M. Lifelong robot learning [J]. **Robotics and Autonomous Systems**, 1995, 15 (12): 25-46.

[13] MITCHELL T M, SHINKAREVA S V, CARLSON A, et al. Predicting human brain activity associated with the meanings of nouns [J]. **Science**, 2008, 320: 1191-1195.

[14] MITCHELL T M. Mining our reality [J]. **Science**, 2009, 326: 1644-1645.

[15] Princeton University. Lecture videos and speaker bios of Princeton centennial celebration of Alan Turing [EB/OL]. (2013-06-14). http://www.princeton.edu/turing/speakers.

[16] Carnegie Mellon University. Read the Web [EB/OL]. [2016-08-05]. http://rtw.ml.cmu.edu/rtw/

index. php.

[17] International conference machine learning. Invited speakers [EB/OL]. (2010-06-24). http://www.icml2010.org/invited.html.

[18] University of Washington. UW CSE speaker abstract of Tom Mitchell [EB/OL]. (2010-10-21). https://www.cs.washington.edu/htbin-post/mvis/mvis?ID=957.

[19] KRISHNAMURTHY J, MITCHELL T M. Which noun phrases denote which concepts? [C] // **Proceeding of ACL**. Stroudsburg: ACL, 2011: 570-580.

[20] MOHAMED T, HRUSCHKA Jr E R, MITCHELL T M. Discovering relations between noun categories [C] // **Proceeding of EMNLP**. Edinburgh: ACL, 2011: 1447-1455.

[21] DALVI B, COHEN W W, CALLAN J. Classifying entities into an incomplete ontology [C] // **Proceeding of the 2013 Workshop on Automated Knowledge Base Construction**. San Francisco: ACM, 2013: 31-36.

[22] WIJAYA D T, TALUKDAR P P, MITCHELL T M. PIDGIN: ontology alignment using web text as Interlingua [C] // **Proceeding of CIKM**. San Francisco: ACM, 2013: 589-598.

[23] WANG W Y, MAZAITIS K, COHEN W W. Programming with personalized pagerank: a locally groundable first-order probabilistic logic [C] // **Proceeding of CIKM**. San Francisco: ACM, 2013: 2129-2138.

[24] VERMA S, JR HRUSCHKA E R. Coupled Bayesian sets algorithm for semi-supervised learning and information extraction [C] // **Proceeding of ECML PKDD**. Bristol: Springer, 2012: 307-322.

[25] DALVI B, COHEN W W, CALLAN J. WebSets: extracting sets of entities from the web using unsupervised information extraction [C] // **Proceeding of the 5th ACM International Conference on Web Search and Data Mining**. Seattle: ACM, 2012: 243-252.

[26] DALVI B, COHEN W W, CALLAN J. Collectively representing semi-structured data from the web [C] // **Proceeding of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction Table of Contents**. Montreal: ACL, 2012: 7-12.

[27] DALVI B, COHEN W W, CALLAN J. Very fast similarity queries on semi-structured data from the web [C] // **2013 SIAM International Conference on Data Mining**. Austin: SIAM, 2013: 512-520.

[28] BALASUBRAMANYAN R, DALVI B, COHEN W W. From topic models to semi-supervised learning: biasing mixed-membership models to exploit topic-indicative features in entity clustering [G] // **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. Prague: Springer, 2013: 628-642.

[29] DALVI B, COHEN W W, CALLAN J. Exploratory learning [G] // **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. Prague: Springer, 2013: 128-143.

[30] MOVSHOVITZ-ATTIAS D, COHEN W W. Bootstrapping biomedical ontologies for scientific text using NELL [C] // **The 11th Workshop on Biomedical Natural Language Processing**. Montreal: ACL, 2012: 11-19.

[31] CHEN X L, SHRIVASTAVA A, GUPTA A. NEIL: extracting visual knowledge from web data [C] // **Proceeding of ICCV**. Sydney: IEEE, 2013: 1409-1416.

[32] SCUDDER H J. Probability of error of some adaptive pattern-recognition machines [J]. **IEEE Transaction on Information Theory**, 1965, 11 (3): 363-371.

[33] FRALICK S C. Learning to recognize patterns without a teacher [J]. **IEEE Transaction on Information Theory**, 1967, 13(1): 57-64.

[34] AGRAWALA A K. Learning with a probabilistic teacher [J]. **IEEE Transaction on Information Theory**, 1970, 16(4): 373-379.

[35] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods [C] // **Proceeding of ACL**. Boston: ACL, 1995: 189-196.

[36] MCCALLUM A, NIGAM K. A comparison of event models for naive Bayes text classification [C] // **Proceeding of AAAI**. Madison: AAAI, 1998: 41-48.

[37] MCCALLUM A, NIGAM K. Employing EM and pool-based active learning for text classification [C] // **Proceeding of ICML**. Madison: Morgan Kaufmann, 1998: 350-358.

[38] WANG L, CHAN K L, ZHANG Z. Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval [C] // **Proceeding of CVPR**. Los Alamitos: IEEE, 2003: 629-634.

[39] KLEIN D, KAMVAR S D, MANNING C. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering [C] // **Proceeding of ICML**. Australia: Morgan Kaufmann, 2002: 307-314.

[40] ZHU X, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using gaussian fields and harmonic functions [C] // **Proceeding of ICML**. Washington D. C.: Morgan Kaufmann, 2003: 912-919.

[41] LIN T, ZHA H B. Riemannian manifold learning [J].

IEEE Transaction on Pattern Analysis and Machine Intelligence, 2008, 30(5): 796-809.

[42] ZHOU Z H, LI M. Semi-supervised regression with co-training style algorithms [J]. **IEEE Transaction on Knowledge and Data Engineering**, 2007, 19(11): 1479-1493.

[43] 管仁初. 半监督聚类的研究与应用[D]. 吉林:吉林大学, 2010.
GUAN Ren-chu. Research and application of semi-supervised clustering algorithms [D]. Jilin: Jilin University, 2010.

[44] CHAPELLE O, SCHÖLKOPF B, ZIEN A. **Semi-supervised learning** [M]. Cambridge: MIT, 2006: 1-30.

[45] GUAN R C, SHI X H, MARCHESE M, et al. Text clustering with seeds affinity propagation [J]. **IEEE Transaction on Knowledge and Data Engineering**, 2011, 23(4): 627-637.

[46] YANG C, BRUZZONE L, GUAN R C, et al. Incremental and decremental affinity propagation for semisupervised clustering in multispectral images [J]. **IEEE Transaction on Geoscience and Remote Sensing**, 2013, 51(3): 1666-1679.

[47] RILOFF E, JONES R. Learning dictionaries for information extraction by multi-level bootstrapping [C] // **Proceeding of AAAI**. Menlo Park: AAAI, 1999: 474-479.

[48]CURRAN J R, MURPHY T, SCHOLZ B. Minimising semantic drift with mutual exclusion bootstrapping [C] // **Proceeding of PACLING**. Melbourne: University of Melbourne, 2007: 172-180.

[49] 王元卓,靳小龙,程学旗. 网络大数据:现状与展望[J]. 计算机学报,2013, 36(6): 1125-1138.
WANG Yuan-zhuo, JIN Xiao-long, CHENG Xue-qi. Network big data: present and future [J]. **Chinese Journals of Computers**, 2013, 36(6): 1125-1138.

[50] WANG L P, WANG J X, WANG M, et al. Using Internet search engines to obtain medical information: a comparative study [J]. **Journal of Medical Internet Research**, 2012, 14(3): e74.

[51] GARDNER M, HUANG K J, PAPALEXAKIS E, et al. Translation invariant word embeddings [C] // **Proceeding of EMNLP**. Lisbon: ACL, 2015: 1084-1088.

[52] WIJAYA D T, MITCHELL T M. Mapping verbs in different languages to knowledge base relations using web text as interlingua [C] // **Proceedings of NAACL**. San Diego: ACL, 2016: 818-827.

[53] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展学报,2016, 53(2): 247-261.
LIU Zhi-yuan, SUN Mao-song, LIN Yan-kai, et al. Knowledge representation learning: a review [J]. **Journal of Computer Research and Development**, 2016, 53(2): 247-261.

[54] GARDNER M, MITCHELL T M. Efficient and expressive knowledge base completion using subgraph feature extraction [C] // **Proceedings of EMNLP**. Lisbon: ACL, 2015: 1488-1498.

[55] RICHARDSON M, DOMINGOS P. Markov logic networks [J]. **Machine Learning**, 2006, 62(1): 107-136.

[56] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data [C] // **Proceedings of ICML**. Washington: Morgan Kaufmann, 2011: 809-816.

[57] NICKEL M, TRESP V, KRIEGEL H P. Factorizing Yago: scalable machine learning for linked data [C] // **Proceedings of the 21st International Conference on World Wide Web**. Lyon: ACM, 2012: 271-280.

[58] SUTSKEVER I, TENENBAUM J B, SALAKHUTDINOV R R. Modelling relational data using Bayesian clustered tensor factorization [C] // **Advances in Neural Information Processing Systems**. Vancouver: MIT, 2009: 1821-1828.

[59] KEMP C, TENENBAUM J B, GRIFFITHS T L, et al. Learning systems of concepts with an infinite relational model [C] // **Proceeding of AAAI**. Boston: AAAI, 2006: 381-388.

[60] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C] // **Advances in Neural Information Processing Systems**. Lake Tahoe: MIT, 2013: 2787-2795.

[61] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] // **Advances in Neural Information Processing Systems**. Lake Tahoe: MIT, 2013: 926-934.

[62] YANG B S, MITCHELL T M. Joint extraction of events and entities within a document context [C] // **Proceedings of NAACL**. San Diego: ACL, 2016: 289-299.