Statistics and Application 统计学与应用, 2020, 9(3), 458-466
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.93049

Hans汉斯

Research Hotspot and Trend of Named Entity Recognition in China

—Analysis of Knowledge Map Based on CNKI

Jiangnan Xu[1], Jiangming Shen[2], Xin Wang[1], Zhiyong Zeng[3*]

[1]Yunnan University Data Operation and Management Engineering Research Center, School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
[2]Enterprise Information Department of China Telecom Corporation Limited Yunnan Branch, Kunming Yunnan
[3]Yunnan University Data Operation and Management Engineering Research Center, School of Information, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 1371441180@qq.com, *zengzhiyong725@163.com

Received: Jun. 1st, 2020; accepted: Jun. 15th, 2020; published: Jun. 23rd, 2020

Abstract

Named entity recognition has made great achievements in China after decades of development. In this paper, CiteSpace is used as an analysis tool to visually analyze the papers on the topic of named entity recognition in CNKI. Through author analysis, research organization analysis, and keyword analysis, this paper discusses the research path and research focus of named entity recognition in China. The results show that a number of influential authors and research institutions have emerged in China. The research path of our country has gone through three stages. At present, the research hotspot in our country is the deep learning method of named entity recognition.

Keywords

Named Entity Recognition, Visual Analysis, CiteSpace

国内命名实体识别研究的热点和趋势

——基于CNKI的知识图谱分析

徐江南[1]，沈江明[2]，王　鑫[1]，曾志勇[3*]

[1]云南财经大学统计与数学学院、云南省高校数据化运营管理工程研究中心，云南　昆明

*通讯作者。

文章引用：徐江南，沈江明，王鑫，曾志勇．国内命名实体识别研究的热点和趋势[J]．统计学与应用，2020，9(3)：458-466. DOI: 10.12677/sa.2020.93049

[2]中国电信股份有限公司云南分公司企业信息化部，云南 昆明
[3]云南财经大学信息学院，云南省高校数据化运营管理工程研究中心，云南 昆明
Email: 1371441180@qq.com, *zengzhiyong725@163.com

收稿日期：2020年6月1日；录用日期：2020年6月15日；发布日期：2020年6月23日

摘 要

命名实体识别经过几十年的发展在我国已经取得了丰厚的成果。本文使用CiteSpace作为分析工具，对中国知网学术期刊库中的以命名实体识别为主题的论文进行可视化分析。通过作者分析、研究机构分析、和关键词分析对我国命名实体识别的研究路径和研究热点进行探讨。研究结果发现，我国已经出现了一批有影响力的作者和研究机构。我国的研究路径经历了三个阶段，目前国内的研究热点是命名实体识别的深度学习方法。

关键词

命名实体识别，可视化分析，CiteSpace

Copyright © 2020 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

命名实体识别是信息提取、问答系统、句法分析、机器翻译等应用领域的重要基础工具。一般来说，命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)。正确率(Precision)、召回率(Recall)和 F1 值(F1 score)常被用来作为命名实体识别的评价指标。1996 年，命名实体评测作为信息抽取的一个子任务被引入 MUC-6 [1]。不仅 MUC 会议，ACE 项目和 CoNLL 会议都进一步推动了命名实体识别的发展[2]。命名实体识别研究的主要方法也经历了从基于规则的方法，到基于统计机器学习的方法，再到基于深度学习的方法的发展路径。我国对于命名实体识别的研究又有自身的特点。汉语的命名实体识别比英语要复杂得多，英语的命名实体往往是首字母大写的并且汉语文本没有类似英语文本中空格之类的显式标示词边界的标示符，分词和命名实体识别互相影响[3]。对我国以往的命名实体识别研究进行分析，有助于总结经验和展望未来的发展方向，更有利于命名实体识别在我国的发展。

2. 数据来源

本文数据来源于中国知网(CNKI)学术期刊库，以命名实体识别为主题对 1996 年至 2019 年的文献进行检索，得到 631 篇中文参考文献，对文献进行筛选、去重最终获取到 628 篇参考文献。

3. 研究方法

CiteSpace 是应用 Java 语言开发的一款信息可视化软件，它可以对特定领域文献进行计量，以探寻出

学科领域演化的关键路径及其知识拐点，并通过一系列可视化图谱的绘制来形成对学科演化潜在动力机制的分析和学科发展前沿的探测[4]。本文采用 Citespace5.5R2 版本对获取到的文献进行研究作者和研究机构进行合作共现分析，对关键词进行关键词共现、关键词突显和关键词聚类的时间线分析。

4. 数据结果及分析

4.1. 总发文量分析

利用 Excel 2016 对 1996 年至 2019 年参与分析的数据进行统计得出图 1。

从总体趋势可以看出国内以命名实体为主题的发文量大致可以分为三个阶段：1) 1996 年至 2005 年为萌芽阶段。发文量较少，还一度出现 0 篇发文量的情况，因为真正将汉语命名实体识别研究作为重要的研究领域，并组织较大规模评测会议，是从 SIGHAN Bakeoff-2006 开始的[1]。2) 2006 年至 2013 年为稳定发展阶段。该阶段每年论文的产出量比较稳定维持在 23 篇左右。该阶段的研究主要集中在统计的机器学习方法。3) 2014 年至 2019 年为快速发展阶段。该阶段论文的发文量出现爆发式增长。近年来源于神经网络模型的深度学习技术成为机器学习领域新的热潮，对于命名实体识别的发展带来强大的发展动力[2]。

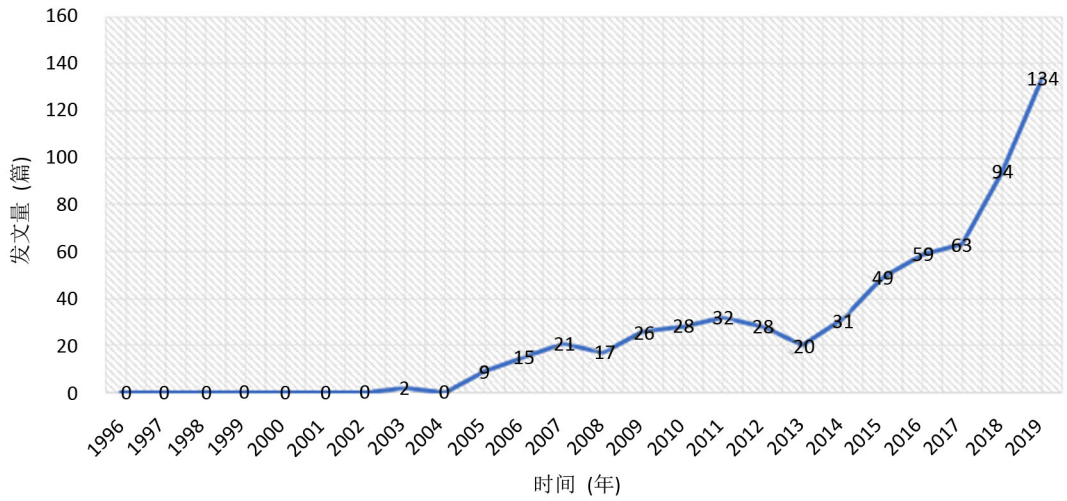

Figure 1. Time distribution of named entity recognition volume in 1998~2019
图 1. 1998~2019 年命名实体识别发文量的时间分布图

4.2. 作者分析

4.2.1. 核心作者分析

使用 CNKI 计量可视化分析中的作者分布并结合 CiteSpace 中的合作共现分析对发文量较多的作者进行统计分析。然而高产作者并不一定是该领域的核心作者[5]。表 1 将利用普赖斯[6]公式确定出的 34 位作者作为核心作者候选人，根据作者的发文量以及 CAJD 的被引频次并用 Excel 计算出每个作者的综合指数[7]。

Table 1. Distribution volume and comprehensive index of core author candidates
表 1. 核心作者候选人发文量及综合指数

排名	作者	发文量	被引频次	综合指数
1	余正涛	13	186	237
2	吕学强	10	470	401

Continued

3	郭剑毅	10	184	210
4	林鸿飞	9	56	116
5	艾山·吾买尔	9	23	93
6	吐尔根·依布拉音	8	23	85
7	卡哈尔江·阿比的热西提	7	23	76
8	朱艳辉	7	6	65
9	周国栋	7	70	108
10	刘挺	7	213	203
11	买合木提·买买提	6	23	67
12	姬东鸿	6	42	80
13	线岩团	6	68	98
14	赵铁军	6	84	108
15	王东波	6	58	91
16	王路路	6	14	61
17	徐啸	6	4	55
18	关毅	5	237	202
19	于江德	5	26	61
20	杨志豪	5	34	66
21	施水才	5	455	348
22	何云琪	4	2	36
23	崔雷	4	20	48
24	艾斯卡尔·艾木都拉	4	50	68
25	邢富坤	4	1	35
26	雷树杰	4	1	35
27	李飞	4	6	39
28	于洪志	4	1	35
29	程学旗	4	57	73
30	严馨	3	9	32
31	顾佼佼	3	36	50
32	姜文志	3	36	50
33	王健	3	22	41
34	王闻慧	3	0	26

根据综合指数法表 1 中核心作者候选人的综合指数大于或等于 100 的为核心作者，可以确定 9 位核心作者：吕学强、施水才、余正涛、郭剑毅、关毅等人。北京信息科技大学的吕学强和施水才《基于层叠隐马尔可夫模型的中文命名实体识别》[8]在 CNKI 上有 365 次的被引次数并且吕学强还注重搜索日志和查询日志中命名实体的识别研究；昆明理工大学的余正涛和郭剑毅不仅尝试使用各种统计机器学习方

法：条件随机场、层叠条件随机场、隐马尔可夫模型，余正涛还对英语、越南语、柬埔寨语不同语种的命名实体识别都有研究；哈尔滨工业大学的关毅一直侧重于电子病历的命名实体研究。

4.2.2. 作者合作共现分析

利用 CiteSpace 进行作者合作共现分析，节点的大小可以看出作者发文量的多少，节点的连线和粗细可以反映在命名实体识别的研究领域作者之间的合作关系和合作强度。

Figure 2. Cooperation and co-occurrence of authors
图 2. 作者合作和共现图谱

由图 2 可见，形成了以吕学强、施水才和余正涛、郭剑毅为核心的两个作者群。也形成了分别以林鸿飞和周国栋为核心的两个作者群，其他核心作者并没有形成合作强度较强的作者合作群。

4.3. 研究机构分析

由图 3 可见命名实体识别的研究主要是学校和研究所，云南省计算机技术应用重点实验室智能信息处理研究所与昆明理工大学信息工程与自动化学院、新疆大学新疆多语种信息技术实验室与新疆大学信息与工程学院存在较强的合作关系，没有形成大规模的合作机构。

4.4. 关键词分析

4.4.1. 关键词共现分析

关键词共现分析可以看到各个关键词之间的联系，从而可以看到各主题之间的联系。在 CiteSpace 软件中进行如下参数设置：Time Slicing From 1998 To 2019 Years Per Slice=1，Node Types=Keyword，TopN=50，Pruning 选择 Pathfinder、Pruning sliced networks 和 Pruning the merged network。并根据关键词共现图谱，把相同意义的关键词进行合并。由图 4 可见条件随机场、深度学习、自然语言处理和信息抽取是高频关键词，也是命名实体识别研究的热点。与命名实体识别相连的最大熵模型、深度学习、条件随机场和隐马尔科夫模型是研究命名实体识别的方法。自然语言处理、文本挖掘、问答系统和知识图谱是命名实体识别被应用到的领域。

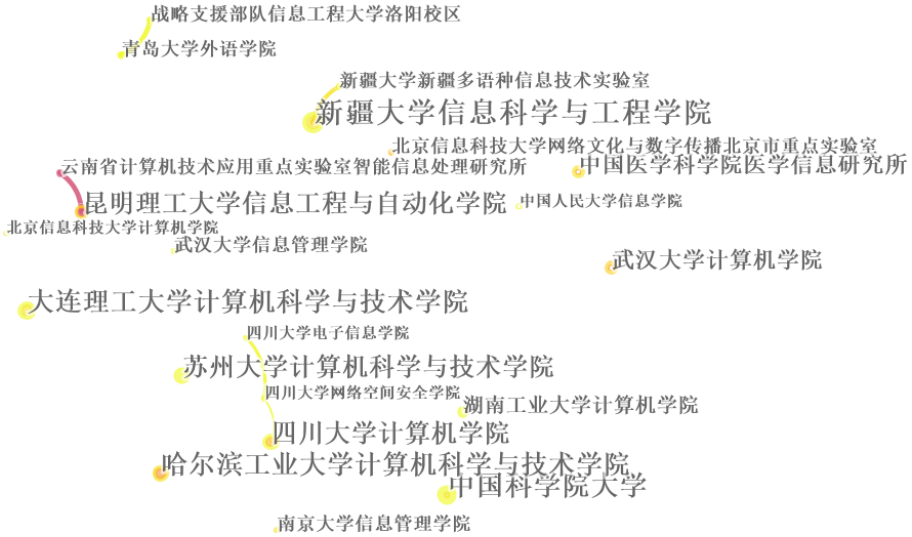

Figure 3. Institutional cooperation and co realization map
图 3. 机构合作共现图谱

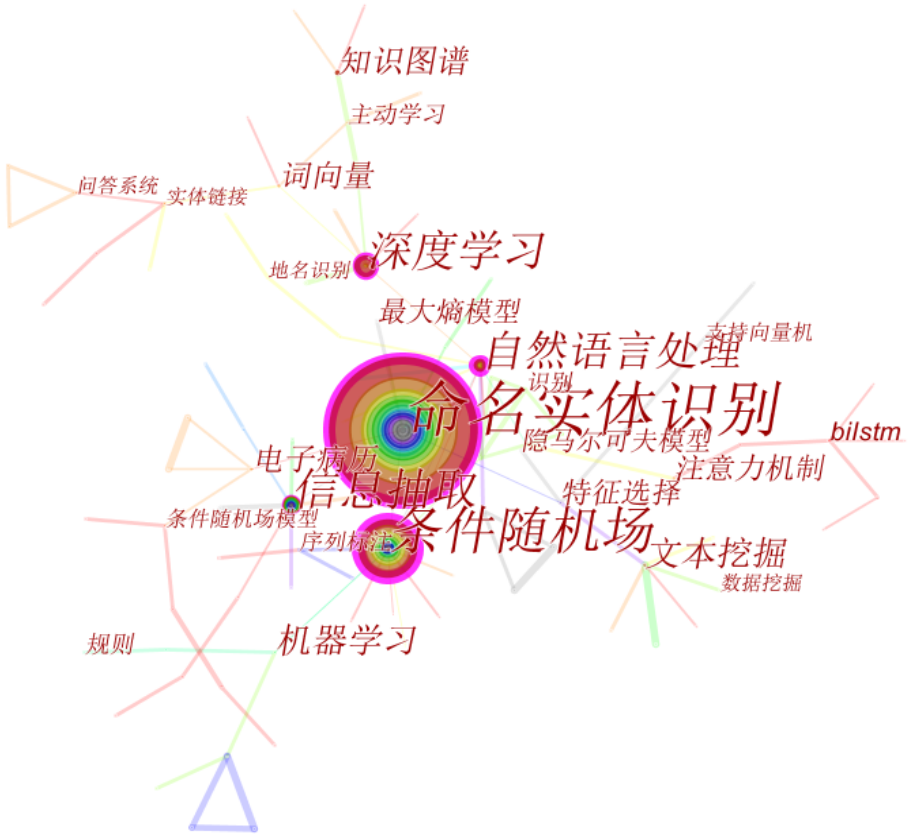

Figure 4. Key words co-occurrence map
图 4. 关键词共现图谱

将关键词按照频次排序并选择其中的前 12 个可得表 2。由表 2 可得以下结论。从表 2 频次上看命名实体识别的研究主要包括条件随机场、深度学习、机器学习、词向量等方法上。命名实体识别也被运用在电子病历的命名实体识别上面。从表 2 中介中心性上看除知识图谱外其余的关键词在关键词知识图谱

中都是关键节点。

Table 2. Keyword statistics
表 2. 关键词统计

关键词	频次	中介中心性
命名实体识别	336	0.86
条件随机场	145	1.03
信息抽取	66	0.28
深度学习	55	0.51
自然语言处理	53	0.9
文本挖掘	24	0.16
机器学习	24	0.19
知识图谱	17	0.07
词向量	15	0.3
电子病历	15	0.22
注意力机制	13	0.19
特征选择	13	0.3

4.4.2. 关键词突现分析

通过关键词突显分析可以看到在命名实体识别研究领域在不同时间段内的受关注的主题。由图 5 可见中文信息处理、计算机应用、隐马尔可夫模型和最大熵模型是从 2006 年至 2010 年的突现词并且是这七个突现词里面强度最大的突显词。可见从 2006 年起国内开始关注中文命名实体识别的研究，并且关注识别生物医学、汽车评论、新闻等文本中的命名实体，而隐马尔可夫模型和最大熵模型是最常被使用到的模型。命名实体识别作为一项基础的任务常被应用于信息抽取从 2005 年到 2015 年都有很高的突显性，可见命名实体识别在信息抽取的任务中一直都收到长时间的关注。命名实体识别可作为一项特征被用于其他任务，对文本特征的选择也会影响命名实体识别的效果，所以特征选择和命名实体识别常常一起被作为关键词。近年来规则常与统计学方法相结合，提升命名实体识别的 F1 值。

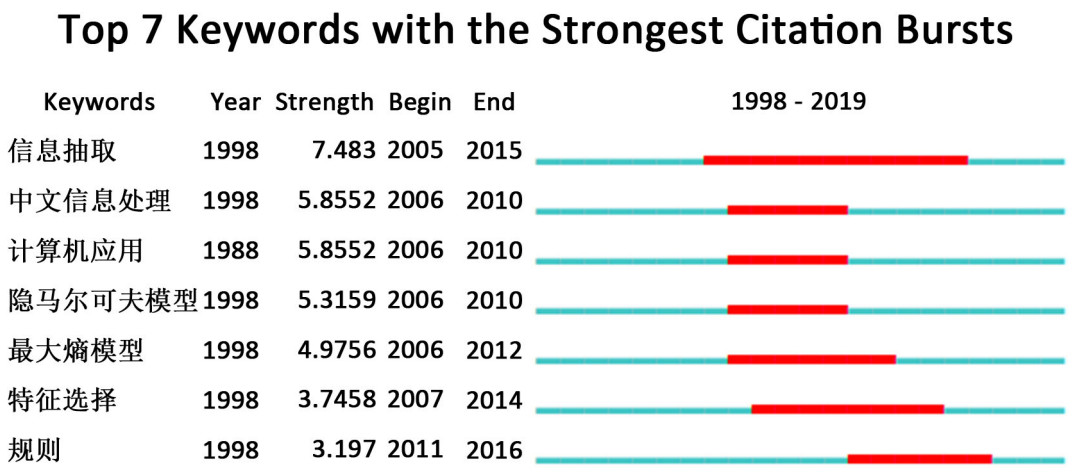

Figure 5. Key words highlighting map
图 5. 关键词突显图谱

4.4.3. 关键词聚类分析

关键词聚类分析可以更好的看出一个领域的研究重点，利用 CiteSpace 对关键词进行聚类并绘制时间线图，可以看到每一个聚类随时间的研究趋势，也可以看出这个领域的研究趋势。图 6 中 Q 值为 0.8253，S 值为 0.526，说明该图谱聚类结构显著，聚类合理。关键词聚类共分为十个类别，#0、#1、#6 和#8 四个聚类主要表现了与命名实体识别研究相关的方法。#2、#3、#4 和#5 是命名实体识别被广泛应用到的领域。#7 和#9 是命名实体识别研究的专业领域。

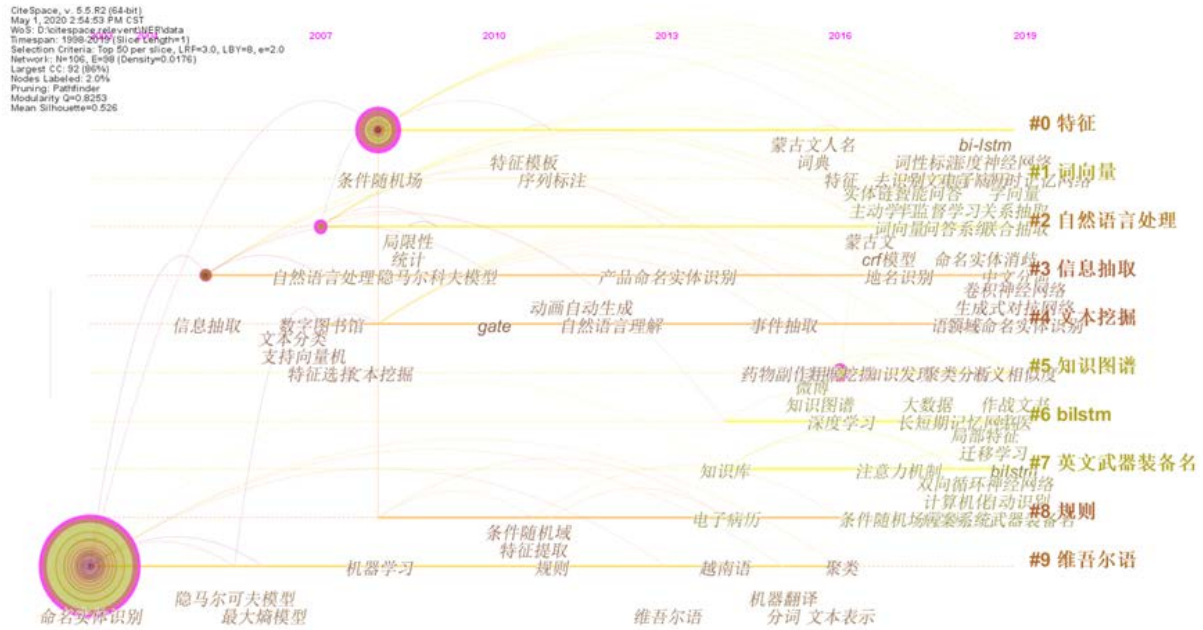

Figure 6. Knowledge map of keyword timeline
图 6. 关键词时间线知识图谱

由图 6 可见，国内对命名实体识别的研究大致分为三个阶段：1) 2003 年至 2013 国内把传统的机器学习方法用于中文文本的命名实体识别，传统的机器学习算法包括：条件随机场、隐马尔可夫模型、最大熵模型等。2) 2013 年至 2016 国内开始解决专业领域和各种语言的命名实体识别问题，语言包括蒙古语、维吾尔语、越南语等。3) 2016 年至 2019 年近年来，源于神经网络模型的深度学习技术成为机器学习领域新的热潮[2]。国内开始把深度学习技术应用在命名实体识别上面，深度学习包括：长短期记忆网络、卷积神经网络、注意力机制等。

5. 总结及期望

运用 CiteSpace 对中国知网检索到的 628 篇文献进行知识图谱绘制和分析有如下发现。

随着国际命名实体识别研究的不断发展，2003 年举办的“863 评测”[2]，汉语命名实体识别作为评测任务首次被提出。更加推动了我国对命名实体的研究，形成了一部分具有影响力的核心作者和一些研究机构。在核心作者分析中，核心作者大部分都是关注传统机器学习方法，而新兴的深度学习方法还没有出现有影响力的作者。所以期望有更多的学者关注深度学习方法，也希望学者团体之间、各个机构之间增加交流与合作。

命名实体识别作为一项基础任务，被应用到了生物、医学、军事等领。涉及到的学科也越来越多，命名实体识别已经成为多个学科关注的研究领域。

总的来说，我国在命名实体研究方面已经取得了丰硕的成果。随着命名实体识别研究技术的成熟，研究重点也转向了解决各专业领域的命名实体识别问题。随着近年来深度学习方法不断受到我国学者的关注，我国对于命名实体识别研究将进一步发展。

基金项目

云南省高校数据化运营管理工程研究中心建设项目。

参考文献

[1] Grishman, R. and Sundheim, B. (1996) Message Understanding Conference—6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics*. Kopenhagen, 5-9 August 1996, 466-471. https://doi.org/10.3115/992628.992709

[2] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[3] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 2009, 23(2): 3-17.

[4] 陈悦, 陈超美, 刘则渊, 胡志刚, 王贤文. CiteSpace 知识图谱的方法论功能[J]. 科学学研究, 2015(2): 242-253.

[5] 曹树金, 吴育冰, 韦景竹, 等. 知识图谱研究的脉络、流派与趋势——基于 SSCI 与 CSSCI 期刊论文的计量与可视化[J]. 中国图书馆学报, 2015, 41(219): 16-34.

[6] 罗式胜. 文献计量学概论[M]. 广州: 中山大学出版社, 1994.

[7] 顾理平, 范海潮. 网络隐私问题十年研究的学术场域——基于 CiteSpace 可视化科学知识图谱分析(2008-2017) [J]. 新闻与传播研究, 2018(12): 57-73, 127.

[8] 俞鸿魁, 张华平, 刘群, 吕学强, 施水才. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87-94.