《智能信息处理》课程作业

基于知识图谱的自然语言处理文献分析

刘琦

作业	分数[20]
得分	

2020 年 12 月 6 日

基于知识图谱的自然语言处理文献分析

刘琦 [1)]

[1)]（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　为了解自然语言处理文献发表情况，并对未来的该领域方向进行预测，利用CiteSpace绘制知识图谱，对发文国家、机构、作者的合作，文献学科分布，关键词共现特征及文献交叉引用进行分析。

关键词　知识图谱；文献分析；CiteSpace；自然语言处理
中图法分类号　　TP36

Literature Analysis of Natural Language Processing Based on Knowledge Map

Liu Qi[1)]

[1)](School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract　In order to analyse the publication of natural language processing (NLP) and predict the future direction of NLP, this paper uses CiteSpace to draw knowledge maps analyzing the cooperation characteristics of the country, institution and author, discipline distribution, keyword co-occurrence characteristics and cross reference.

Key words　knowledge map; literature analysis; CiteSpace; NLP

0. 绪论

知识是信息的一种抽象形式，是构成智能的基础。知识工程的概念由费根鲍姆(Feigenbaum)提出，主要是研究知识获取、知识表示和知识使用的学科。概括地讲，知识工程是研究知识信息处理的学科，它起源于20世纪70年代的专家系统，近年来出现了大规模的知识图谱技术。

2001年，万维网之父蒂姆-伯纳斯-李(Tim Berners-Lee)提出语义网的概念，旨在解决知识的表示和组织形式，维基百科等给知识获取提供了半结构化的信息来源。随着大数据时代的到来，知识库技术突破了规模与质量的瓶颈，2012年谷歌知识图谱的诞生是这一突破的标志性产物，它推动知识工程进入全新阶段。近年来，百度知识图谱和搜狗知立方成为代表性的中文知识图谱，为搜索引擎提供准确和丰富的知识回答提供了核心知识支撑。此外，Freebase等大规模知识图谱为英语和汉语等语言的分析和理解、机器翻译、问答和对话等自然语言处理任务提供了丰富的知识资源。

1. 相关概念

1.1. 本体

本体是对客观的事物以一种形式化的、客观的并且系统化的方式进行描述。本体由哲学领域发起，对现实世界的客观事物进行本质化的描述[1]。它在哲学中的定义为“对世界上客观存在物的系统地描述”, 是客观存在的一个系统的解释或说明, 关心的是客观现实的抽象本质。后来随着在人工智能、计算机以及网络领域中的应用发展, 其定义也被融入了许多新的内容。其中最著名、被

引用最为广泛的定义是由 Gruber 提出的：“本体是概念化的明确的规范说明”。Studer 对本体诸多定义进行概括分析后认为，本体论的概念包括四个主要方面：概念化(conceptualization)：客观世界现象的抽象模型，其表示的含义独立于具体的环境状态；明确(explicit)：概念及它们之间联系都被精确定义；(3) 形式化 (formal)：精确的数学描述,计算机可读；(4)共享(share)：本体中反映的知识是其使用者共同认可的，是相关领域中公认的概念集，它所针对的是团体而不是个体。本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇(术语)和词汇相互关系的明确定义。

1.2. 自然语言处理

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科。在上个世纪 70~80 年代，从语言工程和建立实际应用系统的角度，人们提出了自然语言处理（Natural Language Processing，NLP）的概念，是这一学科方向的内涵得到进一步丰富和扩展。从研究任务的角度，自然语言处理可分为基础技术研究和应用技术研究两部分。其中，基础技术研究包括词法、句法、语义和篇章分析以及知识表示与计算等自然语言处理的基本任务；应用技术研究包括文本分类聚类、信息抽取、情感分析、自动文摘、自动问答与对话和机器翻译等自然语言的处理。

1.3. 知识图谱

知识图谱是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是“实体－关系－实体”三元组，以及实体及其相关属性－值对，实体间通过关系相互联结，构成网状的知识结构[2]。

知识图谱本身是一个具有属性的实体通过关系链接而成的网状知识库。从图的角度来看，知识图谱在本质上是一种概念网络，其中的节点表示物理世界的实体（或概念），而实体间的各种语义关系则构成网络中的边。由此，知识图谱是对物理世界的一种符号表达。

知识图谱的研究价值在于，它是构建在当前 Web 基础之上的一层覆盖网络(Overlay network)，借助知识图谱，能够在 Web 网页之上建立概念间的链接关系，从而以最小的代价将互联网中积累的信息组织起来，成为可以被利用的知识[3]。

知识图谱的应用价值在于，它能够改变现有的信息检索方式，一方面通过推理实现概念检索（相对于现有的字符串模糊匹配方式而言）；另一方面以图形化方式向用户展示经过分类整理的结构化知识，从而使人们从人工过滤网页寻找答案的模式中解脱出来。

2. 图谱制作与分析

2.1. 文献数据来源

如表一，确定检索词的同义词以确保文献检索完全，通过确定上位词与下位词并对他们进行同义词分析，在使用 CiteSpace 进行关键词共现等分析时，进行节点的排出与合并。使用形式概念的知识来确定知识图谱中的概念，使同义词的确定具有依据性。

中文文献从 CNKI 中获取，检索条件为“KY = 自然语言处理 + NLP + natural language processing”，来源选择 SCI 来源期刊，EI 来源期刊，北大核心，年份选择 2008 到 2020 年，共得到 787 条结果。

英文文献从 web of science 核心合集中获取，检索条件为主题为 Natural Language Processing 或是 NLP（将两个主题的检索结果进行 and 组配），年份选择 2015 年到 2020 年，文献类型为 article，review，得到 6557 条结果。

表一 检索词范围表

检索词	同（近）义词	上位词	下位词
自然语言处理	计算语言学	人工智能	机器翻译
	电脑语言学	语言学	问答系统
		深度学习	情感分析
			文本分类
			中文分词
Natural Language Processing	NLP	AI	Machine Translation
	computational linguistics	Artificial Intelligence	Question answering
		Linguistics	Sentiment analysis
		Deep Learning	Text categorization
			Chinese word segmentation

2.2. 关键词共现分析

关键词字体大小对应着相应的记录数目越多，共现网络具有结构性，表现了关键词之间的联系。如图 1，从 web of science 核心数据库的英文文献得到的关键词共现性图谱的结构中可以观察到，有以 ontology，convolutional neural network，sentiment analysis,text mining 为中心的四块。如图 2，从 CNKI 的中文文献得到的关键词共现性图谱的结构中可以观察到，有注意力机制与长短时记忆网络，知识图谱，文本分类与深度学习，语义句法分析四个分块。

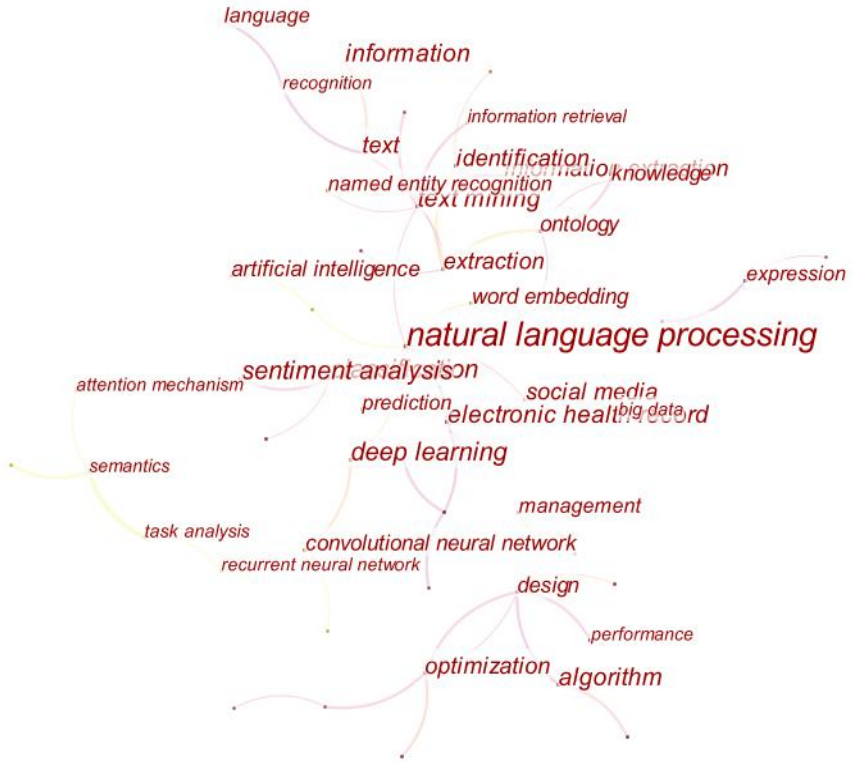

图 1 英文文献中关键词共现性图谱

图 2 中文文献关键词共现性图谱

Top 13 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2010 - 2020
自然语言理解	2010	3.2804	**2010**	2013	
非线性规划	2010	3.3521	**2010**	2011	
本体	2010	3.8747	**2012**	2016	
语义分析	2010	2.6732	**2013**	2014	
主题模型	2010	2.5964	**2015**	2017	
社会媒体	2010	3.0223	**2016**	2017	
文本挖掘	2010	2.5846	**2016**	2020	
卷积神经网络	2010	4.2652	**2017**	2020	
词向量	2010	4.1737	**2018**	2020	
特征提取	2010	3.6082	**2018**	2020	
机器学习	2010	3.0256	**2018**	2020	
深度神经网络	2010	2.8899	**2019**	2020	
神经网络	2010	3.831	**2019**	2020	

图 3 关键词突现

2.3. 关键词突现分析

突现时间的最小单位设为一年，共有 13 个突现点。如图 3，关键词突现中，卷积神经网络、词向量、特征提取的关键词强度大，热度一直持续到 2020 年。

2.4. 机构合作与领域合作分析

图 4 节点与地区字体大小对应着发文数量，中国大陆的发文数量仅次于美国，但中国大陆在机构合作关系图谱中处于边缘位置，与其他机构的合作力度不强。由图 5，可以看到医疗领域与图书情报领域关系密切，生物化学与生物信息联系紧密。

2.5. 文献引用共现分析

图 6 中，能量圈越大的节点说明被引用的次数越多，可见，Mikolov Tomas(2013)，Pennington J(2014)，Mikolov T(2013)，Manning CD(2014)，Collobert R(2011)，LeCun Y(2015)，Srivastava N(2014)，Vaswani A(2017)，Cho K(2014)，He KM(2016)能量圈较大，通过文献引用共现图的结构可以看到文献所引用文献的组合情况

图 4 机构合作关系图

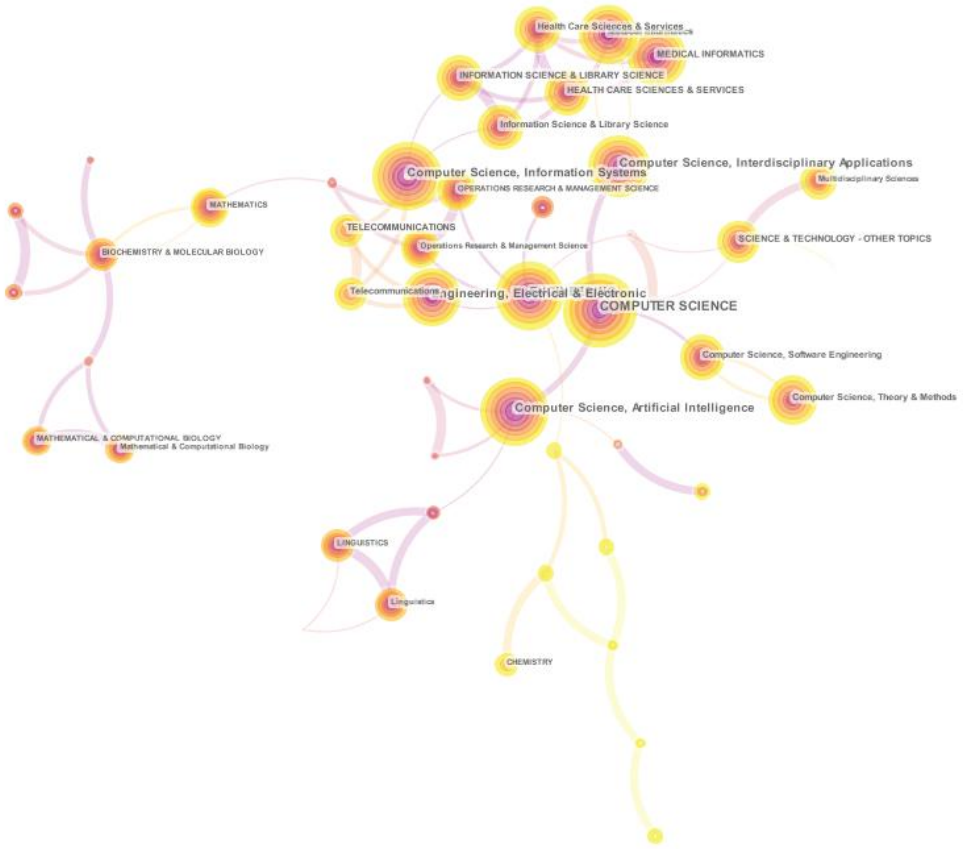

图 5 领域关系合作图

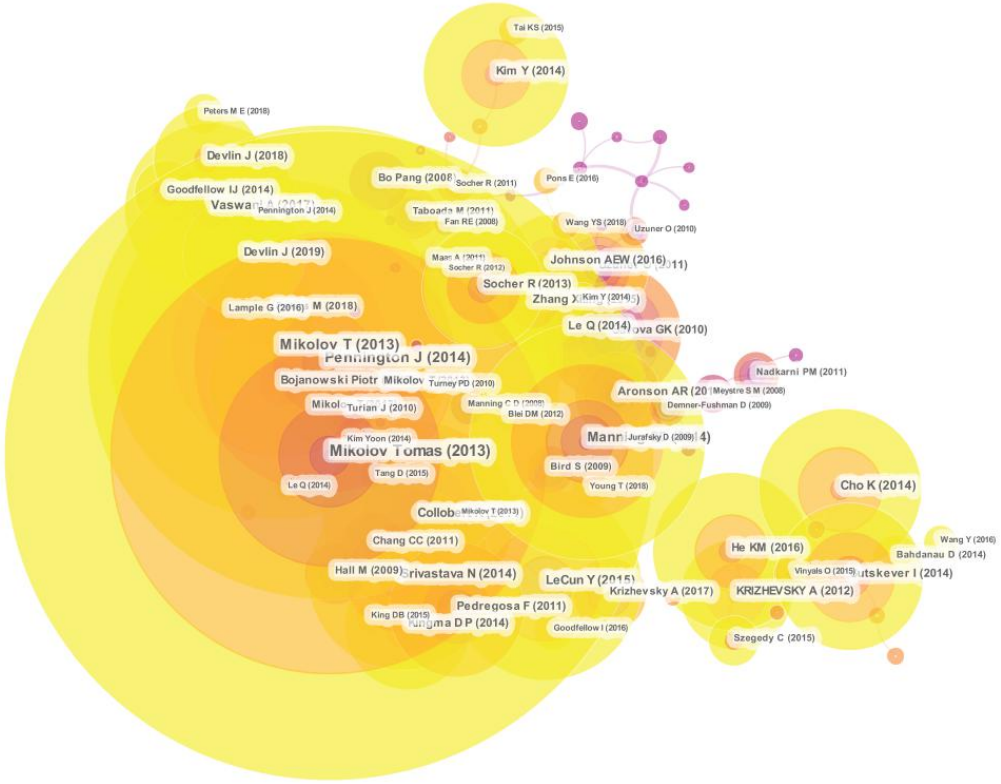

图 6 文献引用共现图

3. 结论

CNKI 中的文献领域集中在计算机科学、语言学、自动化、信息情报，而在 Web of science 核心数据库得到的文献领域中结果占比大的领域多出了医学和保健学。在 CNKI 文献中的突现词卷积神经网络，词向量，特征提取关键词强度大，突现持续到 2020 年。在 Web of science 文献中的突现词为循环神经网络，注意力机制突现持续到 2020 年。从文献引用来看，Mikolov Tomas，Pennington J，Mikolov T，Manning CD，Collobert R，LeCun Y，Srivastava N，Vaswani A，Cho K，He KM 是该研究领域比较有影响力的研究者。通过 web of science 核心数据库的文章分析中机构发文情况来看，中国大陆排在第 2 位，但是与其他机构的合作相对不多。

参考文献

[1] 生佳根. 基于本体的知识获取、管理和应用方法研究[D].南京航空航天大学,2012.

[2] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.

[3] 谢鹏昊,张超,文涛,牛国庆,洪文丹,袁军,沈其荣.基于 Web of science 文献计量分析的青枯病研究进展[J].中国农业大学学报,2020,25(11):62-73.

[4] 刘潇华. 基于知识图谱的高等教育质量研究现状分析[D].湖南大学,2017.

[5] 吕晓赞. 文献计量学视角下跨学科研究的知识生产模式研究[D].浙江大学,2020.