中国科学院文献情报中心
National Science Library, Chinese Academy of Sciences

2020年第1期
总第67期

研究生通讯

文献情报中心

National Science Library

Postgraduate newsletter

中国科学院文献情报中心研究生教育处、中国科学院大学图书情报与档案管理系

北京化学大学

CONTENT
目录

生活小记

卷首语——风

读《明朝那些事儿》走精彩人生

乐生活·思远方

（一）诗与远方篇

（二）健康生活篇

图林论剑

求实讲堂—第四讲

基于SNA的我国知识管理领域作者合作关系研究

知识图谱哪里来—实体关系抽取

精神病学科中文术语识别研究

中国科学院文献情报中心

中心简讯

- 大数据时代的情报学与情报学研究——“情报学前沿问题研究”专家报告会侧记
- 图情档研究生教育现状与未来暨中国科学院文献情报中心研究生教育40周年座谈会召开

光影栏目

雁栖的时光

主编：刘敬仪
执行主编：阮伟南
副主编：郝雪丽，杨晓
美编：任娇菡，徐曾旭林，张敏，阮伟南
封面：阮伟南
封底：阮伟南

风

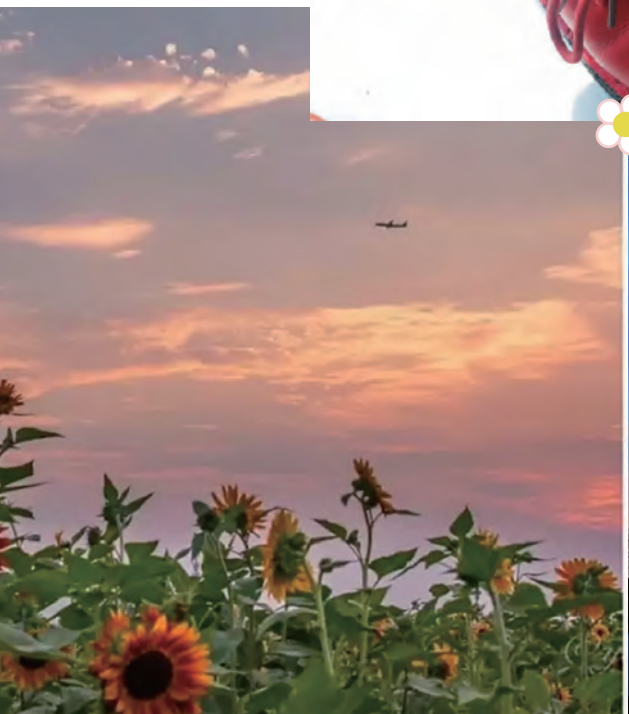
文 | 任娇菡

早晨起来推开房门
衣摆动了一下我知道是风
夜晚时分走在路上
发丝轻轻摇动我知道是风
风从我的指缝间穿过
从我的衣裙间溜走
他流浪了很多个地方
向我诉说着那里的人和事

卷首语

他抚过云见过海
理解山的高耸懂得水的辽阔
像个饱经沧桑的老人历经这世间
我听他讲述的故事听的入神
想和他一起去流浪
转过头他却已经走了
我想
我明天应该还会再遇见他吧
我大步向前走去
衣摆又动了一下

（作者系文献情报中心 2019级硕士研究生）



第四讲

求实讲堂

文|闫群娇

2019年12月18日，由中国科学院文献情报中心情报中心研究生会党支部书记、研究生教育处教育主管张章老师主持，中国科学院文献情报中心情报中心党委书记兼副主任、中国科学院老科学家科普报告团成员、中国科普作家协会理事何林主讲的《请历史记住他们——“两弹一星”研制带给我们的启示》，在同学们的期待中徐徐拉开了序幕。何林书记从国家决策、以身许国、艰苦创业、大师风范四个方面为我们展现了一代老科学家们的艰辛历程、感人故事和爱国情怀。

首先是国家决策部分。何林书记首先向我们展示了23位获得“两弹一星”功勋奖章的老科学家们的全家福，其中21位为中国科学院院士，17人曾在中国科学院工作，科学院也因此与“两弹一星”有了深刻的历史渊源。继而老师为我们讲述了王大珩唱“卡拉OK”，程开甲在浙大西迁中看着受损的图书伤心落泪的故事，表达了老一代科学家们在祖国被侵略的年代中心中无尽的悲痛与屈辱之情，这种悲愤的感情也同时孕育着我国科技事业的发展。紧接着，何林书记开始讲述我国“两弹一星“工作的开端。1945年，人类历史上第一颗原子弹爆炸成功，结合当时的历史背景，毛泽东提出，落后要挨打，发展当自强，并于1955年，中共中央书记处召开发展中国的原子能事业的主题会议，由此拉开了我国原子能事业的大幕。1956年，钱学森向周总理呈交了《建立我国国防航空工业意见书》，同年10月，我国第一个导弹、火箭研制机构国防部五院成立。1957年苏联第一颗人造地球卫星发射成功，标志着人类航天时代的来临。在1958年毛主席宣布：“我们也要搞人造卫星”之后，科学家们意识到，“科学需要学习再学习”，因此中央于1965年决定上马卫星工程。至此，我国的“两弹一星”工作全面展开。正如习近平总书记在中央财经领导小组第七次会议上的讲话“我们要全面研判世界科技

参考文献：
[1]请历史记住他们. 科学时报社编，暨南大学出版社，1999.
[2]“两弹一星”元勋传. 宋健主编，清华大学出版社，2001.
[3]殷玥璨, 徐跃权. 钱学森教授的情报学思想研究[J]. 图书情报工作, 2013(7):34-39.

的发展。紧接着，何林书记开始讲述我国“两弹一星“工作的开端。1945年，人类历史上第一颗原子弹爆炸成功，结合当时的历史背景，毛泽东提出，落后要挨打，发展当自强，并于1955年，中共中央书记处召开发展中国的原子能事业的主题会议，由此拉开了我国原子能事业的大幕。1956年，钱学森向周总理呈交了《建立我国国防航空工业意见书》，同年10月，我国第一个导弹、火箭研制机构国防部五院成立。1957年苏联第一颗人造地球卫星发射成功，标志着人类航天时代的来临。在1958年毛主席宣布：“我们也要搞人造卫星”之后，科学家们意识到，“科学需要学习再学习”，因此中央于1965年决定上马卫星工程。至此，我国的“两弹一星”工作全面展开。正如习近平总书记在中央财经领导小组第七次会议上的讲话“我们要全面研判世界科技创新和产业变革大势，既要重视不掉队问题，也要从国情出发确定跟进和突破策略，按照主动跟进、精心选择、有所为有所不为的方针，明确过我科技创新主攻方向和突破口”，“两弹一星”工作就是国家最高领导层审时度势，从国家利益出发做出的重大战略决策。

下面是老一代科学家们艰苦创业的故事。何林书记给我们讲述了邓稼先带领28名刚毕业的大学生学习基础理论知识，并“请来了三尊‘大菩萨’”指导大家，邓稼先口中的三尊“大菩萨”正是郭永怀、王淦昌、彭桓武，这些大家们秉承着“科学指导，把握方向”、“平等探讨，循循善诱”、“严格要求，激发潜能”，启发、教导着年轻人。这三尊“大菩萨”在研制原子弹的征程中将国家利益放在个人利益之上，王淦昌负责通过试验来设计炸药的装配形式，在“十七号工地”试验成功后，毅然投身大西北，改名王京，经过上千次的试验，终于在1962年“内爆法”的关键环节获得验证；“回国不需要理由，不回国才需要理由”的彭桓武放弃在英国优渥的科研环境回国工作；郭永怀甚至在生命危急时刻依然尽全力保全装有原子弹研制材料的公文包，这些大科学家们完全将自己奉献给了国家。在研制导弹的征途上，钱学森全面抓总，任新民负责发动机，屠守锷负责结构、总体，黄纬禄负责控制、制导，我国相继发射成功了“东风1号”、自主研制的“东风2号”导弹，并在1966年实现了两弹结合。1970年，“东方红一号”卫星发射成功；1974年，首颗返回式遥感探测卫星成功返回地面；1984年，首颗试验通信卫星发射成功。然而这些成果当初的芽，浸透了老一代科学家们奋斗的泪泉与血雨。在我国科技实力和工业基础相对落后的条件下，广大科技工作者发扬艰苦奋斗、奋发图强的拼搏精神，大力协同，联合攻关，谱写了中华民族攀登世界科技高峰的壮丽凯歌。

最后一部分是大师风范。前三部分大师们的风范已可见一斑，何林书记继续举了邓稼先在接受“两弹一星”任务前后在妻子眼中的转变，黄纬禄“对每一个疑点都要打破砂锅问到底，不能把疑点带到天上去”，王大珩与学生姜会林、赵文兴的故事等事例，让我们看到了“两弹一星”元勋在创造卓越的科研成就的同时，也看到了他们在科研作风方面树立起的光辉典范。

1983年钱学森在国防科技情报工作会议上首次明确提出并回答了情报概念的问题：“情报是激活、活化了的知识，或者说是通过资料提取出来的活东西。”钱学森教授在情报学方面也有深刻的研究，在情报科学发展上起了重要的推动引领作用，身为中国科学院文献情报中心一员，我们应该将“两弹一星”这笔宝贵的精神财富继承好、发扬好，为我国科技工作做好服务保障工作，广大科技工作也应当聚焦新时期国家重大需求，攻坚克难，勇于登攀，更好地肩负起科教兴国和创新驱动发展的历史重任，在实现中国梦的伟业中做出无愧于历史和人民的新贡献。

讲座的最后，主持人张章老师对演讲做了总结发言，主讲人何林书记为学生们赠送院士语录书签小礼品，这次讲座圆满结束。

基于SNA的国内知识管理领域作者合作关系研究

赵建/中国科学院文献情报中心

【摘要】本文对近十年内我国知识管理领域的作者合作网络进行分析，旨在发现该领域内作者合作的紧密程度，揭示作者间的知识交流模式与合著规律，为作者的合作交流提供参考。[方法] 本文运用共现分析法和社会网络分析法，以CNKI收录的2009—2018年我国知识管理领域的期刊论文为样本，借助书目共现分析软件提取出作者字段，统计作者频次，形成高产作者共现矩阵，进而利用社会网络分析的常用工具UCINET6软件进行可视化分析，从网络密度、中心性分析和小团体分析三个角度对合著网络进行分析。[结果]经过统计分析，我国知识管理领域139名核心作者的合作网络密度为0.0175；储节旺的程度中心性和中间中心性都居于首位，点度中心度排名前10位的作者中有4位的发文量在所有作者中也排名前10位；合作网络中一共有5个小团体，每个小团体规模较小且作者大部分来自同一单位。[结论]我国知识管理领域作者合作网络密度小，连通性差，不利于知识的分享与传播，该领域作者合作的发展空间较大；程度中心性与作者发文量有一定的相关性，发文量多的作者同他人的合作也相对频繁；作者合作关系受地理位置的影响较大。

【关键词】知识管理；社会网络分析；作者合作；合著网络

1.引言

随着科学技术的迅速发展，科学研究的内容不断深化，研究工具不断更新，科学的发展需要相同和不同学科背景的学者进行交流与合作。科学技术不断发展和科学研究内容不断深化的同时，一批新兴的交叉学科应运而生，这就极大地促进了不同学科之间学者的合作关系。合作关系在科学研究中扮演的角色日益重要，且逐渐成为影响科学生产力的重要因素，对合作关系的研究日益引起学者们的关注。

1.1 研究背景

自管理学大师彼得·德鲁克在1988年提出知识管理的概念以来，经过二十多年的发展，它已经成为学术界和企业管理界极为热门的话题。知识管理是知识经济时代的一种全新的管理，是信息化和知识化的产物。随着人类文明的不断进步，知识的作用不断增加，知识的价值不断提升，人类对知识的管理水平也在不断提升，知识管理正在不断成为社会生活的中心。上世纪90年代，知识管理开始引入中国。自此以后，国内的学者对知识管理进行了深入的研究。储节旺教授探讨了影响知识管理学产生、存在和发展的关键因素，通过对现有成果的分析，发现知识管理学科已经兴起。邱均平教授运用作者共被引分析与社会网络分析的方法,对知识管理领域的被引文献进行多元分析与社会网络分析,找出了该领域的高影响力作者及其所关注的主流领域及热点问题。赵蓉英对知识管理研究的现状进行可视化研究,从而直观地揭示知识管理研究发展的现状。经过近二十年的研究，知识管理在我国得到了极大地发展，形成了一系列的理论和方法。知识管理的跨学科属性决定了该领域存在大量的科研合作，而通过对现有文献的研究发现，我国该领域的合作研究尚十分有限。

1.2 研究目的与意义

本文运用社会网络分析的方法，对近十年内我国知识管理领域的作者合著网络进行可视化分析，旨在发现学者间合作的紧密程度，揭示该领域内学者之间的知识交流模式和合著规律。通过本研究，一方面为新进入该领域的研究人员，提供知识交流的指导；另一方面，对该领域作者之间的知识交流提供参考价值。

1.3 研究方法与研究工具

共现分析法是根据术语在同一篇论文中共同出现的次数来挖掘术语之间联系的方法,一般认为，如果两个术语频繁在同一篇论文中同时出现，往往表明这两个术语之间具有比较密切的联系。社会网络分析方法(Social Network Analysis, SNA) 是对社会网络中各种关系结构及其属性加以分析的一套理论和方法，它主要分析的是行动者之间的关系模式。该方法已经被证实可以应用于作者合作关系分析及其网络结构阐释。社会网络分析法作为成熟的研究方法，可以从多个不同角度构建社会网络进行分析，主要包括:网络密度分析、可达性分析、中心性分析(程度中心性、亲近中心性和中介性)、结构洞分析、凝聚子群分析及核心-边缘等多种分析。

书目共现分析系统(Bibliographic Item Co-Occurrence Matrix Builder，Bicomb) 是由中国医科大学医学信息学系开发的用于抽取、统计文献目录信息的数据分析工具[9]。UCINET(University of California at Irvine NETwork) 是一种功能强大的社会网络分析软件。

2.数据来源及预处理

2.1 数据来源

本文的数据样本来自于CNKI（中国期刊全文数据库），检索策略为“关键词= 知识管理or 篇名= 知识管理”，检索时间段为“2009－2018年”。此次共检索到11264篇论文，除去会议论文、学位论文、无作者的论文以及以编辑部的名义发表的论文后，得到有效样本数据8725篇。选取论文的题名、作者、关键词、摘要、发表时间进行统计分析。

2.2 数据预处理

将得到的8725条有效数据导入Bicomb软件进行初步统计，其中合著论文有3965篇，占总数的45.44%。各年份的论文总数、合著论文数及合著率如表2-1所示，年累计合著率一直保持在45%以上，每年的合著率一直保持在41%以上。

表 2-1 论文合著情况

年份	论文总数	合著论文数	合著率（%）	年累计论文数	年累计合著论文数	年累计合著率（%）
2009	1285	592	46.07	1285	592	46.07
2010	1218	589	48.36	2503	1181	47.18
2011	1145	530	46.29	3648	1711	46.90
2012	1006	454	45.13	4654	2165	46.52
2013	910	414	45.49	5564	2579	46.35
2014	861	394	45.76	6425	2973	46.27
2015	715	297	41.54	7140	3270	45.80
2016	634	292	46.06	7774	3562	45.82
2017	536	229	42.72	8310	3791	45.62
2018	415	174	41.93	8725	3965	45.44
总计	8725	3965	45.44			

我国知识管理领域作者合作度和合作率分别为 1.26 和 45.44%，只有一位作者的论文数占总数的 54.56%，整体合作程度相对较低，作者合作情况分布见表2-2。

表 2-2 作者合著情况分布

每篇论文作者数	论文数	所占比例
15	1	0.01%
10	1	0.01%
8	2	0.02%
7	12	0.13%
6	35	0.40%
5	110	1.26%
4	335	3.84%
3	1118	12.81%
2	2351	26.95%
1	4760	54.56%
总计	8725	100%

（注：合作发文不考虑作者排序，每篇文章的每个作者发文数都记为 1）

2.3 确定核心作者群

由于样本数据包含10952名作者，利用UCINET6 生成的作者合著网络很复杂，不易突出重点，所以本文选择该领域内的核心作者进行研究。普赖斯在他的名著《大科学，小科学》一书中阐释了普赖斯定律，即：全部科学家总人数的平方根，等于撰写了全部科学论文的50%的那些高产科学家的人数。用公式可以表示为：

$$m = 0.749\sqrt{M} \quad (1)$$

其中M为杰出科学家中最高产作者的发文量，m为杰出科学家中最低产作者的发文量。本文10952名作者中，最高产作者的发文量为42，所以杰出科学家中最低产作者的发文量应为4.9篇。

$$\sum_{m+1}^{i} n(x) = \sqrt{N} \quad (2)$$

其中i=M，N为作者总数，n(x)为撰写x篇论文的作者数。根据普赖斯定律，撰写知识管理领域全部论文的50%的高产作者是104名，即发文量在6篇及以上的作者。

表 2-3 核心作者群（139 名）

序号	作者	发文量	序号	作者	发文量	序号	作者	发文量
1	倪节旭	42	30	杨静	10	59	胡汉辉	7
2	郭东强	26	31	刘小峰	9	60	刘建准	7
3	张建华	25	32	李平	9	61	李刚	7
4	何道利	22	33	陈斌	9	62	李鹏	7
5	盛小平	22	34	刘鲁	9	63	李辉	7
6	和金生	18	35	刘伟	9	64	李后卿	7
7	张庆普	18	36	丁玉梅	9	65	倪鸿	7
8	顾新建	17	37	张瑞红	9	66	李莉	7
9	郭春侠	14	38	柯青	8	67	贾玲	7
10	程刚	12	39	葛新权	8	68	郭彤梅	7
11	吕光远	12	40	蔡娜	8	69	刘畅	7
12	刘宏芳	11	41	代凤	8	70	刘娜	7
13	刘希宋	11	42	黄蔚	8	71	刘丹	7
14	汪克夷	11	43	杜维	8	72	陈充	7
15	徐建中	11	44	李娜	8	73	刘洋	7
16	张同健	11	45	陈福集	8	74	陈永兰	7
17	张赛	11	46	林向义	8	75	余翠华	7
18	司有和	11	47	李健	8	76	徐福缘	7
19	张延中	10	48	陈云玲	8	77	王伟	7
20	刘志国	10	49	陈怡	8	78	朱志红	7
21	穆钦丽	10	50	戴巧元	8	79	王建平	7
22	罗洪云	10	51	肖久灵	8	80	姚伟	7
23	李海刚	10	52	杨斌	8	81	孙翠宁	7
24	郭亚军	10	53	王钧	8	82	徐拥军	7
25	姜尔新	10	54	吴庆海	8	83	叶春森	7
26	王平	10	55	徐扬	8	84	王芳	7
27	周荣	10	56	李朝明	7	85	武忠	7
28	张同建	10	57	罗繁明	7	86	王前	7
29	张静	10	58	李纲	7	87	王蕾	7
88	詹湘东	7	106	陈列	6	124	张莉	6
89	杨威	7	107	胡丽原	6	125	张敏	6
90	张帆	7	108	胡树林	6	126	闫上涛	6
91	汪传雷	7	109	陈凤英	6	127	王悦	6
92	王君	7	110	李丽	6	128	王刚	6
93	汪建康	7	111	赵蔚英	6	129	赵静	6
94	刘勇	6	112	王华	6	130	杨洋	6
95	黄林	6	113	张健	6	131	丁家斌	6
96	贾倩	6	114	张少杰	6	132	沈惠敏	6
97	黄敏	6	115	文庆华	6	133	申俊龙	6
98	彭纪生	6	116	张海涛	6	134	王江	6
99	李永先	6	117	张昌	6	135	张新民	6
100	曹霞	6	118	张林	6	136	杨阳	6
101	刘满	6	119	孙振钰	6	137	张鹿	6
102	倪国栋	6	120	张勤	6	138	邱均平	6
103	刘群	6	121	丁勇	6	139	陈国发	6
104	安小米	6	122	朱永	6			
105	李敏	6	123	吴应良	6			

综上所述，将发文量在6篇及以上的作者作为我国知识管理领域的核心作者。经过统计发现，核心作者一共有139名，每位核心作者的发文量如表2-3所示，部分核心作者的共现矩阵如表2-4所示。

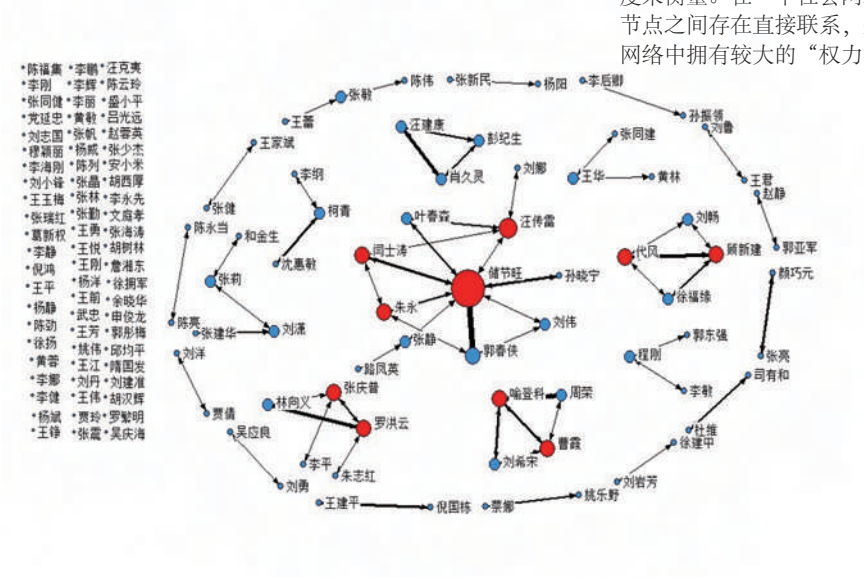
表 2-4 核心作者的共现矩阵（部分）

*	储节旺	郭东强	张建华	喻登科	盛小平	和金生	张庆普	顾新建
储节旺	42	0	0	0	0	0	0	0
郭东强	0	26	0	0	0	0	0	0
张建华	0	0	25	0	0	0	0	0
喻登科	0	0	0	22	0	0	0	0
盛小平	0	0	0	0	22	0	0	0
和金生	0	0	0	0	0	18	0	0
张庆普	0	0	0	0	0	0	18	0
顾新建	0	0	0	0	0	0	0	17

3.可视化分析

3.1 构建作者合作关系网络

将生成的共现矩阵导入UCINET6中进行可视化分析，利用NetDraw生成核心作者的合作网络图，如图所示。在139名核心作者中，71名作者之间有过合作关系，一共有23个子网。其中最大的子网是以储节旺为核心的作者群，包括11名作者。



图中每个圆点代表一位作者，圆点大小表示作者在合作网络中的重要程度，圆点越大，表示作者越重要。其中，储节旺的圆点最大，表示其在领域中处于核心地位；圆点之间的连线表示作者之间有过合作发文，连线的粗细表示他们合作发文的次数，连线越粗，表示他们合作发文的次数越多。其中，储节旺和郭春侠之间的连线最粗，表示他们合作次数最多。有一半的作者没有与其他作者有过合作关系，整体合作水平较低。

3.2 网络密度分析

网络密度是指行动者之间实际联结的数目与他们之间可能存在的最大联结数目的比值。比值越高，这一网络的联结密度就越大。现有研究认为网络密度会影响网络中知识和信息的传播与渗透，密度大的网络，互动性相对比较好，易于知识的分享和传播。合作网络中节点之间的连线越多，该网络的密度就越大。在UCINET6软件中，沿着Network—Cohension—Density路径就可以得出网络密度。经统计，我国知识管理领域近10年核心作者群的网络密度为0.0175，如下图所示。

该合作网络的网络密度较小，核心作者群之间合作的整体结构较松散，连通性不高。同时，这也说明我国知识管理领域的作者合作关系有很大的发展空间，一旦不同子网中的核心作者有了合作，网络的连通性就会增强，促进该领域知识和信息的互动和传播。

3.3 网络中心性分析

中心性是分析社会网络的一个最重要的和最常用的概念工具之一，它是关于行动者在社会网络中的中心性位置的测量概念，反映的是行动者在社会网络结构中的位置或优势的差异。中心性分为三种形式：程度中心性、亲近中心性、中间中心性。由于亲近中心度要求必须是完全相连的图形，我国知识管理领域的合作网络不满足此要求，而且此指标又与程度中心性高度相关，即程度中心性高的人往往亲近中心性也高，所以本文只从程度中心性和中间中心性两个方面进行分析。

3.3.1程度中心性

程度中心性指的是节点在其与之直接相邻的邻居节点当中的中心程度，可以用与之直接相连的节点数，即顶点度来衡量。在一个社会网络中，如果一个节点与很多其他节点之间存在直接联系，那么该节点则处于中心地位，在网络中拥有较大的“权力”。

6t1k

在UCINET6 中，沿着Network—Centrality—Degree路径可以计算出合作网络的程度中心性。表3-1是点度中心度和发文量分别排在前10位的作者对比表。

表 3-1 点度中心度和发文量排在前 10 位的作者

序号	作者	点度中心度	序号	作者	发文量
1	储节旺	36	1	储节旺	42
2	喻登科	20	2	郭东强	26
3	郭春侠	16	3	张建华	25
4	顾新建	12	4	喻登科	22
5	罗洪云	12	5	盛小平	22
6	周荣	11	6	和金生	18
7	肖久灵	11	7	张庆普	18
8	汪建康	10	8	顾新建	17
9	代凤	10	9	郭春侠	14
10	曹霞	9	10	程刚	12

从表3可以看出，点度中心度和发文量都排在第一位的是储节旺，而且喻登科、郭春侠、顾新建三位作者的点度中心度和发文量都排在前10位。这说明作者的发文量和其点度中心度有一定的相关性，点度中心度高的作者，其发文量相对较多，在一定程度上说明发文较多的作者同他人的合作也相对频繁，说明作者的发文量是其在合作网络中地位的重要基础。

3.3.2 中间中心性

中间中心性侧重节点在整个网络的中心程度, 表征的是整个网络的集中或集权程度, 即整个网络围绕一个点或一组点来组织运行的程度。中间中心性指标衡量了一个人作为媒介者的能力，也就是占据在其他两人最短连通距离上重要位置的人。占据这样的位置越多，就越代表他具有很高的中介性，越多的人联络就必须通过他。

序号	作者	中间中心度	标准化的中间中心度
1	储节旺	33.500	0.354
2	汪传雷	9.500	0.100
3	张静	9.000	0.095
4	罗洪云	3.000	0.032
5	张庆普	3.000	0.032
6	张莉	2.000	0.021
7	刘潇	2.000	0.021
8	程刚	1.000	0.011
9	王华	1.000	0.011
10	张敏	1.000	0.011

由上表可知，中间中心度最高的是储节旺，其中间中心度为33.5，标准化中间中心度为0.354，这说明他在合作网络中占有非常重要的位置，拥有较为丰富的资源，其他作者要通过他实现联系。经过统计发现，一共有121名作者的中间中心度为0，占核心作者总数的87%，说明大多数作者并不具备控制信息和资源的能力。整个合作网络的中间中心度指数为0.35%，数值较低，说明整个网络连通性不高，各个核心作者群较为分散。

3.3 小团体分析

计算小团体的方法有两类:一类以节点程度计算,一群相连的节点视为一个小团体,另一类以距离计算,在一定距离内可达的节点为一个小团体。计算小团体的方法有多种：派系（cliques）、n-派系（n-cliques）、n-宗派（n-clan）、k-丛（k-plex）等。本文选择n-派系（任何两点之间在总图中的距离最大不超过n）来计算小团体。

在UCINET6中，沿着Network—Subgroups—N-clique路径可以对小团体进行分析。经过分析发现，将节点间的最大距离n设置为3，小团体节点数目设置为4时，共生成了6个小团体。将出现在多个团体的作者进行合并处理后，进一步归结为5个小团体，和图1基本一致。5个小团体分别是：

团体1：储节旺、汪传雷、闫士涛、郭春侠、朱永、叶春森、张静、刘伟、孙晓宁、路凤英、刘娜，均来自安徽大学。该团体以储节旺为核心，连接汪传雷、闫士涛、朱永和郭春侠等人。通过分析原始数据，该团体作者在2008-2017年共发表知识管理相关论文83篇，其中合著论文有34篇，主要发表于2010-2013年，研究方向主要集中在知识管理学科体系研究和国内图书情报领域知识管理研究方法的共词分析。

团体2：代风、顾新建、徐福缘、刘畅，除徐福缘来自上海理工大学外，其余三位作者都来自浙江大学。在该团体中，代风和顾新建的合作较为紧密，合作发文量为6篇。通过统计分析，该团体作者在2008-2017年共发表了知识管理相关论文27篇，其中有10篇为合著论文，主要发表于2009-2011年，研究方向主要集中在知识评价体系的构建和知识共享方面。

团体3：喻登科、曹霞爱、周荣、刘希宋，均来自哈尔滨工程大学，这说明校友关系或在同一单位工作会大大增加科研合作的机会。该团体作者在2008-2017年共发表知识管理相关论文31篇，其中合著论文15篇。喻登科和周荣的合作次数相对较多，合作发文9篇，占该团体合著论文的60%。

团体4：张庆普、罗洪云、林向义、李平、朱志红。其中，张庆普和罗洪云来自哈尔滨工业大学，林向义和朱志红均来自东北石油大学，李平来自浙江大学。这说明作者合作也与地理位置有关系，有利的地理位置也能促进学者间的合作关系。该团体作者在2008-2017年共发表38篇知识管理相关论文，其中12篇为合著论文。

团体5：张莉、和金生、刘潇、张建华。前作者两位来自天津大学，后两位作者来自郑州大学。该团体作者在2008-2017年共发表知识管理相关论文51篇，其中合著论文仅为5篇，占论文总数的10%，作者之间合作不够紧密。

由以上小团体分析可以得出，相似的研究领域、相同的学习或者工作背景、一定的地理优势都会大大地促进作者间的相互合作。同时，小团体的规模较小，最大的五个小团体的规模分别为11位、5位、4位、4位、4位，这说明作者合作网络中，作者间合作程度有待加强，合作范围不够广泛，不利于该领域的资源的交流和共享。

4.结论与讨论

4.1 研究结论

本文利用社会网络分析的方法，借助于Bicomb和UCINET6 两个工具，选取CNKI所收录的2009-2018年我国知识管理领域的期刊论文作为研究样本，对该领域作者合作关系进行了研究，研究对象为该领域内139名核心作者。经过初步统计，我国知识管理领域的作者合作率不高。为了更为直观地体现该领域作者的合作关系，生成了核心作者的合作网络图，并从网络密度、网络中心性和小团体三个方面进行了分析，得出以下结论：

首先，作者合作网络密度较小，连通性差，信息的交流和共享有待加强，该领域内作者合作的发展空间较大。其次，程度中心性与作者的发文量有一定的相关性，发文量大的作者与其他作者的合作相对频繁，说明作者的发文量是其在合作网络中地位的重要基础。

再次，储节旺的程度中心性和中间中心性都位居首位，处于合作网络中的核心位置，说明其在该领域中处于重要地位，拥有丰富的资源，是其他作者之间合作的“桥梁”。

最后，该领域的作者合作关系受地理位置影响较大，合作网络中每个小团体中的作者几乎都来自于相同的单位。

4.2 研究不足与讨论

本文利用社会网络分析的方法，通过对作者合作发文所建立的真实网络进行分析，研究作者间的合作关系。这仅仅反映了作者间的外在合作关系，没能研究隐藏在真实关系网络背后的潜在合作关系，研究相对片面。可从关键词、共词分析入手，进一步研究隐藏在真实关系网络背后的潜在合作，对我国知识管理领域作者合作关系形成更加全面的认识。

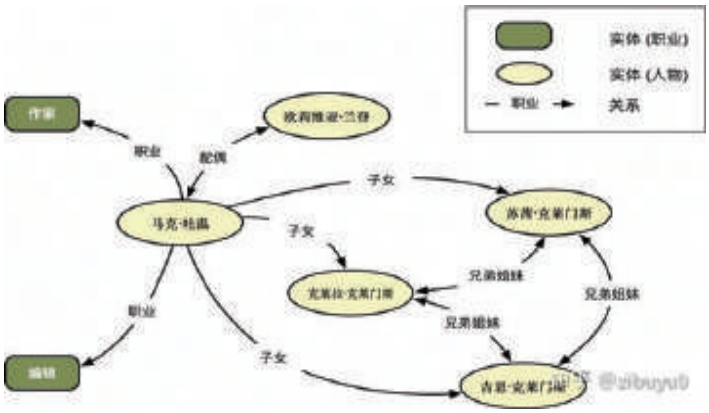
参考文献：
[1]储节旺, 闫士涛, 谈甄. 知识管理学产生、存在与发展的关键因素研究 [J]. 情报杂志, 2012, 31(02): 108-113.
[2]邱均平, 刘艳玲. 近10年我国合著现象的研究进展 [J]. 图书情报工作, 2011, 55(20): 11-14+67.
[3]邱均平, 张晓培. 基于CSSCI的国内知识管理领域作者共被引分析 [J]. 情报科学, 2011, 29(10): 1441-1445.

知识图谱从哪里来：实体关系抽取的现状与未来

■ 韩旭　高天宇　刘知远
（清华大学计算机科学与技术系）

最近几年深度学习引发的人工智能浪潮席卷全球，在互联网普及带来的海量数据资源和摩尔定律支配下飞速提升的算力资源双重加持下，深度学习深入影响了自然语言处理的各个方向，极大推动了自然语言处理的发展。来到2019年的今天，深度学习的诸多局限性也慢慢得到广泛认知。对于自然语言处理而言，要做到精细深度的语义理解，单纯依靠数据标注与算力投入无法解决本质问题。如果没有先验知识的支持，“中国的乒乓球谁都打不过”与“中国的足球谁都打不过”，在计算机看来语义上并没有巨大差异，而实际上两句中的“打不过”意思正好相反。因此，融入知识来进行知识指导的自然语言处理，是通向精细而深度的语言理解的必由之路。然而，这些知识又从哪里来呢？这就涉及到人工智能的一个关键研究问题——知识获取。

知识图谱



现有大型知识图谱，诸如Wikidata、Yago、DBpedia，富含海量世界知识，并以结构化形式存储。如下图所示，每个节点代表现实世界中的某个实体，它们的连边上标记实体间的关系。这样，美国作家马克·吐温的相关知识就以结构化的形式记录下来。

目前，这些结构化的知识已被广泛应用于搜索引擎、问答系统等自然语言处理应用中。但与现实世界快速增长的知识量相比，知识图谱覆盖度仍力有未逮。由于知识规模巨大而人工标注昂贵，这些新知识单靠人力标注添加几无可能完成。为了尽可能及时准确地为知识图谱增添更加丰富的世界知识，研究者们努力探索高效自动获取世界知识的办法，即实体关系抽取技术。

具体来说，给定一个句子和其中出现的实体，实体关系抽取模型需要根据句子语义信息推测实体间的关系。例如，给定句子：「清华大学坐落于北京近邻」以及实体「清华大学」与「北京」，模型可以通过语义得到「位于」的关系，并最终抽取出（清华大学，位于，北京）的知识三元组。实体关系抽取是一个经典任务，在过去的 20 多年里都有持续研究开展，特征工程、核方法、图模型曾被广泛应用其中，取得了一些阶段性的成果。随着深度学习时代来临，神经网络模型则为实体关系抽取带来了新的突破。

神经网络关系抽取模型

模型结构	模型名称	输入特征	F1 (%)
卷积模型	CNN[2]	Word embeddings, Position embeddings	82.7
	MW-CNN[7]	Word embeddings, Position embeddings	82.8
	CR-CNN[8]	Word embeddings, Position embeddings	84.1
循环模型	RNN[3]	Word embeddings, Position embeddings	76.9
	BRNN[9]	Word embeddings, Position embeddings	82.5
	BLSTM[9]	Word embeddings	82.7
		(+)Position embeddings, POS, NER, WordNet, Dependency features	84.3
	ATT-BLSTM[4]	Word embeddings, Position embeddings	84.0
	HRNN[10]	Word embeddings	83.9
		(+)Position embeddings, WordNet, NER	84.3
词法句法模型	MVRNN[5]	Word embeddings	79.1
		(+)WordNet, POS, NER	82.4
	DepLCNN[11]	Word embeddings	84.0
		(+)WordNet, Word around nominals	85.6
	SDP-LSTM[12]	Word embeddings	82.4
		(+)WordNet, POS, GR	83.7
	DepNN[13]	Word embeddings, WordNet	83.0
		(+)NER	83.6
	DRNNs[14]	Word embeddings, WordNet, POS, GR	84.2
		(+)Data Augmentation	86.1
	BRCNN[15]	Word embeddings	85.4
		(+) WordNet, POS, NER	86.3

面向自然语言文本序列已经有很多神经网络类型，例如循环神经网络（RNN、LSTM）、卷积神经网络（CNN）和 Transformer 等，这些模型都可以通过适当改造用于关系抽取。最初，工作 [1，2] 首次提出使用 CNN 对句子语义进行编码，用于关系分类，比非神经网络方法比性能显著提升；工作 [3，4] 将 RNN 与 LSTM 用于关系抽取；此外，工作 [5] 提出采用递归的神经网络对句子的语法分析树建模，试图在提取语义特征的同时考虑句子的词法和句法特征，这个想法也被不少后续工作的进一步探索。这里，我们列出一个表格，总结各类典型神经网络在基准测试数据集合 SemEval-2010 Task-8 [6] 上的效果。

从上表可以看出，这些神经网络模型均取得了优异的实验结果，且相互之间没有显著的性能差异。这是否意味着关系抽取问题就此解决了呢？实际上并非如此。SemEval-2010 Task-8 的任务设定为，对预先定义好的关系类别标注大量的训练和测试样例，样例都是相对简单的短句，而且每种关系的样例分布也比较均匀。然而，实际应用中往往面临很多挑战：

数据规模问题：

人工精准地标注句子级别的数据代价十分高昂，需要耗费大量的时间和人力。在实际场景中，面向数以千计的关系、数以千万计的实体对、以及数以亿计的句子，依靠人工标注训练数据几乎是不可能完成的任务。

学习能力问题：

在实际情况下，实体间关系和实体对的出现频率往往服从长尾分布，存在大量的样例较少的关系或实体对。神经网络模型的效果需要依赖大规模标注数据来保证，存在“举十反一”的问题。如何提高深度模型的学习能力，实现“举一反三”，是关系抽取需要解决的问题。

复杂语境问题：

现有模型主要从单个句子中抽取实体间关系，要求句子必须同时包含两个实体。实际上，大量的实体间关系往往表现在一篇文档的多个句子中，甚至在多个文档中。如何在更复杂的语境下进行关系抽取，也是关系抽取面临的问题。

开放关系问题：

现有任务设定一般假设有预先定义好的封闭关系集合，将任务转换为关系分类问题。这样的话，文本中蕴含的实体间的新型关系无法被有效获取。如何利用深度学习模型自动发现实体间的新型关系，实现开放关系抽取，仍然是一个开放问题。

所以说，SemEval-2010 Task-8 这样的理想设定与实际场景存在巨大鸿沟，仅依靠神经网络提取单句语义特征，难以应对关系抽取的各种复杂需求和挑战。我们亟需探索更新颖的关系抽取框架，获取更大规模的训练数据，具备更高效的学习能力，善于理解复杂的文档级语境信息，并能方便地扩展至开放关系抽取。

我们认为，这四个方面构成了实体关系抽取需要进一步探索的主要方向。接下来，我们分别介绍这四个方面的发展现状和挑战，以及我们的一些思考和努力。

更大规模的训练数据

神经网络关系抽取需要大量的训练数据，但是人工标注这些训练数据非常费时昂贵。为了自动获取更多的训练数据训练模型，工作 [16] 提出了远程监督（Distant Supervision）的思想，将纯文本与现有知识图谱进行对齐，能够自动标注大规模训练数据。

远程监督的思想并不复杂，具体来说：如果两个实体在知识图谱中被标记为某个关系，那么我们就认为同时包含这两个实体的所有句子也在表达这种关系。再以（清华大学，位于，北京）为例，我们会把同时包含「清华大学」和「北京」两个实体的所有句子，都视为「位于」这个关系的训练样例。

远程监督的这种启发式标注规则是把双刃剑，它是自动标注训练数据的有效策略，但其过强的设定不可避免地产生错误标注。例如对于知识图谱中（清华大学，校长，邱勇）这个三元组事实，句子「邱勇担任清华大学校长」可以反映「清华大学」与「邱勇」之间「校长」的关系；但是句子「邱勇考入清华大学化学与化学工程系」以及「邱勇担任清华大学党委常委」并不表达「校长」关系，但却会被远程监督的启发式规则错误地标注为「校长」关系的训练实例。

虽然远程监督思想非常简单也存在很多问题，不过它为更多收集训练数据开启了新的纪元。受到这个思路的启发，很多学者积极考虑如何尽可能排除远程监督数据中的噪音标注的干扰。从 2015 年开始，基于远程监督与降噪机制的神经关系抽取模型得到了长足的发展，工作 [17] 引入了多实例学习方法，利用包含同一实体对的所有实例来共同预测实体间关系。我们课题组林衍凯等人工作 [19] 提出句子级别注意力机制，对不同的实例赋予不同的权重，用以降低噪音实例造成的影响。工作 [20] 引入对抗训练来提升模型对噪音数据的抵抗能力。工作 [21] 则构建了一套强化学习机制来筛除噪音数据，并利用剩余的数据来训练模型。

总结来说，已有对远程监督的降噪方法可以兼顾了关系抽取的鲁棒性与有效性，也具有较强的可操作性和实用性。不过，使用已有知识图谱对齐文本来获取数据训练关系抽取模型，再利用该模型来抽取知识加入知识图谱，本身就有一种鸡生蛋与蛋生鸡的味道。不完善的知识图谱对齐所得到的文本训练数据也将是不完善的，对那些长尾知识而言，仍难以通过这种远程监督机制来得到训练实例。如何提出更有效的机制来高效获取高质量、高覆盖、高平衡的训练数据，仍然是一个值得深入思考的问题。

更高效的学习能力

即使通过远程监督等办法能够自动获取高质量的训练数据，由于真实场景中关系和实体对的长尾分布特点，绝大部分的关系和实体对的可用样例仍然较少。而且，对于医疗、金融等专业领域的专门关系，受限于数据规模的问题可用样例也很有限。而神经网络模型作为典型的 data-hungry 技术，在训练样例过少时性能会受到极大影响。因此，研究者们希望探索有效提升模型学习能力的方法，以更好地利用有限训练样例取得满意的抽取性能。

实际上，人类可以通过少量样本快速学习知识，具有「举一反三」的能力。为了探索深度学习和机器学习「举一反三」的能力，提出了少次学习（Few-shot learning）任务。通过设计少次学习机制，模型能够利用从过往数据中学到的泛化知识，结合新类型数据的少量训练样本，实现快速迁移学习，具有一定的举一反三能力。

过去少次学习研究主要集中于计算机视觉领域，自然语言处理领域还少有探索。我们课题组韩旭同学等的工作 [21] 首次将少次学习引入到关系抽取，构建了少次关系抽取数据集 FewRel，希望推动驱动自然语言处理特别是关系抽取任务的少次学习研究。如下图所示，关系抽取少次学习问题仅为每种关系提供极少量样例（如 3-5 个），要求尽可能提高测试样例上的关系分类效果。

为了更好解决远程监督关系抽取的少次学习问题，我们课题组的高天宇同学等的工作 [22] 提出了基于混合注意力机制的原型网络，同时考虑实例级别和特征级别的注意力机制，在减少噪音标注影响的同时，能更好地关注到句中的有用特征，实现高效少次学习。工作 [23] 提出多级匹配和整合结构，充分学习训练样例之间的潜在关联，尽可能挖掘为数不多的样例中的潜在信息。工作 [24] 则采用了预训练语言模型 BERT 来处理关系抽取中的少次学习问题，基于海量无监督数据训练的 BERT，能够为少次学习模型提供有效的语义特征，在 FewRel 数据上取得了超过人类关系分类的水平。

在对少次学习关系抽取探究的过程中，课题组的高天宇同学等进一步发现两个长期被忽视的方面 [25]：要将少次学习模型用于生产环境中，应具备从资源丰富领域迁移到资源匮乏领域（low-resource domains）的能力，同时还应具备检测句子是否真的在表达某种预定义关系或者没有表达任何关系的能力。为此他们提出了 FewRel 2.0，在原版数据集 FewRel 的基础上增加了以下两大挑战：领域迁移（domain adaptation）和「以上都不是」检测（none-of-the-above detection）。

对于领域迁移挑战，FewlRel 2.0 采集了大量医疗领域的数据并进行标注，要求关系抽取模型在原语料进行训练后，还可以在这些新领域语料上进行少次学习。对于「以上都不是」检测，FewRel 2.0 在原 N-way K-shot 设定（给定 N 个新类型，每个类型给定 K 个训练样本）上，添加了一个「以上都不是」选项，大大增加了分类和检测难度。

初步实验发现，以往有效的模型（包括基于 BERT 的模型）在这两大挑战任务均有显著性能下降。尽管高天宇同学等在 FewRel 2.0 论文中也尝试了一些可能的解决思路：例如对于领域迁移尝试了经典的对抗学习方法，模型性能得到了一定的提升；对于「以上都不是」检测，提出了基于 BERT next sentence prediction task 的 BERT-PAIR 模型，可以在」以上都不是」挑战取得一点效果。但这两大挑战依然需要更多创新探索。

总结来说，探索少次学习关系抽取，让关系抽取模型具备更强大高效的学习能力，还是一个非常新兴的研究方向，特别是面向关系抽取的少次学习问题，与其他领域的少次学习问题相比，具有自身独有的特点与挑战。不论是基于已有少次学习技术作出适于 NLP 和关系抽取的改进，还是提出全新的适用于关系抽取的少次学习模型，都将最大化地利用少量标注数据，推动关系抽取技术的落地实用。

更复杂的文本语境

现有关系抽取工作主要聚焦于句子级关系抽取，即根据句内信息进行关系抽取，各类神经网络模型也擅长编码句子级语义信息，在很多公开评测数据能够取得最佳效果。而在实际场景中，大量的实体间关系是通过多个句子表达的。如下图所示，文本中提到多个实体，并表现出复杂的相互关联。根据从维基百科采样的人工标注数据的统计表明，至少 40%的实体关系事实只能从多个句子中联合获取。为了实现多个实体间的跨句关系抽取，需要对文档中的多个句子进行阅读推理，这显然超出了句子级关系抽取方法的能力范围。因此，进行文档级关系抽取势在必行。

文档级关系抽取研究需要大规模人工标注数据集来进行训练和评测。目前文档级关系抽取数据集还很少。工作 [26，27] 构建了两个远程监督的数据集，由于没有进行人工标注因此评测结果不太可靠。BC5CDR [28] 是人工标注的文档级关系抽取数据集，由 1,500 篇 PubMed 文档构成是生物医学特定领域，且仅考虑「化学诱导的疾病」关系，不一定适合用来探索文档级关系抽取的通用方法。工作 [29] 提出使用阅读理解技术回答问题的方式从文档中提取实体关系事实，这些问题从」实体-关系「对转换而来。由于该工作数据集

Reasoning Types	%	Examples
Pattern recognition	38.9	[1] ***Me Musical Nephews*** is a 1942 one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ... **Relation:** publication date **Supporting Evidence:** 1
Logical reasoning	26.6	[1] "Nisei" is the ninth episode of the third season of the American science fiction television series The X-Files. ... [3] It was directed by David Nutter, and written by Chris Carter, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents ***Fox Mulder*** (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called X-Files. ... **Relation:** creator **Supporting Evidence:** 1, 3, 8
Coreference reasoning	17.6	[1] ***Dwight Tillery*** is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the University of Michigan Law School. [4] ***Tillery*** served as mayor of Cincinnati from 1991 to 1993. **Relation:** educated at **Supporting Evidence:** 1, 3
Common-sense reasoning	16.6	[1] ***William Busac*** (1020-1076), son of William I, Count of Eu, and his wife Lesceline, ... [4] ***William*** appealed to King Henry I of France, who gave him in marriage Adelaide, the heiress of the county of Soissons. [5] Adelaide was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] ***William*** and Adelaide had four children: ... **Relation:** spouse **Supporting Evidence:** 4, 7

是针对这种方法量身定制的，也不那么适用于探索文档级关系抽取的通用方法。这些数据集或者仅具有少量人工标注的关系和实体，或者存在来自远程监督的噪音标注，或者服务于特定领域或方法，有这样或那样的限制。

为了推动文档级关系抽取的研究，课题组姚远同学等[30] 提出了 DocRED 数据集，是一个大规模的人工标注的文档级关系抽取数据集，基于维基百科正文和 WikiData 知识图谱构建而成，包含 5,053 篇维基百科文档，132,375 个实体和 53,554 个实体关系事实，是现有最大的人工标注的文档级关系抽取数据集。如下图所示，文档级关系抽取任务要求模型具有强大的模式识别、逻辑推理、指代推理和常识推理能力，这些方面都亟待更多长期的研究探索。

更开放的关系类型

现有关系抽取工作一般假设有预先定义好的封闭关系集合，将任务转换为关系分类问题。然而在开放域的真实关系抽取场景中，文本中包含大量开放的实体关系，关系种类繁多，而且关系数量也会不断增长，远超过人为定义的关系种类数量。在这种情况下，传统关系分类模型无法有效获取文本中蕴含的实体间的新型关系。如何利用深度学习模型自动发现实体间的新型关系，实现开放关系抽取，仍然是一个「开放「问题。

为了实现面向开放领域的开放关系抽取，研究提出开放关系抽取（Open Relation Extraction，OpenRE）任务，致力于从开放文本抽取实体间的任意关系事实。开放关系抽取涉及三方面的「开放」：首先是抽取关系种类的开放，与传统关系抽取不同，它希望抽取所有已知和未知的关系；其次是测试语料的开放，例如新闻、医疗等不同领域，其文本各有不同特点，需要探索跨域鲁棒的算法；第三是训练语料的开放，为了获得尽可能好的开放关系抽取模型，有必要充分利用现有各类标注数据，包括精标注、远程监督标注数据等，而且不同训练数据集的关系定义和分布也有所不同，需要同时利用好多源数据。

在前深度学习时代，研究者也有探索开放信息抽取（Open Information Extraction，OpenIE）任务。开放关系抽取可以看做 OpenIE 的特例。当时 OpenIE 主要通过无监督的统计学习方法实现，如 Snowball 算法等。虽然这些算法对于不同数据有较好的鲁棒性，但精度往往较低，距离实用落地仍然相距甚远。

最近，课题组吴睿东同学等的工作 [31] 提出了一种有监督的开放关系抽取框架，可以通过」关系孪生网络「（Relation Siamese Network，RSN）实现有监督和弱监督模式的自由切换，从而能够同时利用预定义关系的有监督数据和开放文本中新关系的无监督数据，来共同学习不同关系事实的语义相似度。具体来说，关系孪生网络 RSN 采用孪生网络结构，从预定义关系的标注数据中学习关系样本的深度语义特征和相互间的语义相似度，可用于计算包含开放关系文本的语义相似度。而且，RSN 还使用了条件熵最小化和虚拟对抗训练两种半监督学习方法进一步利用无标注的开放关系文本数据，进一步提高开放关系抽取的稳定性和泛化能力。基于 RSN 计算的开放关系相似度计算结果，模型可以在开放域对文本关系进行聚类，从而归纳出新型关系。

课题组高天宇同学等人工作 [32] 则从另一个角度出发，对于开放域的特定新型关系，只需要提供少量精确的实例作为种子，就可以利用预训练的关系孪生网络进行滚雪球（Neural SnowBall），从大量无标注文本中归纳出该新型关系的更多实例，不断迭代训练出适用于新型关系的关系抽取模型。

总结来说，开放域关系抽取在前深度学习时代取得了一些成效，但如何在深度学习时代与神经网络模型优势相结合，有力拓展神经网络关系抽取模型的泛化能力，值得更多深入探索。

总结

为了更及时地扩展知识图谱，自动从海量数据中获取新的世界知识已成为必由之路。以实体关系抽取为代表的知识获取技术已经取得了一些成果，特别是近年来深度学习模型极大地推动了关系抽取的发展。但是，与实际场景的关系抽取复杂挑战的需求相比，现有技术仍有较大的局限性。我们亟需从实际场景需求出发，解决训练数据获取、少次学习能力、复杂文本语境、开放关系建模等挑战问题，建立有效而鲁棒的关系抽取系统，这也是实体关系抽取任务需要继续努力的方向。

本文总结了我们对实体关系抽取现状、挑战和未来发展方向的认识，以及我们在这些方面做出的努力，希望能够引起大家的兴趣，对大家有些帮助。期待更多学者和同学加入到这个领域研究中来。当然，本文没有提及一个重要挑战，即以事件抽取为代表的复杂结构的知识获取，未来有机会我们再专文探讨。

精神病学科中文术语识别研究

■ 赵旸 朱冬亮 常杲

一、综述

综合目前国内外术语识别的研究，主要有四种思路，基于语言学的方法、基于统计的方法、混合策略方法以及基于深度学习的方法。

1.基于语言学的方法

术语一般以名词或名词短语出现，而对特定领域的术语而言，往往具有特殊的词缀(如以“hypen”为前缀的词常为生物医学术语)和特定的组成模式(如很多术语首字母大写)。鉴于术语这些词性、词缀和词形等语言学方面的特点，可以利用自然语言处理方法，重复使用术语构成的语法和词形模式判断词串是否符合为术语。如Paul等提出利用语言学方法分析术语的层级图(见图1)。其他如Ananiadou等提出基于通用语法的方法，使用统一词形语法、特定词缀字典、拉丁文/希腊文的新古典组合方式表等实现了医学术语识别。但是，这种方法对设定的术语构成模式依赖较大，识别效率有限，在词间关系的识别上尚缺少有力试验的验证。

基于模式规则的方法常依赖人工进行浅层语法分析或领域词典构建模式规则进行概念识别与抽取，对特定语言的词典、标注语料库、语法库等先决资源条件依赖性大，存在规则维护/更新/扩展困难，应用范围有限，可移植性差等问题，尤其对一些新兴词汇、非正式句子、缩写短语、词典中没有的专业术语等识别较差，准确率和召回率低，因而无法大规模地应用于概念识别。

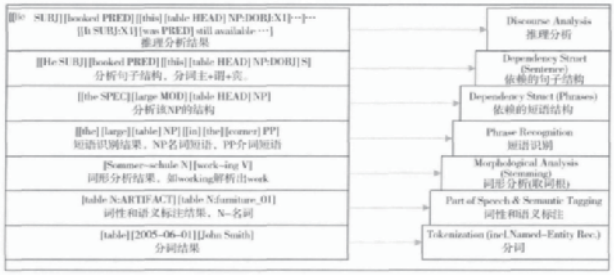


图1 语言学方法分析术语的层级图

2. 基于统计的方法

基于统计分析的方法一般是对大量领域文本数据进行统计分析，将满足统计阈值或条件的字符串序列作为领域概念，常用的统计方法有词频统计、TF-IDF、信息熵、互信息计算等，但该方法存在计算量大，常遗漏低频词，常忽略或缺乏上下文语义分析等问题，因而识别的准确率一直有待提高。为了突破上述局限，随着机器学习和自然语言处理技术的推动，后来大多数研究开始将领域本体概念获取问题规范化表述为概念抽取(Concept Extraction，CE)或术语抽取(Term Extraction，TE)，并归入为一种序列标记的命名实体识别(Named Entity Recognition，NER) 问题，主要采用的就是各种半监督和监督方式混合的机器学习算法及其变体等，聚焦于从领域文本中半自动或自动地获得领域依赖的属性、专门的文本特征、上下文语义信息等以解决上述问题。

Feng 等人分别采用 HMM、ME、CRF 实现命名实体识别，实验结果表明条件随机场识别效果最佳。祝继锋提出利用支持向量机模型来确定中文机构名称的后界，隐马尔科夫模型识别中文机构名称的前缀。结果表明，此结合方案对命名实体识别是有效的，且准确度有一定提高。张玥杰等人在汉语识别实验中利用最大熵算法的特性融入多种特征。Nuo等人将带有实体特性约束的条件随机场与普通的条件随机场进行对比，实验结果表明前者的命名实体识别性能优于后者。龙光宇等人引入词典联合 CRF 的方式对疾病进行命名实体识别，该方式首先需要根据词典提取医学实体相关信息，然后 CRF 利用该信息完成识别任务。都丽婷等人采用条件随机场对临床记录涉及到的病人隐私信息进行识别。

翟菊叶等人采用 CRF 与规则相结合的方法对中文电子病历实体进行识别，结果显示此方法优于CRF。Chen等人设计了特征模板，建立了级联条件随机场模型对中文电子病历中的疾病名，药物名称和症状名称进行识别。S. Doan 等人采用基于支持向量机（SVM）的方法来识别药物相关实体。Tang 等人将结构化的支持向量机算法应用于医疗问题，治疗和测试的实体识别中，实验结果表明此方法优于基于 CRF 的临床实体识别系统。

3．混合策略方法

常用的混合策略方法［21-24］有将层叠条件随机场(CRF)、支持向量机(SVM)算法及几种模式匹配规则结合起来构建的混合模型等; 将基于独热编码词特征表示方法的布朗聚类技术和隐马尔可夫模型(HMM)结合起来实现对未标记语料库的无监督特征表示模型［25-26］等; 将分布式词特征表示方法和随机索引模型结合起来，借助或直接将Wikipedia、WordNet、HowNet、百度百科等的部分文本语料训练生成词嵌入模型以改善概念提取性能［27-28］。这些混合的策略方法常能够最大限度地减少依赖于词汇查找的计算量，并逐步开始能够部分考虑利用上下文语义信息进行领域文本分析，一定程度上提高了识别的准确率和召回率。但传统机器学习方法本质上仍是遵循领域特定的特征工程和分类两个步骤，仍属于高度专用的手工制作系统范畴且需要劳动密集型的专家知识才能实施，需要大量的“经验”(专家知识)和“运气”(人工选取并获得最优特征的过程随机、难以复制且不可控)作为基础，因而难以大规模流行应用起来，自动精准识别领域本体中的概念仍是一项极具挑战意义的研究任务。

4．深度学习

近年来，尤其是2012年以来，深度学习的出现和深度神经网络的激增，在计算机视觉、图像/语音识别等任务中已取得了前所未有的成果。但在NER及其相关扩展任务如领域本体的构建研究中，由于很多领域缺乏大量规范文本、标注语料、基础本体或词库等领域基础资源条件，深度学习的研究应用目前仍主要集中在生物医学领域和通用领域，其他领域还非常少。深度学习的主要优点是能够使用现成的或衍生的各种深度神经网络词嵌入模型或算法从领域文本中自动学习特征，从而避免了繁重且耗时的特征工程，且学习特征的过程是人工、领域、语言非依赖性的，是充分利用嵌入了上下文语义信息的，因而可移植、可重用、可扩展性也强。

有三个核心的优势是将DL应用到NER技术上的原因。首先，NER会收益于非线性变换，这个变化产生了从输入到输出的一个非线性映射，例如对数线性的HMM以及线性链CRF，DL模型是能够学习到复杂隐晦的特征，通过那些非线性的激活函数。其次，DL也省去了设计NER特征的大量时间。那些传统的基于特征工程的方法需要大量的工程技巧以及领域知识。第三点就是，深度神经网络化的NER模型可以使用端到端的模式，通过梯度下降的方法去训练学习。这个特征就使得我们可以设计更加复杂的NER系统。

随着硬件计算能力的发展以及词的分布式表示（word embedding）的提出，神经网络可以有效处理许多NLP任务。这类方法对于序列标注任务（如CWS、POS、NER）的处理方式是类似的：将token从离散one-hot表示映射到低维空间中成为稠密的embedding，随后将句子的embedding序列输入到RNN中，用神经网络自动提取特征，Softmax来预测每个token的标签。

这种方法使得模型的训练成为一个端到端的过程，而非传统的pipeline，不依赖于特征工程，是一种数据驱动的方法，但网络种类繁多、对参数设置依赖大，模型可解释性差。此外，这种方法的一个缺点是对每个token打标签的过程是独立的进行，不能直接利用上文已经预测的标签（只能靠隐含状态传递上文信息），进而导致预测出的标签序列可能是无效的，例如标签I-PER后面是不可能紧跟着B-PER的，但Softmax不会利用到这个信息。

学界提出了DL-CRF模型做序列标注。在神经网络的输出层接入CRF层(重点是利用标签转移概率)来做句子级别的标签预测，使得标注过程不再是对各个token独立分类。常见的有CNN-CRF,LSTM-CRF及BERT-CRF等。

R.Collobert等人采用统一的神经网络模型，并避免特定任务的工程来实现对词性标注、分块、命名实体识别和语义角色标记进行处理。G.Lample等人利用双向长短期记忆网络（Bidirection Long Short-Term Memory, Bi-LSTM）结合CRF 的模型以及基于转换的方法分别对实体进行命名识别，模型采用了从监督语料库中学得的词向量和从未标注语料库中学习的词向量，结果证明模型在四种语言环境下的实体识别都能得到较好的效果。Ma 等人通过使双向长短期记忆网络、卷积神经网络和条件随机场结合，介绍了一种新颖的网络架构，实现端到端的模式进行序列标注任务的处理。A.Bharadwaj等人引入一种注意力神经模型对实体进行识别，在同一种语言环境中，该模型仅使用具有词向量的语言通用语音字符向量进行有监督的学习便取得了良好的表现，并且可以快速适应具有最少数据或没有数据的新语言。张帆等人引入深层神经网络算法对实体进行命名识别，该算法能从数据集中提取更准确的识别信息。金留可针对生物实体命名识别，提出递归神经网络结合预训练词向量的方案，并在实验中调整网络结构以提高算法识别效果。

第一章　绪论

1.1　精神病学的概念与任务

二、实验设计与语料集构建

本研究选取了医学领域中的精神病学科进行中文术语识别研究。

1.语料集的构建

根据前期语料的检索与收集，本研究在LaTeX中文开源医学书籍网站 上下载了精神病学科的书籍，并将其书后附录中的中英文对照索引作为术语词表，而书中的正文作为标注语料。

按照命名实体识别的标注格式，我们首先通过编程处理提取出术语词表，之后基于领域词典的字符串匹配分词方法，对正文内容进行了IOB标注。并按照7：1的比例划分了训练集与验证集。

术语词表共整理了910个术语，标注数据共整理8050个句子。对于测试集的数据，我们摘取了精神病学科的60篇文献摘要，对此进行了人工术语标注，以此作为高质量的开放测试集。

附录 B　中英文对照索引

5-羟色胺综合征	serotonin syndrome

A

阿狄森病	Addison disease
阿尔茨海默病	Alzheimer's disease, AD
阿罗伊瞳孔	Argyll-Robertson's pupils
阿米替林	amitriptyline
阿片	opium
阿普唑仑	alprazolam
安非他酮	bupropion
安定类	ataraxics

B

白痴	idiotism
被害妄想	delusion of persecution
被窃妄想	delusion of being stolen
苯丙胺	amphetamine
苯二氮卓类	benzodiazepines
苯环利定	phencyclidine
苯哌马嗪	phenmetrazine
边缘型人格障碍	borderline personality disorder
表演型人格障碍	histrionic personality disorder
白癜风	willige
丙咪嗪	imipramine
丙戊酸盐	valproate
病理性赘述	circumstantiality

399

2.实验方法的选取

根据前期的调研与学习，本研究选取了机器学习方法（条件随机场CRF）、深度学习方法(双向lstm)、两者混合方法（Bilstm-CRF）以及利用BERT表征词向量的两者混合方法（BERT-Bilstm-CRF）分别进行了中文术语识别实验，并对实验结果进行比较分析。

3.评估指标

本研究通过对识别结果进行评估以判断其识别效果，查准率(Precision，P)，又称精确度，用于衡量被预测为正样本的样本中正确率的高低；查全率（Recall，R），又称召回率，用于衡量实际为正样本的样本中被正确分类的概率；F1值（F1-score）综合了查全率和查准率两项指标，F1值越大，则表示识别的效果越好。本研究通过对四种识别模型的查全率、查准率、F1值进行比较，以分析模型的识别效果。

三、实验结果

1.条件随机场实验结果

在基于条件随机场模型的实验结果中，验证集数据F1值达到0.9357，而测试集的F1值为0.4477。实验结果如表1所示。

2.双向LSTM实验结果

在基于BiLSTM模型的实验结果中，验证集数据F1值达到0.8553，而测试集的F1值为0.3850，验证集和测试集的实验结果均低于CRF模型。实验结果如表2所示。

3.BiLSTM-CRF实验结果

在基于BiLSTM-CRF模型的实验结果中，验证集数据F1值达到0.8395，而测试集的F1值为0.4230，加入CRF的BiLSTM模型在验证集数据的实验结果上有所下降，而在测试集数据的实验结果上有所提升，但仍未超过CRF模型。实验结果如表3所示。

4.BERT-BiLSTM-CRF实验结果

在基于BERT-BiLSTM-CRF模型的实验结果中，验证集数据F1值达到0.8989，而测试集的F1值为0.4680，基于BERT做词向量表征的BiLSTM-CRF模型在验证集和测试集数据的实验结果上均有所上升，且在测试集数据上超过CRF模型。实验结果如表4所示。

表 1 条件随机场模型的实验结果

数据集		查准率（P）	查全率（R）	F1 值
验证集	B	0.9833	0.9262	0.9539
	I	0.9961	0.8670	0.9271
	avg	0.9920	0.8860	0.9357
测试集	B	0.7565	0.3625	0.4901
	I	0.9959	0.2792	0.4361
	avg	0.9443	0.2791	0.4477

表 2 Bilstm 模型的实验结果

数据集		查准率（P）	查全率（R）	F1 值
验证集	B	0.9067	0.8800	0.8931
	I	0.7961	0.8560	0.8203
	avg	0.8477	0.8630	0.8553
测试集	B	0.6508	0.3661	0.4686
	I	0.3958	0.2500	0.3065
	avg	0.5148	0.3075	0.3850

表 3 Bilstm-CRF 模型的实验结果

数据集		查准率（P）	查全率（R）	F1 值
验证集	B	0.9105	0.8468	0.8775
	I	0.8219	0.8025	0.8121
	avg	0.8549	0.8247	0.8395
测试集	B	0.6434	0.3705	0.4703
	I	0.5113	0.2982	0.3767
	avg	0.5763	0.3341	0.4230

表 4 BERT-Bilstm-CRF 模型的实验结果

数据集		查准率（P）	查全率（R）	F1 值
验证集	B	0.9297	0.9080	0.9188
	I	0.8825	0.8825	0.8825
	avg	0.9025	0.8953	0.8989
测试集	B	0.6870	0.3982	0.5042
	I	0.5954	0.3391	0.4321
	avg	0.6412	0.3684	0.4680

四、结果分析

在验证集数据的实验结果中条件随机场模型的识别性能较为优异，而在测试机样本上，BERT-BiLSTM-CRF模型的识别性能最佳。四种模型在验证集样本的实验结果上均表现较好，而在开放测试集上大幅度下降其原因可能有以下几点：

（1）本研究的实验样本数据规模较小，更适用于机器学习方法，而深度学习方法无法通过较少的数据学习到深层次的语言特征；

（2）测试集数据来源于不同体系，其中出现了许多词表中没有标注的术语，增加了模型的识别难度，其结果在条件随机场模型中查全率极低有直接体现；

（3）相比于其他三种方法，BERT表征模型不再采用传统的基于词频、TFIDF等的文本特征，而利用其深层次的神经网络结构对大规模通用文本的开展预训练从而捕获了文本中丰富的语言特征，使识别效果得到飞跃的提升。

在进一步的实验中，我们将进一步扩大词表规模，增加标注样本，以获得更好的实验效果。

读《明朝那些事儿》走精彩人生路

作者：中科院高能所 19 级硕士 李嘉良

记得刚刚上大一时，从高考中解脱出来的我就像一只活泼快乐的喜鹊，对一切新鲜事物都充满着好奇和探索的活力，很多学长学姐在见面会上都告诉我希望我不要辜负大学的青春时光，要好好学习争取大学毕业时考上研究生，去接触更广阔的知识，探索前所未闻的实物。我当时总是会是一笑置之，总觉得考研太难，离我毕业太遥远。然而，随着学期的深入，我的学习遇到了种种的压力，不断的考试、测验和四级。于是压力和抱怨、烦躁的心情开始从心中弥漫。结果大一上学期我的成绩就很不理想，已经有向学渣滑坡的趋势，但是我并不放弃，依然积极努力学习，争取赶上班里的学霸。当我接触到《明朝那些事儿》这本著作时，一开始是抱着消遣了解的心态看这本书，但是后来当我逐渐领悟其中的含义并且加上了对自己大学生活和学习的思考后，一切都变的与众不同。

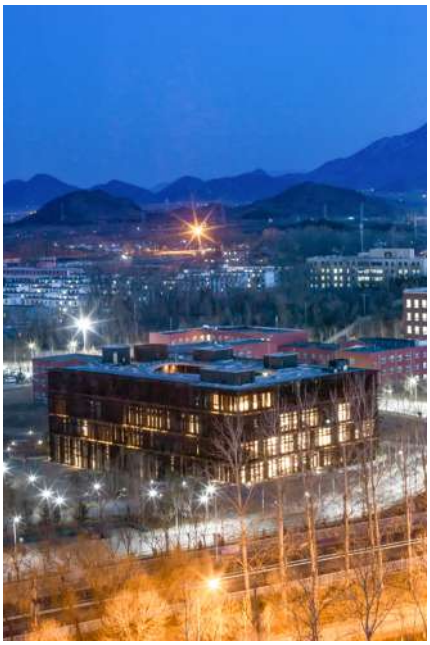
这本著作分为7部，简述从1344年到1644年这三百年间关于明朝的一些故事。但它并不单纯只是历史纪实，它还包含着对人生的思考，做人处事的态度和奋斗、坚忍不拔的精神。

从朱元璋不服命运，用自己坚强的意志反抗元朝腐朽的统治，从一个放牛的孩子经过成百上千的磨难，历经大小无数的战斗先后战胜了陈友谅、张士诚等强敌、率众将领击败北元军队，蜕变成开创大明帝国的皇帝，到崇祯面对清军入侵，民变四起的内忧外患的局面而力挽狂澜，拼命工作而且生活十分节俭，但即便条件是再艰难困苦也绝不投降，毫不退让。它仿佛给自己的学习生活注入了力量，面对学习中的困难和压力，我想起了朱元璋当年为了吃口饭而不被饿死，在皇觉寺中谨小慎微奋发图强，面对那么大的困难与挑战，甚至是死亡的威胁，他都没有向命运低头，一句“百花发时我不发，我若发时都吓杀”表现了他默默努力凌云壮志的豪迈气概，我学到了像朱元璋一样不服自己的学渣命运，奋发图强不惧对手的强大和考试压力，泡在图书馆里学习自己的课程，每门考试都向着接近满分地去准备，用朱元璋的成长来激励自己一定要做一个优秀的人；在生活中和同学之间不断地交流也明白了郑和下西洋的重大意义，也领悟了改革开放苦心之所在。只有对外交流，引进先进技术和思想，才能对我们的成长发展有提高促进作用，闭门造车只能是掩耳盗铃，贻笑大方。同时也学到了王阳明的淡泊名利，宁静致远的心态和哲学思想。在学习之余也感悟了明朝内阁权谋之术，从明朝嘉靖年间从杨廷和、夏言、严嵩、徐阶、高拱以及张居正这些出色的政治大家操持国事、处事之道。张居正作为明朝内阁首辅时开创了一个官员管理政策——考成法，类似于现在政府官员的绩效考核制，在我们学生的眼里就是平时成绩考核制，他让我不仅敬佩几百年前就有了这种管理考核制度，不仅有利于集体管理，还有利于提高官员办事勤政的效率，被现在的各大公司管理体制纷纷借鉴参考；也佩服于海瑞清官谏言，甚至不愿意自己去做轿子，明白了百姓父母官的苦心，更明白勤政爱民的深刻含义；万历时期明朝出兵朝鲜开始抗倭援朝，我体会到了中华民族团结一致的精神，也明白了国防事业对一个国家和民族的重要性，只有国家领土不受外敌的威胁，家国才能安居乐业；同样，万历年间出现的党争也让我领悟了团结的重要性。明朝中后期大臣派系林立，宦官干政，加上文臣内部不团结的因素加速了明朝的内部危机爆发，再赶上后金势力的崛起，明朝后期努尔哈赤以及皇太极的屡屡逼犯明朝边境，农民起义军在国内制造混乱，最终导致内忧外患的严重后果。这个教训让我深刻理解了一个班级、一个团队能够变得更加优秀的前提是内部团结，心系一方.....

后来，我正式开始踏上考研的征途。这本著作它不仅仅是我考研学习的娱乐伙伴，而是一种精神的伙伴，它给我一种坚持的精神力量。它在我连续两天两夜熬夜地设计二级圆柱斜齿轮减速器的工程任务时，它陪伴着我并且告诉我不要怕疲惫和辛苦，胡宗宪、戚继光抗倭历经无数艰难，虽然给明朝的经济带来了损失，许多百姓流离失所，但是他们依靠不屈不挠的精神和智谋最后取得了抗倭的胜利；当我考研复习受挫，信心即将丧失之时，是它让我明白了几百年前燕王朱棣靖难之战也是历经无数波折、失败和生死关头，他几次险些丢了性命，他用了四年时间坐上皇位，五征漠北、七下西洋、迁都北京，组织修撰永乐大典。燕王的置之死地而后生不畏困难的精神让我感触颇深，尽管现在困难重重，专业课压力巨大，考研形势不容乐观，但只要坚定信心目标明确一定能获得属于自己的辉煌。我也想起了崇祯，面对清军步步紧逼，国内民变四起，到处都在打仗。他夜以继日，勤勤恳恳，任劳任怨，每天上朝处理国事十几个小时。登基大位之后他首先除去以魏忠贤为首的罪大恶极的阉党，然后捐献个人的金银财宝支持打仗，一边统筹名将洪承畴、曹文诏、孙传庭等人平定内乱，一边开仓济粮来赈灾荒。崇祯勤政勤俭，自己穿的衣服是皇妃帮他打的补丁，为了不让大臣们看出来，他走路很慢生怕皇袍里面的破衣服飘出来。崇祯面对内忧外患，却依然无怨无悔愿意以一己之力来力挽狂澜，直到明朝结束李自成攻进北京城，也依然没有放弃。是他明确地告诉我，走上了考研这条路，就不能再回头。即便再艰难困苦也要为了信念而战，崇祯虽然没能成功，但是他的气节赢得了所有人的尊重。崇祯面对强敌也敢于亮剑的精神深深地触动了我。当考研已成为我的一种信念之时，它又告诉我杨继盛、杨涟、左光斗为了信念无视东厂锦衣卫特务的威逼利诱，不惜自己的性命而名垂青史，这些忠义之士让我心生震撼。我树立了不破楼兰终不还的决心，不畏惧繁重的学位课，全新全意地学习知识，以坚定的决心和无谓的勇气踏上考研的考场。

最终我成功地录取为研究生，当大学毕业班级聚餐拍摄毕业照之时，我再回首这段拼搏的充实岁月，我不禁感慨自己比刚上大学时要成熟了许多，经历的无数的困难挫折都成为我成长的试金石。我想起了伴随我成长的这本著作在最后快结尾时写道，这本书所讲述的不仅仅是明朝的历史纪实来解我们对历史的好奇心之渴，它还向我们表述了那个时代的人们所表现出来的正义、痛苦、愤怒、热情、刚强、气节、坚持、善良和忠诚等等。一个朝代如此，人亦如此，吸取前人之教训，学习他们的精神，让自己的内心沉淀下来。当我回首这段充实而精彩的奋斗岁月时，我发现这本著作就是我的精神导师，它不仅教会了我面对困难的坚忍不拔的精神，也让我学会拥有了海纳百川的包容胸怀；它既让我领略到明朝文人大学士对信仰的坚韧，也让我明白了做人处事的道理，让我从一个高中的直率男孩变成了有思想有内涵的知识青年。在这本著作的最后一段，作者写徐霞客游记的来源来表达对生活的乐观豁达精神，要按照自己的方式去度过人生。也让我明白了人生的一层含义，我们不需要亦步亦趋的模仿成功人士的道路，也不必去刻意仰慕和攀比他人的惊艳和美丽。我们只需要有自己的思想，要根据自己的实际情况和按照自己的方式去走自己认为最有意义的人生路！

光影栏目 | 雁栖的时光



中丹学院外景

2019年新生开学典礼

供图 | 阮伟南

方寸之间
书香满溢
尽显大千世界

角落·图书馆星空

雁栖湖

风光旖旎，湖水清澈。
春秋两季常有成群的大雁
来湖中栖息，故而得名。

北京日出东方凯宾斯基酒店
如一轮冉冉升起的太阳，象征着和谐统一，故名为日出

古北水镇

古北水镇位于北京市密云区古北口镇，
背靠中国最美、最险的司马台长城，
坐拥鸳鸯湖水库，是京郊罕见的山水城结合的旅游度假景区。

观景台

群山缭绕
雁栖美景
尽收眼底
（完）

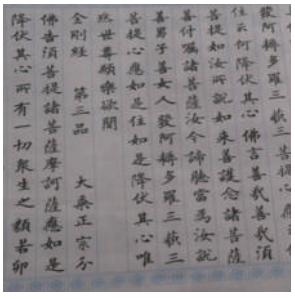
乐生活·思远方

——文情学子的宅家日常之诗与远方篇

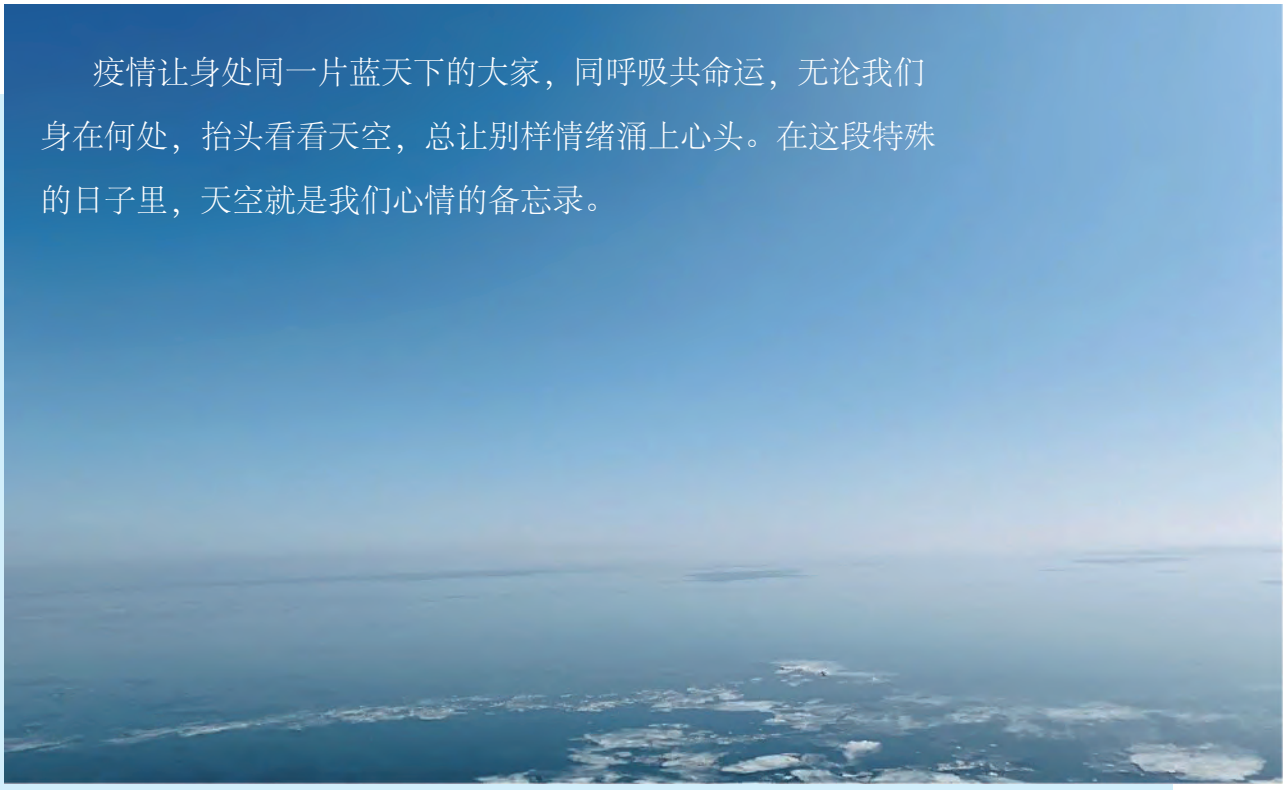
2020年的晚冬，一场来势汹汹的新冠疫情牵动着全国人民的心绪，将我们被迫困在家的小小空间里。居家隔离的日子虽然少了访亲会友的喧闹，但却让人幸得生活中的三寸闲暇，远离社交纷扰，将宝贵的时间留给至亲和自己。在这段艰涩的抗疫日常中，文情学子纷纷发挥着苦中作乐的精神，不仅用美食DIY填满了自己的肚子，也在用“诗与远方”的文艺食粮充实着与病毒作斗争的斗志和勇气。

19级博士研究生 闫群娇

铺开宣纸，重拾画笔，庞娜、闫群娇、王思培三位同学用纸上的一点余白保卫思想防疫战。法和素描曾让儿时的我们“坐立难安”，却不知不觉也成了坚定自我的武器。



居家隔离催生的不只是对自由的想念，还有对多姿生活的勃勃生机。冯凌子同学的手工DIY，让旧物换新面，也让平淡的心情焕然一新。



疫情让身处同一片蓝天下的大家，同呼吸共命运，无论我们身在何处，抬头看看天空，总让别样情绪涌上心头。在这段特殊的日子里，天空就是我们心情的备忘录。

三月的家乡·湖北

3月初春的如期而至也带来了防控降级的好消息。隔离后第一次走出家门，重见自由的天光，重见故乡的美景，

自研会开启“疫情宅家日常素材征集活动”后，来自五湖四海的同学们纷纷发来了自己的生活片段。无论是美食、美景，还是书法、绘画，希望大家都能在这段不寻常的生活里留下珍贵的回忆。

图片/庞娜、闫群娇、王思培、冯凌子、刘欢、廖宇、刘敬仪
文字/严格

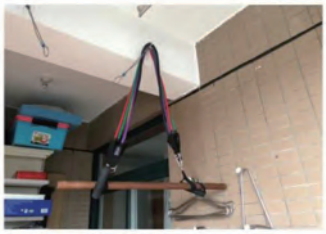
乐生活·思远方

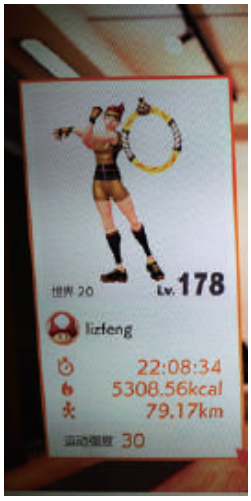
——文情学子的宅家日常之健康生活篇

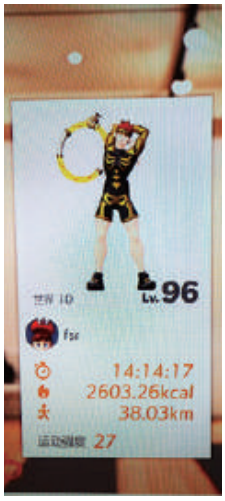
delicious food

FINE FOOD

疫情来袭之初，各省封域，高速封路，举国上下禁足在家——修炼厨艺。蛋糕、油条、凉皮、火锅……各类食谱频频登上各大网站的头条热搜。我中心各位同学居家隔离期间也练就了不少本领。接下来，让我们看看同学们的表演吧！水果糯米糍、部队火锅、清炒西兰花、上汤娃娃菜、沙拉大虾、鸡翅大虾煲……煎炒烹炸煮炖焖，腌卤酱拌生烤蒸，中华小当家今日我当家！学霸龙艺璇同学十八般武艺样样皆精。







18级硕士生阮伟南
——健身全能王

啥，全国的健身房都关门了？不存在的。阮伟南同学把健身房移到了自己的家，拉伸，卷腹，引体向上。生命不息，运动不止，要用汗水捍卫自己每一块肌肉。

18级博士生冯凌子
全家健身环大冒险

冯凌子同学领衔，带动全家在家运动。当电视屏幕变成运动大冒险界面，小弟不和我争游戏了，妈妈也不唠叨我整天看电视了，连守在窗户的狗狗也开始蹦跶了。在游戏中收获健康，在运动中陪伴亲人，何乐不为呢？

白云苍狗 三餐四季 MORE

西红柿，鸡蛋，青椒，土豆，香菇，简单的食材在廖宇同学的手中摇身一变，成为了盘中色香味俱全诱人美食，为家人带去无尽的力量。看他这熟练的刀法，简直就是一个平平无奇的做饭小天才。

在学校，我们的硕博生活或许是单调的，在小小的工位上一坐就是一整天，连用餐也是来回匆匆。青年人要能沉下心来做学问，同时也要认真对待生活。宅家期间，自律学习，自在生活。让我们用健康的体魄，昂扬的斗志，一起期待开学相聚时吧~！

图片/张颖、龙艺璇、廖宇、闫群娇、阮伟南、冯凌子
文/郝雪丽

居家小记

作者：廖宇

在家的生活和在学校生活最大的不同可能就是计时单位的不同，在学校都是用天和周计时的，在家都是按月计时的。在家时间过得飞快，也早已按耐不住回学校、回单位的心了。但是家里也有家里的乐趣，每天在家除了学习之外做的最多的事情就是做饭和带小侄女。

讲起来做饭这个事情，我做多了就发现了一个有趣的事情。自己做的饭总觉得越来越难吃了，反而是别人做的饭就很好吃。想起来真是特别奇怪，然后我就百思不得其解，这究竟是因为啥呢？难道是自己做饭偷吃吃饱了？好像也不是那么回事儿。还是自己做的饭菜缺乏了神秘感，因为自己都知道每道菜里都放了些啥，看似有道理总归还有些牵强。我妈妈猜测是不是油烟吸多了，影响了食欲？这似乎有道理了。于是我就进行了简单的实验，第一次我做饭的时候戴着口罩，照常偷吃一点菜，然后自己做炒菜完成全过程，我发现还是觉得菜就那样。然后，我第二次就让我哥哥给我放调料，不戴口罩，照常偷吃，结果自我感觉有点变化，多吃了一些。第三次，我没有偷吃，自己炒菜，也没有戴口罩，结果发现感觉菜品变化不大。难道是别人做菜带来的神秘感，让自己食欲大增？我扪心自问了一下，可能是不会放调料吧。

在家的生活和在学校生活最大的不同可能就是计时单位的不同，在学校都是用天和周计时的，在家都是按月计时的。在家时间过得飞快，也早已按耐不住回学校、回单位的心了。但是家里也有家里的乐趣，每天在家除了学习之外做的最多的事情就是做饭和带小侄女。

跟小侄女一起玩儿的日常真的是哭笑不得，她想让我买雪糕给她吃的时候，一般都会给我前奏，先夸我，"幺爸你好帅，你会眶娃娃（带孩子）"，然后就进入主题，"幺爸，今天有太阳，可以吃雪糕了（她爸妈规定了没太阳2不许吃雪糕）"。当她不需要我的时候，就让我走，说不喜欢我了。最让我开心的事情是我给她看了一段彭于晏的视频，然后真诚发问"宝宝你觉得他帅还是我帅？"，她说，"幺爸帅！"。那小孩是不会撒谎的。

最后，希望大家身体健康，热爱生活，珍惜眼前！

大数据时代的情报学与情报学研究

——“情报学前沿问题研究”专家报告会侧记

10月21日，由中国科学院文献情报中心（以下简称“中心”）、中国科学院科技战略咨询研究院（以下简称“战略院”）联合主办的“情报学前沿问题研究”专家报告会在中国科学院学术会堂举行。

会议邀请到武汉大学资深教授马费成、中山大学咨询管理学院院长龙乐思（Miguel Baptista Nunes）两位专家作学术报告。中心和战略院的师生员工共100余人聆听了此次报告。报告会由中国科学院大学图书情报与档案管理系主任初景利教授主持。

马费成教授作了题为“大数据中的知识关联”学术报告。马教授从行业大数据转换为知识大数据的大背景引入，提出智慧数据的重要性，从而引出知识关联的理论与应用价值，并指出知识图谱为知识关联的应用提供了条件。

他从理论和应用两个方面对大数据中的知识关联进行阐释。在理论方面，他介绍了大数据知识关联的内涵、特征、四大关联模式。在应用方面，他从银行股权机构、知识服务平台两个案例分析知识关联的应用价值。马费成教授结合现场操作，为听众展现了此平台的界面与功能。通过此次跨学科的大数据关联的研究，指出作为数据全公开的金融风险研究始终是研究热点，通过穿透式查询进行数据的控制及分析也是此次研究的创新点，以及多学科的交叉研究的难度很大，需要长时间的磨合。

龙乐思教授作了题为“Qualitative Research Theory and Practice in Information”的报告。龙院长以“if it does not have statistics it is not science!!!”为切入点开始此次报告。他阐述了“what is research？”，以及社会科学与自然科学的区别，对社会研究的本体论与认识论进行了简要概述。他对演绎法与归纳法进行了比较阐述。通过一幅画、咖啡、在学研究生三个例子对比分析了定性数据与定量数据的区别，指出定性数据作为一种描述性数据，其分析难度更高。随后，他区分了定量研究与定性研究，对于同一研究对象，两种研究方法会给出不同的结论。

两位专家的报告站位高，富有前瞻性，阐释了大数据时代情报学发展的重要领域与研究方向。报告内容丰富，讲解清晰，理论结合实际，具有重要的引领和指导作用和意义。

图：刘琪
文：杨晓，郝雪丽

（责任编辑：郝雪丽）

图情档研究生教育现状与未来

暨中国科学院文献情报中心研究生教育40周年座谈会召开

今年是中国科学院文献情报中心（以下简称“中心”）研究生教育创立40周年。10月21日上午，中心在7D会议室隆重举办“图情档研究生教育现状与未来暨中国科学院文献情报中心研究生教育40周年座谈会”。

中心主管部门领导发展规划局副局长黄晨光、中国科学院大学副校长牛晓莉、中国科学院大学经济管理学院院长汪寿阳出席了此次座谈会。会议还邀请到吴慰慈、马费成、冯惠玲、戴国强、张久珍、方卿、苏新宁、张斌、柯平、耿骞、李玉海、龙乐思（Miguel Nunes）、王晰巍、李广建14位学界嘉宾参会。中心领导班子成员、各部门领导、图书情报与管理系与研究生教育处相关负责人、中心老专家孟广均、孟连生与校友50余人与会。

座谈会的前半场由文献情报中心副主任刘细文主持。首先，文献情报中心主任刘会洲研究员致欢迎辞。会洲主任代表中心党政班子，对各位领导和学界嘉宾的到来表示热烈的欢迎，对长期关心、支持中心研究生教育工作的各级领导、业界专家、同行同事及国内外校友致以诚挚的感谢与崇高的敬意。会洲主任对中心40年来的研究生教育工作和取得的成就作了简要汇报。他表示中心未来要加大对研究生培养的投入，优化师资队伍，不断推动研究生教育改革，继续探索具有中国科学院图情档研究生教育特质的研究生培养路径与模式。

中国科学院发展规划局副局长黄晨光代表中国科学院发展规划局对中心研究生教育工作40年取得的成果表示祝贺，并对中心的研究生教育工作表示肯定与感谢。他认为研究生培养点的发展，既需要卓越的带头人引领方向，也需要办学平台、科研环境的强强联合，方能保持学科发展的良好势头。希望中心在新的工作历程中不忘初心、再接再厉，更好地发挥院所协同的平台优势，为图情档研究生教育事业高质量发展贡献更大力量。

中国科学院发展规划局副局长黄晨光代表中国科学院发展规划局对中心研究生教育工作40年取得的成果表示祝贺，并对中心的研究生教育工作表示肯定与感谢。他认为研究生培养点的发展，既需要卓越的带头人引领方向，也需要办学平台、科研环境的强强联合，方能保持学科发展的良好势头。希望中心在新的工作历程中不忘初心、再接再厉，更好地发挥院所协同的平台优势，为图情档研究生教育事业高质量发展贡献更大力量。

随后北京大学资深教授吴慰慈、武汉大学资深教授马费成、中国人民大学一级教授冯惠玲三位资深专家相继发言。

吴慰慈教授表示中科院图书情报与档案管理学科开拓发展的40年，令人敬佩，值得庆贺。吴教授十分肯定中心历届领导对研究生教育工作的重视程度、诸位导师对研究生培养的认真态度，40年来中心一直保持着优秀的学术作风和办学水平。他还总结出中心研究生教育在发展中形成身独有的六个新特点，他认为中心这种依托本院优势，以特色驱动发展，着力找准定位，加强专业技能型和实践应用型人才的研究能力培养模式，是值得提倡的。他表示图情档一级学科的硕博士学位授权点来之不易，呼吁业界要珍惜前辈的奋斗成果，长久传承精益求精的教育精神。

马费成教授表示武汉大学和中心在研究生教育工作方面长期保持着良好紧密的联系，参与过中心研究生培养制度制定、学科建设等工作。他表示中心研究生培养的特色非常鲜明，这些离不开中科院在图情档领域发展进程中具备的三大条件：强大的导师队伍、丰富的教学资源和特色的服务文化。

冯惠玲教授祝贺中心图情档专业教育事业进入不惑之年，她认为历经40年风雨兼程，国科大图情档系进入知目标、知方向、知使命、知路径的发展新阶段。她特别指出中国科学院大学在图情档学科培养方面最大的特色就是知实践，图情档学科具有很强的场景特征，知实践显得尤为重要。中国科学院大学有知实践的条件和自觉，对环境的变化感知最敏锐、最准确，能更好地和实践对接，中心研究生教育能够避免同质化办学，拥有先天的异质特点。最后，冯惠玲教授期待中心研究生教育知者不惑、仁者不忧、勇者不惧，将图情档教育办出特色、办出水平、办出精彩。

接下来，中国科学院经管学院院长汪寿阳、科睿唯安中国区总裁郭利为中国科学院大学经济管理学院与Clarivate科睿唯安研究生实习基地揭牌。

汪寿阳院长代表中国科学院大学经济管理学院对中心的研究生教育工作给予赞扬，充分肯定了中心研究生培养的特质。同时感谢科睿唯安对经济管理学院和文献情报中心的支持，让研究生能够了解企业的实际和更多的学科前沿问题。在图情档未来的发展中，汪院长希望能够以重大项目为抓手，积极争取发展机会，关注科研信息多元化发展趋势，与信息技术、数字处理、新媒体等专业合作，加强学科共建。

科睿唯安中国区总裁郭利介绍了Clarivate科睿唯安与中科院的合作情况，她表示公司与中心的合作将从研究生教育切入，面向未来，希望能为中科院研究生设立和提供更多企业实践项目的机会。

会议后半场由中国科学院大学图书情报与档案管理系主任初景利主持。在以“图情档研究生教育现状与未来”为主题进行的开放研讨中，各位专家从不同角度出发，就当前图书情报与档案管理研究生的培养重点与培养思路进行了探讨和交流。

中国人民大学信息资源管理学院院长张斌教授从三个角度展开论述。一是时代发展的变化。他提到数字时代的管理形态，逐步从“纸质形态”过渡到“数字形态”，在这种变化中，图情档学科的理论体系、教材体系要积极适应变革。二是社会需求的变化。当今社会的新机遇、新领域，对学科之间的交叉与融合度特别高。在互联网、新媒体的这些新领域中，图情档学科要在坚守传统领域的基础上，积极地适应社会需求。三是研究生教育对象和重心的变化。图情档研究生的培养数量超过了本科生，要转变培养模式，提高培养规格，提升培养质量。

南开大学商学院信息资源管理系教授、长江学者柯平教授感叹图情档一级学科建立的坎坷历程与光荣使命，追忆图情档教育变革的重要时期。他赞赏中心研究生教育独特的学、研、馆结合的人才培养模式，对培养学生面向实践的能力发挥了重要的作用，实现了从理论到实践的无缝衔接。

南京大学信息管理学院教授、长江学者苏新宁提出在未来新的发展环境下应该怎样培养图情档研究生的关键问题。他表示首先要培养实践型研究生，要牢记我们的使命，坚守图情档学科教育内核；其次，在当前环境下，要做“引领”型人才，真正让图情档人才在各个相关领域发挥更重要的引领作用。

武汉大学信息管理学院院长、长江学者方卿教授则对图情档学科名称问题进行了讨论。华中师范大学信息管理学院院长李玉海认为图情档的学科建设是由社会需求推动。

中心前辈孟广均先生对中心隆重举办研究生教育 40 周年的会议感到十分欣慰，作为中心老教师代表，他向中心研究生教

吉林大学管理学院学院副院长王晰巍就研究生教育中课程和教材问题发言，提出建立院校联盟机制的畅想，以此实现学生和青年教师双层次的校际交流和人才培养的良性发展。

科睿唯安中国区业务总监宁笔对中心支撑科研、服务师生的重要作用表示敬佩，对奋斗在图情一线的教育者表示感谢。

孟连生研究员从自身研学经历谈到，自己是中心研究生教育工作的参与者与见证者。在从事图情工作几十年的工作经验中，他深刻体会到中心在研究生培养中始终强调实践与科研相结合的独特优势，为中心在人才培养中获得的成绩感到自豪。

北京师范大学政府管理学院副院长耿骞对中心研究生教育40周年表示祝贺，他谈到北师大和文献情报中心的人才交流密切，感谢中心为北师大输送的图情档专业人才。希望图情档学科的未来的发展更加注重教育实践，高校之间能有更好的协同合作。

中山大学资讯管理学院院长龙乐思（Miguel Nunes）教授谈到图情档学科的引领式作用需要承担学科变化带来的职责、解决学硕和专硕培养教育中存在的问题，以此来面对社会信息需求带来的挑战和机遇。

北京大学信息管理系主任张久珍教授认为好的教育是由好老师、好学生构成，在图情档教育发展的征程中要留下好的学生，图情人要具备危机感和使命感，与时间赛跑，不断奋斗。

中国科学技术信息研究所所长戴国强研究员提出图情档教育教学方面应实现突破和跨越，期待图情档领域在大数据算法领域有所开拓。

中心研究生教育走过了40年的不凡历程，得益于中科院、国科大、文献中心、国内专家和全体师生的大力支持，得益于国家科技与教育创新发展的大好环境。40年已经属于历史。不忘初心，座谈会只是一个标识，中心研究生教育（图书情报与档案管理）在国科大经管学院与文献情报中心科教融合的大环境下，将扬帆远航再出发，走特色发展道路，为国家培养更多更好的图情档高层次专业技术人员。

THE END

图：阮伟南
文：郝雪丽，齐铮

（责任编辑：郝雪丽）