近二十年统计学领域研究的可视化分析

南雪琪

（西北师范大学　数学与统计学院，甘肃 兰州　730070）

摘　要：随着计算机技术的发展，日益扩大了传统的和先进的统计技术的应用领域，促使统计科学和统计工作发生了革命性的变化，统计理论和实践深度和广度方面也不断发展。以CSSCI为数据源，对2000-2018年发表的关于统计学领域的相关论文进行比较研究，使用Ucinet6、SPSS22.0、Gephi0.8.2、CiteSpace和VOSviewer软件分别从高产作者、高频关键词和期刊三个方面进行可视化分析研究近二十年来统计学在我国的主要应用领域和研究方向及热点。

关键词：统计学；可视化分析；近二十年；研究热点

基金项目：国家自然科学基金项目（11861075）

中图分类号：C81　　**文献标识码：**A

文章编号：1674-537X（2020）09.0015-07

DOI:10.16722/j.issn.1674-537x.2020.09.004

引言

统计学的英文 statistics 最早源于现代拉丁文 statisticum collegium（国会）以及意大利文 statista（国民或政治家）。德文 Statistik，最早是由 Gottfried Achenwall 于 1749 年使用，代表对国家的资料进行分析的学问，也就是“研究国家的科学”。在十九世纪统计学在广泛的数据以及资料中探究其意义，并且由 John Sinclair 引进到英语世界。统计学起源于研究社会经济问题，在两千多年的发展过程中，统计学至少经历了“城邦政情”“政治算数”和“统计分析科学”三个发展阶段。所谓“数理统计”并非独立于统计学的新学科，确切地说：它是统计学在第三个发展阶段所形成的所有收集和分析数据的新方法的一个综合性名词。概率论是数理统计方法的理论基础，但是它不属于统计学的范畴，而属于数学的范畴。而现代统计学的理论基础概率论始于研究赌博的机遇问题，大约开始于 1477 年。数学家为了解释支配机遇的一般法则进行了长期的研究，逐渐形成了概率论理论框架。在概率论进一步发展的基础上，到十九世纪初，数学家们逐渐建立了观察误差理论，正态分布理论和最小平方法则。于是，现代统计方法便有了比较坚实的理论基础。[1] 统计学是应用数学的一个分支，主要通过利用概率论建立数学模型，收集所观察系统的数据，进行量化分析、总结，做出推断和预测，为相关决策提供依据和参考。它被广泛的应用在各门学科之上，从物理和社会科学到人文科学，甚至被用来工商业及政府的情报决策之上。随着数字化的进程不断加快，人们越来越多地希望能够从大量的数据中总结出一些经验规律从而为后面的决策提供一些依据。统计学专业不是仅仅像其表面的文字表示，而是包含了调查、收集、分析、预测等，应用的范围十分广泛。因此，此领域的研究成果也比较丰硕。但是，纵览这些研究发现缺乏对这一领域的可视化分析。从这个角度出发，本文通过一系列专业软件对统计学这一领域做出了简明的可视化分析，以便读者能从宏观角度更为直接的了解这一领域，帮助大家更好的学习、研究和预测。可视化分析概括的说就是利用计算机图形学和图像处理技术，将数据转换成图形或图像，可直观清晰地提供了解学科发展状况、跟踪学科研究热点、选择科学方向等知识图谱 [2]。可以发现，对统计学这一学科的可视化分析有助于更深层次的探讨它的有用价值和未来更有意义的研究。

一、数据来源及处理

（一）数据来源

表1：研究数据来源

	内容
数据来源	中文社会科学引文索引（CSSCI）
检索篇名	统计学
时间跨度	2000-2018
检索结果	352篇论文
检索时间	2019年4月27日

本文基于中文社会科学引文索引（Chinese Social Sciences Citation Index，简称 CSSCI）数据库（是最具学术权威性的引文信息源），CSSCI 遵循文献计量学规律，采取定量与定性评价相结合的方法，从全国 2700 余种中文人文社会科学学术性期刊中精选出学术性强、编辑规范的期刊作为来源期刊。以“统计学”为篇名检索从 2000 年到 2018 年间发表的文章，匹配条件为“精确”，检索到符合要求的文献共计 325 篇。数据下载日期均为 2019 年 4 月 27 日。

（二）数据处理。

可视化分析综合运用计算机图形学、图像等技术，将采集的不可见或难以显示的数据映射为可感知的图形或符号等，其目的是以更直观的方式洞悉蕴含在数据中的现象和规律。随

着大数据时代的到来，一些公司和高校相继开发了具有很强适用性和易用性的可视化分析软件，并被广泛应用于情报学、科学计量学和管理学等学科研究中。[3]本文使用的可视化分析软件有Ucinet6（University of California at Irvine NETwork）、SPSS22.0、Gephi0.8.2、CiteSpace和VOSviewer。Ucinet6是目前最流行的社会网络分析软件。[4]本文运用Ucinet6绘制关键词的共现网络图谱。SPSS22.0具有图形输出、数据管理和数据分析等功能，受到世界上许多有影响的报刊杂志的高度评价。[5]本文运用SPSS22.0对高频关键词进行聚类分析。 Gephi0.8.2是一款开源复杂网络分析软件，其复杂网络探测和动态图形处理的技术居于世界一流水平。[6]本文运用Gephi0.8.2绘制高频作者–研究主题的2–模网络图谱。

二、结果分析

（一）研究基础分析

分析参考文献可以有效弄清某一领域的知识研究基础。高频参考文献之间的内在关系通过共被引关系显现出来。运用Vosviewer 对高频参考文献进行聚类分析，得到高频参考文献密度视图（见图 1）。图中圆圈代表参考文献，圆圈越大，表明被引频次越高；亮度越高，说明共被引次数越多。

图 1 显示，高频参考文献可分为三个大的聚类：《数理统计学简史》（聚类 1）、《大数据与统计新思维》（聚类 2）和《统计学》（聚类 3）。

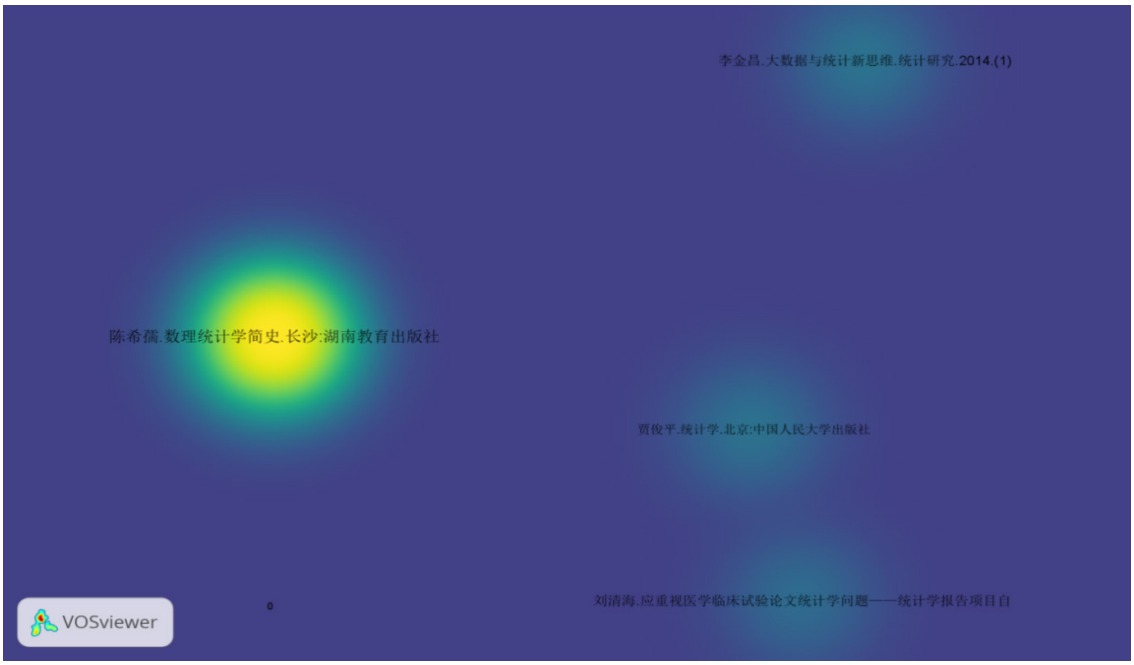

图1：高频参考文献密度视图

1、聚类 1

《数理统计学简史》是经陈希孺编著，2002 年由湖南教育出版社出版的一本图书。该书主要论述了自 17 世纪迄今数理统计学发展的简要历史。内容包括：概率基本概念的起源和发展，伯努利大数定律和狄莫旨二项概率正态逼近，贝叶斯关于统计推断的思想，最小二乘法与误差分布－高其正态分布的发现过程，社会统计学家对数理统计方法的主要贡献等知识点。

2、聚类 2

《大数据与统计新思维》是由浙江工商大学教授李金昌所写。随着信息时代的来临大数据正在改变着人们的行为与思维，这篇文章基于对大数据的理解，认为统计思维需要发生三个方面的改变，即要改变认识数据的思维、收集数据的思维和分析数据的思维。其中，数据分析思维又要在统计分析过程、实证分析过程、推断分析逻辑等方面发生变化，同时统计分析评价的标准也要有所调整。围绕这些变化，作者在文章中提出“需要从八个方面去积极应对大数据，以促使统计学跟上时代的步伐”。

3、聚类 3

《统计学》是由贾俊平、何晓群和金勇进合作编著，经由中国人民大学出版的图书。该书作为高等院校经济管理类专业本科生统计学课程的教材，也可作为 MBA 的教材或参考书，对广大实际工作者也极具参考价值，且《统计学》是一本很经典的统计学优秀教材。

（二）载文机构分析

由于学术期刊上刊登了大量学科前沿及其研究热点的文章，可以全面反映本学科的发展的现状与趋势，在科学技术活动的文献交流起着非常重要的作用，是一种正规的学术交流媒介，同时也为决策部门提供了客观的评价我国科学活动的基础数据。所以对于载文机构的研究很有必要。本文运用 CiteSpaceIII 绘制载文机构知识图谱（见图 2）。图中圆圈大小代表机构发文量的多少，圆圈越大，表明该机构发文越多；连线代表各机构间的合作，连线越多，表明机构间合作越频繁。

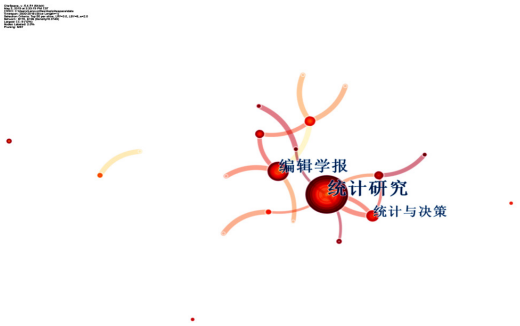

图2：2000–2018年载文机构知识图谱

图 2 中，表明在 2000–2018 年间统计学领域发文量最多的三个期刊首先是《统计研究》，其次是《编辑学报》，最后是《统计与决策》。且三个机构之间的合作关系也很密切。

1、《统计研究》

本书于 1984 年创刊，以月为出版周期的期刊，由中华人民共和国国家统计局主管、中国统计学会和中华人民共和国国家统计局统计科学研究所主办、中华人民共和国新闻出版总署正式批准中国国内外公开发行的学术性期刊。主要栏目是统计基本理论问题、统计理论方法与应用、经济分析与统计分析、经济核算问题研究等。2004 年获国家期刊奖百种重点期刊，2012–2015 年被评为中国最具国际影响力学术期刊，2014 年获得中国人文社会科学期刊评价报告统计学类权威期刊。

2、《编辑学报》（Acta Editologica）

本书是由中国科学技术协会主管、中国科学技术期刊编辑学会主办的学术性期刊。创刊于 1989 年，出版周期为双月刊。主要发表有关科技期刊编辑出版理论与实践问题研究的文章。是我国信息与知识传播类核心期刊和中国科技核心期刊，其主要栏目有理论研究、编辑工程与标准化、经营管理、期刊现代化、人才培养、办刊之道、学术争鸣、期刊评价、他山之石、编辑感悟、有问必答、谬误辨析、编余雅兴、消息等。于 2014 年中国人文社会科学期刊评价报告获得新闻学与传播学类核心期刊。

3、《统计与决策》

本书是由湖北省统计局主管，湖北省统计局统计科学研究所主办，出版周期是半月刊。此期刊的宗旨是立足统计理论，关注经济热点，推介决策方法，传递学术信息。期刊的特色是观点新颖、内容务实、风格泼辣、统计与决策结合和理论实务并重。主要的读者对象是院校师生、科研院所研究人员和统计工作者。主要栏目包括理论新探、决策参考、经济纵横、工作视点、统计观察、财经论坛和当代统计人。获奖情况：中国学术期刊（光盘版）全文收录期刊，中国期刊网全文收录期刊，全国中文核心期刊，中文社会科学引文索引（CSSCI）来源期刊，中文科技期刊数据库来源期刊。此期刊的研究方向是统计前沿理论，探究统计新方法，力主从统计的独特视角解析经济社会中的热点与难点问题，推介决策理论与方法，崇尚数量实证分析。

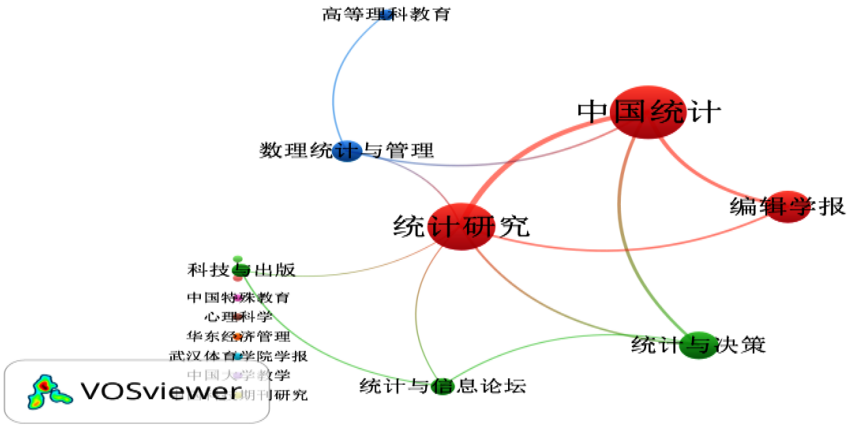

图3：载文机构知识图谱

图 3 中，中国统计的节点最大，是六大载文机构中最中心的节点。其和统计研究、编辑学报连线最粗，这表示中国统计和这两个机构的合作密切。但是统计研究与其他机构的连线最多，说明统计研究这一期刊和其他的机构之间的合作是最丰富的。

（三）研究主体分析

研究主体是一群围绕某一研究领域共同开展研究的学术共同体。通过对研究主体的分析，可以看统计学领域研究的高产第一作者和所在第一机构。第一作者的发文量反映了他们在统计学这一研究领域的学术贡献度。运用 Ucinet6 绘制第一机构的共现网络图谱（如图 1），运用 VOSviewer 软件绘制了第一作者的聚类（见图 4）。对于高频作者所在第一机构的研究有助于发现在统计学这个领域研究的主力军及其作者与机构之间的合作联系。

图 4 显示，从网络节点看，贺铿的圆点最大，表明其发文量最多，具有较大的学术影响力和凝聚力；从网络关系看，该网络是一个非联通网络，由 9 个相互独立的子网络构成，整体连通性不好。

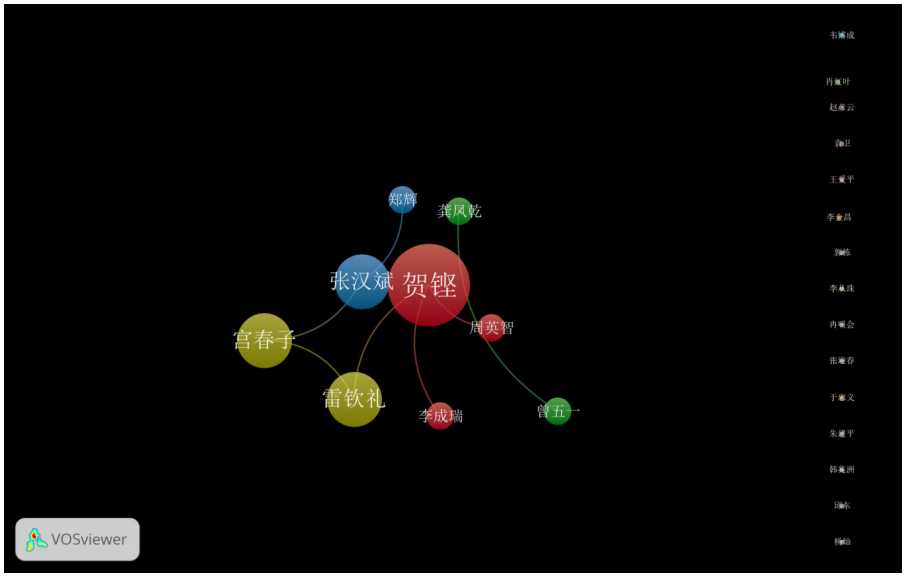

图4：高频作者知识图谱

（四）研究前沿分析

研究热点是指在某一时间段内，有内在联系的、数量相对较多的一组论文探讨的研究主题。分析高频关键词可以揭示某一领域的研究热点，因为高频关键词是被研究者集中研究的主题内容[7]。关键词可以根据齐普夫第二定律[8]（Zipf’s second Law：$T=\frac{-1+\sqrt{1+8I_1}}{2}$，其中 T 为高频词和低频次的分界频次，$I_1$ 为出现一次的关键词的数量）进行提取。根据齐普夫第二定律，计算得出 T=5，即频次高于 5 的关键词为高频关键词，共 34 个。为了体现高频关键词的分散和集中特征，运用 Ucinet6 绘制高频关键词共现网络图谱（见图 5）。

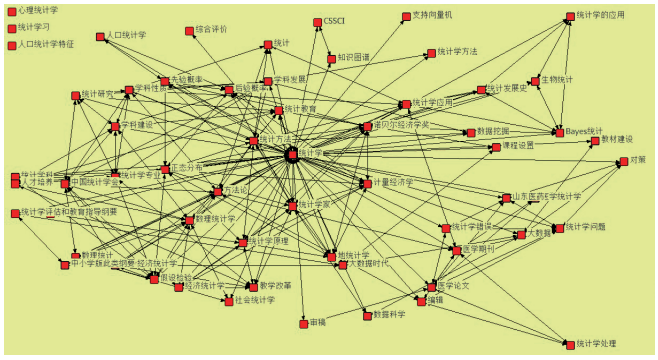

图5：高频关键词共现网络图谱

运行 Ucinet6 绘制高频关键词共现网络图谱（见图 5）。该网络图谱由 57 个节点及相互间的连线构成，每一个节点代表一个关键词，节点越大，反映该节点所代表的关键词与其他关键词共同出现在一篇文献中的次数越多。且节点较密集，节点之间连线稠密，且箭头指向较多，说明该关键词受关注度较高。

1、关键词共现网络图谱分析

对高频关键词共现网络图谱（图 5）进行分析可以得到如下结果：

（1）高频关键词涉及到诸多研究方向，如计量经济学和医学；

（2）高频关键词涉及到学科研究，如学科建设和学科性质；

（3）高频关键词涉及各类概念论专业研究，如正态分布和后验概率。

可见，我国统计学领域的研究内容多、范围较广、专业性强，这说明该领域研究难度大、任务重。结合高频关键词频次可知，统计学研究主要集中在数理统计、经济学的领域，同时还延伸到心理学和医学领域，这表明统计学的研究及其应用相对来说还是比较狭隘的。

2、高频关键词多维尺度分析

对高频关键词进行多维尺度的分析，能更进一步明确关键词之间的联系。聚类分析是根据关键词之间的相似性，将关键词进行聚类。而战略坐标分析能呈现聚类的关键词（热点主题）的研究发展状况。战略坐标的横轴（X 轴）表示向心度，纵轴（Y 轴）表示密度，原点是向心度和密度的平均值。向心度是用来衡量一个热点主题和其他热点主题的相互影响程度。向心度越高，表明该热点主题与其他热点主题之间有密切的联系，是研究者争相研究的主题，处于研究领域的中心位置；向心度越低，表明该热点主题与其他热点主题之间没有建立良好的沟通关系，是研究者不太重视的主题，处于研究领域的边缘位置。密度是用来衡量各个热点主题内部高频关键词之间联系的强度，密度越高，表明该热点主题内部联系紧密，已经形成了一定的研究规模，研究趋向成熟；密度越低，表明热点主题内部联系松散，研究规模小且尚不成熟，需要研究者投入较多的时间和精力。根据高频关键词聚类分析结果和高频关键词共现矩阵，计算每个聚类的向心度和密度，绘制出热点主题的战略坐标图。将高频关键词共词矩阵导入 SPSS22.0 中得到高频

关键词多维尺度聚类图谱（见图 6）。

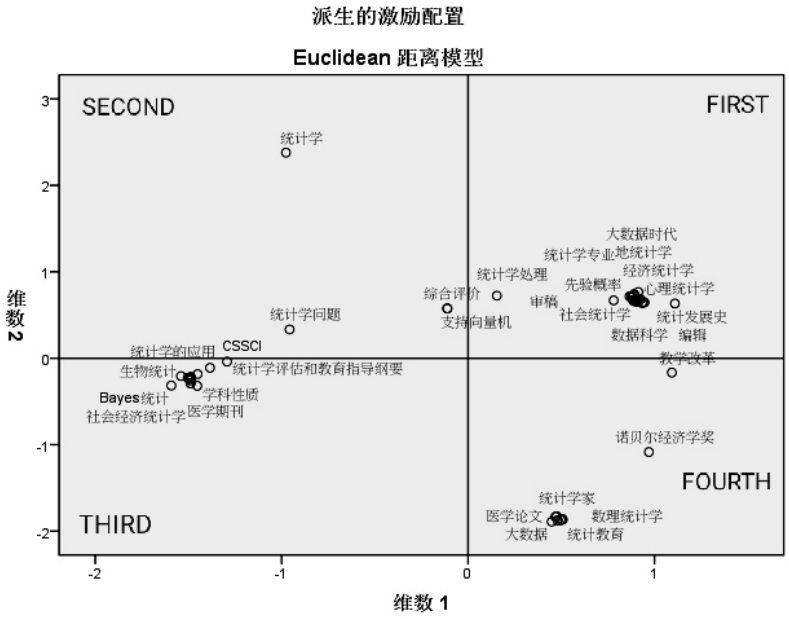

图6：高频关键词多维尺度聚类图谱

（1）向心度

从向心度来看，位于 X 轴右侧的热点主题的向心度高，热点主题所在的位置越偏右，该类热点主题的向心度越高。图 6 显示，热点主题所处位置从左向右依次是统计学的应用、统计学的处理、经济统计学，表明统计学的研究还是基于概率的统计研究，对其他方向的研究较低。

（2）密度

从密度来看，位于 Y 轴上方的热点主题的密度高，热点主题所在的位置越偏上，该类热点主题的密度越高。图 6 显示，热点主题所处位置从上至下依次是统计学问题、医学论文，表明统计学问题成熟度最高，医学论文成熟度最低。

（3）四大象限

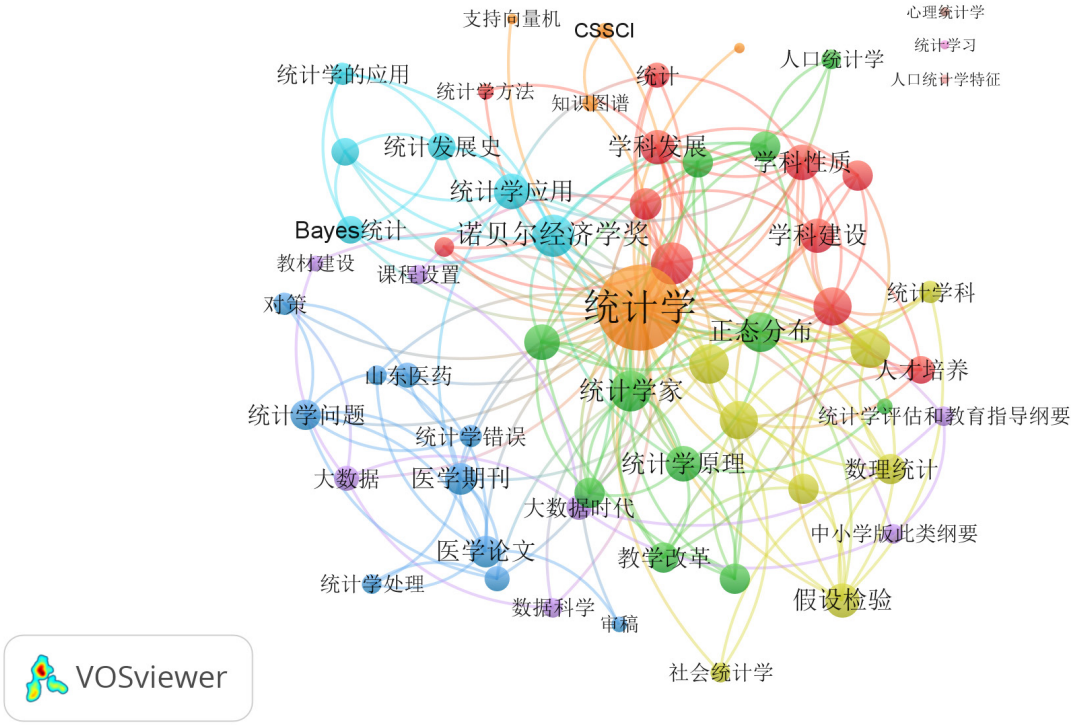

图7：高频关键词知识图谱

位于第一象限的热点主题，密度和向心度都高，表明该象限的热点主题不仅内部联系紧密，而且与其他热点主题存在较强的联系。第一象限的热点主题类似植物“结的果”。就像果实能在社会中产生积极作用一样，该主题的研究也能在社会中产生很大的影响。图 6 显示，第一象限热点主题分布集中，表明我国统计学领域研究产生了一定的社会效益。

位于第二象限的热点主题，密度高，但向心度低，表明该象限的热点主题内部联系紧密，但与其他热点主题联系较为松散，并不被广大研究者所重视。第二象限的热点主题类似植物“开的花”。就像花与花的授粉后才能结果一样，第二象限的热点主题研究需要博采众长。因此，位于该象限统计学问题研究为其他方向的研究铺垫了很好的基础。

位于第三象限的热点主题，向心度和密度都低，表明该象限的热点主题不仅内部联系松散，具有不稳定性，而且与其他热点主题的联系也不紧密，对外交往能力差。第三象限的热点主题类似植物“萌的芽”，生命力较脆弱，需要积极培育。图 6 显示，第三象限热点主题种类较广，表明我国还需要好好地培育其他方面的研究热点，才能形成可持续研究、多方位研究的良好局面。

位于第四象限的热点主题，密度低，但向心度高，表明该象限的热点主题与其他热点主题联系紧密，对外交往活跃，但内部联系松散，还未自成一体。第四象限的热点主题类似植物“长的叶”。就像叶具有的光合、蒸腾和呼吸作用是促进植物生长一样，第四象限的热点主题需要博观约取。因此，位于该象限的研究热点不仅从医学、经济方面拓宽，还要在后续的发展中继续壮大研究方向。

运用 VOSviewer 绘制出高频关键词的网络知识图谱（见图 7）

（4）2- 模关系

2- 模关系网络图谱是用来描述热点主题与其他因素之间的共现关系，能挖掘热点主题与其他因素共现关系的深层次结构。且不同作者的研究主题也不尽相同。

热点主题 - 高影响力作者 2- 模关系。发文量高且被引频次高的作者为高影响力作者。首先通过普赖斯定律（Prices Law）取高产作者和高被引作者的交集，得出高影响力作者共 20 位，分别是周英智、刘海清、王爱平、冉明会、邱东、肖红叶、邢菊、杨灿、张迎春、李金昌、朱建平、雷钦礼、赵彦云、曾五一、宫春子、韦博成、张汉斌、郑辉、郭栋、贺铿。运用 Gephi 绘制高产第一作者 - 研究主题 2 模网络图谱（见图 8），深入探讨两者间的关系。图中节点之间的连线代表热点主题与高影响力作者之间的共现频次，连线越粗，表明该作者对这一类热点主题的研究越多，也说明高影响力作者往往能就多个热点主题开展研究。图 8 的结论与图 7 的结果一致：越是热点的主题，越能吸引高影响力作者；越是影响力大的作者，越能捕捉到热点主题。

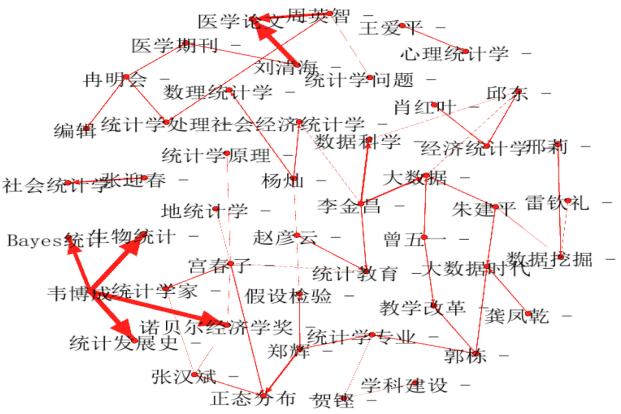

图8：高频作者–研究主题2模网络图谱

三、结论与展望

通过对 2000–2018 年统计学专业研究高频参考文献、载文机构、高频作者和高频关键词进行可视化分析，得到以下结论：

（一）高频参考文献

高频参考文献主要还是统计学主流期刊和专著，根据引用量最多的是由陈希孺编著的《数理统计学简史》，这本书是 2002 年湖南教育出版社出版的一本图书。此书自身可读价值高，不仅对于本专业本科生或研究生甚至对统计学专业感兴趣的其他人士的适用性很强。其次就是对于大数据的研究以及一些概率论方面的文章都是主要参考文章。可见，对于统计学的研究学者们对于统计的基础内容都是很看重的，所以想了解相关内容的人士可以选择一些相对基础的书本或者是文献进行初步学习。另外本书作者陈希孺从事数理统计学的基础理论研究工作，在参数估计、非参数统计、线性回归和统计大样本理论方面有一些成果，在国内外有一定的影响，其成果曾获国家和中科院的自然科学奖。

（二）载文机构

载文机构主要集中在我国主流统计期刊，其中从发文量可以看出来《统计研究》不仅是中国最具影响力的学术期刊，还是整个统计学载文机构的领军期刊。从学科性质来看主要还是数学学科占据大头。了解核心期刊具有重要的意义。核心期刊集学术性、综合性、前瞻性为一体，载文质量高，被看作是学科研究的重要参考。就编者而言，可以从核心期刊吸取经验。就读者而言，树立核心期刊意识，可以明确价值取向，提高阅读档次。期刊的状况是与一个国家的市场经济发展状况相联系的，我国的市场经济目前处在一个高速发展的阶段，市场正趋于成熟，但同其相联系，我们的期刊市场秩序则需要提高期刊的有效阅读量，使得期刊市场能有效地发挥市场机制的作用。另一方面，我国的期刊虽然有八千多种，但是办得好的并不多。许多党政部门、事业单位、企业等都有自己的期刊，但是办刊专业人员缺乏、财政支持力度不大，有些期刊仅靠收取作者的版面费生存。我国的期刊产业有很大的发展空间。中国的市场很大，还有很多潜在的读者群可供开发。我们对报纸、电视、

互联网的研究已经很多，但是对期刊关注的还不够。其实，我们对期刊的评判标准应当有所改变，不一定单看发行量或者是期刊的影响因子，期刊杂志的文章深度才是最重要的。

（三）论文作者

论文作者是推动学科发展的力量。以贺铿、张汉斌为代表的高频作者为统计学研究做出了卓越贡献，但是贺铿等 9 人的合作并不是很密切。故而激发学者们之间的合作，以此促进更多学者开展学术的合作。在当前这个快速发展的网络信息时代，便捷的交流方式不该被忽视。对于高校的师生而言，更是要打破高校的地域界限，开展跨地域和跨高校的交流合作，这将是培养后备人才的必然之路，后起之秀也必节节高。为我国统计行业的多维度、多方向的发展起到领军带头作用，让这一学科的研究不仅只局限在片面的讨论，或者是某一领域的超越发展，而是可以和更多领域的权威人士相互交流合作，使得研究不止在书面领域快速的发展，而且可以为普通人的学习生活提供参考和启发，从而达到更有深度的发展，真正达到让学术的成果可以服务人民、造福人民的目的。不然一切的研究惠及不到大众，所有的高谈阔论都只是“纸上谈兵”。

（四）研究热点

研究热点是指在某一时间段内，有内在联系的、数量相对较多的一组论文所探讨的专题。通过图示可以看到现阶段对于统计学的研究主要集中在学科和方法，更多的是对于社会科学和大数据的研究，在医学领域的研究较少，其他领域几乎没有。作为信息时代，数据是我们最容易获得的，对于数据的整理分析可以帮助到很多方面的预测和决策，比如我们的机器学习软件通过样本数据的采集可以很好的预测被测数据的结果，甚至在现代医学的各项疾病诊断预测等方面有广泛的应用。此外，对于教育行业甚至很多电商工业的高端运行都是基于大数据的研究。因此我觉得有必要学者们可以适当的将视角延伸至其他学科，实现各个学科的融会贯通。譬如教育行业可有针对性的分析各个学科或者对于年级的不同用统计学大数据统计分析的手段更好的制定学习计划，实现真正的“因材施教”。

参考文献：

[1]贺铿.关于统计学的性质与发展问题[J].统计研究,2001(9):3-7.

[2]姜春林,等.《中国科技期刊研究》研究热点及其演进知识图谱[J].中国科技期刊研究,2008(6):954-958.

[3]陈为等.数据可视化的基本原理与方法[M].北京:科学出版社,2013.11.

[4]刘军.整体网分析讲义:Ucinet软件实用指南[M].上海:格致出版社,2009.

[5]李兴绪,等.SPSS经济统计分析[M].北京:中国统计出版社,2008.

[6]Gephi Makes Graphs Handy[EB/OL].https://gephi.github.io/features/2015-03-05.

[7]李文兰,杨祖国.从关键词的变化看中国图书馆学研究主题的发展[J].图书情报工作,2004 (12):115-118.

[8]张松.基于词频g指数的共词聚类关键词选取研究[J].现代教育技术,2013(10):53-57.