第四届全球技术挖掘与第十九届科学技术指标国际会议参会报告

（Report of 4th Annual Global TechMining Conference & 19th international Conference on Science and Technology Indicators）

胡正银

一、会议概况

第四届全球技术挖掘(4th Annual Global TechMining Conference, 4th GTM)与第十九届科学技术指标国际会议(19th international Conference on Science and Technology Indicators, 19th STI) 这两个会议于 2014 年 9 月 2 日-9 月 5 日在荷兰莱顿大学联合召开。其中，4th GTM 会议由美国佐治亚理工学院技术政策与评估中心(Technology Policy and Assessment Center, TPAC)与技术挖掘论坛(VP Institute)举办，时间是 9 月 2 日。19th STI 会议在欧洲指标开发者网络(European Network of Indicator Developers, ENID)的支持下，由莱顿大学科技研究中心(Centre for Science and Technology Studies, CWTS)主办，时间是 9 月 3 日-9 月 5 日。

首届 GTM 会议于 2011 年 9 月在美国亚特兰大举行，该会议每年举行一次，会议时间均为一天。会议每次都选择与亚特兰大科学与创新政策会议(Atlanta Conference On Science and Innovation Policy)或 STI 会议联合举行。会议目标是：在利用科学、技术、商业文献进行技术挖掘的情报分析专家、软件专家及学科领域专家之间建立起跨学科沟通的平台。本次会议共收到 50 多篇论文，参会者来自美国、中

国、巴西、俄罗斯、荷兰、西班牙、芬兰等多个国家。会议正式讨论部分共分为：方法研究(Development in Methods)、主题模型(Topic Modeling)、专利(Patents)、科技创新合作影响(Impacts of Collaboration in S,T&I)、新数据(Novel Data)、新方法(Novel Methods)、科技演化评价(Evaluating the Evolution of S&T)多个 Session 与专利映射(Patent Mapping)一个 Workshop。会议针对技术挖掘新方法、新数据及专利技术挖掘等进行了重点研讨。此外，本次会议除了传统的 Poster 展示环节外，还设计了一个每人 5 分钟的 Power Talks 环节，尽可能为参会者提供多边交流的平台。

本次 GTM 会议的主题发言人是科技战略公司(SciTech Strategies, Inc.)总裁 Kevin Boyak 博士，报告主题是："Methods, Metrics, and Missions that Matter"。Kevin 博士用生动形象的方式，对本次 GTM 会议投稿文章进行了分析，发现大部分文章讨论的焦点是 Methods 或 Metrics。Kevin 博士则认为技术挖掘是一项任务驱动的科学或应用，单纯的 Methods，Metrics 研究应该转向到以 Mission 为导向的研究上来，应该采用终端用户听的懂、能理解的语言与方式进行技术挖掘研究，这样技术挖掘才能有更好的发展空间与未来。本次 GTM 会议，相对前三次，无论是参会人数、还是会议内容都丰富很多。更多信息请参见(http://www.gtmconference.org/pages/program.html)。

STI 会议是科技指标领域一个历史悠久的会议，专注于科技指标设计、应用及其评估，今年是第 19 届。本次会议的主题是："Context Counts: Pathway to master big and little data"。本次会议共分为：

“Exploring Citation Impact, Research Evaluation, Social Sciences and Humanities, Behavior of Scientists, Usage Data and Citations, Studies of Funding Systems, Innovation Studies, Careers and Trajectories, University-Industry Collaboration, Research Teams, Altmetrics and Applications, Gender Studies, New Algorithms, Research Evaluation Methods, Data Sources, Patent Analysis, Evaluation of Individuals, Disciplines and Interdisciplinary, Field Analysis, Citation Analysis Normalization”等 20 多个 Session。在 Session 期间，会议还以专家访谈的形式，就诺干议题设计了 Special Event，如：“Access to Data, Quality Standards for Evaluation Indicators: Any Chance for the Dream to Come True?”等。

主题发言(Keynote)是会议内容与主旨的风向标。本届 STI 会议主题发言相当丰富，总共设计了四个 Keynote 环节。第一个 Keynote 由丹麦 University of Copenhagen 的 Peter Mahler-Larsen 教授做题目为“Constitutive Effects of Performance Indicator Systems”的演讲。Peter 教授从多个方面分析了影响指标性能的因素，其中特别提到了“同行评议(Peer Review)”也存在广泛且不可克服的弊端，认为指标的动态性是保证指标长久有效的重要因素之一。第二个 Keynote 是 STW 技术基金会执行主席 Eppo Bruins 博士作“Quality Indicators for the Applied and Technological Sciences”演讲。Eppo 博士从 STW 实践出发，阐述了针对应用与技术领域的质量指标选取、设计原则与实践。第三个 Keynote 是 Elsevier 信息计量学研究小组的 Henk Moed 博士作

“Metrics-Based Research Assessment”演讲。Henk 博士从 Metrics 的潜力与局限、多维度的研究评估与大数据环境下信息计量三个角度阐述了主题，令人印象十分深刻。佐治亚理工学院 Diana Hicks 教授作了题为“Grand Challenges”的主题演讲。Diana 教授从更广阔的视角给出了信息计量面临的巨大挑战，如：“AltMetrics, Gender Studies, Individual Level Evaluation, Standards”等。总体来说，本次 STI 会议内容丰富，日程安排紧凑，信息含量相当大。更详细信息可以在 STI 官方网站上找到(http://sti2014.cwts.nl/Program)。

二、参会活动

2.1 4th GTM 会议

我与院中心王小梅老师、兰州中心马建霞老师一起参加 4th GTM 会议，并作了“Automatic classification of patents oriented to problems and solutions: a case study on large aperture optical elements”与“Exploring Potential Patent Portfolios: An Integrated Approach Based on Topic Identification and Correlation Analysis”两个口头报告。

第一个报告“Automatic classification of patents oriented to problems and solutions: a case study on large aperture optical elements”被安排在“Topic Modeling” Session，是紧接着主题演讲之后的第一个口头报告。报告完毕后，提问踊跃。在 Alan Porter 教授引导性的提问之后，荷兰代尔夫特理工大学的 Scott 教授，TRIZ 专家西班牙瓦伦西亚理工大学 Gomila 教授，芬兰 VTT 的资深科学家 Arho 博士等分别就 Topic Model 的矩阵构成；相比传统 TRIZ，个性化分类体系的优点；

如何自动构建中观 Problems & Solutions 层；该方法的应用范围”等感兴趣的问题进行了提问与交流。该 Session 中，另外两个报告分别是芬兰 VTT 的 Arho 博士所作的“Garbage in, garbage out: Impact of sequence matching based text cleaning and phrase identification on unsupervised text mining”与兰州中心马建霞老师所作的“Emerging Topic Detection based on LDA Combined with Bibliometrics Indices”。Arho 博士是 VTT“Innovations, Economy, and Policy”组的技术负责人，主要从事面向科学家的技术创新情报分析。Arho 博士的论文重点介绍了通过对文本语料库进行前预处理与后验分析，来提高语义文本处理的准确率，避免“Garbage in”与“Garbage out”。Arho 博士的研究具有较强的实践意义，在他的研究中，将自然语言处理、LDA 主题模型、对象识别融为一体，能较好的关键语义单位的识别率与辨识度。马建霞老师的论文结合 LDA 主题模型与文献计量学指标(novelty index, published volume index, cited volume index)来对 emerging 主题进行早期探测。通过对“Machine Learning”数据集的实证研究，该方法可以较好的进行领域 emerging 主题的探测。该方法可以广泛应用到爆发性技术主题预测、话题发展趋势等领域。

第二个报告“Exploring Potential Patent Portfolios: An Integrated Approach Based on Topic Identification and Correlation Analysis”被安排在“Novel Methods” Session，由我代替张娴进行大会口头报告与交流。报告结束后，反响热烈，效果很好。俄罗斯科学院的 Dorodnicyn 计算中心 Vladimir 教授针对“如何确定主题的标签”进行了提问与交流。

芬兰 VTT 的 Arho 博士则具体咨询了本报告分析的数据是否只包括中科院数据。在得到肯定的答复后，Arho 博士建议可以考虑将除中科院之外其他类似机构的数据也纳入分析范围，并认为该报告分析思路值得借鉴与推广。该 Session 另外两个报告“Peaks, Slopes, Canyons, Plateaus: Identifying Technology Trends throughout the Life Cycle”与“Future of Sustainable Military Operations under Emerging Energy and Security Considerations”都是来自俄罗斯国家研究大学统计研究与经济知识所(Institute for Statistical Studies and Economics of Knowledge，National Research University)。前一个报告针对现有自动挖掘技术趋势(Technology Trend, TT)方法中存在一些固有的缺陷，如：结果过于宽泛或无价值；难以识别处于弱信号阶段的技术趋势等，提出了一个基于“黑盒(black box)”原则的技术趋势预测方法。该方法采用技术趋势本体与技术趋势指标，基于 Gartner’s Hype Cycle 来构建模型，通过识别趋势指标与本体启发式推理相结合来解读技术发展趋势。论文采用了绿色能源领域进行实证研究，认为该方法具有较好的领域无关性与通用性。该报告得到了与会者热烈回应，个人认为值得深入研究。后一个报告，则是基于场景的技术挖掘研究。论文将场景分析与技术挖掘结合起来。场景不同，需求各异，技术挖掘关注的焦点、指标、方法都会有差异。论文以军事活动对能源的依赖性为例，分析了在不同能源供应场景下，军事活动的发展历程。在此基础上，分析了新兴能源与安全边界的场景下，重大军事活动未来可能趋势。

除了作报告 Session 之外，我们还参与了“Patent”，“Evaluating the

Evolution of S&T”与“Patent Mapping”三部分会议内容。其中，重点关注的报告包括：瓦伦西亚理工大学 Gomila 教授的结合 Semantic TRIZ 与 TechMining 分析技术的系统演化、Alan Porter 教授的集成路线图，专利与文献来发现新兴技术、北京理工大学周潇博士的采用文本挖掘方法衡量技术创新路径、荷兰 Rathenau 研究所 Edwin 博士的从多维视角可视化专利组合的方法及英国知识产权局 Peter Evans 博士关于对专利数据错误解读常见情况分析。“Patent Mapping” workshop 环节介绍了一种基于跨引用模式(Cross-Citation Patterns)，对 IPC 层次模式进行解聚(Disaggregate)与重新聚集(re-aggregate)的专利映射模式。在该 workshop 环节针对“Nano-Enabled Drug Delivery”专利具体映射效果进行了开放式讨论。在实际专利技术挖掘中，不同知识组织体系之间的“Patent Mapping”经常涉及，本次会议讨论的虽然只是 IPC 层级之间的 Mapping，但其映射思路可以借鉴应用到其它方面。

GTM 会议期间，除了与院中心王小梅老师，兰州中心马建霞老师外，我还与佐治亚理工的 Alan Porter 教授、Jan Youtie 教授、TRIZ 专家 Gomila 教授、VTT 的 Arho 博士、北京理工大学周潇博士、同济大学刘玉仙博士等进行了不同程度的交流。其中，Gomila 教授与 Arho 博士都表示了希望能就双方感兴趣的研究课题，进一步深入合作的意愿。

2.2 19th STI 会议

本次 STI 会议日程安排相当紧凑，内容翔实。除了 Keynote 与 Special Events 环节外，我重点参加了“Exploring Citation Impact”，

“Usage Data and Citations”，“Innovation Studies”与“Altmetrics Applications”等 Session 及 CitNetExplorer 特邀报告环节。印象特别深刻的是 Altmetrics。Altmetrics 在国外研究已经如火如荼，但对我而言是一个全新的方向。刘玉仙博士认为相比于其他的 metrics, altmetrics 的实质是把科研过程所产出的数据都考虑进来的计量学；国际科学计量学与信息计量学学会(ISSI)主席罗纳德·鲁索认为 altmetrics 这新指标尚未落地，属于软指标，易受操控、主观性强[1]。汤姆逊路透的专家在本届 STI 会议上作了“Measuring Emerging Scientific Impact and Current Trends: A Comparison of Altmetric and Hot Papers Indicators”的报告，报告中给出了 Altmetrics 框架，并将其与引证指标进行了对比分析。在“Altmetrics Applications”Session 环节，Zohreh 博士等基于 Mendeley 读者关系进行了广泛的 altmetrics 分析。Peter Kraker 等则基于 altmetrics 对领域知识演化可视化描述进行了研究。另外还有学者进行了 altmetrics 是否存在性别偏向等研究。

另一个重点关注的会议内容是基于文献 download 的计量分析。基于文献 download 的计量分析是近几年才出现的一个新的研究方向。文献 download 量与其 citation 之间，存在一定的关联性，但也有较大的区别。Wolfgang 博士在其论文“Cross-national preferences and similarities in downloads and citations of scientific articles: A pilot study”中，从实证角度出发，系统分析了科研论文 download 量与 citation 之间关联性，结果表明两者呈现良好的正相关性。Gali Halevi 博士在其

[1]翟自洋.由信息计量学新词 altmetrics 的翻译想到的.
[EB/OL].(2013)[2014-09-04].http://blog.sciencenet.cn/blog-630081-679433.html

论文“Usage patterns of scientific journals and their relationship to citations”中系统比较了 download 指标与 citation 指标之间的异同。

会议以特邀报告的形式了介绍了莱顿大学 CWTS 开发的大尺度科技文献引证关系可视化分析软件 CitNetExplorer。该软件支持百万量级文献、千万量级引证关系量级数据分析。在数据方面，CitNetExplorer 支持 Web of Science 数据的直接导入，支持 Pajek 数据格式导出。在可视化方面，除了支持引证网络放大缩小外，还支持非直接引证关系可视化。在分析方面，CitNetExplorer 也有较大改进。如支持文献数据交互式裁剪、引证网络直观拓展分析、大量支持发现核心文献集、文献路径的算法。个人感觉 CitNetExplorer 是一款非常好用、扩展性极强的文献可视化分析软件。

此外，在 STI2014 会议上还介绍了一个 STI 会前 workshop 主要内容。这个 workshop 在 9 月 2 日召开，主要讨论的议题包括：“计量指标的归一化(Normalization of Bibliometric Indicators)、引证影响力指标(Indicators of Citation Impact)、引证网络权重与字段归一化(Weighting Citation Networks and Field-Normalization)、期刊影响因子指标(Journal Impact Indicators) 、指标比较(Indicator Comparison)、技术因子(Technological Factor) 、计量分析中的统计推理(Statistical Inference in Bibliometric Analysis)”等。根据个人兴趣，我重点关注了“Statistical Inference in Bibliometric Analysis”与“Technological Factor”，并与相关专家进行了交流。

由于 STI 会议内容比较分散，很多 Session 对我来说，都是全新

的议题。因此，在 STI 会议期间，我以参会聆听、学习为主，交流较少。主要与院中心杨立英老师、王小梅老师、兰州中心马建霞老师、佐治亚理工 Jan Youtie 教授、同济大学刘玉仙博士、莱顿大学 CWTS 的 Nees 博士(CitNetExplorer 的主要开发人员)等就参会感受与心得等进行了交流。

三、心得感受

本届 GTM 与 STI 会议涵盖技术挖掘、科技指标领域的诸多方面内容。通过参会，向国际同行展示了我们在相关领域的研究工作，了解了国际同行在上述各方面的研究前沿和进展。在联系老朋友、结识新朋友的同时，也就未来可能的合作空间与方向进行了交流与探讨。总体来说，受益匪浅。

具体而言，有以下五个方面心得与感受：

1、技术挖掘应该面向任务：SciTech Strategies 公司总裁 Kevin Boyak 博士在 GTM 会议上作了“Methods, Metrics, and Missions that Matter”的主题报告。在报告中，Kevin 博士认为技术挖掘是一项任务驱动的科学或应用，应该采用终端用户听的懂、能理解的语言与方式进行技术挖掘研究。对 Kevin 博士的观点，我深表赞同。技术挖掘目的：面向**特定的用户**，根据**特定的任务**，挖据**特定的数据**，得出**特定的结论**。从本质上来说，TechMining 就是任务驱动的。挖掘数据的方法，分析数据的指标都应该是围绕特定任务服务的。从这个意思上来说，“TechMining is an art as well as science”。因此，在我们今后的工作中，应该根据任务来挑选方法、指标及模型；没有最好的方法、指

标及模型，只有最适合我们任务的方法、指标及模型。

2、技术挖掘相关概念有待厘清与统一：在国内学术界，谈到技术挖掘时，对框架(Framework)、模型(Model)、方法(Method)、算法(Algorithm)等概念没有明确的区分与界定，有时甚至混为一谈。在本次会议期间，通过与相关专家进行交流，梳理了国外同行对这些概念的定义与理解，供国内同行探讨。首先，Framework 是指通用的框架与流程，它从较宏观的层面定义了技术挖掘工作标准化的流程。在每一个流程环节，只涉及抽象方法或步骤，不涉及具体的技术与算法。其次，Model 是针对特定技术挖掘任务，基于 Framework 定义的个性化分析方案。对一个技术挖掘模型来说，一般包括数据集(Data Set)、算法(Algorithm)、参数配置(Parameter Configuration)等内容。而 Method 是一个较通用的概念，既可以是 Framework 中定义的抽象方法，也可以是 Model 中具体的细化技术方法。再次，Algorithm 是模型中采用的具体方法。最后，针对技术挖掘不同角度，会有不同的 Model。对于一个具体的企业级技术挖掘应用来说，一般会从政策、经济、技术、环境等角度进行技术挖掘。这些不同的 Model 共同组成一个 Enterprise Panel。

3、语义 TRIZ 与专利映射是中、微观层面技术挖掘的关键技术：中观、微观层面的专利技术挖掘更多关注专利具体技术层面信息。语义 TRIZ 是在专利日益膨胀的背景下，对传统 TRIZ 的语义扩展。它基于 TRIZ 思想，采用现代语义技术，运用数据挖掘模型，对专利中隐性技术信息进行语义知识表示。专利映射则是基于特定标准，在不

同知识组织体系之间，对专利信息进行知识重组。专利映射有助于打通专利、科技文献、商业信息之间连接的通道，在更广泛维度发现专利技术的价值。基于语义 TRIZ 与专利映射的技术挖掘，能深化专利技术挖掘程度、拓展专利技术挖掘范围、更新专利技术挖掘视角，是专利技术挖掘热点。

4、科技指标目的究竟是什么？：在 STI 会议的主题报告演讲中，丹麦哥本哈根大学 Peter 教授从多个方面分析了影响指标性能的因素，特别提到了及时是“同行评议(Peer Review)”也存在广泛且不可克服的弊端，认为指标的动态性是保证指标长久有效的重要因素之一。在其后的多个报告及 Special Event 讨论中，多个专家都反复强调：“Indicators are used to mine information, not to judge”。对科技指标的目的讨论，终究是智者见智，仁者见仁。如果说指标完全不应该具备评价(judge)功能，那么个人认为这个学科的价值或应用前景也要大打折扣。但是科技指标不能滥用，不能简单机械的使用，则是广大同行的共识。

5、科学计量学逐渐走出象牙塔：在国外，科学计量学不再简单是学界在象牙塔中研究的对象，而出现了像 STW 这样专门的基于科技计量的信息分析服务公司。STI 会议出版 Daily Issue 更是以“Bibliometricist Leading Science in the Right Direction”为题进行报道本次 STI 会议。总体感觉，至少在欧洲，科学计量学的应用前景十分广阔。