网上科技信息深度分析方法及应用①

丛培民

(中国科学院 计算机网络信息中心，北京 100864)

摘 要：对网上科技信息进行专业化聚合，实现定制分享和态势分析，营造有利于知识发现的研究环境，已成为知识创新和科技管理所面对的重要课题。因此，在分析科技工作者对科技领域公共信息搜索需求的基础上，借鉴行业搜索模式，设计运用垂直搜索引擎等技术手段聚合网上科技信息资源，构建科技信息资源库和共享平台。探讨为科技工作者提供网上科技资源搜索以及科技信息深度分析服务的方法，为知识创新服务。

关键词：信息聚合；知识发现；数据挖掘；信息共享

Depth Analysis Method of Internet Scientific Information

CONG Pei-Min

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100864, China)

Abstract: To professional polymerization the internet technology information and realize public data sharing, and to create knowledge-benefit environment through internet S&T information trend analyzing, has become a subject that scientific management and knowledge innovation must face. So, based on analyzing the scientific workers' demand for public information search in S&T area, draw lessons from industry search mode, design and apply vertical search engine technology polymerizes internet technology information resources, constructs S&T information resources database and network public data sharing platform. It also explores and discusses, through the establishment of S&T information network situation analysis, contributing to the knowledge innovation.

Key words: resource syndication; knowledge discovery; data mining; infomation sharing

互联网快速发展，知识积累越来越容易、学术交流越来越便捷、数据分析越来越深入、信息利用越来越广泛，从而使科技工作者在网上获取科学思想的需求变得越来越迫切，使科技管理者在网上发现科技动态的需求变得越来越重要。其态势表明，信息共享越充分，信息积累就越丰富，科学思想就越活跃，知识发现就越有可能基于互联网。聚合公共科技信息资源，推进科技信息共享，研究科技信息深度分析方法，促进知识发现和科学发展，则是值得深入探讨的课题。

1 互联网环境下的知识创新面临挑战

1.1 信息共享已上升为国家科技发展战略

国际社会对公共领域科研数据共享的重视由来已久且日益迫切。很多国际组织在近十几年来，都要求成员国做出对公共领域科学研究数据“完全与公开”政策的承诺[1]。欧美一些国家很早就意识到数据共享平台建设对推动科学发展的重要性，并积极推动相关方面的研究。2011 年 2 月 11 日出版的《科学》杂志在社论中指出：数据推动科学发展[2]，则深入阐述了信息共享在科技发展中的重要作用。我国科技界和科技管理部门也认识到科学数据共享与提高科技资源利用的重要性。自 1994 年起，孙枢院士等科学家就开始呼吁科学数据共享[3]。但是，由于相关制度建设相对滞后，法律法规还不健全，使得“由于缺乏健全的数据管理制度，项目结束后，科学数据、资料和相关信息或依然处于离散分布而丢失损毁”[4]。因此，在《国家中

① 收稿时间:2011-08-08;收到修改稿时间:2011-09-08

长期科学和技术发展规划纲要（2006-2020 年）》指导下，科技部联合有关机构，启动了国家科技基础条件平台建设，在国家层面上进行了有益探索。

1.2 互联网信息分析已开始在社会管理方面显露优势

互联网世界日益成为公共信息和舆论的形成与传播空间，公共机构对于与其利益攸关的网络信息需求也日益凸现。一些部门运用网络搜索等相关技术，开展了一系列的网络舆情分析研究与服务，对于提升社会管理水平产生了积极作用。也有一些机构开展了专门领域的舆情分析，比如国家质检总局每日食品安全的网络舆情报告，也成为管理决策的重要依据。

1.3 在互联网信息中发现知识正在成为一种可能

如今，科技工作者表达科学观点、传播科学思想通过网络变得越来越容易。科技管理者利用网络搜集意见和建议也成为管理创新不可或缺的反馈机制。一方面，各类科研计划的制订及其管理实施的合理性必须建立在科技工作者的普遍性意见和建议之上；另一方面，各部门和机构可以通过网络及时准确地了解本机构和本系统的外界评价以及国内外相关竞争性机构的最新动向。网络无疑是了解这些信息的最有效和快捷的渠道，但如何有效搜集和分析相关网络舆情并高效反馈，同时对舆论的方向建立起一种符合中国国情的引导机制则成为科技管理者关注的问题。对科技工作者而言，在这其中，不乏新的科学思想闪光，以及人的知识，智慧，经验，技能通过互联网转换成实际收益的互联网新模式[5]。通过信息聚合以及深入分析方法发现知识则成为可能。

2 信息搜索技术应用状况分析

2.1 公共信息搜索及行业搜索服务状况

以谷歌和百度为代表的互联网信息搜索服务商的出现，帮助大众初步解决了如何从海量信息中查找所需信息之类的需求。但是，其服务切入点侧重于政治、经济、文化、生活等大众需求，对于特定行业需求的垂直型搜索功能相对薄弱，相关服务滞后或效率很低，大大妨碍了行业类信息的传播和使用[6]。

谷歌针对行业类信息的垂直搜索功能开发虽有一定的投入，如 Google Scholar 满足了中初级科技人员对于科技信息的即时查询，并和大型科技信息数据库形成了业务互动。但随着谷歌公司自 2010 年以来在中国业务的变故，包含 Google Scholar 在内的搜索服务功能变得非常不稳定，访问抵达率降低。目前，百度在中国搜索市场份额已趋于垄断地位。但受竞价排名规则的局限，百度搜索结果的客观性受到影响[7]。随着行业垂直搜索需求的出现，目前已出现了一些成功案例。如电子商务领域的淘宝网搜索。可以预见，随着互联网的发展，垂直搜索在网络信息检索中的地位日渐重要。垂直搜索将会更加流行，同时对网络生活的方方面面也将产生更为深刻的影响[8]。

2.2 科技信息深度搜索与分析方法依然空白

相比之下，面向科技领域公共信息深度搜索服务，特别是中文科技信息，目前处于相对空白的状态。在百度、谷歌提供的搜索服务中，科技领域公共信息往往被淹没在数以千计的搜索结果中，难以满足广大科技工作者对科技类信息进行聚合和筛选的特定需求。究其原因，其一，系由于此类服务很难产生商业上的广告回报，商业资本不愿意进入此领域，其公益属性导致了此类服务一直阙如。其二，此类信息的价值取决于专业化程度，专业化程度越高对搜索技术和服务的要求也越高，极具挑战。

3 科技信息深度搜索技术及其应用研究

3.1 以建立信息共享平台为研究基础

建立一个专业化的科学信息采集与信息共享平台，进行全面、系统、标准化、实时性的数据积累，并将结果和数据开放，使数据分享过程简单化、数据利用率最大化。通过专业化信息聚合，提供信息搜索、共享服务。从元数据和元知识机制及其自然产生机制出发，在科研过程中的信念产生、概念形成、工具选择、策略采用、团队形成、网络构建、技术动向以及社会文化背景等要素的基础上，对正在形成中的科技知识进行动态分析，帮助科技工作者把握学科发展和知识创新的可能方向，使其形成对科技信息的交互、动态与精细的加工能力和趋势分析，从而为寻找创新生长点、打破学科界限、开辟新的研究路径乃至动态调整知识创新发展战略和方向提供切实有效的支持。

主要研究内容有：

① 科技信息搜索引擎：依托垂直搜索引擎等技术手段聚合互联网中海量科技信息资源，根据受众需求对采集数据进行专业化分类处理，建立科技信息资源目录体系，形成标准化的信息资源索引，使其满足用户的个性化需求并能动态反映科技的最新进展。

② 科技信息源深度探查：通过互动服务实现科技思想的碰撞，研究深入探查信息的技术方法，建立信息的可靠性、准确性、动态性标准，并对信息的所有权、使用权和管理机制进行初步探索。

③ 网络科技信息深度分析与反馈机制：建立对科技类网站及论坛、博客、微博等信息态势的深度分析与反馈机制。通过对突发事件舆情采集、敏感事件舆情监控、失真报道与网络谣言监控、科学道德舆情监控、违法违规行为监控等个性化网络配置，发布科技舆情分析及舆论引导建议，为科研管理与决策服务。

网络信息不同于传统的书刊及论文，很多信息并非成果的最终呈现，而更多地反映了正在进行中的科研细节和尚不成熟的知识，其公共性或产权属性也较为复杂。在方法论上寻求创新，如引入交互本体论[9]、数据可视化和知识管理中的交互式规划等方法，为知识发现与分析服务。

3.2 以提供科技信息深度分析为主要目标

以建立科技领域垂直搜索引擎和公共信息数据库，并提供相关的科技信息态势及舆情分析为目标，实现对分散在互联网中各类科技资源进行深度发现、采集、加工和再发布，为快速实现科技资源的聚合和共享奠定基础。特别是在数据资源的处理方面，自动采集结合专业人员整理、干预和加工，形成科技信息的多维分类、深度标引。既能够向用户提供一键式信息搜索，又能够提供用户个性化的资源导航、学科导航、科学专题、热点聚焦、前沿探索等服务，还能使包括本学科、交叉学科和跨学科等最广泛的研究领域的学者受益。

通过网络科技信息深度分析与反馈机制旨在改进科技管理、营造创新文化。一方面，通过科技界的信息互动促进科技管理的科学化，提高科技管理决策的民主程度；另一方面，使公众对科学及科学家、科学家对科学热点、年轻一代科技工作者对科学及科学家等的认知普遍提高，在全社会和科技界营造一种有利于创新的科学文化。

3.3 科技信息深度搜索和分析平台逻辑架构设计

按照上述目标，科技信息深度搜索与分析平台逻辑架构如图 1 所示。

3.4 深度搜索的关键技术方法

① 大规模分布式信息采集技术：应用大规模分布式信息采集技术，重点对索引页的识别、采集调度、维护等方面的算法进行研究和配置，提高检索效率。同时，根据运行情况提高相关技术的标准化程度，并使之具有更强的适应性。

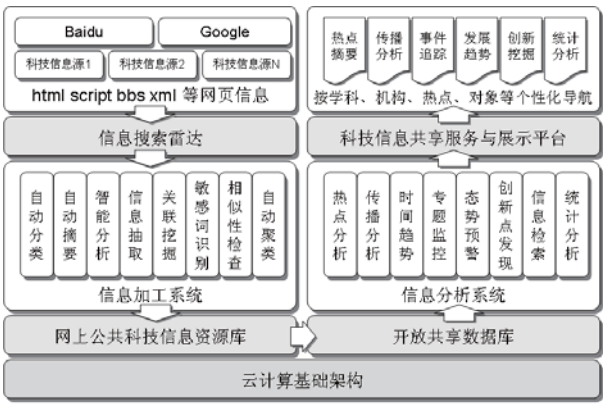

图 1　科技信息深度搜索与分析平台逻辑架构图

② Deep Web 采集技术：采用两种技术路线实现对 Deep Web 有效、健壮的采集，一是采用脚本引擎技术，使采集工具能够从使用 JavaScript 的网页中成功解析出链接和内容等信息；二是采用链接分析控制技术，控制采集工具对于深层网络的收集轨迹和策略。

③ 自动分类等智能处理技术：专业化的科技信息搜索平台是专门针对科技资源搜索需求定制的，信息分类导航、学科导航等更加细化。对信息聚合优化策略包括利用主题词、同义词、近义词作为自动分类的基础。同时，也将利用元数据和元知识理论研究的最新成果，从知识创新的自然形成的维度进行优化搜索，采取人机结合、默会知识与可表征知识结合[10]、专家直觉与数据挖掘结合、学科内与交叉学科和跨学科结合等启发式优化策略。

3.5 平台集成的技术功能

① 信息搜索智能化：根据信息资源类型、格式及服务对象进行多级导航，方便用户选择浏览和搜索；提供个性化订阅和推送，满足高端用户的个性化需求，推送方式采用在线、邮件等方式；提供智能检索，提供丰富的检索组合表达式解析；提供结果分析聚合，自动支持检索结果的分类统计和浏览，显示检索结果的分布情况,告诉各类别的命中记录数等。同时，及时搜集所提供服务与用户需求的关联度和用户的满意度，通过实时动态反馈机制改进搜索模式和服务品质。

② 信息加工自动化：对搜索到的网页进行自动分析和标引，包括信息内容提取、格式自动转换、属性自动标引、内码自动转换，采用元数据自动提取技术，

分析并标注信息属性。对信息进行自动过滤、自动分类、自动摘要、自动排重、自动聚类等。

③ 系统平台虚拟化：在云环境中搭建系统平台，系统资源集中管理、动态扩充。实现智能管理；数据采用分布采集，可灵活扩展；系统资源集中管理，在应用层实现 web 远程状态查询及任务执行，方便灵活。

3.6 数据积累与服务的技术方法

① 运用数据采集系统实现资源聚合：实现对各种信息资源的自动采集，包括文档、图片、视频、动画等各种格式的文献、成果、专利、报告、观点、看法等科技资源。采集系统具有快速采集能力，支持各种网站发布形式和资源格式，并对采集内容进行基本的过滤和属性抽取，对非结构化的网络数据进行元数据化，存储到数据库中。

② 运用数字资源加工系统进行预处理：数字资源加工系统充分利用文本挖掘技术，实现对文本内容的智能挖掘和分析，同时形成专业人员对数据资源检查、标引等深度加工及质量控制的技术环境和分工协作体系。通过智能数据加工技术和人工专业整理相结合，对采集的信息资源进行按学科、格式、服务对象等角度的多维分类、排重、关联等加工，以形成深度立体的资源整合，提供个性化的资源搜索和导航服务。预处理后的信息资源，则更加有利于共享和利用。

③ 运用海量存储及检索引擎为数据共享服务：网络资源搜索服务具有海量数据存储能力和动态、灵活的扩展能力。对于以非结构化数据为主要形态和服务形式的科技资源，其存储和管理采用专业的目录索引数据库系统。对于图片、音视频等数据对象的存储和服务，形成本地存储与源网络地址相结合的形式，为用户提供服务。

④ 智能化搜索服务和立体化资源展现：在资源采集、整合及加工基础上，不但提供一键式搜索服务，而且根据用户类型对科技资源的需求，按学科、机构、新闻热点、服务对象、资源格式、资源类型等进行个性化智能导航，信息搜索服务需要形成立体、多维展示和关联，形成具有科技行业垂直搜索引擎的服务特点和应用价值。

⑤网络舆情分析与舆论引导咨询服务：在上述工作基础上，辅以相应的人工干预，通过网络舆情形成机制分析，运用网络说服技术的最新成果，根据用户或机构的个性需求，分别提供信息定制服务和网络舆情分析和舆论引导咨询服务。

4 可行性分析

4.1 网络环境的普及性

据 CNNIC《第 28 次中国互联网络发展状况统计报告》披露，截至 2011 年 6 月底，中国网民已突破 4.8 亿,普及率持续提高。通过网络查找科技资源和开展研究，已经是科技工作者所具有的基本能力和研究方法，也是科技工作者整理资料、开展研究和提高自身科学水平的手段。

4.2 科技信息发布和服务的广泛性

科学数据资源库建设和网站建设已经在我国广大科研院所普及，各种科技信息资源等以网络形式广泛存在，这为科技信息资源获取创造了条件。

4.3 搜索引擎技术的成熟性

当前，搜索引擎技术和应用相对成熟。除几大互联网搜索引擎向用户提供日常搜索服务外，还有众多的行业搜索引擎，如新华网搜索、人民网搜索、国家门户网站的政府信息资源搜索引擎等。从搜索内容看，已经实现了对新闻、文献、地图、结构化数据等各类格式内容的搜索服务。为深度搜索平台建设奠定基础。

4.4 知识产权的推广性

互联网信息本身即具备公共性。科学家在互联网上提供的各种信息均是可传播的；科研机构建立的各种科技信息资源库也具备公共性。在数据共享平台中聚合的此类信息，在很大程度上是对其知识产权的一种推广，从而在更广泛的环境中让需求者获知，给知识产权的推广带来效益。

5 结语

综上所述，建立科技信息深度分析平台，从技术上是可行的。其创新点在于：第一、运用垂直搜索引擎技术，实现科技信息资源的聚合，促进分散的科技信息资源优化重组，拓宽科技信息资源开发利用空间，帮助科研工作者提高信息获取的效率。第二、发挥多学科交叉渗透和跨学科的优势，促进学科的建设与发展。学科发展需要长时间积累，特别是学术思想的形成，学科方向的凝练和人才梯队的组建。第三、保证和提高科研质量是科学发展永

(下转第 249 页)

Bayesian filtering, Dept. Eng., Univ. Cambridge, UK, Tech.Rep, 1998.

15 Comaniciu D. Ramesh V. Meer P. Kernel–based object tracking, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(5):564–577.

16 Hager GD. Dewan M, Stewart CV. Multiple kernel tracking with SSD. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

17 Shi J, Tomasi C. Good features to track. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1994: 593–600.

18 Collins TR, Liu Y, Leordeanu M. Online selection of discriminative tracking features. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(10):1631–1643.

19 Grabner H, Bischof H. On–line boosting and vision. Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition, CVPR 2006. 2006.

20 Liu R, Cheng J, Lu HQ. A Robust Boosting Tracker with Minimum Error Bound in a Co-Training Framework. ICCV 2009. 2009.

21 Levin A, Viola P, Freund Y. Unsupervised improvement of visual detectors using co-training. IEEE International Conference on Computer Vision (ICCV). 2003: 626–633.

22 Breiman L. Random Forests. Machine Learning, 2001, 45(1):5–32.

23 Saffari A, Leistner C, Santner J, et al. On-line Random Forests. 3rd IEEE ICCV Workshop on On-line Computer Vision. 2009.

24 Leistner C, Godec M, Saffari A, et al. Online Multi-View Forests for Tracking. Proc. of Symposium of the German Association for Pattern Recognition (DAGM). 2010.

25 Sato K, Aggarwal J. Temporal spatio-velocity transform and its application to tracking and interaction. Comput. Vision Image Understand. 2004,96(2):100–128.

26 Chen Y, Rui Y, Huang T. Jpdaf based hmm for realtime contour tracking. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2001: 543–550.

27 Bertalmio M,. Sapiro G, Randall G. Morphing active contours. IEEE Trans. Patt. Analy. Mach. Intell., 2000,22(7): 33–737.

(上接第 239 页)

恒的主题。通过科技信息资源的聚合，为学科交融、形成优势以及充分利用科技信息资源提供了前提条件。第四、我国科技资源依然存在着配置不均衡的现象，通过科技改革和资源流动等方法，需要长时间逐步改进，而通过搜索引擎技术实现的科技资源深度搜索，则面向广大用户提供各取所需的快捷服务，能够促进科技资源的均衡分配。

参考文献

1 孙鸿烈.认清科学数据的战略地位.光明日报,2004.3.12.

2 Hanson B. Making Data Maximally Available, Science, Feb. 11,2011

3 刘闯.我国科学数据共享机制建设研究.国土资源信息化. 2004,1.

4 张景勇.科技部:我国科学数据亟需改变“单兵作战”现状.新华网,2003.2.6.

5 刘锋,张玲玲,顾基发.知识管理在互联网中的应用.ISKSS, December 2007,4(4).

6 王文钧,李巍.垂直搜索引擎的现状与发展探究.情报科学, 2010,(3):477–480.

7 互联网实验室:互联网行业垄断调查及对策研究报告.(2010).2011.

8 郑凯明,李义杰.垂直搜索引擎及其应用价值.信息技术, 2008,(4):45–47.

9 袁昱明,施建华,陶培根.本体在网络学习主客体构建和交互中的应用.中国远程教育,2009,(12).

10 叶德营.默会知识与程序性知识的比较研究[硕士学位论文].杭州:浙江大学,2008.