新型搜索引擎畅想

黄建年[①②]

（①南京财经大学图书馆　南京 210046）（②南京农业大学人文学院 南京 210095）

【摘要】网络搜索需求增加，导致了搜索引擎的大量涌现。论文认为在网络世界将会出现九种新型的搜索引擎，它们分别是零次文献搜索引擎、潜在文献搜索引擎、知识发现搜索引擎、大型元搜索聚类引擎、专业学术型聚类引擎、学术趋势搜索引擎、概念类比联想搜索引擎、解疑答难型搜索引擎、教学研究平台搜索引擎。

【关键词】搜索引擎；聚类；潜在文献；学术趋势；知识发现；概念类比；零次文献

【中图分类号】G354

Reflections on New Search Engine

Huang Jian-nian[③④]

(③Nanjing University of Finance and Economics, Nanjing 210046, China)

(④Nanjing Agricultural University, Nanjing 210095, China)

【Abstract】Quick increment of need on internet information resources leads to a rush of search engines. This article introduces some new type of search engines which is appearing and will appear. These search engines includes as follows: grey document search engine, invisible web search engine, knowledge discovery search engine, clustering meta search engine, academic clustering search engine, conception comparison and conception analogy search engine, consultation search engine, teaching and studying search engine.

【Keywords】search engine; clustering; invisible web; academic trend; knowledge discovery; concept comparison; grey document

0 引言

笔者曾于2000年就搜索工具的发展趋向作过初步研究，提出了九大发展趋势[1]，时至今日，其中的绝大部分已经或者正在变为现实。可是，网络发展日新月异，网络信息的迅速增加，导致了网络搜索需求的不断扩大，同时也带来了搜索引擎品种的不断增加和新型搜索引擎的不断涌现。这些搜索引擎的发展会出现怎样的新征候和新特征？对图书馆工作、情报工作有怎样的启示？带着这些问题，畅游于网络世界之中，意图通过探寻网络世界中的蛛丝蚂迹，以求找出新一代搜索引擎的典型特征和可能出现的趋向，从而为图书馆学情报学的发展探索一种新的思路，也期望为传统图书馆学情报学理论在网络世界的回归、应用与再现树立信心。经过认真的研究与谨慎的探析，笔者认为下列搜索引擎的出现不仅不可避免，而且还会出现数量众多的变种。

1 零次文献搜索引擎

何谓零次文献？维客[2]认为，零次文献是一种特殊形式的信息源，主要包括两个方面的内容：

- ► 形成一次文献以前的知识信息，即未经记录、未形成文字材料，是人们的“出你之口，入我之耳”的口头交谈；
- ► 是未经正式发表的原始文献，或未正式出版的各种书刊资料，如书信、手稿、记

录、笔记和包括一些内部使用通过公开正式的订购途径所不能获得的书刊资料。

零次文献一般是通过口头交谈、参观展览、参加报告会等途径获取，不仅在内容上有一定的价值，而且能弥补一般公开文献从信息的客观形成到公开传播之间费时甚多的弊病。

在网络世界中，这样的定义显然有些老化。实际上，现在零次文献的范围在扩展，博客、播客、视客、论坛上的资料纷纷进入人的视野，成为人们参考利用的资料，具有典型的零次文献特征。如何搜集和利用这些资料将成为搜索引擎研究和发展的重要内容。从其历史来观察，这类搜索引擎也一直在发展和壮大，从最初的论坛搜索引擎，到现在的博客搜索引擎，无论是技术还是内容均出现了很大的变化。最近又出现播客（PODCAST）和视客(Video Blog)搜索引擎，也就是分别对音频、视频资料进行采集与加工，并提供利用的专用搜索引擎。另外，一些大型综合搜索引擎也在逐步增加有关内容的搜索和开发，也纷纷推出了这些文献的内容，比如百度博客搜索[3]、新浪博客搜索[4]、爱问博客搜索[5]等等。

这类搜索引擎的发展的出现，无疑使我们对网络上出现的大量零次文献有了更加全面的了解。但是，网络世界新生事物不穷，所以可以预期这类搜索引擎的发展应该是最热闹和最有意思的。没有了零次文献，大约网络的诱惑力会大大降低。所以，作为新一代搜索引擎，零次文献搜索引擎可能会向两个方面发展：1）向全面搜集某一类零次文献方向的专题型搜索引擎发展；2）向完全新型的零次文献搜索引擎方向发展。只要网络存在，新型的零次文献就会不断产出，这种搜索引擎就会发展。所以，此类搜索引擎的发展应该作为研究的重点。又因为此类搜索引擎往往采用最新的技术来运行和操作，因而，此类引擎又往往会成为新搜索技术利用的试验场，与此同时这些搜索引擎也会提出新的搜索理论和方法，关注此类引擎也很有必要。

2 潜在文献搜索引擎

Web 网按其分布状况可以分为“表层网”（Surface Web）和“深层网”（Deep Web，也有称 Invisible Web，Hidden Web）。目前对不可见网页（Invisible Web）的研究已经越来越为广大研究者所注意，但是还没有突破性的技术和方法出现。所以有关此类资料的搜索和加工，就成这此类搜索引擎的重点任务。

据 Bright Planet 公司技术白皮书《The Deep Web-Surfacing the Hidden Value》，Deep Web 资源容量约为 Surface Web 的 500 倍。该白皮书指出，当时最大的搜索引擎只索引了 Surface Web 中的 16％信息量，而如果算上那些无法被传统搜索引擎索引的 Deep Web 中的信息，那么一般搜索引擎只能搜索 0.03％的 Web 信息。可见，研究和挖掘 Deep Web 对于提高搜索覆盖率和准确率有着非常重要的意义。[6]据研究全球网页大约 1000 亿网页，而收录最多的 GOOGL 也只收录了 120 亿网页，可见已经被收录的网页只是冰山的一角。尽管不同的来源其数据量略有不同，但是，未知信息数量巨大应该无庸置疑。如何挖掘未知信息、未收录信息应该成为搜索引擎下一步研究的重点。

据一份调查（IDG.NET）显示，现在网络上非文本信息与文本信息的比例是 3:1[7]，也就是说互联网上有近 75%的信息是以非文本格式存放的，而目前搜索引擎收录和处理的主要内容是文本信息。换句话说，75%左右的非文本格式的信息处理得不是很到位，或者说没有处理。

不仅如此，即使对文本信息的处理搜索引擎也存在不小的问题。目前搜索引擎的强项主要在于对静态网页的处理，而对动态网页处理能力则相对较差。比如搜索引擎对动态网页处理时出现的黑洞问题，就引起了很多搜索引擎公司的注意，也加大了研究的力度。但是，遗憾的是目前世界上还没有一家主要的搜索引擎宣称完全支持动态网页，因为大

多数负责搜索网页的蜘蛛软件不敢去碰动态网页，怕被变化无穷的动态系统黑洞吸进去出不来。然而，网站使用动态网页生成工具乃是大势所趋，ASP、PHP、JSP 等编程工具日益流行，解决动态网页查找的问题已经是人心所向[8]。百度公司日前号称支持动态文本处理，实际上也仅能处理部分相对不变的动态网页，而对于哪些网址不断变化的网页（比如 VOD 系统为了防止下载采用的动态网址）基本上还是无能为力。所以，未来关于这类搜索引擎的研究应该会有一个较大的突破。

3 知识发现搜索引擎

搜索引擎到底是以查全（RECALL）为目标，还是以发现（DISCOVERY）为目标？这一问题的准确定位对于搜索引擎的发展有相对大的指导作用。国外有多位学者关注这一问题的研究，如C.L.Bernier、H.H.Wellisch等人，德国化学信息专家Robert Fugmann先生详细回顾了这一历程[9]，明确将搜索引擎分为以RECALL型搜索引擎和DISCOVERY型搜索引擎，这样的划分非常有意义。

传统的第一代和第二代搜索引擎主要是以提供各类资料为其最高目标，也就是是以 RECALL 为目标。而随着网络信息的增加，搜索引擎提供的结果往往成千上万，已经超出了用户的处理能力，所以用户希望搜索引擎能够提供更加精确的结果，并且帮助用户发现未知信息。有基于此，未来的搜索引擎应该更多的以提供用户新知识和新信息为主，它不仅在于提供用户已知的消息，关键是要向用户提供未知的知识和资料，将搜索过程转变成发现之旅。

发现之旅也许对用户个人拥有的信息的处理显得更加有意义。目前桌面型搜索发展速度十分迅速，既反映了用户处理自身拥有信息的愿望，也说明了搜索引擎功能和市场的细化过程。所以未来几年同样应该是桌面型搜索引擎发扬光大的时代，搜索引擎厂商应该更多地关心用户自身的需求，帮助用户提高自己的认识水平和发现新知的能力。

此类搜索引擎，主要可以采用 2 种知识组织方法：

- 树形知识体系表达法，就是通过知识树来表现用户的资料和认知结构。换句话说，就是提供知识树，来发现知识树的生长合理与否，从而找出已有知识体系中的空白点。
- 可视图形表达法，可以借鉴 VIVISIMO 提供的知识图来实现这一目标。所有重要的概念和术语通过知识图来形象展示知识体系，从而通过概念或者术语的粒度发现用户自己的重点研究领域和未来的知识增长点，有针对性进行相关研究。

无论是哪种方式，均应该在用户已知信息和未知信息之间做出差别性的标记（比如通过颜色、字体等方式表示），以方便和提醒用户新知识和新资料的发现。

4 大型元搜索聚类引擎

网页数量的迅速增加使得用户对搜索结果的选择成为一种负担，因此，整理搜索结果，提高其准确性成为一种时代的呼声。顺应这一潮流，目前出现了很多基于普通搜索引擎的动态聚类搜索引擎，如 Vivisimo[10]、Kartoo[11]、Clusty[12]、BBMAO[13]等。一般来说，这些搜索引擎本身没有自己的数据库，而对根据用户需求临时从普通搜索引擎中抽取相关资料，然后通过特定的聚类算法，对搜索结果进行聚类后再提交给用户。

目前此类搜索引擎的规模还很小，还没有成为传统搜索引擎的竞争对手，一方面是其本身的依附性质，决定了此类搜索引擎很难形成自己独立的地位和优势。另一方面，此类搜索引擎对于传统搜索引擎的使用又增强了传统搜索引擎的强势地位。再次，聚类搜索引擎目前所能够采用的聚类技术和算法还存在种种弊端，对于搜索结果的聚类也还存在若干不准确、不切实际的因素，所以，仍然有待于进一步完善。

因为采用的文本聚类算法相对来说比较简单,所以目前大多数聚类搜索引擎以数量有限的传统搜索引擎（通常在 5 个以下）为其资料来源，这样能够保证聚类搜索引擎响应时间较短的特点。但是，正如前文所说，即使是收录最全的搜索引擎也只收录网页资源的极小部分，所以，将若干搜索引擎的资源进行资源整合后，再进行聚类就很有必要。

将来聚类搜索引擎可以在以下几个方面得到发展：1）探索新的高级聚类算法；2）采用分布式并发聚类方法；3）研究网页去重技术，很有可能要借鉴传统图书馆学的去重理论；4）基于 XML 的元数据技术的大规模应用，类似于传统图书馆学实践中的在版编目（CIP），由网页供应商自动进行网页元数据标注，以加快网页加工和聚类的速度。

5 专业学术型聚类引擎

专业数据库越来越多，也得到了越来越多的关注。一般情况下，传统的专业数据库主要采用分类主题法来组织相关资料，但是，随着数据库文献数量的增加，检索结果的大量出现已经影响了用户的使用。为了解决这一问题，目前一些厂商研究和推出了专业型聚类搜索引擎，自动对结果进行聚类，比如中国期刊网就提供了这样的功能，在检索结果页面，提供对结果的自动聚类，默认为 6 个，如果需要更多的结果可以点击“更多”按钮，出现更多的聚类结果。笔者利用“自动聚类”作为关键词，在检索结果页面出现 6 个类目。如果点击“更多”，出现如下 20 个类目：聚类(177)、聚类分析(115)、聚类算法(81)、模糊聚类(65)数据挖掘(53)、自动分类(40)、算法(32)、神经网络(32)、自动文摘(23)、模糊聚类分析(23)、文本聚类(21)、向量空间模型(21)、图像分割(20)、特征提取(19)、模式识别(19)、自动识别(19)、自动标引(19)、聚类中心(18)、文本分类(17)、搜索引擎(15)[14]。笔者通过分别点击相关类目，发现相关性程度较高。经过仔细分析，笔者发现，准确的原因主要在于 CNKI 拥有强大的期刊论文库，也就是充分利用了作者对每篇论文的聚类成果：篇名、文摘和关键词三个方面。

此类搜索引擎因为依托于自身所拥有的庞大数据库,所以,可以结合原有的归类方法,进行适当的聚类，实现分类与聚类的完善结合，充分发挥专业数据库的专业学术性质。

目前此类搜索引擎采用专业术语，以面向学术用户为主，而普通用户使用起来还存在一定的难度，笔者以为，此类搜索引擎将会更多地利用知识组织系统（KOS）的成果,检索时实现自然语言与受控语言并用，从而使专业数据库更进一步面向普通用户。

6 学术趋势搜索引擎

了解所关注学科的最新进展，关注有关基金项目的情况简报，跟踪学术同行的学术动态，从而了解学术发展的最新趋向，这是每一位学者必须关注的内容之一。目前在这方面,一些搜索引擎进行了一些努力,如 CNKI 推出了学术趋势搜索[15]、Google 推出了 Google Scholar 搜索[16]等，但是，这样的举措目前仍然只是一些尝试，离真正的实用还有很大的距离。目前此类搜索引擎的功能仅仅找到专业的学术资料，比如各类期刊论文、课件等的使用情况，更多地基于计量学、统计学的成果，离自动聚类找出学术趋势，或者说提供有关学术趋势的内容还比较少见。将来这类引擎主要功能应该更多地着眼于综合或者归纳功能的应用，这种应用与传统文献中的综述有些类似，更多地具有人工智能的成份。

将来的发展可以从以下几个方面取得进展：1）对学术会议信息与相关会议文献的搜集与加工；2）利用文献计量学、网络计量学的理论对结果进行可视化呈现；3）按地区和时间等因素进行分析和研究某一学科或者领域的进展，提供最新学术研究动向；4）按行业、按系统、按行政区划进行研究，了解各系统、各地区、各行业的学术动向。

7 概念类比联想搜索引擎

在学术研究中概念的类比和联想往往带来科学研究的突破。如何分析和研究其中的机制和规律，也应该是未来搜索引擎的重要内容。同一概念在不同的学科中应用不同，其成熟度也不尽相同，了解其它领域的应用方式和方法往往会给我们很多启示。通过不同领域概念以及相关概念的对比，既可以及时发现不同领域的共同规律，也可以了解本领域的发展趋向。在科技史上，物质与能量守恒定律是由 7 个国家的 10 多位不同学科的科学家同时独立地发现，说明科学的共生性，也说明了科学的移植性和可借鉴性。所以，利用概念类比原理制作的搜索引擎无疑也会受到更多的关注。

初步设想如下：由用户指定学科，然后提出检索词，分别进行相关检索，返回检索结果后，再进行可视化显示，分别进行对照比较。至于分析对照的学科目前原则上不宜太多，最好控制在 4 个以内，这样能够保持较好的响应速度。当然随着计算机与网络技术的发展，学科数量可能适当增加。但是，考虑到单一用户所熟悉的学科领域终究有限，能够同时熟悉 4 个以上学科的专家亦为数不多，所以目前的建议仍然以 4 个以内的学科为主。

实际上，这种搜索引擎的设计并不困难，可以充分利用现有可视化搜索引擎的成果，比如 VIVISIMO[17]、CLUSTY[18]等，只不过增加相应的学科范畴，使之更加细化、更加专业而已。如果这样的搜索引擎早日出现，则将极大地提高科学研究的主动性和针对性。

8 解疑答难型搜索引擎

解疑答难是日常生活中不可或缺的一部分内容，所以，对此类搜索引擎的需求也是与日剧增，比如百度知道[19]一面世就受到了世人的关注，该系统目前可以解决的问题总数已经达到 14，201，196 条，为常见问题的解答提供了一个新的解决途径。但是，这样的解决方案仍然完全是民间的，完全来源于网络。换句话说，现有的此类搜索系统，更多地依托网民，也就是答案的权威性如何还有待考证，专业性的学术研究人员对于利用这样的系统仍然心存疑虑。

而 ABOUT.COM[20]这样的网站则相对专业化，基本采用了学科组织体系，是一个百科全书，类似于《十万个为什么》。但是，这一网站因为没有充分利用互联网提供的巨大信息源，所以，有大量的问题（比如计算机故障等）又无法找到答案。如何实现专家型系统与业余型解疑答难系统的集成，应该成为未来几年解疑答难型搜索引擎重点关注的领域。

实际上，解疑答难型搜索引擎，现在还出现了一些新的变化，它不仅能够解答知识型问题，而且还对日常生活（比如找工作、找房子、旅游等）中各种问题进行解答。为了满足某种搜索需求，还出现了专家在线型的搜索引擎，实时解答用户问题。

目前的做法有两种：1）通过专家系统，也即知识库来解答用户的提问，类似于图书馆工作的 FAQ 系统；2）通过在线专家回答相关问题，角色和作用类似于图书馆的咨询馆员。一般来说，大众型的搜索引擎多采用第一种方式，实行免费服务，而专业学术型的解疑答难型搜索引擎则以采用第 2 种方式为主，实行有偿服务。

9 教学研究平台搜索引擎

目前的搜索引擎主要仍然以大众生活信息和学术研究信息为主要研究对象，而对于教学型信息的搜索、利用以及教学平台的建设则仍然付之阙如。此类信息目前主要集中在各种 BBS 之中，但是，还没有形成相对独立的体系，更没有形成以专门的学科体系来组织，所以，此类信息的搜索有相当的挖掘价值。

此类搜索引擎可以形成一个系列，比如课件搜索、案例搜索、教案搜索、学习工具搜索、常见参考书搜索等等。可以预见，未来几年这种搜索引擎会成为新的增长点。

搜索引擎与研究平台的集成系统的研制也应该成为重点。用户利用搜索引擎返回的结果，可以自动形成特定的参考文献格式，从而成为自己的专用数据库，换句话说，形成自己的个人学术研究搜索引擎，以备进一步的研究和学习。

EBSCO 等数据库提供了参考文献的直接导出功能[21]，这样的功能应该来说是初级功能，高级功能应该是用户如果直接复制相关资料可以直接生成参考文献，作为脚注或者尾注附在用户研究论文之中。这种做法，既可以部分解决论文中的假注和伪注情形，也可以减少人工生成参考文献过程中可能出现的误差。

【参考文献】

【1】黄建年.网络搜索工具旳发展趋向[J].图书情报工作，2000，(2)：34－36

【2】零次文献 [EB/OL] .[2007－04－10].http://www.wiki.cn/w/index.php?title=%E9%9B%B6%E6%AC%A1%E6%96%87%E7%8C%AE&redirect=no

【3】百度博客搜索 [EB/OL] .[2007－04－10].http://blogsearch.baidu.com/

【4】新浪博客搜索 [EB/OL] .[2007－04－10].http://search.blog.sina.com.cn

【5】爱问博客搜索 [EB/OL] .[2007－04－10].http://blog.iask.com/

【6】深层网络：揭示网络中隐藏的价值 [EB/OL] .[2007－04－10].http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf

【7】信息发现与 Invisible-Web[EB/OL].[2007－04－10].http://blog.loverty.org/ 2004/02/invisible-web.html

【8】认识中文搜索引擎的十大误区 [EB/OL] .[2007－04－10].http://www.canco-soft.com/include/ShowDetail.asp?series_id=339

【9】Fugmann, Robert. Learning the Lessons of the Past [EB/OL] .[2007－04－10].http://www.chemheritage.org/events/asist2002/14-fugmann.pdf

【10】VIVISIMO [EB/OL] .[2007－04－10].http://www.vivisimo.com/

【11】KARTOO [EB/OL] .[2007－04－10].http://www.kartoo.com

【12】CLUSTY [EB/OL] .[2007－04－10].http://www.clusty.com

【13】BBMAO [EB/OL] .[2007－04－10].http://bbmao.com/

【14】CNKI.CNKI 知识搜索 [EB/OL] .[2007－04－10].http://search.cnki.net/SearchResult.aspx?searchword=%E8%87%AA%E5%8A%A8%E8%81%9A%E7%B1%BB

【15】CNKI 学术趋势 [EB/OL] .[2007－04－10].http://trend.cnki.net/

【16】Google Scholar [EB/OL] .[2007－04－10].http://www.lib.tju.edu.cn/resource/database/probation/200512200002.htm

【17】VIVISIMO [EB/OL] .[2007－04－10].http://www.vivisimo.com/

【18】CLUSTY [EB/OL] .[2007－04－10].http://www.clusty.com

【19】百度知道 [EB/OL] .[2007－04－10].http://zhidao.baidu.com/

【20】ABOUT.COM [EB/OL] .[2007－04－10].http://about.com/

【21】直接导出 [EB/OL] .[2007－04－10].http://web.ebscohost.com/ehost/newfeatures?vid=4&hid=17&sid=aba596d0-a1f4-4bd6-99db-bbc632a095ba%40sessionmgr2