元搜索引擎的发展悖论及建议

王 雁 杰

(北京大学信息管理系 北京 100871)

摘 要 通过对部分有代表性的元搜索引擎和独立搜索引擎的对比研究,发现元搜索引擎的返回结果在数量和覆盖范围上远比独立搜索引擎小,进而提出元搜索引擎面临着一个"发展悖论",最后给出元搜索引擎走出困境的方法建议。

关键词 元搜索引擎 独立搜索引擎 发展悖论

1 问题的提出

元搜索引擎是在独立搜索引擎的基础上建立起来的可以同时或分时查询多个搜索引擎(含独立搜索引擎或其他元搜索引擎)的网络信息查询系统。其英文 Meta - search Engine 意为普通搜索引擎之后或之上的搜索引擎。

元搜索引擎的关键技术环节主要包括用户请求的识别转化、成员搜索引擎的选择以及结果的汇总输出。首先须将用户的检索需求转化成为成员搜索引擎能够识别的格式,然后发放到适当的成员搜索引擎,这样可以有效避免成员引擎的不相关调用所引起的系统负担以及有效降低对结果筛选的难度。对成员搜索引擎的选择主要有三种方式:粗略描述的方法、统计的方法和基于学习的方法。在汇总输出过程中,元搜索引擎往往通过种种算法(例如加权、全局相关性、用户自由选择、基于学习的方法等等)限制从各成员搜索引擎的返回结果数量,并采取一系列诸如去重、聚类以及多种排序等技术对结果进行后处理。

定位于解决独立搜索引擎的不足,元搜索引擎产生于上世纪 90 年代末。据统计,当前任何一个独立搜索引擎对网络信息的覆盖率都小于 1/3,而且在所谓的"信息时代",信息膨胀的速度远远超过了搜索引擎检索范围的扩张速度,比如 1997 年的统计数字表明,最大的搜索引擎对网络信息覆盖面接近 30%,到 1999 年这一数字已降至 16%,检索范围相对狭小的问题日益明显。元搜索引擎由于其可以调用多个独立搜索引擎以及能够处理、利用独立搜索引擎返回的结果,所以人们把对更大范围网络资源进行检索的期望落在元搜索引擎上,这基本上是元搜索引擎产生的初衷。但事实上,在同样检索请求下,元搜索引擎检索返回的结果数量比单独使用其子独立搜索引擎返回的数量少得多,而且通过对结果的一一验证,发现其结果覆盖范围也较其子独立搜索引擎小很多,不妨看一组对当前一些著名元搜索引擎与独立搜索引擎的对比研究得出的数据(见表 1)。

表 1 一些著名元搜索引擎与独立搜索引擎的对比

表格	检索词	Arnold Schwarzenegger	information management	search engine comparison	Olympic Games	Institutional change
元搜索引擎	Dogpile	97	118	90	94	89
	Profusion	34	39	19	40	22
	Metacrawler	101	133	97	99	88
独立搜索引擎	Google	1020000	13100000	2590000	2920000	3870000
	Infoseek	341000	6060000	1090000	1100000	1420000
	百度	10400	756000	1890	19100	4630

注:表格中的数字为对应检索词的返回结果数量。上述数据于 2004 年 3 月 20 日上午取得。

从表 1 中可以发现两个事实:a. 经元搜索引擎检索得出的结果数量远远小于独立搜索引擎;b. 对应于不同的检索请求,独立搜索引擎结果数量的变动幅度远大于元搜索引擎(通过比较结果数量的方差可得),即,元搜索引擎的返回结果数量较为稳定。

造成这样结果的原因是:元搜索引擎对于从各独立搜索引擎中返回的结果进行严格的数量限制,由于所含有的子独立搜索引擎数目固定,所以结果数量较小,而且变化幅度也不大。对多个元搜索引擎的实证研究表明:元搜索引擎从每个子独立搜索引擎接纳数量很小的结果,从而导致最终检索结果数量远远低于独立搜索引擎,例如,Dogpile 只接受 5～10 条从各独立搜索引擎中返回的结果。检索结果数量差距如此之大没有理由使人相信元搜索引擎能够覆盖较大范围的网络资源。

这样就产生了一个悖论:既然元搜索引擎不能解决独立搜索引擎检索范围狭小的难题,那么元搜索引擎的存在又有何意义呢?

2 元搜索引擎的困境

上述悖论的内生逻辑不难理解:在当前的技术条件下,元搜索引擎往往面临着哈姆雷特式的两难困境:一方面,若不严格限制从各独立搜索引擎中返回的数量,由于从各独立搜索引擎返回的结果中有很多重复的网页,加总后数量更是极其庞大,那么在后处理(去重、排序、聚类等等)时,系统将面临重大的负担,增加了检索时间;另一方面,若是采用限制从独立引擎中返回的数量的方法,由于目前元搜索最为切实可行的抽取方式就是从各独立搜索引擎检索结果的前几条抽取(因为经过独立搜索引擎的排序,前几条往往比较相关),而如果各独立搜索引擎技术比较成熟的话,那么对于一个话题,其前几条往往有很多是相同的,再经过元搜索引擎的去重,结果所剩无几。最终,元搜索必须对这两种情况进行权衡,但其最终数量终将远远低于独立搜索引擎的结果数,范围也往往较小(因为普遍来说只是覆盖了各独立搜索引擎的前几条结果的范围),但这显然违反了元搜索建立的初衷。在这种情况下,元搜索引擎应如何改变其弱势地位呢?

当前对搜索引擎的改进主要有创建新型网络结构体系、改变搜索引擎工作模式、运用传统信息组织与信息检索技术和优化检索结果后处理技术四种方式。但出于难度上的考虑,目前搜索引擎行业又将重点放在优化检索结果后处理技术上(因为前三种涉及基本模式的改变,属于较深层次的技术跃迁),因此当前几乎所有搜索引擎的前三种条件(网络结构体系、搜索引擎工作模式和信息组织与信息检索技术)都基本相同,所以目前衡量搜索引擎质量好坏的主要

指标就是检索结果的后处理技术。

对比几年前后不同的网页也得出类似的结果：目前元搜索引擎技术的主要改进基本上定位于对结果的后处理，比如 Dogpile 新增了按照相关度进行排列，增添了对检索结果的自动聚类。但这样的改进并不涉及元搜索引擎与普通独立搜索引擎的本质区别，随之而来便产生了新的疑问：元搜索引擎既然做好了后处理，为什么不变成独立搜索引擎呢？理由如下：元搜索引擎与独立搜索引擎相比，元搜索没有自己的 robot 和没有建立自己的 URL4，但需要将用户的检索式经过一定的转化发放到适合的引擎中，也需要对各引擎返回的结果进行整合。元搜索引擎处理检索词和整合检索结果的成本往往可能大于制作 robot，而且这种难度随着独立搜索引擎的发展而愈演愈烈(显而易见，当独立搜索引擎功能单一，比如只有关键词检索时，元搜索只需按图索骥将用户检索式直接提交到各独立搜索引擎即可，但当独立搜索引擎支持布尔逻辑检索或是分字段进行检索时，那么转化检索式进行提交的难度便大大增加)。所以对于那些后处理技术做得较为成功的元搜索引擎来说，是否会考虑转变为一个独立搜索引擎呢？比如说 Ask Jeeves 从前本是业内著名的元搜索引擎，但自从收购了一家独立搜索引擎(Teoma)后，摇身一变成为独立搜索引擎的杰出代表(收购自然可以获得 robot 等独立搜索引擎的技术)。

3　摆脱困境的可能途径

以上关于困境的讨论基于对目前现状的一点观察，即：现在不管是元搜索引擎还是独立搜索引擎，几乎所有的搜索引擎都是综合性搜索引擎，因此对于用户来说，这些搜索引擎的差异并不明显，换句话说就是搜索引擎都在争夺同样的用户群，所以在讨论中才能大胆地假定成本的减小就一定使利润增加(前文论证了元搜索引擎处理检索词和整合检索结果的成本往往高于制作 Robot)，如果用户群有所不同，那么元搜索引擎转变为独立搜索引擎就会失去原来自身的用户群，有可能使得收益下降，此时即使成本下降，但利润的高低也未可知，元搜索引擎未必存在动力转向独立搜索引擎。不妨设想，当网络搜索引擎有了较为明显的分化，尤其是大量专科性搜索引擎的出现，元搜索引擎作为综合性的搜索引擎，其存在是有重要意义的。

利用博弈论的知识进行一下简单的分析：假设这样一个极端状态：市场中所有的独立搜索引擎都被元搜索引擎所使用，那么如果有一家独立搜索引擎退出的话，那么对于用户来说，由于其登录元搜索引擎界面的成本与独立搜索引擎相若，而元搜索引擎集成了多个独立搜索引擎，对比仅仅一家独立搜索引擎，用户有理由认为元搜索引擎返回的结果远远优于单独一家独立搜索引擎的结果，所以会优先选择登录元搜索引擎，对于该独立搜索引擎来说，脱离就意味着丧失用户，因此最优的战略就是依然被元搜索引擎所使用，这种战略对于所有的独立搜索引擎都适用，于是这便是一个纳什均衡(即元搜索引擎和成员独立搜索引擎都没有意愿脱离这个状态)，这个状态是较为稳定的。在这个稳定的状态下，一方面独立搜索引擎专业化加深了数据挖掘的程度，另一方面因为不同独立搜索引擎返回结果重复较少，减轻了元搜索引擎系统整合结果的负担，有可能挖掘较大范围的网络信息资源，这个稳态是网络信息资源组织发展到较高阶段的体现，因此从全局利益来看，元搜索引擎主导专业化独立搜索引擎的状态是一个较好的状态。

以上给出了理想状态存在性的说明，但并没有说明如何到达这种理想状态。一种可能的解释是：由于网络资源高速增长，使得任何一个独立搜索引擎都没有能力检索网络上的全部资源，而且由于行业竞争的加剧，一些搜索引擎便会进行重新的市场定位，这样的定位便产生了用户群的分化，使得元搜索引擎逐渐变得较为有优势，而元搜索引擎的机理又注定它是要做综合性搜索引擎的，于是在激烈的市场竞争中此消彼长，原有的一些综合性的独立搜索引擎又会在市场中重新定位，产生专业化，又会产生新一轮元搜索引擎增长高潮，慢慢的如上文所说：对于用户来说，其登录元搜索引擎界面的成本与独立搜索引擎相若，但由于元搜索引擎集成了多个独立搜索引擎，对比一家独立的搜索引擎，用户对使用元搜索引擎效益期望远远大于那一家独立的搜索引擎，因此不会去登录那家独立搜索引擎，对于此独立搜索引擎来说，用户再次逐渐减少，如此循环便可以实现元搜索引擎主导专业化独立搜索引擎的状态。

胡誉耀给出了元搜索引擎的另一种发展途径：元搜索引擎一般没有自己独立的数据库，而数字图书馆虽有属于自己的独立数据库，但就整体而言，它是局部性、区域性的，甚至很多都是专业化的，而且只能是海量数据库的一个极小构成部分。元搜索引擎技术可以将这些分布式的局部数字图书馆进行有效的整合，且能够在满足用户需求的情况下保持原来各局部图书馆特色优势。元搜索引擎与数字图书馆的结合不仅能大大优化数字图书馆的信息查询与搜索功能，同时也会成为元搜索引擎发展的重要推动力，这同样是一个"双赢"的局面。

另外，比较各个元搜索引擎和独立搜索引擎可以发现元搜索引擎大都集成了许多的功能，例如 Dogpile 的 Web page，White page 和 Yellow page，而独立搜索引擎往往仅仅是比较纯粹的检索工具，例如 Baidu 和 Google。元搜索引擎变得越来越像综合性的信息服务中介，这是目前比较明显的一个发展方向。

4　结　语

元搜索引擎作为新兴出现的搜索引擎，其作为整体与独立搜索引擎相比依然处于相对劣势，其前景依然扑朔迷离，当前国外的元搜索引擎都努力地尝试延伸到不同的领域(例如 Dogpile 的 White page 和 Yellow page)，无论结果成功与否，只有通过这样不断的尝试，不断地进行重新定位和调整，元搜索引擎才能有较好的定位。本文论证了存在一个元搜索引擎占主导地位的稳定状态，但是这种状态的实现很大程度上取决于今后网络整体发展的趋势，而且整个分析过程是一个极其简化的论述，其间忽略了复杂的技术变迁等等，我们都在拭目以待元搜索引擎如何走上这样一条通向理想状态的"康庄大道"。

参 考 文 献

1　胡誉耀．元搜索引擎在数字图书馆中的运用．图书与情报，2003；(5)
2　王　铮，胡永杰．元搜索引擎的设计与实现．河北师范大学学报(自然科学版)，2001；(2)
3　朱茂盛，王　冠，程学旗．元搜索引擎及其实现．计算机工程，2002；(11)
4　张健奕．搜索引擎的新发展——元搜索引擎．河南图书馆学刊，2002；(2)
5　李广建，黄　崑．元搜索引擎及其主要技术．情报科学，2002；(2)
6　刘海凯，黄碧云，张　畅．元搜索引擎 Profusion．情报科学，2002；(9)
7　张俭恭，陈定权，吴振新．关于搜索引擎与元搜索引擎的讨论．现代图书情报技术，2002；(2)
8　宋玉兰，杨高波．Internet 元搜索引擎评析．宁波高等专科学校学报，1999；(2)
9　刘柏嵩，韩惠琴．支持用户信息需求的新一代元搜索引擎研究．图书情报工作，2002；(4)
10　李　明．中文元搜索引擎万纬搜索研究．现代图书情报技术，2003；(5)
11　楼松亭，张惠惠．中文电子期刊的元搜索引擎．情报科学，2003；(11)

(责编：勃王京)