搜索引擎特性分析

张爽　李彭城　张硕　（北京理工大学计算机科学与工程系　100081）

摘要：本文分析阐述了搜索引擎共有特性及各自不同特点。通过分析、综合几个主流引擎的共性和特性，揭示了搜索引擎内在本质。使用户对搜索引擎的工作原理有一更深刻的认识，澄清误解，并可针对其不同特性恰当的选择适于自己目标的搜索引擎。

关键词：搜索引擎　搜索工具

本文目的是分析阐述搜索引擎共有的特性，并通过介绍几个主流搜索引擎在这些特性方面的不同特点，使您对搜索引擎的工作原理有一个更清楚的认识，并澄清对搜索引擎的一些误解和含糊之处。

一、搜索引擎共性分析

1. 拥有的网页数量

不论是搜索者还是网络管理员，其最关心的问题可能都是搜索引擎的网页数据库中究竟已经存储了多少网页。的确，这个问题很重要。如果一个搜索引擎的网页数据库中存储的网页数量很多，对一个搜索者来说，他会查询到更多的网页，而对于网络管理员来说，他的站点中也会有更多的网页被收录。但并不能以此来判断一个搜索引擎的好坏。因为，搜索引擎在搜集网页中使用的策略是不同的，有的搜索引擎在找到你的站点之后，会采集所有能访问到的网页，而有的搜索引擎 只会抽样采集，收集它们认为具有代表性的网页，还有一些搜索引擎只会采集你站点中的那些主页。这种不同的策略是出于不同的考虑，各有利弊。对所有网页都采集的搜索引擎会使你得到更多的网页，但很可能结果清单中的一些网页都是出自同一个站点，从另一个角度来说又影响了你搜索的广泛性。而那些抽样采集或只采集主页的搜索引擎，虽然网页数量少，但结果清单中出现同一个站点的网页的现象就会比较少，从这个角度来看是扩大了搜索结果。所以，当使用搜索引擎时，要有针对性地选择，而不要一概而论。在后面的搜索引擎特性表中，列出了一些主流搜索引擎目前所拥有的网页数量，以供读者参考。

目前为止还没有包含了 Internet 上所有的网页的搜索引擎。这里有多种原因。Internet 上目前所拥有的网页数量已经上亿计，而且还在增长。因此，搜索引擎要想拥有全部网页，首先就要给自己准备一个专门定制的特大硬盘，其次再给自己配备最快速的 CPU 和其他硬件，并且还要有一个同样高效的程序，以便在搜索者查询时能及时反馈回结果。当然，最辛苦的还是蜘蛛，它要不停地去访问已收集的网页，以保持网页数据库中总是最新内容。同时，蜘蛛自己也需要升级，因为它们中的一些已不适应网页中出现的新事物，如：一些蜘蛛无法识别帧格式、图象连接等，这就会使其丢掉一些网页，而且 Web 技术仍在发展，每当你的网页中应用新技术时，都有可能成为蜘蛛难以逾越的障碍。由此可知，搜索引擎要想包含所有的网页确实还有许多困难。

2. 蜘蛛的爬行速度

这也是一个需要关心的问题。蜘蛛的爬行速度是以它一天能访问多少个网页来衡量的。如果蜘蛛的速度越快，那么，它所维护的网页数据库中的内容就总能保持一个较新的面貌。这个数据指标也被列在本文后面的表格中。

但这并不衡量网页数据库刷新速度的唯一标准，因为一些蜘蛛已经变得聪明，它们会根据测得的网页的变化频率来制定重新访问此站点的时间。若你的网页天天更新，那么，这些蜘蛛可能就会天天光顾你的网页，如果你的网页半年更新一次，那么，它们就一年去你那儿两次。所以，在表格中的刷新速度一项中，网页数据库中网页被刷新的时间可能是一个范围。

3. 刷新速度

系指网页数据库中内容被刷新的频率。由上述可知，搜索引擎的蜘蛛重新光顾你的网页是有一定时间间隔的，而且不同的网页间隔的时间也不一样，每个搜索引

擎都有自己的策略，它们可能偏爱一些流行的站点，而冷落一些无名的“隐士”。所以，在它们的网页数据库中不同网页被更新的频率也是不一样的。因此，可能在你的结果清单中，会出现这样的网页，当你访问该网页时发现并无你查询的关键词，而且内容也与结果清单中列出的概述不符。因为你是对网页数据库查询，而不是直接在 Internet 之上查询，该网页在数据库中并未得到更新，有的可能过了几天，而有的可能已经过时了几个月。

4. 提交网页

Internet 上的站点肯定都是与 Internet 相连接的。所以从理论上讲，当蜘蛛爬行于 Internet 之上时，它应该是可以顺着站点之间的连接爬到你的站点的，也就是说你的站点和网页肯定会被蜘蛛发现。但是从上面的介绍可知，现论和现实是有差距的，有可能你的站点和网页会由于某种原因而被蜘蛛遗忘或无法到达。

因此现实的做法是，主动向搜索引擎送上你的网页，即提交网页。大部分主流搜索引擎在它的界面网页中都提供了这个功能，你可以通过这个功能将你站点或网页的 URL(地址)传送给搜索引擎，这样它的蜘蛛就会把对你的拜访提前到一个较近的日程。当然，如果你的站点有一系列的网页，那么，你只需提交其中的主页或几个关键网页即可。在你站点中的其他网页，蜘蛛会顺着你提交的网页中的连接自动找到。但是，搜索引擎对该站点中被提交的网页和该站点中未被提交的网页的处理速度是不一样的。被提交的网页可能会在近期内即被访问，并纳入网页数据库，而未被提交的网页则被列入蜘蛛的日程表，一段时间之后再去访问。在搜索引擎特性表的提交网页和未提交网页两行中已经分别列出了它们被处理的速度。同时，此处的提交与主题向导中的提交有所不同，虽然都是把站点介绍给它们，但是，提交之后的处理会有所不同。主题向导会由工作人员来浏览并处理你提交的站点或网页，而搜索引擎是由自动程序“蜘蛛”去完成这项工作。

5. 蜘蛛爬行深度

该项指标用于指出搜索引擎采集网页的策略是无限制采集还是抽样采集。如果是无限制采集，那么蜘蛛会采购站点中的所有东西，如果是抽样采集，那么蜘蛛只会采集部分网页。当然，如果有蜘蛛不可逾越的障碍(如：网页中的帧格式、图象连接等有可能阻碍一些蜘蛛)，也会影响蜘蛛的爬行深度。该指标也在搜索引擎特性表中列出。

6. 帧格式和图象连接

前面提到过，一些蜘蛛在爬行中可能会遇到一些障碍，帧格式和图象连接就有可能成为这些障碍，因为一些蜘蛛就象老的浏览器一样不识别网页中的新事物，所以，这些蜘蛛就无法顺着帧格式和图象连接爬得更深，从而会遗失你站点中的大量网页。在下面特性表中也列出了这些搜索引擎的蜘蛛对帧格式和图象连接的识别能力。

7. 站点密码

如果你的站点是需要密码和口令才能进入的站点，如：一些收费站点，那么，蜘蛛能否进入呢？一些搜索引擎的蜘蛛是能够进入这样的一个站点的。同时不用担心它们是“黑客”，它们的目的只是搜索你站点内网页的信息，以便搜索者能够发现你的存在，其实这是一种免费宣传，而访问者要想进入还是需要口令和密码的。当然，如果你不想任何人知道你的站点，那也有一个办法，就是将你的站点和 Internet 断开。我想凡是连到 Internet 上的站点都是希望能够进行交流的，而蜘蛛就是为此而设计的。至于哪些搜索引擎的蜘蛛能够进入口令站点，可在特性表中查到。

8. 忽略单词

搜索引擎的网页数据库中存储的只是网页中的文本形式的信息，这些文本是由单词组成的。其中，并非对所有单词搜索引擎都会给予重视，因为有许多单词是一些没有实际意义的单词，如：冠词 a、the 等。这些单词几乎所有文章都会出现，如果以此作为查询关键词进行搜索的话，那么肯定会造成“灾难”性的后果。因此，搜索引擎基本上都会忽略这些普遍使用且没有实际意义的单词，当查询词中出现这些单词时，引擎会对其忽略然后进行搜索。而且，有一些搜索引擎对一些有实际意义，又比较普遍的单词也会忽略，如：web、new 等类似的单词。例如你输入的查询关键词是 web developing，那么这些搜索引擎会忽略单词 web。当然，如果你必须包含这些单词作为关键词的一部分的话，你可以使用搜索引擎提供的一些功能，如：使用引号将其作为词组来查询，输入“web developing”，这样搜索引擎就会找出所有含有 web developing 的网页。在特性表的忽略单词一行中，显示了哪些搜索引擎会忽略单词。

9. Internet“污染源”

对于一些用不道德手段欺骗搜索引擎的“假冒伪劣”

网页，所有搜索引擎一经发现都会毫不留情的给以惩罚。这些“假冒伪劣”网页是那些使用诸如：重复使用单词等恶劣手段欺骗搜索引擎以获得高相关度等级的网页，它们严重影响了 Internet 使用者通过搜索引擎进行查询的质量。因此，搜索引擎一旦检测出这些网页，轻则大幅度降低其相关度等级，重则将其从网页数据库中删除且永不收录。从特性表中可以看出搜索引擎对这些“污染源”所持的态度。

10. Meta 标志

许多人都认为所有搜索引擎都支持 Meta 标志，但实际上，只有部分搜索引擎支持 Meta 标志。因为有些搜索引擎认为 Meta 标志并不可靠。而且有些也只是支持 Meta 标志中的一部分，即它们把 Meta 标志中定义的概述仅当文本收集，并用它来控制出现在结果清单的概述。特性表中也列出了各搜索引擎对 Meta 标志的识别程度。

表 1 部分搜索引擎特性表

搜索引擎	AltaVista	Excite	HotBot	InfoSeek	Lycos
拥有网页数量（百万）	大 (100)	大 (55)	大 (80)	中 (30)	中 (30)
每天可访问的网页数量	1 千万	3 百万	最多 1 千万	不详	6 百万到 1 千万
网页刷新的时间间隔	一天到三个月	一周到三周	一天到两个星期	一分钟到两个月	一周到两周
提交网页	一天	三周	一到二天	不超过一分钟	一到二周
未提交网页	一到三个月	三周	二周	一到二个月	一到二个周
蜘蛛爬行深度	无限制，穷尽	无限制，穷尽	无限制，穷尽	抽样	抽样
识别帧格式	不识别	不识别	不识别	识别	识别
识别图象连接	识别	不识别	不识别	识别	不识别
口令保护站点	不进入	进入	不进入	进入	进入
忽略单词	忽略	忽略	忽略	不忽略	忽略
Internet 污染源	惩罚	惩罚	惩罚	惩罚	惩罚
meta 标志	识别	不识别	识别	识别	部分识别
概述	Meta 标志或网页头几行	待考察	Meta 标志或网页头几行	Meta 标志或网页体前 200 个字符	基于内容创建
蜘蛛名称	Scooter	Architext Spider	Slurp the Web Hound	Side winder	T-Rex

注：该表格中的数据有时间性，随着搜索引擎的发展会有所变动。

11. 蜘蛛名称

每个搜索引擎都会使用一个蜘蛛帮助它们收集网页，大部分蜘蛛都有名称，在特性表中提供了这些搜索引擎的蜘蛛的名称，通过检查日志文件可知哪个蜘蛛访问了你的站点。

二、搜索引擎特性

在搜索的过程中，每个搜索引擎都有自己的特色。下面，针对一些引擎的特性作一简略分析。

如果你已有比较贴近搜索目标的搜索词，需要一个精确的搜索，此时，建议选用 AltaVista 因为它具有巨大和快速的全文本索引，可以进行具体准确的搜索。

在搜索的短语中，如果含有一些普遍存在的或一些没有实际意义的单词，例如 New Orleans、Vitamin C，其中，New 是一个普遍存在的单词，C 是一个没有实际意义的单词，但他们都不能被忽略，这时就需要用 Infoseek，Infoseek 可帮助搜索在其他引擎中忽略的一些单词。

用 Magellan、Lycos、TOP 5% Home、WebCrawler 可以考察、评估站点。

如果只想浏览，或寻找一个比较普遍的话题，Yahoo 或 A2Z 会合适一些，Yahoo 或 A2Z 均有能对站点作简短描述的主题树目录。

三、结论

通过对上述搜索引擎的分析、综合、给出了不同搜索引擎之间的优缺点，在应用中，针对需求的不同恰当地选择适于自己目标的搜索引擎，以便更有的放矢，游刃有余。

参考文献

[1] Alfred and Emily Glossbrenner. Search engines for the world wide web. Peachpit Press, 1998

（来稿时间：1999 年 3 月）