

这是《文献管理与信息分析》课程内容的一部分。

对于科研工作者来说，除了掌握快速的收集信息和有效管理信息的能力之外，还需要有一定的信息分析能力。譬如，检索某个研究方向，结果文献有上千篇，此时我们该如何对待这些文献？精炼检索条件很可能会导致有价值的文献被排除在外。在交叉科学盛行的今天，如果了解其它领域的进展情况，由于缺乏相应的专业知识，如何判断哪些文献是有重要参考价值的文献？这些问题的解决都需要我们具备一定的文献信息分析能力。

有人说，在 WOS 里按照引用次数排序，引用次数最多的必定参加价值更大，其实不然，后面会有解释。

我所知的文献信息分析软件有基于引文的分析软件 histcite，基于内容分析的 refviz、omniviz、Tda 等，还有分析研究前沿的 citespace 等。

这里先简要介绍一下引文分析软件 histcite 它的功能和用法。

histcite = history of cite，意味引文历史，或者叫引文图谱分析软件。该软件系 sci 的发明人加菲尔德开发，能够用图示的方式展示某一领域不同文献之间的关系。可以快速帮助我们绘制出一个领域的发展历史，定位出该领域的重要文献，以及最新的重要文献。

软件的最新试用版本可以从 [www.histcite.com](http://www.histcite.com) 网站下载到。

软件的试用非常简单，但如何从软件给出的图谱中得出有价值的信息，以及不同图谱展示的内在含义，需要我们不断揣摩和理解。

下面先介绍一下如何使用软件；然后介绍一下软件里涉及的一些概念；最后给出几个个示例。

## 一、软件使用

1、从 histcite 网站下载软件，安装之后，点击 histcite 图标即可开启软件；软件打开后的窗口类似 IE 的界面；

2、数据的获取；histcite 目前用于分析的文献信息只能来源于 web of science 数据库；在 wos 数据库进行检索后，在页面的底端选择需要导出的数据记录，由于 wos 目前只支持每次导出 500 条记录，如果检索结果超过 500 条需要分多次导出。选择导出 的文献记录之后，第二步，一定要选择输出全记录，并且要包含引文信息；第三步将需要的文献保存成文本文件。一般来说，如果文献记录少于 500 条，分析的意义不是很大。合适的数据量个人认为在几百到几千条记录之间比较合适。

3、从 file 菜单下点击 add file，导入上述保存的数据；如果有多个文本文件，可以重复执行导入；

4、数据导入后，软件会自动进行分析。初学者可以不去追究各种按钮的含义。在 tool 菜单下，选择 graph maker，然后新的界面点击左上角的 make graph 按钮。软件会根据默认的条件作出一张引文关系图来，来展示当前数据库中重要文献之间的关联。

5、读图 作出图之后，理解图谱才是关键。一般默认会画出 30 篇文献之间的关联。图上有 30 个圆圈，每个圆圈表示一片文献，中间有个数字，是这篇文献

在数据库中的需要。圆圈的大小表示引用次数的多少，圆圈越大表示受关注越多。不同圆圈之间有箭头相连，箭头表示文献之间的引用关系。多数情况下，你会看到最上面有一个圆圈较大，并有很多箭头指向这篇文章。那么这篇文章很可能就是这个领域的开山之作。

以上就是软件的简单使用过程。

## 二、软件功能和基本概念

将数据导入到软件之后，文献会自动排列在软件的主界面。文献的排序方式可以按日期，可以按杂志或按作者进行排序。

文献记录的上方还有一些蓝色字体的按钮。这些词都是可以点击的，并进行相应分析。如点击 authors，软件会列出所有作者，并将每位作者的文献数、引用次数等信息列出来。这些命令较容易理解，不多做介绍。

在默认窗口的右侧，有 LCS、GCS、LCR、CR。下面分别解释一下这几个功能。

GCS 是 global citation score，即引用次数，也就是你咋 web of science 网站上看到的引用次数。如果你点击 gcs，软件会按照 GCS 进行排序，此时的结果与你在 wos 网站按被引频次排序的结果是一样的。

CR 是 cited references，即文章引用的参考文献数量。如果某篇文献引用了 50 篇参考文献，则 CR 为 50。这个数据通常能帮我们初步判断一下某篇文献是一般论文还是综述。

LCS 和 LCR 是 histcite 里比较重要的两个参数。

LCS 是 local citation score 的简写，即本地引用次数。与 gcs 相对应，gcs 是总被引次数。lcs 是某篇文章在当前数据库中被应用的次数。所以 LCS 一定是小于或等于 GCS 的。

一篇文章 GCS 很高，说明被全球科学家关注较多。但是如果一篇 GCS 很高，而 LCS 很小，说明这种关注主要来自与你不是同一领域的科学家。此时，这篇文献对你的参考意义可能不大。举个例子，2003 年发表在 nature 上的两篇文章 P1 (GCS:580,LCS:12) 和 P2(GCS:36,LCS: 24)。第一篇文章 gcs 很高，lcs 很低，说明关注这篇文章的绝大部分作者与你关注的方向不同。而第二篇文章 gcs 较低，但 LCS 比第一批要高，即很多引用 p2 的文章都在当前数据库，也即与你的研究方向相关。所以，p1 p2 相比，p2 应该更贴近你的研究方向，参考价值更大。

在第一部分的介绍中，make graph 时，默认是按 LCS 排序的，也可以选择按 gcs 排序。你可以比较一下这两者的差异，一般 LCS 作图，得到的关联较丰富，而 gcs 作图往往文献之间没什么关联。这就回答了上面开始提出的一个问题，为什么按引用频次排序往往不是很有参考价值的原因。

LCR 与 CR 对应是 local cited references，是指某篇文献引用的所有文献中，有多少篇文献在当前数据库中。如果最近有两篇文章，p1 p2,都引用了 30 篇参考文献，其中 p1 引用的 30 篇文献中有 20 篇在当前数据库，p2 只有 2 篇文献在当前数据库。此时，p1 相对更有参考价值，因为它引用了大量和你的研究相关的文献。

根据 LCS 可以快速定位一个领域的经典文献，LCR 可以快速找出最新的文献中哪些是和自己研究方向最相关的文章。

注；引文有些不规范导致引文分析结果偏差，这里暂不做讨论。感兴趣的朋友可以参考引文相关的理论文献或书籍。

### 三、部分结果示例

这里选择了一个我上课的例子，和学生作业中的例子。更多示例请大家参考中科大生命科学实验中心论坛上学生提交的作业。

<http://biotech.ustc.edu.cn/forum>

第一个例子是关于合成生物学（synthetic biology），2000 年之后发展起来的、并在近年广受关注的学科。在 wos 里以 synthetic biology 进行检索，2009 年 10 月 30 日可以获得 2297 篇参考文献。下载后导入 hiscite，按 ICS 作图；结果如下：

从第一个图可以看出 704 这篇文献应该是一片开创性的工作，或是一篇重要的综述。后面的 1077 1134 1089 这三个圆圈较大，说明这几篇文献受到广泛关注，具有较重要的地位。

图 2 就是相同的数据库按 GCS 做图得到的结果，可以看到这些文献之间没什么关联。所以在 wos 中，按引用频次排序，即使排在前面的文献对您也未必有多大参考价值。

图 3 是学生提交的作业，原作者的分析附后，基本上是正确的。

( 学生作业中对上图的解释：在 web of science 上检索主题包含 G-quadruplex 和 detection 的文章，一共 108 篇，用 histcite 作图，该图表明有关将 G-quadruplex 应用于各种检测的研究工作主要从 05 年 Nagatoishi S;发表在德国应化和 He F 发表在 JACS 上的全文开始，逐渐有大量的文章发表，而前述两篇文章的 LCS 是最高的，其可读性也是非常大，从 09 年的数据来看，对同年发表的文章的引用次数大大增加，表明该领域在 09 年的发展迅速。从 CR 的分析来说，编号为 64,55,和 1 的几篇文章引用文献的数量非常大，应该是综述性文章，对初涉此领域的人也有较大的可读性。)

LCR 分析结果

可以看出，排在前 10 位的都是近期非常具有参考价值的文献。

#	<a href="#">Date</a> / <a href="#">Author</a> / <a href="#">Journal</a>	<a href="#">LCS</a>	<a href="#">GCS</a>	<a href="#">LCR</a>	<a href="#">CR</a>
1	2009 Agapakis CM, Silver PA  Synthetic biology: exploring and exploiting	0	0	29	103

	<p>genetic modularity through the design of novel biological networks</p> <p>MOLECULAR BIOSYSTEMS. 2009; 5 (7): 704-713</p>				
2	<p>2179 Purnick PEM, Weiss R</p> <p>The second wave of synthetic biology: from modules to systems</p> <p>NATURE REVIEWS MOLECULAR CELL BIOLOGY. 2009 JUN; 10 (6): 410-422</p>	1	3	29	126
3	<p>2284 Weber W, Fussenegger M</p> <p>The impact of synthetic biology on drug discovery</p> <p>DRUG DISCOVERY TODAY. 2009 OCT; 14 (19-20): 956-963</p>	0	0	27	72
4	<p>1699 O'Malley MA, Powell A, Davies JF, Calvert J</p> <p>Knowledge-making distinctions in synthetic biology</p> <p>BIOESSAYS. 2008 JAN; 30 (1): 57-65</p>	5	6	21	78
5	<p>1969 Leonard E, Nielsen D, Solomon K, Prather KJ</p> <p>Engineering microbes with synthetic biology frameworks</p>	1	2	21	81

	TRENDS IN BIOTECHNOLOGY. 2008 DEC; 26 (12): 674-681				
6	2165 Picataggio S  Potential impact of synthetic biology on the development of microbial systems for the production of renewable fuels and chemicals  CURRENT OPINION IN BIOTECHNOLOGY. 2009 JUN; 20 (3): 325-329	0	0	21	60
7	1451 Drubin DA, Way JC, Silver PA  Designing biological systems  GENES & DEVELOPMENT. 2007 FEB 1; 21 (3): 242-254	20	32	20	107
8	2008 Tanouchi Y, Pai A, You LC  Decoding biological principles using gene circuits  MOLECULAR BIOSYSTEMS. 2009; 5 (7): 695-703	0	2	17	78
9	2184 Bhalerao KD  Synthetic gene networks: the next wave in biotechnology?  TRENDS IN BIOTECHNOLOGY. 2009 JUN; 27 (6):	0	1	16	66



	368-374				
10	<p>2132 Deplazes A</p> <p>Piecing together a puzzle An exposition of synthetic biology</p> <p>EMBO REPORTS. 2009 MAY; 10 (5): 428-432</p>	0	0	15	38

结语;

文献信息分析只是为我们提供了不同的视角去了解别人的工作,让我们能够更快的找到最有价值的信息,但任何分析都不能代替我们阅读文献。

如果你想了解更详细的软件使用方法,可以下载《文献管理与信息分析》课程中相应的教学录像。也可以关注我后续将推出的软件教程。