●邢美园 袁 明 苏开颜

平均累积报道时差统计法
——MEDLINE CD-ROM 报道时差实例分析

摘 要 “平均累积报道时差统计法”用于数据库报道时差的统计,可快捷精确地统计较长年限的累积时差,消除了手工抽样统计由于统计年限较短带来的统计误差。统计结果可直接用于相同专业的不同种类检索工具之间平均报道时差的评价。公式 2。表 3。参考文献 3。

关键词 平均累积报道时差统计法 MEDLINE CD-ROM 数据库评价
分类号 G350

ABSTRACT The Statistical Method of Average Cumulative Inclusion Lag, a method for the statistics of database inclusion time lag, can be used easily for the statistics of cumulative lag in a long period of time, and eliminates errors caused by short statistical period in manual sampling. The results of the statistics can be used for the evaluation of different retrieval tools in the same field. 2 formulas. 3 tabs. 3 refs.

KEY WORDS Statistical Method of Average Cumulative Inclusion Lag. Medline. CD-ROM. Database evaluation.
CLASS NUMBER G350

文献检索工具的报道时差,即原始文献从正式发表到检索工具中反映出来的时间间隔,是评价检索工具质量的主要指标之一。报道时差的长短直接影响到最新文献的利用。长期以来国内统计文献检索工具的报道时差主要是以手工抽样逐篇统计,计算量十分庞大,统计年限也有限。因此,建立一种快捷而又较为准确的统计检索工具报道时差的方法十分必要。本文用一种暂称为“平均累积报道时差统计法”统计了美国 MEDLINE 数据库的平均报道时差,并探讨了其适用范围和影响因素。

1 “平均累积报道时差统计法”公式推导

目前,大部分检索数据库的主要文献源为定期出版的期刊文献。这些期刊的出版周期稳定,报道量也较稳定,且它们的文献在 1 年内也是平均分布的。因此我们假设:

检索工具的数据每月更新,日期取月中;检索工具来源期刊的出版日期为每月月中,则每月的出版日期为 n - 0.5 (n 为累积月份,不按年份的 12 个月取值,取值范围为 1,2,3,……n);全年期刊文献发表日期的中间值为每年年中。

由假设得 1 月份和 12 月份的出版日期为 0.5 月和 11.5 月,则全年的中间值为:

$$(0.5 + 11.5) \div 2 = 6$$

由此可得出某月更新数据的滞后月数为:该月的出版累积月减去全年期刊文献发表日期的中间值。表达式为:

$$n - 0.5 - 6 = n - 6.5$$

报道时差为所有文献被收录滞后月数的总和与收录总文献数之比。设某数据库每月收录文献数 Vn,则至第 n 月的总文献数为 $\sum Vn$;每月收录文献的滞后月数为 $(n - 6.5) \cdot Vn$,则至第 n 月所有文献被收录时滞后月数的总和为 $\sum (n - 6.5) \cdot Vn$。

该数据库的平均累积报道时差(Cn,单位为月)为:

$$Cn = \frac{\sum (n - 6.5) \cdot Vn}{\sum Vn} = \frac{\sum n \cdot Vn}{\sum Vn} - 6.5 \qquad (公式\ 1)$$

公式 1 中,$\sum Vn = V1 + V2 + V3 + \cdots\cdots Vn$

其含义是某年发表的文献在某数据库中自该年第 1 个月起至第 n 个月每月收录文献的累积和:

$$\begin{aligned}\sum n \cdot Vn &= V1 + 2V2 + 3V3 + nVn \\ &= (V1 + V2 + \cdots\cdots + Vn) \\ &\quad + (V2 + V3 + V4 + \cdots\cdots + Vn) \\ &\quad + \cdots\cdots + (V_{(n-1)} + Vn) + Vn\end{aligned}$$

其含义是某年发表的文献在某数据库中自该年第 n 月收录的文献数至第 1 个月份收录文献数累积和的累积和。

那么平均累积报道时差(Cn)的概念也可用以下文字来表达:平均累积报道时差(Cn)等于某年发表的文献在某数据库中第 n 月收录的文献数至第 1 个月份收录文献数累积和的累积和与该年第 1 个月起至 n 个月每月收录文献的累积和的除积减去 6.5。

同理,如果该数据库是周更新数据,其平均累积报道时差(Cw,单位为周)为:

$$Cw = \frac{\sum w \cdot Vw}{4.3 \cdot Vw} - 28.16 \qquad (公式\ 2)$$

w 为某数据库数据更新的周数(w = 1,2,3,……w)。

2　实例验证

下面以月更新类型数据库 MEDLINE CD-ROM 为例,验证公式 1。

数据来源:1990～1999 年的 MEDLINE CD-ROM,2001 年版,美国银盘公司出版。

(1)统计每年发表的文献在发表后各年年底时的平均报道时差。

a. 分别用“ud = 199001”、“ud = 199002”、“ud = 199003”……“ud = 199912”,查出每年度 MEDLINE CD - ROM 中每月输入数据库的记录数量。

b. 分别统计出 1990 年 1 月～1999 年 12 月期间每年发表文献的每月输入数据库记录数量的累积和。

c. 分别统计出 1990 年 1 月～1999 年 12 月期间每年发表文献的每月输入数据库记录数量累积和的累积和。

d. 按公式 1 计算每年发表的文献在发表后各年年底时的平均报道时差。

(2)统计每年发表的英语文献在发表后各年年底时的平均报道时差。

a. 分别用“la = english and ud = 199101”、“la = english and ud = 199102”……“la = english and ud = 199912”,查出每年度 MEDLINE CD-ROM 中每月输入数据库的英语文献记录数量。

b. 按公式 1 计算每年发表英语文献在发表后各年年底时的平均报道时差。

(3)统计每年发表的非英语文献在发表后各年年底时的平均报道时差。

a. 分别用“ud = 199101 and la = non-english”、“ud = 199102 and la = non-english”……“ud = 199112 and la = non-english”,查出每年度 MEDLINE CD-ROM 中每月输入数据库的非英语文献记录数量。

b. 按公式 1 计算每年发表的非英语文献在发表后各年年底时的平均报道时差。

(4)统计每年美国和非美国出版的文献在发表后各年年底时的平均报道时差。

a. 分别用“ud = 199101 and cp = united-states”、“ud = 199102 and cp = united-states”……“ud = 199112 and cp = united-states”,查出每年度 MEDLINE CD-ROM 中每月输入数据库的美国出版的文献记录数量。

b. 按公式 1 计算每年美国出版的文献在发表后各年年底时的平均报道时差。

c. 以每年度 MEDLINE CD-ROM 中每月输入数据库的记录数量减去每月输入数据库的美国出版的文献记录数量,得出每月输入数据库的非美国出版的文献记录数量。

d. 按公式 1 计算每年非美国出版的文献在发表后各年年底时的平均报道时差。

计算结果见表 1～表 3。

表 1　每年发表的文献在发表后各年年底时的平均报道时差　(单位:月)

	文献被 MEDLINE 收录的年份									
发表年份	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1990	2.16	5.61	5.99	6.05	6.09	6.13	6.14	6.16	6.16	6.17
1991		2.43	5.81	7.06	7.15	7.25	7.27	7.29	7.30	7.31
1992			2.25	5.05	5.35	5.45	5.46	5.47	5.49	5.51
1993				1.91	5.15	5.42	5.48	5.52	5.58	5.60
1994					2.22	5.71	5.91	6.07	6.24	6.27
1995						2.32	5.83	6.47	6.73	6.79
1996							3.15	6.80	7.11	7.20
1997								2.26	5.64	5.87
1998									2.09	5.01
1999										2.19

表 2　每年发表的英语和非英语文献在发表后各年年底时的平均报道时差　(单位:月)

		文献被 MEDLINE 收录的年份									
发表年份		1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1990	E	2.03	5.21	5.52	5.55	5.58	5.59	5.59	5.61	5.62	5.62
	NE	2.89	8.79	9.34	9.46	9.60	9.59	9.76	9.76	9.77	9.77

续表

发表年份		文献被 MEDLINE 收录的年份									
		1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1991	E		2. 30	5. 57	5. 76	5. 83	5. 88	5. 89	5. 91	5. 92	5. 92
	NE		3. 29	7. 24	7. 73	8. 07	8. 42	8. 49	8. 50	8. 54	8. 55
1992	E			2. 18	5. 66	5. 87	5. 93	5. 93	5. 94	5. 95	5. 98
	NE			2. 75	6. 63	7. 14	7. 38	7. 51	7. 53	7. 58	7. 58
1993	E				1. 82	4. 83	5. 01	5. 06	5. 11	5. 17	5. 20
	NE				2. 60	7. 04	7. 83	7. 90	7. 97	7. 98	7. 98
1994	E					2. 17	5. 51	5. 68	5. 82	5. 98	6. 02
	NE					2. 67	7. 01	7. 41	7. 65	7. 85	7. 85
1995	E						2. 28	5. 65	6. 30	6. 55	6. 61
	NE						2. 67	7. 00	7. 88	8. 17	8. 27
1996	E							3. 04	6. 67	6. 92	6. 99
	NE							4. 06	7. 76	8. 49	8. 63
1997	E								2. 22	4. 64	4. 77
	NE								2. 67	6. 63	7. 40
1998	E									2. 06	4. 90
	NE									2. 89	6. 33
1999	E										2. 03
	NE										2. 76

注:E —为英语文献,NE —为非英语文献。

表 3　每年美国和非美国出版的文献在发表后各年年底时的平均报道时差　　(单位:月)

发表年份		文献被 MEDLINE 收录的年份									
		1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1990	A	1. 78	4. 38	4. 61	4. 64	4. 67	4. 68	4. 68	4. 69	4. 69	4. 69
	NA	2. 25	6. 56	7. 03	7. 10	7. 17	7. 22	7. 24	7. 26	7. 27	7. 27
1991	A		1. 97	4. 73	4. 82	4. 87	4. 88	4. 89	4. 89	4. 90	4. 90
	NA		2. 94	6. 70	7. 05	7. 16	7. 34	7. 37	7. 40	7. 42	7. 43
1992	A			1. 91	4. 07	4. 22	4. 23	4. 23	4. 24	4. 26	4. 28
	NA			2. 61	5. 85	6. 28	6. 42	6. 46	6. 47	6. 50	6. 51
1993	A				1. 63	4. 16	4. 39	4. 54	4. 57	4. 62	4. 63
	NA				2. 21	5. 98	6. 38	6. 46	6. 52	6. 58	6. 61
1994	A					1. 93	4. 79	4. 92	5. 05	5. 25	5. 25
	NA					2. 55	6. 51	6. 77	6. 95	7. 10	7. 14
1995	A						2. 03	4. 96	5. 33	5. 64	5. 71
	NA						2. 65	6. 60	7. 45	7. 68	7. 73
1996	A							2. 72	6. 05	6. 29	6. 36
	NA							3. 67	7. 52	7. 88	7. 97

续表

发表年份		文献被 MEDLINE 收录的年份									
		1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1997	A								1.99	4.08	4.19
	NA								2.56	5.60	5.94
1998	A									1.71	4.06
	NA									2.52	5.85
1999	A										1.82
	NA										2.58

注:A —为美国出版文献,NA —为非美国出版文献。

3　讨论

3.1　关于文献检索工具报道时差的统计年限

从表 1～表 3 可见,每年发表的文献在 MEDLINE 数据库中随着统计年限的加长,其平均报道时差呈单调递增,但递增幅度逐渐减少。一方面,这是由于在 MEDLINE 数据库每月输入的新记录中有一部分是往年发表的文献,且随着间隔年份拉长所追加的记录数量逐渐减少之故。从理论上讲,统计年限越长结果越精确;从实际统计结果看,统计年限起码应 ≥5 年($P < 0.01$),以往手工抽样统计当年或两年内的报道时差的结果显然偏小。另一方面,要求从 5 年或 5 年以上年限的检索工具中用手工挑选以往某一年发表的文献,再逐篇统计,工作量十分浩大,实际操作将非常困难。因此,与手工统计相比,"平均累积报道时差统计法"用于文献检索工具报道时差统计,不但结果精确而且将大大提高工作效率。

从表 1～表 3 中还可见,不同年份发表的文献其平均报道时差也有差异。如果仅仅统计了其中某一年发表文献的报道时差,并将这某一年的时差代表该检索工具与另一种检索工具中另一年发表文献的报道时差相比,其结论显然不科学,而这种不科学的比较在以往评价不同种类检索工具的报道时差时是常见的。

3.2　每年发表的文献在发表后各年年底时的平均报道时差

从表 1 可见,若不论语种和出版国别,当统计年限 ≥5 年时,1990～1995 年期间不同年份发表的文献在 MEDLINE 中的平均报道时差为 5.46～7.31 个月。这与文献中有关 MEDLINE 或其主体部分 INDEX MEDICUS 的报道时差为 1 个月左右[1]、2～3 个月[2]、3～6 个月[3]相比有较大差异。

3.3　每年发表的英语文献和非英语文献在发表后各年年底时的平均报道时差

从表 2 可见,当统计年限 ≥5 年时,1990～1995 年期间不同年份发表的英语文献在 MEDLINE 中的平均报道时差为 5.11～6.61 个月,非英语文献的平均报道时差为 7.51～9.77 个月。同一年发表的英语文献的平均报道时差比非英语文献的平均报道时差短约两个月左右,差异显著。

3.4　每年美国和非美国出版的文献在发表后各年年底时的平均报道时差

从表 3 可见,当统计年限 ≥5 年时,1990～1995 年期间不同年份美国出版的文献在 MEDLINE 中的平均报道时差为 4.23～5.71 个月,非美国出版文献的平均报道时差为 6.46～7.71 个月。同一年出版的美国文献的平均报道时差比非美国出版文献的平均报道时差短约 2～3 个月左右。

4　结论

(1)应用"平均累积报道时差统计法"统计 1990～1999 年期间每年发表的文献在 MEDLINE 数据库中的平均报道时差,统计年限应 ≥5 年;1990～1995 年期间,不同年份发表的文献在 MEDLINE 中的平均报道时差为 5.46～7.31 个月,不同年份发表的英语文献在 MEDLINE 中的平均报道时差为 5.11～6.61 个月,非英语文献的平均报道时差 7.51～9.77 个月;同一年发表的英语文献的平均报道时差比非英语文献的平均报道时差短约两个月左右,不同年份美国出版的文献在 MEDLINE 中的平均报道时差为 4.23～5.71 个月;非美国出版文献的平均报道时差为 6.46～7.71 个月;同一年出版的美国文献的平均报道时差比非美国出版文献的平均报道时差短约 2～3 个月。

(2)"平均累积报道时差统计法"用于数据库报道时差的统计,可快捷精确地统计较长年限的累积时差,消除了手工抽样统计由于统计年限较短带来的统计误差。由于该方法可方便地统计检索工具中不同年份、不同语种发表的、不同国家或地区出版的文献的平均报道时差,故其统计结果可直接用于相同专业的不同种类检索工具之间平均报道时差的评价。

参考文献

1　徐延香,戴惠珍.医学情报检索.南京:南京大学出版社,1992
2　陈界.医学文献检索.北京:中国科学技术出版社,1994
3　朱允尧,朱象喜,崔竹金.医学文献检索.杭州:北京医科大学中国协和医科大学联合出版社,1992

邢美园　袁　明　浙江大学图书馆湖滨分馆工作。通讯地址:杭州市。邮编 310031。

苏开颜　浙江大学图书馆工作。通讯地址:杭州市。邮编 310027。

(来稿时间:2002-12-02)