(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105447313 A
(43)申请公布日 2016.03.30

(21)申请号 201510823761.1

(22)申请日 2015.11.23

(71)申请人 成都云堆移动信息技术有限公司
地址 610041 四川省成都市高新区府城大道西段399号5栋1单元12层1-3号

(72)发明人 王飞 张国鸿 张何君

(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340
代理人 郭霞

(51)Int.Cl.
G06F 19/00(2011.01)

权利要求书3页 说明书7页

(54)发明名称

电子文件阅读数非自然增长识别方法

(57)摘要

本发明公开了一种电子文件阅读数非自然增长识别方法，包括以下步骤：对电子文件公开的原始数据进行采集及阅读曲线绘制；数据预处理；对已绘制的阅读曲线进行趋势分析、特征分析，最后进行非线性拟合，同时将横坐标的时间转换为自然数序列，建立实时曲线模型；第一次计算；第二次计算；第三次计算；第四次计算；根据上述四次计算得到的非自然增长形态的或然率，得到最终的非自然增长形态的综合或然率C_f；根据非自然增长形态的综合或然率C_f判断电子文件阅读数非自然增长概率。通过本发明能够精确判断电子文件阅读数非自然增长概率打下坚实的基础，实现对日益增长的电子文件阅读数非自然增长进行较为准确的监测和识别，有利于助推网络市场的健康发展。

CN 105447313 A

1.一种电子文件阅读数非自然增长识别方法，其特征在于：包括以下步骤：

(1)原始数据采集及阅读曲线绘制：对电子文件的公开阅读数进行实时监测，实时监测的时间间隔可以为一个或多个，定时采集相应的阅读数，最终绘制出电子文件的实时阅读曲线；

(2)数据预处理：通过数据归整和清洗，将原始数据处理为每个相同时间间隔的时间点均有数据与之对应的序列，最终得到包括序列X、更新时间T和阅读数R这三列的数据M；

(3)对已绘制的阅读曲线进行趋势分析、特征分析，最后进行非线性拟合，同时将横坐标的时间转换为自然数序列，建立实时曲线模型如下：

$$y = a - \frac{b}{e^{\frac{cx}{600}}}$$

其中c代表时间间隔；

(4)第一次计算：利用实时曲线模型对坐标系X-Y进行拟合，根据拟合度计算出本次计算的非自然增长形态的或然率C_1；这里的X代表数据预处理后的序列，Y代表阅读数R；

(5)第二次计算：依次判断夜间非自然增长形态和白天非自然增长形态，并根据夜间非自然增长形态和白天非自然增长形态计算得到本次计算的非自然增长形态的或然率C_2；

(6)第三次计算：结合第二次计算的数据，对相邻阅读增量进行差值处理，得到本次计算的非自然增长形态的或然率C_3；

(7)第四次计算：结合第二次计算的数据，计算阅读曲线斜率角度，最后得到本次计算的非自然增长形态的或然率C_4；

(8)根据上述四次计算得到的非自然增长形态的或然率，得到最终的非自然增长形态的综合或然率C_f；

(9)根据非自然增长形态的综合或然率C_f判断电子文件阅读数非自然增长概率，C_f值越大，电子文件阅读数非自然增长概率越高，反之越低。

2.根据权利要求1所述的电子文件阅读数非自然增长识别方法，其特征在于：所述步骤(4)中第一次计算的具体方法为：

先求出将$X1 = \frac{1}{e^{\frac{cx}{600}}}$一项视为自变量X1，然后利用线性方程y＝a-b*X1拟合求出a、b和预测值Y′，并计算得出曲线拟合度R^2，其计算公式为：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i)^2}$$

根据以下公式计算出本次计算的非自然增长形态的或然率：

$$C_1 = \frac{1}{1+\exp(10*R^2-6)};$$

所述步骤(5)中第二次计算的具体方法为：

如果处理后的数据M包括两个或两个以上时间间隔的数据，则数据M中取出其中一个时间间隔的数据，组成新的数据；首先求出相邻时间点的阅读数R的差值，此为相等时间间隔的阅读增量r_Δ，即后一个阅读数减去前一个阅读数的差，由此值形成ID列，从而得到相等时

间间隔阅读增量占总阅读的比率rate＝r_{Δ}/max(R)及其对应的时间点t和序列号x，并得到数据列(ID,t,y,rate)，其中y即为r，表示阅读量R中的元素；首先判断夜间非自然增长形态：设置电子文件发布后第一个凌晨从2:00至7:00，且其阅读增量比率阀值为3％，此后凌晨1:00至7:00的阀值为1.5％，若电子文件第一个凌晨2:00至7:00和其它凌晨1:00至7:00的阅读增量N的比率rate超过对应的阀值，则将其记录在向量H中，根据以下公式得到夜间非自然增长形态的或然率：

$$C_n = \frac{\sum_{i=1}^{m} H_i}{\sum_{i=1}^{k} N_i}$$

其次判断白天非自然增长形态：去除电子文件凌晨时刻的数据，首先判断发稿前2-4个白天时刻数据的阅读增量占总阅读的比率rate是否有大于或等于0.3的数据，若有，则判断白天非自然增长形态的或然率为0.8，即C_d＝0.8，若没有，则C_d＝去掉白天前4个时刻和夜间的其它时刻的阅读增量比率最大值；

最后，根据公式C_2＝C_n+C_d得到本次计算的非自然增长形态的或然率C_2；若C_2大于或等于1，则都取为1.0，若C_2小于1，则取其实际值；

所述步骤(6)中第三次计算的具体方法为：

根据第二次计算的ID列，即r_{Δ}值，对相邻阅读增量进行差值处理，即后一个r_{Δ}减去前一个r_{Δ}之差，得到相等时间段下的增量差y2，选出y2列前5个点的最高值max，再从y2的第五个点以后的所有点中找出所有满足y2>max/3这一范围的值，该值至少有两个，如只出现一个则视为自然增长形态，不在此算法识别范围内，求出这几个数的平均值P，则本次计算的非自然增长形态的或然率C_3的计算公式为：

C_3＝((P-max/3)/(max/2-max/3))＝(6P-2max)/max

若C_3大于或等于1，则为非自然增长形态，当C_3介于(0,1)之间则存在非自然增长形态的可能性，当C_3＝0，则为自然增长形态；

所述步骤(7)中第四次计算的具体方法为：

根据第二次计算所得数据列(ID,t,y,rate)、ID列即r_{Δ}值，以及处理后的数据Y即R，根据以下公式计算得到阅读曲线斜率角度：

$$\text{degree} = \frac{\tan^{-1}\left(\frac{y}{\max(Y)\cdot 0.06}\right)\cdot 180}{\pi}$$

对degree四舍五入成整数degree1，首先判断当degree1大于或等于20，且相邻两个degree1的差值小于3度时，将二者的位置和数值分别记录在数据框location_1和value_1中，Value_1的列数i则为连续出现相近阅读增量的最大次数，当i>＝6时，则cheat_line_1＝1.0，当i＝5时，则cheat_line_1＝0.8，当i＝4时，则cheat_line_1＝0.5，当i<＝3时，cheat_line_1＝0.0；

记录调整后degree1出现连续小于等于3度的位置(location_2)和值(vaue_2)，Value_2的列数j则为连续出现斜率角度小于等于3度的最大次数，根据j得出其首次出现小于等于3度的度数的位置k，k即为阅读量不再出现大幅度增长的最早时刻，若K>＝24，则cheat_line_2＝0.0，否则cheat_line_2＝(24-k)/24，根据以下公式本次计算的非自然增长形态的或然率C_4：

C_4＝cheat_line_1+cheat_line_2

当C_4>＝1.0时，则取C_4＝1.0；当C_4<1.0时，则C_4＝cheat_line_1+cheat_line_2；

所述步骤(8)中非自然增长形态的综合或然率C_f的具体计算方法为：

取四种算法的最高分并赋予权重0.8，再求出其余算法得分的均值并赋予权重0.2，则四种算法的综合得分即非自然增长形态的综合或然率C_f的计算公式为：

$$C_f = 0.8 * \max(c_n) + 0.2 * \frac{\sum_1^4 c_n - \max(c_n)}{3}$$ 。

电子文件阅读数非自然增长识别方法

技术领域

[0001] 本发明涉及一种电子文件阅读数监测方法，尤其涉及一种电子文件阅读数非自然增长识别方法。

背景技术

[0002] 随着网络技术的深入发展，电子文件网上在线阅读数量越来越大，比如微信、微博、网络新闻、网络小说等，每天都有成千上万的用户在阅读。对于在线阅读来说，阅读数量的多少及其增长速度是体现该阅读内容是否具有吸引力的重要参考指标，对于商家来说，阅读数量的多少及其增长速度更是关系经济收益的重要信息，也正是这个原因，所以部分媒体或商家通过一些非正常手段来提高阅读数量，或在短期内实现爆发式增长，即实现非自然增长，以达到谋取暴利的目的。显然，这种行为是不利于网络市场健康发展的，但目前尚没有合理的手段能够实现对这种阅读数非自然增长进行较为准确的监测和识别，制约了网络市场的健康发展。

发明内容

[0003] 本发明的目的就在于为了解决上述问题而提供一种电子文件阅读数非自然增长识别方法，这种方法能准确识别电子文件阅读数非自然增长情况。

[0004] 本发明通过以下技术方案来实现上述目的：

[0005] 一种电子文件阅读数非自然增长识别方法，包括以下步骤：

[0006] (1)原始数据采集及阅读曲线绘制：对电子文件的公开阅读数进行实时监测，实时监测的时间间隔可以为一个或多个，定时采集相应的阅读数，最终绘制出电子文件的实时阅读曲线；

[0007] (2)数据预处理：通过数据归整和清洗，将原始数据处理为每个相同时间间隔的时间点均有数据与之对应的序列，最终得到包括序列X、更新时间T和阅读数R这三列的数据M；

[0008] (3)对已绘制的阅读曲线进行趋势分析、特征分析，最后进行非线性拟合，同时将横坐标的时间转换为自然数序列，建立实时曲线模型如下：

[0009] $$y = a - \frac{b}{e^{\frac{cx}{600}}}$$

[0010] 其中c代表时间间隔；

[0011] (4)第一次计算：利用实时曲线模型对坐标系X-Y进行拟合，根据拟合度计算出本次计算的非自然增长形态的或然率C_1；这里的X代表数据预处理后的序列，Y代表阅读数R；

[0012] (5)第二次计算：依次判断夜间非自然增长形态和白天非自然增长形态，并根据夜间非自然增长形态和白天非自然增长形态计算得到本次计算的非自然增长形态的或然率C_2；

[0013] (6)第三次计算：结合第二次计算的数据，对相邻阅读增量进行差值处理，得到本次计算的非自然增长形态的或然率C_3；

[0014] (7)第四次计算:结合第二次计算的数据,计算阅读曲线斜率角度,最后得到本次计算的非自然增长形态的或然率C_4;

[0015] (8)根据上述四次计算得到的非自然增长形态的或然率,得到最终的非自然增长形态的综合或然率C_f;

[0016] (9)根据非自然增长形态的综合或然率C_f判断电子文件阅读数非自然增长概率,C_f值越大,电子文件阅读数非自然增长概率越高,反之越低。

[0017] 作为优选,所述步骤(4)中第一次计算的具体方法为:

[0018] 先求出将$X1=\frac{1}{e^{\frac{cx}{600}}}$一项视为自变量X1,然后利用线性方程y=a-b*X1拟合求出a、b和预测值Y′,并计算得出曲线拟合度R^2,其计算公式为:

[0019] $$R^2=1-\frac{\sum_{i=1}^{n}(Y_i-Y_i')^2}{\sum_{i=1}^{n}(Y_i)^2}$$

[0020] 根据以下公式计算出本次计算的非自然增长形态的或然率:

[0021] $$C_1=\frac{1}{1+\exp(10*R^2-6)};$$

[0022] 所述步骤(5)中第二次计算的具体方法为:

[0023] 如果处理后的数据M包括两个或两个以上时间间隔的数据,则数据M中取出其中一个时间间隔的数据,组成新的数据;首先求出相邻时间点的阅读数R的差值,此为相等时间间隔的阅读增量r_Δ,即后一个阅读数减去前一个阅读数的差,由此值形成ID列,从而得到相等时间间隔阅读增量占总阅读的比率rate=r_Δ/max(R)及其对应的时间点t和序列号x,并得到数据列(ID,t,y,rate),其中y即为r,表示阅读量R中的元素;首先判断夜间非自然增长形态:设置电子文件发布后第一个凌晨从2:00至7:00,且其阅读增量比率阀值为3%,此后凌晨1:00至7:00的阀值为1.5%,若电子文件第一个凌晨2:00至7:00和其它凌晨1:00至7:00的阅读增量N的比率rate超过对应的阀值,则将其记录在向量H中,根据以下公式得到夜间非自然增长形态的或然率:

[0024] $$C_n=\frac{\sum_{i=1}^{m}H_i}{\sum_{i=1}^{k}N_i}$$

[0025] 其次判断白天非自然增长形态:去除电子文件凌晨时刻的数据,首先判断发稿前2-4个白天时刻数据的阅读增量占总阅读的比率rate是否有大于或等于0.3的数据,若有,则判断白天非自然增长形态的或然率为0.8,即C_d=0.8,若没有,则C_d=去掉白天前4个时刻和夜间的其它时刻的阅读增量比率最大值;

[0026] 最后,根据公式$C_2=C_n+C_d$得到本次计算的非自然增长形态的或然率C_2;若C_2大于或等于1,则都取为1.0,若C_2小于1,则取其实际值;

[0027] 所述步骤(6)中第三次计算的具体方法为:

[0028] 根据第二次计算的ID列,即r_Δ值,对相邻阅读增量进行差值处理,即后一个r_Δ减去前一个r_Δ之差,得到相等时间段下的增量差y2,选出y2列前5个点的最高值max,再从y2的第五个点以后的所有点中找出所有满足y2>max/3这一范围的值,该值至少有两个,如只出现

一个则视为自然增长形态，不在此算法识别范围内，求出这几个数的平均值P，则本次计算的非自然增长形态的或然率C_3的计算公式为：

[0029] $C_3=((P-max/3)/(max/2-max/3))=(6P-2max)/max$

[0030] 若C_3大于或等于1，则为非自然增长形态，当C_3介于(0,1)之间则存在非自然增长形态的可能性，当$C_3=0$，则为自然增长形态；

[0031] 所述步骤(7)中第四次计算的具体方法为：

[0032] 根据第二次计算所得数据列(ID,t,y,rate)、ID列即r_Δ值，以及处理后的数据Y即R，根据以下公式计算得到阅读曲线斜率角度：

[0033] $$degree=\frac{\tan^{-1}(\frac{y}{\max(Y)\cdot 0.06})\cdot 180}{\pi}$$

[0034] 对degree四舍五入成整数degree1，首先判断当degree1大于或等于20，且相邻两个degree1的差值小于3度时，将二者的位置和数值分别记录在数据框location_1和value_1中，Value_1的列数i则为连续出现相近阅读增量的最大次数，当i>=6时，则cheat_line_1=1.0，当i=5时，则cheat_line_1=0.8，当i=4时，则cheat_line_1=0.5，当i<=3时，cheat_line_1=0.0；

[0035] 记录调整后degree1出现连续小于等于3度的位置(location_2)和值(vaue_2)，Value_2的列数j则为连续出现斜率角度小于等于3度的最大次数，根据j得出其首次出现小于等于3度的度数的位置k，k即为阅读量不再出现大幅度增长的最早时刻，若K>=24，则cheat_line_2=0.0，否则cheat_line_2=(24-k)/24，根据以下公式本次计算的非自然增长形态的或然率C_4：

[0036] $C_4=$cheat_line_1+cheat_line_2

[0037] 当$C_4>=1.0$时，则取$C_4=1.0$；当$C_4<1.0$时，则$C_4=$cheat_line_1+cheat_line_2；

[0038] 所述步骤(8)中非自然增长形态的综合或然率C_f的具体计算方法为：

[0039] 取四种算法的最高分并赋予权重0.8，再求出其余算法得分的均值并赋予权重0.2，则四种算法的综合得分即非自然增长形态的综合或然率C_f的计算公式为：

[0040] $$C_f=0.8*\max(c_n)+0.2*\frac{\sum_1^4 c_n-\max(c_n)}{3}$$ 。

[0041] 本发明的有益效果在于：

[0042] 本发明通过对海量的自然增长的阅读曲线进行分析，总结出自然增长阅读曲线的规律，建立曲线模型，并根据模型找出样本曲线中不符合自然增长曲线模型的点和时间段，计算出曲线非自然增长的或然率，即电子文件阅读数非自然增长的或然率，从而为精确判断电子文件阅读数非自然增长概率打下坚实的基础，实现对日益增长的电子文件阅读数非自然增长进行较为准确的监测和识别，有利于助推网络市场的健康发展。

具体实施方式

[0043] 下面以对海量公众号的监测和识别为例，对本发明的具体方法进行说明：

[0044] 一种电子文件阅读数非自然增长识别方法，用于对海量公众号的阅读数进行识别，包括以下步骤：

[0045] (1)原始数据采集及阅读曲线绘制:对海量公众号的实时阅读数进行实时监测,实时监测的时间间隔可以为一个或多个,这里优选两个时间间隔,10分钟和1个小时,定时采集相应的阅读数,最终绘制出电子文件的实时阅读曲线;

[0046] 原始数据样例如下:

[0047]

更新时间（T）	阅读量（R）	点赞量（Z）。
2015-10-16 15:00:00	1748	85
2015-10-16 16:00:00	4604	88
2015-10-16 17:00:00	4626	88

[0048] 上述数据还列出了点赞量Z,表示本方法还可以对点赞量进行监测和识别。

[0049] (2)数据预处理:通过数据归整和清洗,将原始数据处理为每个相同时间间隔的时间点均有数据与之对应的序列,最终得到包括序列X、更新时间T和阅读数R这三列的数据M;

[0050] 本步骤存在的理由是:原始数据存在的问题:阅读数更新时间点不规范,如以10分钟间隔更新的阅读数,更新时间点可能为13分;时间点可能出现重复或断点,如同一时间点可能出现两条以上记录或某一时间点无数据。

[0051] 下面举例说明上述数据预处理的具体处理流程:

[0052] 对原始数据更新时间点进行标准化处理,即对更新时间点进行四舍五入处理,使其间隔时间为标准的10分钟或60分钟,如更新时间点为12:12:53,经标准化处理后为12:10:00,14:18:35经标准化处理后为14:20:00;

[0053] 对重复记录进行剔重处理,保留阅读数最大值,对断点采用均值补缺法补齐数据,即断点处阅读量取最近上一时间和下一时间的中间值,算法如下:

[0054] 时间序列中连续两条记录为:2015-10-15 14:12:00|2015-10-15 14:33:00,则中间出现断点,缺失2015-10-15 14:20:00的记录;

[0055] 模型建立:

[0056] 最近上一时间点:t_p

[0057] 最近上一时间点阅读量:r_p

[0058] 最近下一时间点:t_n

[0059] 最近下一时间点阅读量:r_n

[0060] 时间间隔(分):t_s[10,60]

[0061] 断点个数:M

[0062] 断点时间点:t_{bm}

[0063] 断点阅读数:r_{bm}

[0064] 计算过程:

[0065] $$M = \frac{t_n - t_p}{t_s} - 1$$

[0066] $t_{bm}=t_p+m*t_s$ $m\in[1,M]$,m为自然数

[0067] $r_{bm} = \text{round}(\frac{t_n - t_p}{M+1}) + t_p;$

[0068] 按时间序列顺序排序，以自然数列[1,2,3...]对每一行进行序列编号。

[0069] 经过上述处理，最终数据M有四列：

[0070]

序列（X）	更新时间（T）	阅读量（R）	点赞量（Z）。
1	2015-10-16 15:00:001748	85	
2	2015-10-16 16:00:004604	88	
3	2015-10-16 17:00:004626	88	。

[0071] (3)对已绘制的阅读曲线进行趋势分析、特征分析，最后进行非线性拟合，同时将横坐标的时间转换为自然数序列，建立实时曲线模型如下：

[0072] $$y = a - \frac{b}{e^{\frac{cx}{600}}} \qquad c \in \{10,60\}$$

[0073] 其中c代表时间间隔。

[0074] (4)第一次计算：利用实时曲线模型对坐标系X-Y进行拟合，根据拟合度计算出本次计算的非自然增长形态的或然率C_1；这里的X代表数据预处理后的序列，Y代表阅读数R；

[0075] 上述第一次计算的基本思想为：通过大数据分析，建立自然增长形态的数据模型，将样本与此模型进行比对，通过匹配的程度判断非自然增长形态的或然率。

[0076] 上述第一次计算的具体方法为：

[0077] 先求出将$X1 = \frac{1}{e^{\frac{cx}{600}}}$一项视为自变量X1，然后利用线性方程y＝a-b*X1拟合求出a、b和预测值Y′，并计算得出曲线拟合度R^2，其计算公式为：

[0078] $$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i)^2}$$

[0079] 根据以下公式计算出本次计算的非自然增长形态的或然率：

[0080] $$C_1 = \frac{1}{1+\exp(10*R^2-6)}。$$

[0081] (5)第二次计算：依次判断夜间非自然增长形态和白天非自然增长形态，并根据夜间非自然增长形态和白天非自然增长形态计算得到本次计算的非自然增长形态的或然率C_2；

[0082] 上述第二次计算的基本思想：将样本阅读曲线按时段分为白天与夜间两部分，分别对这两个时段与模型进行比对，如果在各自时段内不符合模型的规律，则通过匹配的程度判断非自然增长形态的或然率；

[0083] 上述第二次计算的具体方法为：

[0084] 如果处理后的数据M包括两个或两个以上时间间隔的数据，则数据M中取出其中一

个时间间隔的数据，组成新的数据；首先求出相邻时间点的阅读数R的差值，此为相等时间间隔的阅读增量r_{Δ}，即后一个阅读数减去前一个阅读数的差，由此值形成ID列，从而得到相等时间间隔阅读增量占总阅读的比率rate＝r_{Δ}/max(R)及其对应的时间点t和序列号x，并得到数据列(ID,t,y,rate)，其中y即为r，表示阅读量R中的元素；首先判断夜间非自然增长形态：设置电子文件发布后第一个凌晨从2:00至7:00，且其阅读增量比率阀值为3％，此后凌晨1:00至7:00的阀值为1.5％，若电子文件第一个凌晨2:00至7:00和其它凌晨1:00至7:00的阅读增量N的比率rate超过对应的阀值，则将其记录在向量H中，根据以下公式得到夜间非自然增长形态的或然率：

[0085] $$C_n = \frac{\sum_{i=1}^{m} H_i}{\sum_{i=1}^{k} N_i}$$

[0086] 其次判断白天非自然增长形态：去除电子文件凌晨时刻的数据，首先判断发稿前2-4个白天时刻数据的阅读增量占总阅读的比率rate是否有大于或等于0.3的数据，若有，则判断白天非自然增长形态的或然率为0.8，即C_d＝0.8，若没有，则C_d＝去掉白天前4个时刻和夜间的其它时刻的阅读增量比率最大值；

[0087] 最后，根据公式C_2＝C_n+C_d得到本次计算的非自然增长形态的或然率C_2；若C_2大于或等于1，则都取为1.0，若C_2小于1，则取其实际值。

[0088] (6)第三次计算：结合第二次计算的数据，对相邻阅读增量进行差值处理，得到本次计算的非自然增长形态的或然率C_3；

[0089] 上述第三次计算的基本思想为：将样本的增长速率与此模型进行比对，如果数据异常，与模型差异较大，即存在大起大落，或瞬时增量非常大的情况，则存在非自然增长形态；

[0090] 上述第三次计算的的具体方法为：

[0091] 根据第二次计算的ID列，即r_{Δ}值，对相邻阅读增量进行差值处理，即后一个r_{Δ}减去前一个r_{Δ}之差，得到相等时间段下的增量差y2，选出y2列前5个点的最高值max，再从y2的第五个点以后的所有点中找出所有满足y2>max/3这一范围的值，该值至少有两个，如只出现一个则视为自然增长形态，不在此算法识别范围内，求出这几个数的平均值P，则本次计算的非自然增长形态的或然率C_3的计算公式为：

[0092] C_3＝((P-max/3)/(max/2-max/3))＝(6P-2max)/max

[0093] 若C_3大于或等于1，则为非自然增长形态，当C_3介于(0,1)之间则存在非自然增长形态的可能性，当C_3＝0，则为自然增长形态。

[0094] (7)第四次计算：结合第二次计算的数据，计算阅读曲线斜率角度，最后得到本次计算的非自然增长形态的或然率C_4；

[0095] 上述第四次计算的基本思想为：将样本的增长量与此模型进行比对，如果增长过于均匀，则存在非自然增长形态；

[0096] 上述第四次计算的具体方法为：

[0097] 根据第二次计算所得数据列(ID,t,y,rate)、ID列即r_{Δ}值，以及处理后的数据Y即R，根据以下公式计算得到阅读曲线斜率角度：

[0098] $$degree = \frac{\tan^{-1}(\frac{y}{\max(Y)\cdot 0.06})\cdot 180}{\pi}$$

[0099] 对degree四舍五入成整数degree1，首先判断当degree1大于或等于20，且相邻两个degree1的差值小于3度时，将二者的位置和数值分别记录在数据框location_1和value_1中，Value_1的列数i则为连续出现相近阅读增量的最大次数，当i>＝6时，则cheat_line_1＝1.0，当i＝5时，则cheat_line_1＝0.8，当i＝4时，则cheat_line_1＝0.5，当i<＝3时，cheat_line_1＝0.0；

[0100] 记录调整后degree1出现连续小于等于3度的位置(location_2)和值(vaue_2)，Value_2的列数j则为连续出现斜率角度小于等于3度的最大次数，根据j得出其首次出现小于等于3度的度数的位置k，k即为阅读量不再出现大幅度增长的最早时刻，若K>＝24，则cheat_line_2＝0.0，否则cheat_line_2＝(24-k)/24，根据以下公式本次计算的非自然增长形态的或然率C_4：

[0101] C_4＝cheat_line_1+cheat_line_2

[0102] 当C_4>＝1.0时，则取C_4＝1.0；当C_4<1.0时，则C_4＝cheat_line_1+cheat_line_2。

[0103] (8)根据上述四次计算得到的非自然增长形态的或然率，得到最终的非自然增长形态的综合或然率C_f；其具体方法为：

[0104] 取四种算法的最高分并赋予权重0.8，再求出其余算法得分的均值并赋予权重0.2，则四种算法的综合得分即非自然增长形态的综合或然率C_f的计算公式为：

[0105] $$C_f = 0.8 * \max(c_n) + 0.2 * \frac{\sum_1^4 c_n - \max(c_n)}{3}$$ 。

[0106] (9)根据非自然增长形态的综合或然率C_f判断电子文件阅读数非自然增长概率，C_f值越大，电子文件阅读数非自然增长概率越高，反之越低；针对自然增长形态的判断，结论并非“是”与“否”，在实际运用中，可结合具体的业务模型，利用决策树算法对结果进行处理。

[0107] 上述实施例只是本发明的较佳实施例，并不是对本发明技术方案的限制，只要是不经过创造性劳动即可在上述实施例的基础上实现的技术方案，均应视为落入本发明专利的权利保护范围内。