基于用户兴趣分析的网页生命周期建模

王勇[1]，刘奕群[1]，张敏[1]，马少平[1]，茹立云[2]

（1. 智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹），清华大学计算机系，北京 100084；
2.搜狐公司研发中心，北京 100084）

摘 要： 网页在其生命周期内的活跃程度会随时间发生变化。有的网页只在特定的阶段有价值，此后就会过时。从用户的角度对网页的生命周期进行分析可以提高网络爬虫和搜索引擎的性能，改善网络广告的效果。利用一台代理服务器收集的网页访问量信息，我们队网页的生命周期进行了研究，给出了用户兴趣演变的模型。这个模型有助于更好地理解网络的组织与运行机理。

关键词： 用户行为分析；网页生命周期；网络日志挖掘

Modeling Lifetime of Web Pages Based on User Interest Analysis

Wang Yong[1], Liu Yiqun[1], Zhang Min[1], Ma Shaoping[1], Ru Liyun[2]

(1. State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084;
2. Sohu Inc. R&D center, Beijing 100084)

Abstract: The activeness of a web page varies during its lifetime. Some pages are valuable only in a specific period, and then become obsolescent. Web page lifetime analysis from users' perspective is important to enhance the performance of web crawlers and search engines, and to improve the efficiency of web advertising. With page view data collected by a proxy server, we were able to perform large scale analysis in web page lifetime. A model is given to describe user interest evolution based on an experiment conducted with the page view data of more than 36,000,000 web pages for two months. The model is the foundation to better understand how the web is organized and operates.

key words: user behavior analysis; web page lifetime; web log mining

基金资助：得到国家重点基础研究（973）(2004CB318108)、自然科学基金（60621062, 60503064, 60736044）和 863 高科技项目（2006AA01Z141）资助

作者简介：王勇(1982-)，男，山东济南人，硕士 email:wang-yong05@mails.tsinghua.edu.cn

1 简介

最近一段时间，互联网上的网页数量飞速膨胀。尽管有人怀疑这种指数性的增长能否维持[1]，但在可预见的未来一段时间内，这种趋势仍将维持。网页质量的衡量标准对于用户选择网页起到了很重要的作用。当前广泛采用的网页质量衡量标准大都是基于链接的，比如 PageRank [2]和 HITS [3]。实践验证这些算法能够收到较好的效果，但是仍然有一些问题。第一：它反映的是网站管理员的喜好，而不是最终用户的，而且很容易被垃圾误导，因为垃圾制造者很容易制造出一些指向垃圾页面的链接。第二，与存在了很长时间的网页相比，只有少量链接指向新发布的网页，新网页的价值容易被低估。第三，很多新闻类的网页的质量随着时间波动，基于连接的算法无法迅速体现出这种变化。上述不足都可以通过对用户行为进行分析而得到弥补。首先，用户行为数据能够直接反应用户的喜好，避免受到网站管理员偏向的影响。用户行为数据是由无数用户的行为积累而成，垃圾制造者很难施加影响。其次，用户行为能够立刻发送到服务器，服务器可以以比较小的延迟获取用户的即时行为。用户行为对于网络的变化反应迅速，可以引导服务器发现新出现的受关注的网页。第三，每个用户动作都记下了时间戳，能够体现出网络的动态性。如果一个网站或者一个网页的质量提高了，它就能吸引到更多的用户。

一般说来，被大量用户频繁访问的网页质量比较高。所以网页的质量可以利用用户关注的程度进行评估，而不是链接关系。网页质量并非总是一成不变。一个报道总统选举结果的网页在发布之初质量很高，但是一周之后，当所有人都知道选举结果的时候，很少有人会再次访问这个网页，这个网页的价值就大大降低了。所以动态的网页质量评估是很重要的。每天都有大量网页出现。与此同时，每天也有大量网页过时、消亡。网络爬虫应该避免下载过时的网页，搜索引擎也不需要把这样的网页收入索引，广告商也没有必要在这样的网页上发布广告。对网页生命周期的研究有助于区分过时的网页。在本文中，我们关注用户对于网页的兴趣的演变过程，找出这种演变的一些特征，并利用这些特征对网页是否有时效性进行分类。

大多数研究网页生命周期的人认为，当一个网页被发布出来，它就诞生了，当它被修改或者删除时，它就死亡了[4][5][6][7]。与这种网页生命周期的定义不同，我们认为从用户的角度定义网页生命周期更加合理，因为仅当一个网页被用户访问时，它的存在才是有意义的，它的重要性应该由吸引到的用户的注意力来衡量。

首先提出一些定义来描述网页生命周期。一个网页被发布出来的时间是它的发布日；它首次被访问的时间是它的激活日；如果在足够长的一段时间里，没有人访问这个网页，那么这段时间的第一天是它的休眠日；它被从服务器上删除的时间是它的死亡日。网页的实际生命是其激活日到休眠日之间的一段时间。用户的访问是网页存活的唯一标志。过去通过一个网页是否可被访问来判断其存活性，是因为这样比较容易收集数据。但是现在，搜索引擎可以收集起大量的用户访问信息，使得从用户的角度对网页的生命周期进行分析成为可行。

网页的访问量与用户兴趣密切相关，而且是由后者决定的。如果用户对于一个网页非常感兴趣，它的访问量就会很高。用户兴趣是一个相对抽象的概念，无法直接获取。然而，它可以被访问量反映出来。网页的访问量在其生命周期内波动，用户兴趣曲线可以从访问量数据中获得。

本文组织大纲如下：第二节描述了基于传统定义对于网页生命周期的研究工作。第三节

给出获取用户兴趣曲线的方法。后面的部分是用户兴趣的应用。第四节基于用户兴趣演变的不同特征，把网页分为两类：有时效性网页和无时效性网页。第五节探讨有时效性的网页的一些特征。最后一部分是结论和将来的工作。

2 相关工作

很多研究者认同当网页可以被访问，它就是存活的。他们用网络爬虫周期性地下载网页，新发现的网页被标识为刚诞生，如果一个网页无法被访问，就标识它死亡。研究者持续数个月或者数年观察互联网，得出一些结论。

网页生命周期符合对数分布[8]，这意味着互联网上的大部分网页生命很短暂，大量网页诞生后一个月之内就消亡了。然而还有少量网页的生命很长，长到几年。网页生命的半衰期（半数网页死亡所需的时间）是两个月。

[4][9][10] 假设网页被修改或者删除是随机且独立的，那么网页的寿命就是独立同分布的，修改、删除事件可以用泊松过程模拟。给定一个网页 p 和它的修改概率 λ，它的寿命小于时间 t 的概率是

$$\Pr(T \le t) = 1 - e^{-\lambda t} \qquad (t>0) \qquad (1)$$

很多网络爬虫能够自适应地调整对网页寿命的预测[10][11][12]。它们起初以相同的时间间隔反复下载网页，当发现网页发生了变化，则缩短对其寿命的预期，并以更小的时间间隔检查其更新。如果在连续的几次下载中，网页都保持不变，则延长它的预期寿命。这样一来，爬虫运行时间越久，它对于网页寿命的预测就越准确。

不同域的网页的寿命有很大不同。.com 域的网页更新最为频繁，.edu 和.gov 的网页更新明显慢得多。

大多数网页都诞生不久，部分是由于更新频繁，部分由于每天都有大量新网页发布出来[1]。在传统的网页寿命定义中，网页的状态是二值的：存活和死亡。在我们的提出定义中，网页的生命状态是连续的，其活跃程度可以用访问量来衡量。基于这个定义，我们可以研究网页的活跃程度及其演变规律。

3 获取用户兴趣曲线

用户对网页的兴趣可以分作两类：主动兴趣和被动兴趣。如果一个用户对网页感兴趣，主动去访问它，这个用户对网页有主动兴趣。如果用户预先并不知道这个网页，仅仅是因为这个网页被推荐给用户而被用户访问，这样，用户对网页有被动兴趣。导航页上的链接、搜索引擎和网页上的广告都可以把网页推送给用户。这两类用户兴趣都可以带来访问量，所以它们被通称为用户兴趣。

网页的访问量是由用户兴趣决定的，然而二者并不精确一致，访问量还会受很多其他因素的影响，比如是否是周末等等。所有的随机因素综合在一起可以看作是白噪声，对用户兴趣有影响。访问量就是由用户兴趣和白噪声叠加而成的。为了获取用户兴趣曲线，我们可以建立坐标系，其中横坐标是时间，纵坐标是访问量。给定一个网页每天的访问量数据，就在坐标系上有若干连续的点，对这些点进行拟合，就可以消除白噪声的影响，恢复出用户兴趣曲线。

一般说来，很少有用户知道一个新诞生的网页。一段时间以后，越来越多的用户知道这

个网页，对它感兴趣并且访问它。后来，有的网页过时了，无法再吸引用户访问，然而另外一些网页的内容则不会过时，它能吸引到的用户量保持相对恒定。用户兴趣变化的趋势，一般从零开始上升，到达最大值之后，要么下降，要么保持相对恒定。对数正态分布的概率密度函数可以很好地表示这两种趋势。具体做法如下：

给定一个网页 p，用函数 f(x)来拟合连续 n 天的访问量数据(x_i, y_i) (1≤i≤n, $x_i = i$ 是日期, y_i 是第 i 天的访问量)，其中

$$f(x) = A \times f_{\ln(x)} = A \times \frac{1}{\sqrt{2\pi}\sigma(x-b)} \times e^{-\frac{(\ln(x-b)-\mu)^2}{2\sigma^2}} \tag{2}$$

参数(A, b, μ, σ)可以描述在这 n 天中的用户兴趣。在第 i 天的用户兴趣可以计算为

$$z_i = f(i) = A \times \frac{1}{\sqrt{2\pi}\sigma(i-b)} \times e^{-\frac{(\ln(i-b)-\mu)^2}{2\sigma^2}} \tag{3}$$

我们收集了 75,112,357 个网页从 2006 年 11 月 13 日到 2007 年 1 月 11 日共计 60 天的访问量数据，去掉其中总访问量小于 60 的网页，还剩 975,151 个网页。用上述方法对每一个网页的访问量数据进行拟合，得到其用户兴趣曲线。图 1 和图 2 是其中两个网页的用户兴趣曲线。

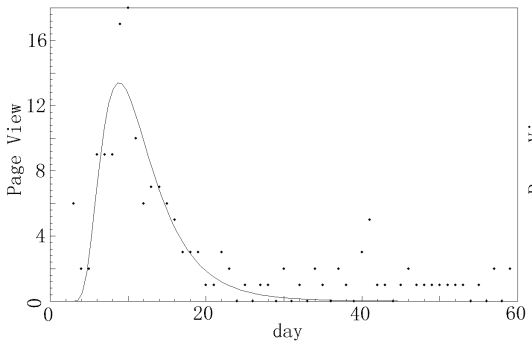


图. 1 有时效性网页用户兴趣的演变

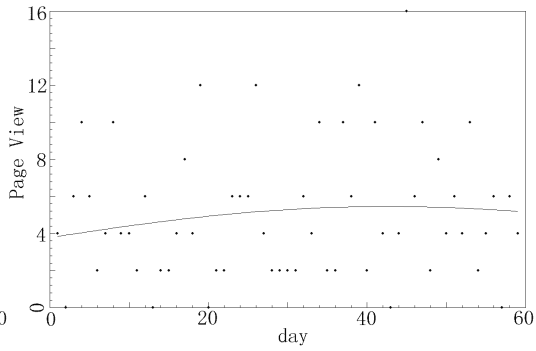


图. 2 无时效性网页用户兴趣的演变

有的网页的访问量有多个高峰。比如报道央行加息的新闻的网页在诞生之初会吸引到很多用户，然后用户减少。到了下次加息时，很多用户想查阅以前加息时的情况，这个网页又能够吸引到比较多的用户。但在本文中，我们主要关注网页在几周或者几个月内的生命状态。在这样比较短的时间内，用户兴趣一般不会有多个高峰。对数正态分布的概率密度函数足够描述没有或者只有一个高峰的用户兴趣。

4 区分有时效性的网页和无时效性的网页

如图 1、图 2 所示，有的网页的用户兴趣有明显的上升和下降，另外一些网页的用户兴趣保持相对恒定。网页可以以此分为有时效性的网页和无时效性的网页两类。这两类网页的定义如下：

有时效性的网页：用户兴趣随时间变化的网页。这类网页的内容是有时效性的，比如新闻报道。典型的有时效性的网页的用户兴趣曲线如图 1 所示。

无时效性的网页：用户兴趣不随时间变化的网页。这类网页的内容不会过时，比如介绍感冒防治的网页。典型的无时效性的网页的用户兴趣曲线如图 2 所示。

数据拟合可以将访问量数据中的噪音去掉。这步操作之后，有时效性的网页的用户兴趣曲线的拱形保留了下来，而无时效性的网页的访问量的无规律的抖动被消除了，其用户兴趣曲线近乎水平。所以用户兴趣曲线波动的程度可以作为网页是否有时效性的区分标准。给定网页在 n 天中的用户兴趣数据($z_1, z_2, \ldots, z_n$)，其相对标准差(RSD)可以描述用户兴趣波动的程度。用户兴趣相对标准差的计算方法如下：

$$RSD_p = \frac{s}{\bar{z}} = \frac{\sqrt{\sum_{i=1}^{n}(z_i-\bar{z})^2/(n-1)}}{\bar{z}} \quad (4)$$

其中 z_i 是第 i 天用户对这个网页的兴趣，S 是 z_i 的标准差，$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ 是 n 天中用户兴趣的平均值。

有时效性的网页和无时效性的网页可以根据其用户兴趣的波动程度，用一个阈值 C 进行区分。如果 $RSD_p > C$，网页 p 就是有时效性的，反之亦然。我们手动标注了 40 个有时效性的网页和 40 个无时效性的网页，计算出来它们的 RSD，得出 RSD = 3 是合适的区分两类网页的阈值。

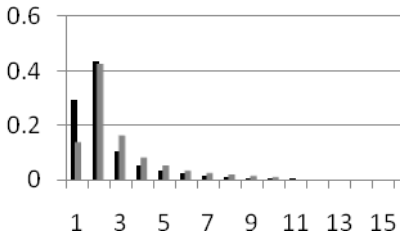
我们用上述方法对 975,151 个网页进行自动分类，其中 200,517 被分为有时效性，774,634 被分为无时效性。

网页的日均用户兴趣 $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ 可以用来评估网页的质量。有时效性的网页和无时效性的网页吸引到的用户兴趣是不一样的，统计结果显示，用户更关注无时效性的网页。在图 3 中，横坐标是日均用户兴趣，纵坐标是网页的比例，黑色矩形代表有时效性的网页，灰色矩形代表无时效性的网页。日均用户兴趣小于 2 的网页中，有时效性的网页的比例高于无时效性的网页。日均用户兴趣大于 2 的网页中，无时效性的网页的比例比较高。

5 有时效性的网页的性质

5.1 用户兴趣达到最大值所需的时间

令拟合函数的一阶导数 $f'(x) = 0$，解方程得到 $x = e^{\mu-\sigma^2} + b$，意味着有时效性的网页诞生 $e^{\mu-\sigma^2}$ 天后，用户兴趣达到最大值。对 200,517 个有时效性的网页的用户兴趣达到最大值所需时间进行统计，结果见图 4，其中横坐标是时间，纵坐标是网页的比例。



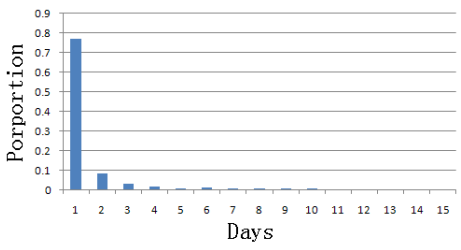


图 3 有时效性和无时效性的网页吸引的用户兴趣

图 4 有时效性的网页用户兴趣达到最大值所需的时间

图 4 中显示，大约 77%的网页在诞生之后的第一天，用户兴趣就达到了最大值，其余的网页的用户兴趣也大都在一周内达到最大值。这意味着大部分有时效性的网页的用户兴趣都是单调递减的，新的网页比老的网页更有价值。

5.2 有时效性的网页满足 80%用户兴趣所需的时间

令$\Phi_{ln}(x)$为对数正态分布函数的累积分布函数，$\Phi_{ln}^{-1}(x)$是$\Phi_{ln}(x)$的逆函数，则$\Phi_{ln}^{-1}(0.8)$是一个网页积累 80%用户兴趣所需的时间。以图 5 中用户兴趣曲线为例，b 是网页诞生的时间。在图 5 中，x 满足$\Phi_{ln}(x-b)=80\%$，灰色区域的面积是由用户兴趣曲线和 x 轴围成的图形面积的 80%。这意味着网页诞生后 x – b 天，所有会访问这个网页的用户中的 80%已经访问过了。尽管还有 20%用户将要访问它，但这些访问很稀疏地散布在网页生命的长尾中，此时的网页的价值已经不高了。x – b 的分布的统计结果见图 6。

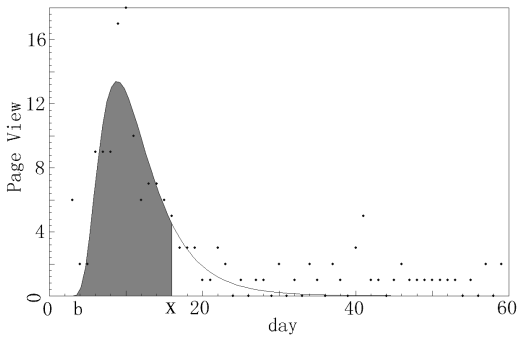


图. 5 积累的用户兴趣

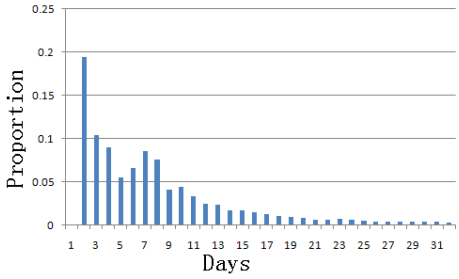


图. 6 满足 80%用户兴趣所需的时间

从图 6 中可见，有时效性的网页生命大都很短暂。超过 80%的有时效性的网页的活跃期不超过两个星期。有时效性的网页的生命符合 80-20 原理，即网页在很短的时间内获得大部分的访问量，此后尽管网页还储存在服务器上很长时间，已经很少有用户访问了。

6 结论和未来工作

用合适的函数对访问量数据进行拟合，就可以得到用户兴趣曲线。根据用户兴趣曲线变化的不同特征，网页可以分为有时效性的网页和无时效性的网页两类。有时效性的网页的用户兴趣会随时间衰减，而无时效性的网页的用户兴趣则保持相对恒定。有时效性的网页会过时，而无时效性的网页则不会。一个典型的有时效性的网页在诞生后不久用户兴趣就达到最大值，在较短的时间内获得大量的访问，此后用户兴趣逐渐下降，网页失去了价值。

将来，我们打算找出高质量网页的用户兴趣曲线的特征。我们猜测高质量的网页的用户兴趣曲线应该有一些特征，比如诞生之初就迅速达到最大值，平均用户兴趣应该比一般网页高，并且用户兴趣在在较高的水平上可以维持较长的时间。如果上述猜测是正确的，我们就可以通过用户兴趣曲线预测网页的质量。

PageRank 一般是周期性地计算的，新下载的网页在下一轮计算之前就没有 PageRank 数据，在排序的时候就会被低估。同一个网站发布的网页很有可能会有类似的用户兴趣曲

线。用一个网站已有网页的用户兴趣曲线，可以得到这个网站的特性，由此预测新发布的网页的用户兴趣曲线，进而根据它的年龄得到它当前的质量。

很多用户访问搜索引擎来访问新闻类的网页。假如新发布的新闻网页没有被搜索引擎收录，用户就会转到其他搜索引擎。为了满足用户的这种需求，网络爬虫应该监视新闻网站，在新闻网页发布后的第一时间把它下载下来。现在的新闻网站列表大都是手动生成的，难以避免主观性，而且对于新出现的新闻类的网站也不够敏感。对有时效性的网页和无时效性的网页进行自动分类之后，我们就可以得到一个网站中有时效性的网页的比例，如果这个比例高于某个阈值，这个网站就是新闻类的网站。这样就可以自动地生成新闻网站列表，供网络爬虫应该重点监视。

7 参考文献

[1] Brewington B, Cybenko G. How Dynamic is the Web[A], Proceedings of WWW9 –9th International World Wide Web Conference (IW3C2), pp. 264-296.

[2] Page L, Brin S, Motwain, R., and Winograd T. The Pagerank Citation Algorithm: Bringing Order to the Web[A]. In 7th World Wide Web Conference (1998).

[3] Kleinberg J. Authoritative sources in a hyperlinked environment[A]. In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 668, 1998.

[4] Dhyani D, Wee K, and Bhowmick S. A Survey of Web Metrics[A], ACM Computing Surveys, 2002

[5] Edwards J, McCurley K, Tomlin J. An Adaptive Model for Optimizing Performance of an Incremental Web Crawler[A]. Proceedings of the 10th international conference on World Wide Web

[6] Dalal Z, Dash S, Dave P, et al. Managing Distributed Collections: Evaluating Web Page Changes[A]. Movement, and Replacement, 2004 Joint ACM/IEEE Conference

[7] Ashman H. Electronic Document Addressing: Dealing with Change[A]. ACM Computing Surveys, 2000

[8] Gomes D. Modelling Information Persistence on the Web[A]. Proceedings of the 6th international conference on Web engineering

[9] Fetterly D, Manasse M, Najork M, Wiener JL. A Large-Scale Study of the Evolution of Web Pages[A]. Proceedings of WWW03, pp. 669-678.

[10] Cho J, Garcia-molina H. Effective Page Refresh Policies for Web Crawlers[A]. ACM Transactions on Database Systems

[11] Cho J, Garcia-Molina J. Estimating Frequency of Change[A]. ACM Transactions on Internet Technology, 2003

[12] Najork M, Heydon A. High-Performance Web Crawling[A]　Kluwer Academic Publishers Norwell,2002

[13] Ntoulas A, Cho J, Olston C. What's New on the Web? The Evolution of the Web from a Search Engine Perspective[A]. Proceedings of the 13th international conference on World