基于用户浏览图的网页质量评估方法的比较分析*

薛宇飞　刘奕群　张敏　马少平　茹立云
智能技术与系统国家重点实验室
清华大学信息科学与技术国家实验室
清华大学计算机系　北京 10084
E-mai: xueyufei@126.com

摘　要：面对海量繁杂的网络数据环境，网页质量评估成为互联网搜索引擎面临的主要技术挑战之一，当前针对互联网网页评估的主要研究思路是基于网络超链接结构的分析完成。然而，Web2.0、搜索引擎结果优化(SEO)，网络作弊等现象的出现严重影响了互联网超链接分析的可靠性。为此，基于用户互联网访问日志构建用户浏览关系图成为互联网网页质量评估的重要研究方向。本文基于海量规模真实网络用户行为数据和网页质量评估数据，对基于用户浏览关系图结构分析的几种主要网页质量评估算法进行了比较与分析，实验结果说明，将传统链接结构分析算法应用于用户浏览关系图，可以取得较好的网页质量评估效果。

关键词：用户浏览图，PageRank，TrustRank，BrowseRank

Analysis and Comparison to Web Page Quality Evaluation Algorithms Based on User Browsing Graph

XUE Yufei, LIU Yiqun, ZHANG Min, MA Shaoping, RU Liyun
State Key Lab of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China P.R.
E-mai: xueyufei@126.com

Abstract: Web page quality analysis is one of the top challenges for practical commercial search engines. State of the art quality analysis algorithms are mostly based on hyperlink structure analysis. However, the this kind of algorithm doesn't work well due to the change in Web structure caused by Web 2.0, search engine optimization (SEO) and Web spam. Therefore, user browsing graph has been paid much attention to as an important way in page quality estimation. In this paper, we compare 3 kinds of link analysis algorithms (PageRank, TrustRank and BrowseRank) on a user browsing graph which is constructed with large scale Web access log data. Experimental results show that traditional link analysis algorithms perform well on user browsing graph for the task of page quality estimation.

Keywords: User browsing graph, PageRank, TrustRank, BrowseRank

* 本文相关工作得到国家重点基础研究（973）(2004CB318108)、自然科学基金（60621062, 60503064）和863 高科技项目（2006AA01Z141）资助.

1 引言

对网页质量进行评估是搜索引擎系统中一项非常重要，同时也很具有挑战性的工作[5]。网页质量的评估结果不仅会影响搜索结果的排序，同时也会用于指导搜索引擎系统中的索引以及网页抓取。

目前，网页质量评估的主要方法都是基于对互联网上的链接结构分析的。PageRank[1]、TrustRank[2]、HITS[3]等评估方法的广泛应用都证明了通过链接结构分析进行网页质量评估是一种切实可行的方法。这些基于网页链接分析的评估都用到了 Web 上的两个假设：一是推荐假设，即若网页 A 上有一个链接指向网页 B，即表示网页 A 的作者认为网页 B 的内容是值得推荐的；另一个是内容相关假设，即 A、B 两个页面之间存在链接关系，表明这两个页面的内容在某种程度上存在关联。目前商业搜索引擎以及研究界所使用的链接分析算法，都是基于这两个假设的。

随着 Web 2.0 的发展，互联网上的信息量迅速增长，信息来源日趋多样。SNS 网站、博客等新的 Web 形式不断涌现，Web 的主要功能已经不再仅仅是传播知识和信息，而在很大程度上成为了一个表达个人情感，沟通人际关系的平台。这就使得链接的推荐假设和内容相关假设越来越不可靠。

同时，由于基于推荐假设的网页质量评估的方法被广泛使用，互联网上出现了越来越多的垃圾网站。许多垃圾网站之间通过大量的相互链接，企图通过这样的作弊手段，使它们在网页质量评估算法中获得更好的结果，以在搜索引擎的检索结果中得到更高的排名。大量的垃圾信息成为 Web 环境面临的一个重要的问题，也对搜索引擎技术提出了巨大的挑战。搜索引擎必须通过改进页面质量评估的方法，解决大量的恶意链接对基于链接分析的页面评估算法带来的干扰。

近年来，Web 研究者越来越重视对用户行为的分析[6][7][8]。从对大量用户行为数据的统计中，可以总结出用户对不同页面的偏好程度，这种信息同样可以反映出一个页面的质量。基于这个观点，[4]提出了基于用户浏览关系图（User Browsing Graph）的概念，在此关系图中，只有曾经被用户访问过的超链接关系和网页节点才会被记录。基于用户浏览图，他们提出一种新的页面质量评估的算法，称为 BrowseRank[4]。实验结果证明，基于用户浏览关系图的 BrowseRank 算法能够比在原始链接关系图上运行的 PageRank 算法取得更好的页面质量评估效果。

与 PageRank 算法相比，BrowseRank 除了使用了不同的网页关系图外，还在模型上有一定的区别。PageRank 算法模拟用户在网页之间不断浏览的过程，但并不考虑用户在网页上停留时间的长短，而只关心用户会跳转到哪里。BroweRank 算法中，使用了连续时间的马尔可夫过程模型，除了考虑用户跳转的去向外，还考虑了用户在页面上的平均停留时间。此外，在 BrowseRank 算法中还使用了网页的重置概率（Reset Probability）信息。它表示一个页面被用户在浏览器地址栏中键入的概率。

随后，[9]的工作对用户浏览关系图进行了更深入地分析，他们认为从用户浏览的数据中提取的页面关系图比原始链接关系图更适用于网页质量评估任务。由于经过了用户行为的过滤，这个图要比链接关系图规模减小，记录的信息也更加准确。

综上所述，用户点击关系图比传统的链接关系图更符合链接结构算法所基于的推荐假设和内容相关假设，而基于用户点击关系图专门设计的 BrowseRank 算法也能够取得更好的页面质量评估效果。但是，这些工作中并没有解决这种评估效果的提升是来自于 BrowseRank 的算法改进，

还是来自于用户点击关系图的问题。

在本文的研究中，我们希望通过对大规模网络用户行为数据的分析，解答以下问题：BrowserRank 算法和基于用户浏览图的 PageRank 算法效果有着多大的差异？用户浏览图的使用和连续时间马尔可夫过程模型的引入，是如何影响 BrowseRank 算法的结果的？

2 用户浏览记录、用户浏览图与链接关系图

近年来，许多商业搜索引擎都通过开发浏览器工具条的方式吸引更多的用户。Google、百度、Live Search、搜狗等搜索引擎都有相应的产品。用户可以在工具条上直接输入查询的关键词，并获得检索结果。同时，这些工具条还提供屏蔽弹出的广告窗口或访问加速功能。为了向用户提供增值服务，搜索引擎的工具条也在用户许可的情况下匿名地收集用户的浏览历史。

用户浏览历史会被记录在服务器端的日志中，一般每条日志中包含以下内容如表 1 所示：

表 1 用户浏览历史日志的内容

日志记录字段	说明
Session ID	随机生成的用户的浏览器会话的 ID
Source URL	用户点击的链接所在的网页 URL
Destination	用户点击的链接所指向的新页面 URL
Time Stamp	用户点击链接的时刻

从记录以上内容的日志中，我们可以获得用户浏览的情况的详细记录。根据这些记录，我们能构造出这样一个有向图：以网页为结点，以用户从源地址到目的地址的点击作为边，以用户点击次数作为边的权值。同时，通过同一用户在相邻访问的两个网页的时间之差，还可以得到用户在前一网页上停留的时间；还可以通过源地址的情况计算用户访问的目标页面的重置概率。利用这些信息，我们可以计算网页的 BrowseRank 值，也可以使用其中一部分信息，在一个有向无权图上按照 PageRank 或 TrustRank 算法计算相应的重要性值。

在我们的前期工作[9]中，我们尝试比较了通过用户浏览历史构建的用户浏览图与相应的链接关系图的差异。他们的工作是在同一搜索引擎抓取到的网页的链接关系图中，将用户浏览图中出现的结点所组成的子图抽取出来，与用户浏览图对比。对比发现，用户浏览图中的边数仅为链接关系图边数的 7.6%。而用户浏览图中的边只有约 1/4 在链接关系图中出现了。而链接关系图中超过 98%的边在用户浏览图中没有被点击过。以上分析都表明，通过用户浏览历史获得的用户浏览图与通过爬虫抓取到的链接关系图有着显著的差异。

工作[9]还研究了用户浏览图随时间变化的特点。由于 Web 上每天都会有大量的新网页产生并被用户浏览，因此随着使用的用户浏览记录时间段的延长，用户浏览图会不断扩大。工作[9]中的实验是从某天开始使用逐天增加的用户浏览历史日志构建用户浏览图，观察图每天的变化。通过观察发现，最初几天，图中每天新出现的结点和边较多。从大约第 15 天开始，每天新出现的边和结点的数量占总数量的比例趋于固定。因此，如果想得到比较可靠的链接关系图，至少应该使用两个星期以上的用户浏览历史日志。

3 PageRank 算法与 BrowseRank 算法

PageRank 和 BrowseRank 都是基于以网页为结点，链接为边的图的网页质量评估算法。它们无论在思想上还是在算法本身上都有相似之处，但同时也有许多不同。

PageRank 算法的思想中，有几个基本的假设：一是第一节中所说的链接的推荐假设；二是马尔可夫性质假设，即用户在浏览过程中访问的下个页面只与当前页面有关，与之前曾经访问过的页面无关；三是时间齐次性，即用户的浏览行为与浏览动作发生的时间点无关。

BrowseRank 算法摒弃了在 Web 2.0 时代已经不够可靠的推荐假设，但是仍然延用了后面两个假设。同时，由于 BrowseRank 使用的是用户访问的历史数据，因此它还假设不同的用户浏览会话的访问行为是独立的。

PageRank 算法的直观表达是模拟一个用户在 Web 上不断地随机浏览的过程。而一个网页的 PageRank 值则是表示当这样的过程所进行的时间趋于无穷时，用户当前恰好在访问这个页面的概率。在 PageRank 的模型中，这个值只与网页链接图有关，而与用户具体的访问行为无关。

BrowseRank 算法为了更好地通过用户行为估计页面的质量，使用了连续时间马尔可夫过程模型。算法中通过一个考虑噪声的模型来估算用户在各个页面上的平均停留时间，以此判别页面好坏的一个依据。另外，BrowseRank 算法中引入了重置概率的概念，基本想法是，一个 URL 被用户在地址栏直接输入的次数越多则这个页面的质量较高的可能性越大。

4 实验结果

我们使用了 2008 年 9 月至 10 月间 30 天的用户浏览历史日志，构建了一个用户浏览图。利用这个用户浏览图，我们实现了网站级别 BrowseRank 算法和基于用户浏览图的 PageRank 和 TrustRank 算法。随后对这些网页质量评估的结果作了对比测试。

第一项测试是二分类问题中常用的 ROC/AUC 测试，用于评价这些算法的结果是否能很好地区分出高质量的网站以及垃圾网站。第二项测试是多个有序网站对顺序比较。

4.1 ROC/AUC 测试

我们的测试数据是两个标注人员标注过的 2630 个网站。我们从标注过的网站中提取了被标为高质量、低质量、垃圾网站、非法网站的部分，并按标注时间的先后分成了三个子集。在三个集合上，我们分别进行了两种分类方式的测试。第一种方式是将标注为高质量的作为一类，其余作为另一类；第二种方式是将未被标注为垃圾网站的作为一类，垃圾网站为另一类。通过这样的测试，分别比较不同的算法对高质量网站和垃圾网站的区分能力。

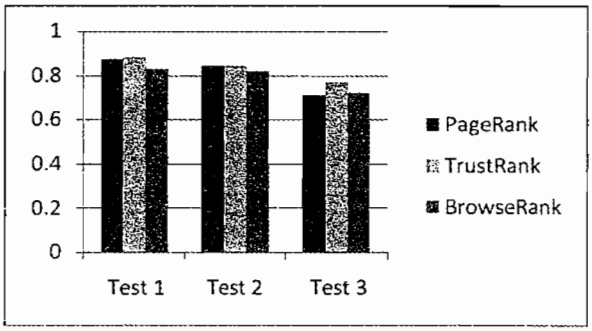
区分高质量网站的 ROC 曲线测试得到的 AUC 值如表 2 及图 1 所示：

表 2 区分高质量网站的 ROC/AUC 测试结果

	PageRank	TrustRank	BrowseRank
Test 1	0.8695	0.8807	0.8263
Test 2	0.8441	0.8417	0.817

	PageRank	TrustRank	BrowseRank
Test 3	0.7078	0.7662	0.7214

图 1 区分高质量网站的 ROC/AUC 测试结果

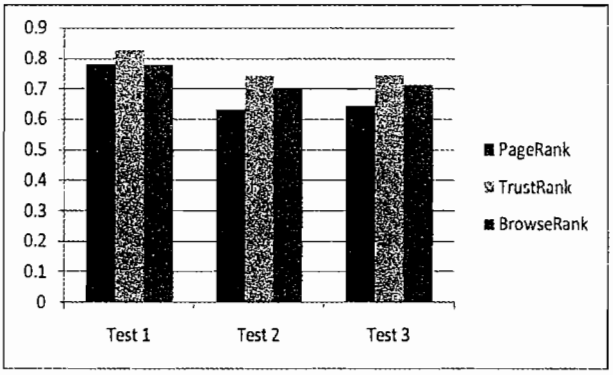


针对垃圾网站的区分测试结果如表 3 和图 2 所示：

表 3 区分垃圾网站的 ROC/AUC 测试结果

	PageRank	**TrustRank**	**BrowseRank**
Test 1	0.7803	0.8267	0.7794
Test 2	0.6294	0.7406	0.7011
Test 3	0.6445	0.7440	0.7131

图 2 区分垃圾网站的 ROC/AUC 测试结果



在区分高质量网站的三组测试中，TrustRank 有两组表现最优，最高 AUC 值为 0.8807；而 BrowseRank 有两组表现最差。在区分垃圾网站的三组测试中，TrustRank 算法都是表现最好的，而 PageRank 算法除了在第一组测试中比 BrowseRank 略好，在其余两组测试中都是最差的，最低的 AUC 值仅有 0.6294。从以上结果中可以看出，TrustRank 算法在区分不同质量网站的任务中表现比较出色，而 BrowseRank 算法对垃圾网站的区分能力较强。而对于高质量网站的识别，PageRank 算法的效果要略优于 BrowseRank。我们可以推测，这一现象与 BrowseRank 算法本身

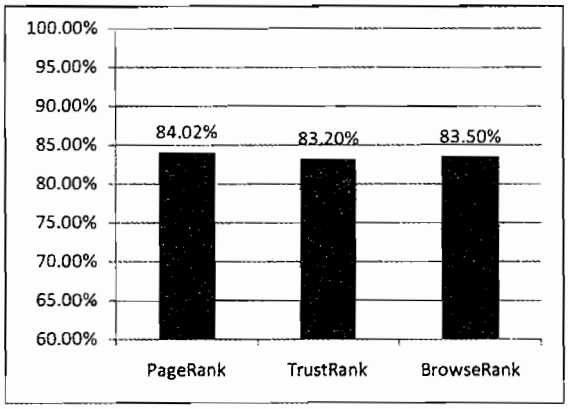
是有关系的。用户在浏览过程可能会在一些质量不高的网站停留，而在访问到垃圾网站时，则会很快离开或关掉。BrowseRank 算法中使用到了停留时间信息，因此垃圾网站会被较好地区分出来。

4.2 有序网站对测试

第二个测试是基于一些标注过的成对的网站。标注人员对 795 对网站进行了手工标注，挑出两个网站之中他们认为较好的站点。每个网站对中的两个站点都是相同领域，内容关联较大的网站，这些网站的质量都比较高。

我们对比三种算法对网站的排序与这些标注的网站对的优劣的相符程度。对与每个算法，我们都将这些网站对逐个与算法给出的网页质量评价做对比，得到各算法的评价结果在这些测试网站对上的准确率，结果如图 3 所示：

图 3 有序网站对测试结果



从图中可以看出，它们的结果非常相近，并不能明显地比较出算法的好坏。但是，虽然它们的准确率相近，但它们判断正确的样例集却有较大差别。工作[4]的结论认为 BrowseRank 排名倾向于将 Web 2.0 站点排到比较靠前的位置。在我们的成对网站测试中，我们同样发现了类似的现象。例如，BrowseRank 算法经常会将一些论坛类站点排在相同主题的门户网站之前，从而造成对网站对的错误判断。而这种情况在 PageRank 和 TrustRank 算法中很少出现。

5 结论与未来工作

从前期工作[9]中，我们发现在使用同一算法的情况下，使用用户浏览图或经用户浏览信息过滤的链接关系图比使用传统的链接图得到的效果更好。而 BrowseRank 算法的结果与在用户浏览图上使用 PageRank 算法得到的结果差异并不显著。通过对有序网站对的测试，我们可以看到，BrowseRank 算法在对质量较好的网站进行排序时，效果与用户浏览图上的 PageRank 和 TrustRank 相近。而对于垃圾网站判别的任务，BrowseRank 算法要好于基于用户浏览图的 PageRank 算法，但不如 TrustRank 算法。

BrowseRank 算法使用了比 PageRank 更多的用户浏览信息，但在测试结果中并没有表现出明

显的优势。我们考查BrowseRank算法中比基于用户浏览图的PageRank算法中多使用的信息——平均停留时间和重置概率，试图分析其中的原因：

BrowseRank中将平均停留时间作为评价网页质量评价的一个信息，可能是由于作者认为用户不会在低质量的页面上停留过长时间。但实际上，用户在网页或网站上停留的时间长短并不能直接和网页质量联系起来。用户的停留时间与网页的内容类型具有很大关系。例如，当用户在视频网站上观看视频时，他的停留时间会比较长。而当用户阅读门户上最新的新闻简讯时，停留的时间则很短。这种时间的差异并不能说明网页质量的优劣。

重置概率是BrowseRank算法中的另一项重要指标。[4]作者认为重置概率越大，相应的页面越可能是高质量页面。这个假设本身是非常合理的，但是通过搜索引擎工具栏记录的用户浏览历史，很难准确判断哪些访问是通过用户输入URL访问的。在我们所使用的用户浏览历史中，并没有明确地记录每一次访问的动作，而记录了源地址和目的地址。对于点击正在浏览的网页上的超链接而形成的浏览记录，源地址会是一个URL；而对于用户通过地址栏输入新的URL而形成的访问记录，源地址会被记录为空。但是，对于浏览器外部调用的URL，源地址也会记录为空。这种情况对于重置概率的估计会产生一定影响。

未来的工作可以从以下方面入手：

(1) 进一步研究用户浏览图的性质，寻找更有效的反映网页质量的特征，改进已有的网页质量评估算法。

(2) 研究用户在网页上的停留时间与页面类型、页面内容的关系。构建更合适的网页质量与停留时间关系的模型，并将其应用在识别垃圾网站的工作中。

参 考 文 献

[1] Brin S. and Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane.

[2] Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. 2004. Combating web spam with trustrank. In Proceedings of the Thirtieth international VLDB Conference. Vol. 30. 576-587.

[3] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of ACM 46(5), 604-632.

[4] Liu, Y., Gao, B., Liu, T., Zhang, Y., Ma, Z., He, S., and Li, H. 2008. BrowseRank: letting web users vote for page importance. In Proceedings of the 31st ACM SIGIR Conference. pp. 451-458.

[5] Henzinger, M.R., Motwani, R. & Silverstein, C. (2003). Challenges in Web Search Engines (pp. 1573-1579). In the 18th International Joint Conference on Artificial Intelligence.

[6] Fuxman, A., Tsaparas, P., Achan, K., and Agrawal, R. 2008. Using the wisdom of the crowds for keyword generation. In Proceeding of the 17th WWW Conference. 61-70.

[7] Bilenko, M. and White, R. W. 2008. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In Proceeding of the 17th WWW Conference. 51-60.

[8] Liu, Y., Cen, R., Zhang, M., Ma, S., and Ru, L. 2008. Identifying web spam with user behavior analysis. In the 4th international Workshop on Adversarial information Retrieval on the Web. AIRWeb '08. 9-16.

[9] Yiqun Liu, Yijiang Jin, Min Zhang, Shaoping Ma and Liyun Ru. User Browsing Graph: Structure, Evolution and Application. Late breaking result session in Second ACM International Conference on Web Search and Data Mining (WSDM 2009).