一种基于页面时间的排序算法

刘素芹，　硕　珺，　李兴盛，　孟令芬

(中国石油大学(华东)计算机与通信工程学院　山东 青岛 266555)

摘要：在传统排序算法的基础上，提出了增加页面时间因素的页面时间排序算法，使用户能更快地得到满意的查询结果. 通过与传统相关度算法的分析比较，证实了此方法的可行性，该算法使得检索结果在查准率方面有了较大提高.

关键词：查准率；无应答网页；页面时间排序算法

中图分类号：TP 393　　**文章编号**：1671－6841(2009)01－0015－04

0　引言

检索时往往不是因为搜集到的信息数量不够多，而是不能更快地找到合适的数据. 根据iProspect 的调查报告，2006 年高达 90%的用户只点击搜索结果页前 3 页里的结果. 根据iResearch 的《个人门户发展趋势研究报告》，高达 57.9%的网民表达了对搜索引擎结果中冗余信息多的不满. 这些数据说明，用户对搜索引擎的要求越来越高，他们希望花在寻找结果上的时间越来越少. 因此，排序查找到的结果比搜索本身更为重要，研究搜索引擎排序算法的改进逐渐成为热点问题. 随着研究的深入，越来越多的人意识到排序质量不能令人满意的原因不是网页提供的信息太少，而是可用的信息太少，或所用的信息不恰当[1].

如何解决非应答或无应答网页是现在的排序算法所应关注的重要问题. 排序所查找到结果的相关度算法就是希望能够帮助用户更快地找到更符合目标的网页，以达到更好的查准率. 然而传统的排序算法却无法解决一些基于提问式的无应答或非应答网页，不能为用户带来准确有用的结果. 作者通过在传统网页相关度算法的基础上增加页面时间因素，解决了无应答和非应答页面的问题，提高了查准率.

1　传统排序算法的局限性

1.1　传统网页的相关性排序原则

所谓相关性，就是指网站信息符合搜索条件的程度. 搜索引擎在判断关联性时，基本上是根据网页中关键词的“匹配/位置/频次”原则，即网站内容中的字词、词组或短语与用户输入的关键词越匹配，出现的次数越多，则该网站的关联程度越高，在搜索结果中排名也越靠前[2]. 相关性计算是检索系统中最重要的环节，子检索机采用多种相关性计算方法[3].

传统的信息检索大多是基于全文检索，其相关性都是基于词频统计的，即用户输入检索词句时，搜索引擎就会去找那些检索词所在的网页[4]. 然而，这样的排序算法有很大的局限性. 首先，检索词的匹配不一定就是查找到准确文档的保证，这和用户输入的检索词内容有关. 更值得注意的是，对于大量的提问应答式的网页，很可能出现检索词句完全匹配并且检索词出现在重要位置，但是没有合适的回答，不能给用户提供有用的信息. 其次，每个人都可以随心所欲地在网上发表各种内容，词频相同的 2 个网页，质量可能相差很远[4]. 而按照传统的网页排序算法，这 2 个网页的排序应该是一样的，这显然是不合理的.

1.2　基于提问式的非应答或无应答网页

在搜索引擎上搜索总会发现这样的情况，就是搜索到合适问题出现地方，却没有答复或没有合适答复，

收稿日期：2008-10-10

作者简介：刘素芹(1968—)，女，副教授，博士，主要从事计算机网络及高性能计算研究，E-mail：liusq@upc.edu.cn.

典型的就是提问无应答网页.所谓基于提问式的非应答或无应答网页,就是针对某些问题用户在某些论坛或者网站上输入问题,但是没有合适的答案或没有回答的情况.在检索时也常常发生这样的状况:因为输入的查询词句和网页中提问的问题匹配,这些提问应答式网页的排序就比较靠前,但是,很多这样的网页都是非应答或无应答网页.

1.3 用户反馈的问题

在资源选择模块中加入日志信息、反馈机制,充分利用用户反馈信息能够提高信息检索性[5].用户兴趣挖掘是个性化服务的核心,只有当用户的'个性'信息能很好地被系统'理解',才可能实现理想的个性化服务[6].但经常看见的是用户反馈基本没有起作用,原因有2个方面:一是一般的用户反馈不够人性化.用户并没有必要花费时间也没有义务评价内容,于是用户也不想花费多余的时间在关闭网页前做这样的工作,相应地有些网站也就将这些形同虚设的反馈不计入网页重要度排序了;二是这种方法不够通用和可靠,并没有要求所有人都要做,这样即使将这种反馈计入评价中,除非高级用户指定权值,否则,权值很难确定.

作者提出了一种基于页面时间的相关度排序算法,解决了基于提问式的非应答或无应答网页的问题,弥补了用户反馈的问题.通过分析和实验可以看出,此种方法不仅具有实用性和可行性,也使得排序结果具有更加优良的排序质量.同时,这种算法是按照在传统排序算法上增加页面时间的思想而提出的,可以方便地被移植.

2 页面时间算法

2.1 算法的思想

一般来说,用户只会对有价值的网页才会仔细研究内容,对于没有答复或没有合适答复的网页,用户一般会马上关闭网页.所以页面时间算法的主要思想就是有用的网页会比无用的网页的生存期长.算法中的数据定义如下.

定义1 对于一个用户检索时,确定搜索引擎抓取到的所有网页集合S,对于S中的任何一个网页用R_i表示.

定义2 用户需要检索的目标网页集合G,G中为检索到有效结果而习惯打开网页的个数用N表示,G是S的子集.

如果没有用户特别指定,默认N值为10,这是因为一般1个检索页面检索到的网页为10个,而调查结果显示,越来越多的用户只点击搜索结果页第1页里的结果.

定义3 用户打开网页的个数用n表示,$n \leqslant \mathrm{N}$.

定义4 当用户打开一个页面时,记录2个时间:打开页面的时间,用t_{start}表示;关闭页面的时间,用t_{end}表示.

定义5 网页从打开到关闭所花费的时间称为页面的生存期,t_{ij}表示第i个页面第j个用户的生存期.

$$t_{ij} = \sum_{j=1}^{\mathrm{users}} \sum_{i=1}^{n} (t_{\mathrm{end}} - t_{\mathrm{start}}) . \tag{1}$$

定义6 users代表参考反馈的用户数目.

算法的主要实现步骤如下:

1) 用户输入关键词句检索时,确定检索到的所有网页集合S和用户想要检索的目标网页集合G,确定G的个数N,如果用户没有特殊说明,N的值默认为10.

2) 确定users个数,默认为100.

3) 对于打开的n个网页,$n \leqslant \mathrm{N}$,

for($0 < j < \mathrm{users}$)

for($0 < i < n+1$)

利用每个页面的打开时间t_{start}和关闭时间t_{end},按照式(1)对G中的每一个网页R_i累加users个用户页面的生存期.

4) 对于每一个网页R_i,$i \in [1,n]$,查看页面的生存期t_{ij},对n个检索到的结果按照t_{ij}由大到小进行排

序.在同样的查询重要度下,t_{ij} 值越大,其重要度越大.

这样,有用的网页就会更快地被用户查询到,无应答或非应答网页的问题就解决了.

2.2　算法的理论分析

首先,可行性和可移植性.打开一个网页时经常会有登录时间这样的信息:"您上次登陆的时间是""您上次最后登陆的时间是",即可以获得网页打开和关闭的时间信息,表明此想法是可行的,而且可以从网页上直接获取时间信息加以利用,而不用为网页增加更多的元素,这样就更有利于广泛地使用和方便地被移植.

其次,用户透明和自动反馈.知道网页打开和关闭的时间,再算出网页利用时间,不必用户完成,它对用户是透明的,不需要用户主动反馈信息,并实际增加了用户反馈功能,从而更好地实现了查准率的目标.

第三,实用性.这种方法理论上可以去除无效的网页,比如说没有答复或没有合适答复的网页,一般发现这种情况就会立即关闭该网页,而对有价值的网页才会仔细研究内容,对用户有用的网页也会更快地被用户查询到.

3　实验结果及分析

为了证明页面时间算法的有效性,做了如下实验.

在评价指标方面,通常用户更关心查准率,而在大量数据中更快得到有用的数据,并不需要得到所有的有效页面,所以又定义了最快查准率作为评价指标,查准率越大越好,而最快查准率越小越好.

$$\text{查准率}(p) = \frac{\text{检索到的相关信息量}(W)}{\text{检索到的信息总量}(M)} \times 100\%, \tag{2}$$

$$\text{最快查准率}(p_f) = \frac{\text{最先检索到相关信息打开的页面个数}(n_f)}{\text{用户习惯检索到结果打开页面的个数}(N)} \times 100\%. \tag{3}$$

利用在搜索引擎中抓取的 5 个问题,确定抓取到的检索结果 S 和需要检索结果 G,记录第 i 个页面 j 个用户的生存期 t_{ij}.分别利用传统的排序方法(pagerank 算法)和页面时间算法(pagetime 算法)对结果相关度排序.为了简化问题,取 $M = 10, N = 10$,users $= 5$,得到的结果如表 1 所示.

表 1　Pagerank 和 pagetime 算法对 5 个检索词的实验结果

Tab.1　The experimental results of five key words using pagerank and pagetime algorithms

关键词句	提问应答式网页个数	n_f		p		p_f	
		pagerank	pagetime	pagerank	pagetime	pagerank	pagetime
问题 1	6	6	1	0.2	0.6	0.6	0.1
问题 2	9	1	1	0.9	0.9	0.1	0.1
问题 3	8	9	1	0.1	0.4	0.9	0.1
问题 4	—	0	0	0	0	0	0
问题 5	5	2	1	0.7	0.8	0.2	0.1

可以看出,提问应答式页面是普遍存在的现象,而无用的提问应答式页面也很常见.尤其从第 1 个和第 3 个查询词可以看出,如果利用传统的排序方法,用户分别需要点击 6 个和 9 个网页才能找到满意的结果,这样用户有可能因失去耐心而漏掉有用的结果.如果利用页面时间算法对所得结果进行排序,再进行查找,则通常情况下只要一次就可以找到满意的结果.分析以上数据可以得出:

1) 无用的提问应答式网页的频繁出现,干扰了用户的正常检索.

2) 页面时间算法能够使用户更快、更有效地找到满意结果,提高了查准率.

3) 页面时间算法更好地缩短了查找到有用结果的时间.

4) 这种算法也有一定的局限性.如果用户点击页面能够找到合适查询,那么这种方法就有优越性,可是如果用户没有找到合适的页面,这种算法就不是这么有效了.值得注意的是,本文在相同的关键词下使用不同的算法得出检索结果,但是查询时关键词的匹配程度也值得进一步研究.

5) 直接将生存期用于排序可能忽略了页面大小对阅读时间的影响.页面大的一般阅读时间也会相对较长,但是并不一定就有效.

4 结束语

相关度算法是查准率的保证,传统相关度算法无法解决无效的提问应答式页面的问题,用户反馈也不能解决这一问题.作者提出的页面时间算法对于查询时常出现的无应答和非应答页面有很好的效果,实验结果表明,此种算法具有很好的实用性,有效地提高了网页的查准率,使用户能够更快地得到满意的结果.由于没有考虑到页面大小对阅读时间产生的影响,而且实验数据仅仅选用了5个主题10个用户作为测试,不能完全说明问题,对它们的研究以及找到更好的改进方案都将成为下一步研究的目标.

参考文献:

[1] 田甜,倪林.基于Pagerank 算法的权威值不均衡分配问题[J].计算机工程,2007,33(18):53-55.
[2] 搜索引擎直通车.网站排名基本原则[EB/OL].http://www.se-express.com/submit/submit.htm.
[3] 李智超,熊风,富羽鹏,等.分布式大规模文本检索系统[J].广西师范大学学报:自然科学版,2007,25(2):179-180.
[4] 过仕明.Pagerank 技术分析及网页重要性的综合评价模型[J].图书馆论坛,2006,26(1):79-81.
[5] 许静芳,李星.可扩展的分布式信息检索的设计与实现[J].清华大学学报:自然科学版,2005,45(S1):1844-1845.
[6] 李村合,杨献峰,张培颖.基于Web 挖掘与相关反馈的多层次用户兴趣挖掘算法[J].微计算机应用,2007,28(9):911-912.

A Rank Algorithm Based on the Pagetime

LIU Su-qin, SHUO Jun, LI Xing-sheng, MENG Ling-fen
(*College of Computer and Communication Engineering, China University of Petroleum(East China), Qingdao* 266555, *China*)

Abstract: Taking the time of the webpage into the consideration, the pagetime rank algorithm based on the traditional rank algorithm is proposed. The pagetime rank algorithm can improve the effectiveness of the inquiry and help users to get the satisfactory answers more quickly. Compared to the traditional rank algorithm, the new algorithm shows good feasibility and necessity, and can enhance the precision ratio.

Key words: precision ratio; no response to the website; pagetime rank algorithm