基于PageRank的网站服务质量影响因素的重要性排序算法

齐玉东 何 诚 袁 伟
(海军航空工程学院 烟台 264001)

摘 要 通过德尔菲法进行网站服务质量影响因素的筛选,在分析各个因素之间的关系时,借鉴PageRank的投票思想,将影响网站服务质量的因素间的相互影响视作网页间的相互链接,产生影响视为投票,利用各个因素之间存在的内部关系计算影响力值,最终得到影响网站服务质量的因素重要性排序,为网站服务质量的评价提供了参考和借鉴。

关键词 PageRank,影响因素,重要性排序

中图法分类号 TP391 **文献标识码** A

Algorithm of Importance Ranking for Influencing Factors of Website Service Quality Based on PageRank

QI Yu-dong HE Cheng YUAN Wei
(Naval Aeronautical Engineering Institute, Yantai 264001, China)

Abstract In this paper, factors were filtered through the Delphi method. But in the analysis of the relationship between the various factors, PageRank voting ideas look the interaction of the factors which affect the quality of web service as a link with each webpage, and this method uses this interaction as votes and calculates the affected value by internal relationship of each factors. The sequence of factors which affects the quality of web service by importance will be got eventually, and this method offers an example and reference for evaluation of quality of the web service.

Keywords PageRank, Influencing factors, Importance ranking

1 引言

近年来,网站服务质量评价成为了信息科学领域关注的热点研究问题之一。许多研究机构和学者对此进行了大量的研究工作,从不同的角度提出各种测评方法,如e-SERVQUAL,WebQual,.comQ等,并对网站的服务质量进行了测评[1]。

赵杨[2]在信息可用性、信息充足性、专业性、易用性和交互性5个层面,综合运用多维标度分析法和层次分析法,从直观层面和精准层面对高新技术行业信息中心网站的信息服务质量进行了实证测评。邓仲华等[3]构建了信息资源云服务的服务质量评价指标,但由于没有对指标进行权重赋值、检验和修正,因此并没有构建出完整意义上的评价指标体系。刘志亮[4]通过研究不同生命周期IT服务质量的关键影响因素,提出基于企业IT服务提供者的视角下的具有多维多层的质量测度模型,并采用总体分散技术对模型进行了检验和修正。

综上所述,目前学者和研究机构所做的服务质量评价工作,从对象上看,主要集中在政府办公、电子商务、图书馆等领域;从内容上来看,涵盖人员、技术、资源、过程、服务等方面,但在评价指标上没有形成统一标准;从评价方法上看,有德尔菲法、层次分析法、主成分分析法、灰色关联度法和熵值法等。但目前所做的工作中并没有具体的、针对软硬件功能的服务质量评价模型,对各项业务应用服务也缺少有效的度量手段和质量监控。学者的研究在评价范围、评价方法、指标选取、调查手段等方面都存在较大差别,从定性角度进行评价的较多,定量研究的则较少。

2 网站服务质量影响因素的筛选

影响网站服务质量的因素有很多种,涉及到信息服务提供过程的方方面面,包括人员、过程、技术和资源等。本文侧重于从基础设施资源,即软硬件层面进行研究。首先,必须针对评估对象构造一个科学、合理的评价指标体系,这需要对评价对象大量相互关联、相互制约的复杂因素之间的关系层次化、条理化,并能够区分它们各自对评估目标影响的重要程度,以及对那些只能定性评估的因素进行恰当、方便的量化处理。

本文采用德尔菲法来筛选影响网站服务质量的因素。通过进行3轮德尔菲专家咨询,归纳监测项,得到部分影响因素,参考借鉴ITSS对信息系统信息服务质量影响因素的分析,最终选取的影响因素如表1所列。

齐玉东(1973-),男,博士,副教授,主要研究方向为指挥信息系统;何 诚(1993-),男,硕士生,主要研究方向为网络作战辅助决策;袁 伟(1991-),男,硕士生,主要研究方向为军用信息系统,E-mail:674690154@qq.com。

表 1　网站服务质量影响因素关系表

1	网络及网络设备	网络设备整体运行状态	34	存储	读、写缓存分配比例情况	67	数据库	数据库连接性监测
2		各项硬件资源开销状况	35		数据读、写命中率情况	68		周期性关键设备主备切换/应急演练
3		链路健康状况	36		存储硬盘空间使用情况	69		侦听连接正常性监测
4		链路端口工作稳定性	37		介质读、写正常性监测	70		数据库正常登陆监测
5		链路负载百分比	38		端口访问监测	71		SQL 执行正常性监测
6		路由条目变化	39	操作系统	CPU 使用情况	72		表空间正常访问监测
7		设备机身、板卡或模块的工作情况	40		内存使用情况	73		表读写正常性监测
8		CPU 使用峰值情况	41		磁盘使用情况	74		客户端连接监测
9		内存使用峰值情况	42		网络端口状态和流量	75		过期归档日志清除
10		主要端口的利用率	43		光纤端口状态和流量	76	中间件	中间件运行状态
11		链路的健康状态	44		重要文件系统空间使用情况	77		主要进程运行状态
12		无效 IP 包	45		日志情况	78		应用服务运行情况
13	服务器	整体运行状态	46		CPU 使用峰值情况	79		通信网络连接情况
14		电源工作状态	47		内存使用峰值情况	80		日志是否有报错信息
15		CPU 工作状态	48		硬盘使用情况	81		服务器业务 CPU 使用峰值情况
16		CPU 使用峰值情况	49		系统配置文件备份	82		服务器业务内存使用峰值情况
17		内存工作状态	50		IO 读写情况	83		服务器业务会话连接数情况
18		内存使用峰值情况	51		磁盘读、写正常性监测	84		连接正常性监测
19		硬盘工作状态	52		输入、输出设备情况	85	应用资源	应用的请求和反馈响应时间
20		网络端口工作状态	53		临时文件清理	86		资源消耗情况
21		资源分配情况和策略	54		端口访问监测	87		进程状态
22		文件系统空间使用情况	55	数据库	主要进程运行情况	88		服务或端口响应情况
23		IO 读写情况	56		数据库日志是否有异常	89		会话内容情况
24		网络流量情况	57		数据库日常备份是否正常	90		数据库连接情况
25		存储的链路性能监测	58		数据流网络流量情况	91		存储连接情况
26	存储	控制器工作状态	59		数据库 CPU 使用情况	92		作业执行情况
27		电源工作状态	60		数据库内存使用情况	93		应用的请求和反馈响应情况
28		数据存储介质工作状态	61		数据库表空间使用情况	94		关键进程及资源消耗检查
29		接口卡工作状态	62		数据库锁情况	95		主机操作系统的漏洞扫描
30		数据存储介质空间使用情况	63		数据库会话数	96		补丁检查
31		读写速率情况	64		数据库 BUFFER 等命中率情况	97		启动或停止服务或进程
32		读写命中率情况	65		数据库等待事件情况	98		增加或删除用户账号
33		IO 读写速率情况	66		网络通信正常性监测	99		建立或终止会话连接

3　网站服务质量影响因素的重要性排序

3.1　PageRank 算法与网站服务质量因素

PageRank 算法基于这样一个假设，即网页的质量和重要性可以通过其他网页对其链接量来衡量。如，把从 A 页面到 B 页面的链接解释为 A 页面给 B 页面投票，根据投票来源（甚至来源的来源，即链接到 A 页面的页面）和投票目标的等级来决定新的页面等级。一个页面的“得票数”由所有链向它的页面的重要性决定，到一个页面的链接相当于对该页投一票，一个页面的 PageRank（PR）是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到的[7]。显然，一个拥有较多链入的页面容易获得较高的等级，相反，如果某个页面不存在任何其他页面指向它，那么它就没有等级[8]。

在网站运维的实际过程中，每个因素都可以对其他因素产生影响，如 IP 包传输时延、IP 包丢失率、IP 包误差率等因素会对网络稳定性产生影响，进而会对客户端的连通性产生影响；数据库连通性会对 SQL 能否正常执行、表空间能否正常访问、表能否正常读写产生影响。借鉴 PageRank 的投票思想，通过分析各个因素之间的关系，产生影响即视为投票，最终将各个因素的投票结果进行排名，从而得到影响网站服务质量的关键指标。

3.2　网站服务质量影响因素的重要性排序

为方便表述，定义某因素被其他因素影响为影响力传入，定义某因素影响其他因素为影响力传出。影响力传入会提高影响力值，而影响力传出对影响力值没有影响，即如果因素 A 对因素 B 有影响，那么因素 A 的影响力值 $IV(A)$ 会累计到 $IV(B)$ 上，但是 $IV(A)$ 不会下降。如果因素 A 只对因素 B 有影响，即只有一条影响力传出，那么因素 A 的影响力值全部“贡献”给 B，如果还有其他影响力传出，且传出数总和为 $L(A)$，则其他因素分得因素 A 的影响力值为 $IV(A)/L(A)$。由此，可以得到因素 i 在其他因素影响下的 $IV(i)$：

$$IV(i)=\sum\left(\frac{IV(j)}{L(j)}\right)$$

考虑偶然概率 d，即存在影响关系的因素之间被影响的概率为 d；之前不存在影响关系的因素之间也有 $(1-d)$ 的概率产生影响，如果一共有 N 个因素，在概率平均的情况下，则有 $(1-d)/N$ 的概率产生影响。得到修正后的公式：

$$IV(i)=d\sum\left(\frac{IV(j)}{L(j)}\right)+\frac{(1-d)}{N}$$

如图 1 所示，有 A,B,C,D,E,F 6 个因素，其中，A 有 2 个传入；B 有 4 个传入，1 个传出；C 有 1 个传入，2 个传出；D 有 1 个传入，3 个传出；E 有 1 个传入，2 个传出；F 有 1 个传出。

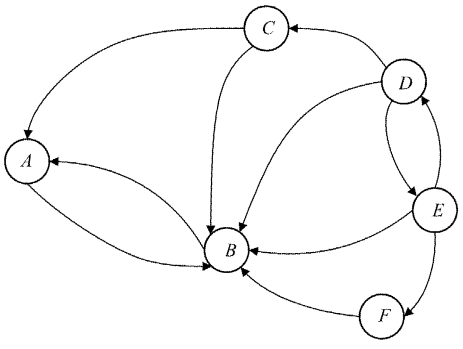

图 1　因素间的相互关系示意图

首先，对这张关系示意图中构造的传入关系进行表示，如表 2 所列。

表 2 传入关系表

影响因素	被影响因素
A	B
B	A
C	A,B
D	B,C,E
E	B,D,F
F	B

建立关系矩阵 R：

$$R=\begin{bmatrix}0&1&0&0&0&0\\1&0&0&0&0&0\\1&1&0&0&0&0\\0&1&1&0&1&0\\0&1&0&1&0&1\\0&1&0&0&0&0\end{bmatrix}$$

将关系矩阵 R 倒置后，将每个元素除以各自行的所有要素和，则得到转移矩阵 M：

$$R=\begin{bmatrix}0&1&1/2&0&0&0\\1&0&1/2&1/3&1/3&1\\0&0&0&1/3&0&0\\0&0&0&0&1/3&0\\0&0&0&1/3&0&0\\0&0&0&0&1/3&0\end{bmatrix}$$

设初始时每个因素的 IV 为 $1/N$，这里即是 $1/6$，按 $A-F$ 的顺序将各因素的 IV 表示为向量 v。

$v=[1/6\ \ 1/6\ \ 1/6\ \ 1/6\ \ 1/6\ \ 1/6]^{T}$

M 的第一行分别是 A,B,C,D,E 和 F 对因素 A 的影响力传入，而 v 中的元素分别是 $A-F$ 当前的 IV。将矩阵 M 和向量 v 进行向量运算，所得向量的第一个元素就是因素 A 的最新 IV 值；同理，Mv 的结果就分别代表 $A-F$ 新的 IV：

$v=[1/6\ \ 19/36\ \ 1/18\ \ 1/18\ \ 1/18\ \ 1/18]^{T}$

重复迭代，当 $v=[0.3796\ \ 0.5957\ \ 0.0062\ \ 0.0062\ \ 0.0062\ \ 0.0062]^{T}$ 时已经趋于收敛。将影响力值从高到低排序，结果如表 3 所列。由表 3 可知，当影响力传入越多时，对影响力值的贡献就越大，影响力值也越高，符合图 1 的描述；并且通过计算可以发现，各因素的影响力与 IV 的初始值无关，虽然当 v 的初始值不同时，Mv 收敛时的值也不同，但是各因素的影响力值排序保持不变。

表 3 收敛后的影响力值排名表

排名	影响力值	影响因素	影响力传入数
1	0.5957	B	5
2	0.3796	A	2
3	0.0062	C	1
4	0.0062	D	1
5	0.0062	E	1
6	0.0062	F	1

3.3 实验及结果分析

对表 1 中的 99 个影响因素的相互关系构建关系矩阵，借助 MATLAB 的矩阵运算功能，实现对转移矩阵及影响力值向量的计算。其代码如下：

```
R=xlsread('99factors.xls');%关系矩阵文件
%将 xls 文件中存储的数据转换为矩阵存储到变量 R 中
P=[];%定义空矩阵 P
[n,n]=size(M);%提取 P 中的行和列的维度，即影响因素的数量
for i=1:n
P=[P;R(i,:)/sum(R(i,:))];%将 R 的数据除以每行数据之和
end
d=0.7;%偶然概率 d
sigma=0.001;%定义收敛阈值
e=ones(n,1);%n 维全 1 向量
M=d*P'+(1-d)*((e*e')/n);
x=ones(n,1)/n;%各因素的影响力初始值
v=M*x;%影响力值计算公式
while(norm(v-x)>=sigma);%收敛判断
x=v;
v=M*x;
end
v %收敛后得到 v,并输出
```

其中，d 的取值见参考文献[9]。计算结果如表 4 所列。

根据计算结果，影响力值排序的顺序为：操作系统输入、输出设备情况，操作系统磁盘使用情况，服务器 CPU 使用峰值情况，操作系统内存使用情况，服务器内存工作状态…，即在以上诸多影响因素中，操作系统输入、输出设备情况对网站服务质量的影响最大，其次是操作系统磁盘使用情况，再次是服务器 CPU 使用峰值情况，以此类推。由于影响力值的初始值不同，最终收敛时的值也不同，但影响力值的排序与初始值无关，因此此计算只能通过影响力值的大小得到其重要性排序，若要得到准确的影响力值，就必须输入相对准确的影响力初始值。

表 4 各因素影响力值排序表

影响力值排序	因素序号	影响力值	影响力值排序	因素序号	影响力值	影响力值排序	因素序号	影响力值
1	52	0.0122	34	95	0.0104	67	48	0.0097
2	41	0.012	35	27	0.0103	68	58	0.0097
3	16	0.0119	36	98	0.0103	69	71	0.0097
4	40	0.0119	37	8	0.0102	70	22	0.0096
5	17	0.0116	38	10	0.0102	71	36	0.0096
6	57	0.0116	39	68	0.0102	72	63	0.0096
7	5	0.0115	40	83	0.0102	73	73	0.0096
8	88	0.0115	41	56	0.0101	74	79	0.0096
9	28	0.0113	42	64	0.0101	75	30	0.0095
10	92	0.0113	43	67	0.0101	76	53	0.0095
11	46	0.0112	44	87	0.0101	77	75	0.0095
12	18	0.0111	45	42	0.01	78	50	0.0094

（续表）

影响力值排序	因素序号	影响力值	影响力值排序	因素序号	影响力值	影响力值排序	因素序号	影响力值
13	94	0.0111	46	59	0.01	79	66	0.0094
14	14	0.011	47	69	0.01	80	3	0.0093
15	25	0.011	48	78	0.01	81	54	0.0093
16	84	0.011	49	82	0.01	82	55	0.0093
17	35	0.0109	50	9	0.0099	83	65	0.0093
18	89	0.0109	51	11	0.0099	84	70	0.0093
19	15	0.0108	52	13	0.0099	85	90	0.0093
20	31	0.0108	53	24	0.0099	86	12	0.0092
21	38	0.0108	54	77	0.0099	87	45	0.0092
22	19	0.0107	55	80	0.0099	88	61	0.0092
23	72	0.0107	56	97	0.0099	89	96	0.0092
24	1	0.0106	57	33	0.0098	90	29	0.0091
25	23	0.0106	58	43	0.0098	91	49	0.0091
26	32	0.0106	59	47	0.0098	92	76	0.0091
27	74	0.0106	60	60	0.0098	93	91	0.0091
28	6	0.0105	61	62	0.0098	94	99	0.0091
29	7	0.0105	62	81	0.0098	95	51	0.009
30	39	0.0105	63	2	0.0097	96	20	0.0089
31	93	0.0105	64	4	0.0097	97	86	0.0089
32	37	0.0104	65	26	0.0097	98	21	0.0088
33	85	0.0104	66	44	0.0097	99	34	0.0088

结束语　本文将影响网站服务质量的各个因素之间存在的内部关系用矩阵进行表示，迭代计算影响力值，最终得到影响网站服务质量的因素重要性排序，避免了以往层次分析法等方法主观性较强的缺点。实验结果表明，依据影响因素间内部关系计算得到的重要性排序，能反映出各因素对网站的真实影响，对网站服务质量的维护有一定的参考价值。在进行重要性排序时，本文的初始值设为相同，而事实上很多因素的影响初值并不相同。在下一步的研究中，可以结合工程经验对初始值进行设定，避免因为影响力传入数目多而提高其重要性。

参 考 文 献

[1] LANDRUM H, PRYBUTOK V R. A service quality and success model for the information service industry[J]. European Journal of Operational Research, 2004, 156(3): 628-642.

[2] 赵杨. 行业信息中心网站信息服务质量评价[J]. 情报资料工作，2009(6)：32-37.

[3] 邓仲华，汪宣晟，李志芳，等. 信息资源云服务的质量评价指标研究[J]. 图书与情报，2012(4)：12-15.

[4] 刘志亮. 企业 IT 服务质量评价模型及其应用研究[D]. 武汉：华中科技大学，2013.

[5] 卢军. 非量化质量管理体系评价方法初探[J]. 轻工科技，2015(7)：147-148.

[6] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: Bringing order to the Web[J]. Stanford Digital Libraries Working Paper, 1998, 9(1): 1-14.

[7] 宫秀文，张佩云. 基于 PageRank 的社交网络影响最大化传播模型与算法研究[J]. 计算机科学，2013(S1)：136-140.

[7] 庞红美，刘宏志. 基于 PageRank 算法的信息工程安全监理风险评估研究[J]. 计算机安全，2014(8)：17-20.

[9] 邵晶晶. 基于 PageRank 排序算法改进的若干研究[D]. 武汉：华中师范大学，2009.

（上接第 60 页）

[11] KE J J, HU J Z. Fault feature extraction method based on Manhattan distance and stochastic neighbor embedding [J]. Application Research of Computers, 2015, 32(10): 2992-2995.

[12] WANG L F, WANG Y, et al. Application of Chebyshev local collocation method to trajectory optimization [J]. Journal of Harbin Institute of Technology, 2013, 45(5): 95-100.

[13] XIE J Y, XIE W X. Several Feature Selection Algorithms Based on the Discernibility of a Feature Subset and Support Vector Machines [J]. Chinese Journal of Computers, 2014, 37(8): 1704-1718.

[14] YU Y Y. Multi-model Estimation Based on Jaccard Distance and Conceptual Clustering [J]. Computer Engineering, 2012, 38(10): 22-26.

[15] YANG H F, LI G J. Novel antenna selection algorithm based on Tanimoto similarity [J]. Journal of Systems Engineering and Electronics, 2008, 19(3): 624-627.

[16] CHEN D L, SHEN Y T, et al. A Measure Model of Similarity for Finding the Best Coach [J]. Journal of Northeastern University (Natural Science), 2014, 35(12): 1697-1700.

[17] WU D, TENG Y P. Word Segment and Search Techniques for Chinese Information Search Engines [J]. Journal of Computer Applications, 2004, 24(7): 128-131.

[18] XIAO W, TANG D K, et al. Knowledge push based on Lucene and collaborative filtering algorithm [J]. Journal of Changchun University of Technology (Natural Science Edition), 2016, 37(5): 503-506.

[19] HE W. The Research for Fast Exact String Matching Algorithm [D]. Hefei: Hefei University of Technology, 2010.

[20] http://baike.baidu.com/item/%E6%93%8D%E4%BD%9C%E7%B3%BB%E7%BB%9F/192?sefr=enterbt.

[21] WANG Z. Analysis of producer and consumer problem algorithm [J]. Journal of Jilin Province Economic Management Cadre College, 2008, 22(3): 78-81.