



大型 IP 网络时延的主成分分析

韦烜, 黄晓莹

(中国电信股份有限公司研究院, 广东 广州 510630)

摘要: 网络时延是评估网络性能的关键指标之一。主成分分析 (PCA) 是数据挖掘领域常用的一种多变量分析和降维算法。通过对大型 IP 网络时延进行 PCA 分析, 旨在挖掘网络时延的深层原因及网络各节点间的相互依赖关系, 并搭建一个科学合理的网络时延评价体系, 最终得到 IP 网络建设、优化改造的有效建议。对历史网络时延进行离线分析只是主成分分析方法的一种初步应用, 今后可结合网络拓扑结构、现网流量流向、路由、距离等相关因素, 将主成分分析方法应用到针对网络流量、网络时延、网络丢包等网络性能的实时在线监测分析中, 进一步提升网络运营的效率和质量。

关键词: 网络时延; 主成分分析; 相关系数矩阵; 方差贡献率

中图分类号: TN915

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2021059

Principal component analysis of time delay in large IP network

WEI Xuan, HUANG Xiaoying

Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Network time delay is one of the key indexes to evaluate network performance. Principal component analysis (PCA) is a kind of multivariable analysis and declination algorithm commonly used in the field of data mining. Based on PCA analysis of time delay in large IP networks, aiming to find out the deep reason of time delay and the interdependencies among nodes of the network, a scientific and reasonable network time delay evaluation system was built, and effective suggestions for IP network construction and optimization were finally got. The off-line analysis of the historical network delay is only a preliminary application of the PCA. In the future, PCA can be applied to the real-time on-line monitoring and analysis of the network performance, such as network traffic, network delay, network packet loss, etc., in combination with the network topology, current network traffic direction, routing, distance and other related factors, thus the efficiency and quality of network operations can be further improved.

Key words: network time delay, principal component analysis, coefficient matrix, proportion of variance

1 引言

随着各类新型业务和应用的不断涌现, 通信

网络向全业务、全 IP 化演进成为大势所趋, 身负承载重任的 IP 网络面临着巨大挑战, 其承载性能的好坏直接影响业务开展, 并最终决定网络的核

心竞争力。网络时延是衡量网络通信性能优劣的关键指标之一,对实时性要求敏感的交互式音/视频流业务对IP网络的时延性能与服务质量提出了更高的要求。在此背景下,IP网络的时延性能评测和优化迫在眉睫,如何科学合理地评估现有网络的性能指标,并以此为基础对现有网络进行优化改造、建设更优质的新网络以满足上层业务的需要、提高用户体验,成为网络规划建设工作中需要重点考虑的问题。

目前,针对IP网络时延的研究主要集中于时延的原因分析、时延的测量方法、时延对业务的影响、通信技术层面如何降低时延等方面,而对于如何客观评价整网的时延性能、时延与网络结构之间的深层关系、如何通过网络结构的优化改善时延性能等方面的研究尚比较欠缺。本文尝试通过主成分统计方法对大型IP网络的时延进行研究分析,旨在挖掘网络时延的深层原因及其与网络各节点间的相互依赖关系,并搭建一个相对合理的网络时延评价体系。

2 网络时延概述

2.1 网络时延及其构成

IP网络时延通常指一个报文或分组从网络的一端传送到另一端所需要的时间,网络时延的构成如图1所示。它主要由节点时延和传输时延构成,其中节点时延又包括处理时延、排队时延和发送时延。

处理时延主要指检查数据分组头部并决定将

该分组发往何处所需要的时间,同时还包含检查比特差错等其他因素所需要的时间;排队时延是数据分组等待链路上前面数据分组传输完毕所需要的时间,一个特定数据分组的排队时延取决于先期到达的正在排队等待链路传输的数据分组的数量;发送时延是路由器将数据分组的所有比特发送到链路所需要的时间,它等于数据分组长度除以链路传输速率。传输时延是数据分组从路由设备一端传输到另一端所需要的时间,它等于两台路由器之间的距离除以链路传输速率,而传输速率取决于链路的物理媒质(光纤、双绞铜线等),广域网中传输时延为毫秒量级。

通常,发送时延与传输时延是网络时延的主要组成部分,对于报文长度较大的情况,发送时延是主要矛盾;对于报文长度较小的情况,传输时延是主要矛盾。而对于大型的IP网络而言,由于端到端距离跨度长达数千千米甚至上万千米,此时传输时延将成为最需要关注的主要矛盾。

2.2 网络时延对通信性能的影响

网络时延是衡量网络通信性能优劣的关键指标之一,一般而言,时延越小越好,但时延只能降低却不可能消除,并且受通信设备处理能力 & 网络拓扑、传输媒质等物理条件的限制,网络时延只能在一定范围内进行优化和改善。

不同的业务和应用对实时性要求不同,从而对网络时延的要求也不同。例如交互型直播视频类业务,对时延的要求较高;而文件传输、邮件等应用,对时延的要求则较低。在进行网络设计

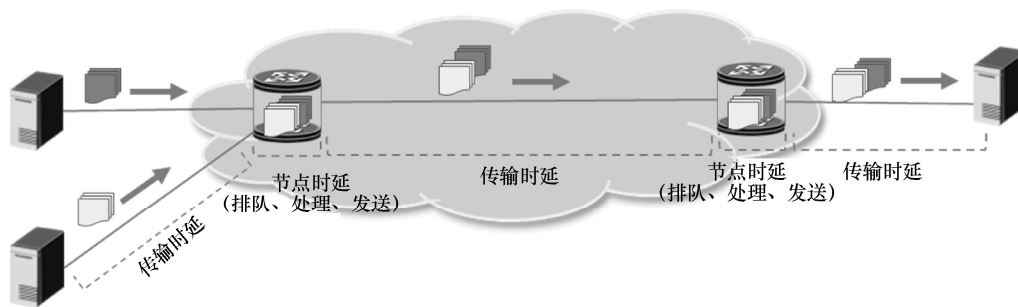


图1 网络时延的构成



和优化时,通常会考虑不同业务的需求及其业务等级,优先保证高优先级业务的时延性能。一个设计不合理的网络,网络时延容易出现异常或超限的情况,由此可能造成网络拥塞、数据丢包甚至网络瘫痪,从而影响数据传送的效率和有效性。

3 PCA 基本原理

主成分分析 (principal component analysis, PCA),是数据挖掘中一种常用的降维算法,是一种多变量的统计方法。PCA 通过正交变换将一组可能存在相关性的变量重新转换为一组线性不相关的综合变量以代替原来的变量,其基本思想是将原来 n 维特征映射到 p 维特征上 ($p < n$),这个重新构造的相互独立 p 维特征则称为主成分。

PCA 提取较少数量的主成分以减少数据空间的维度,被提取出的主成分必须能够最大限度地包含原始数据的特征,同时最小限度地损失原始数据的信息量,以期在降维的同时发现原始数据的最主要特性及高纬度背后更丰富的内涵。PCA 每一个主成分的贡献率可用其方差来度量,第 1 主成分 Comp.1 是所有变量的线性组合中方差最大的,其次是第 2 主成分 Comp.2,依次类推, n 维变量最多具有 n 个主成分,一般提取方差累计贡献率在 80% 以上的前 p ($p < n$) 个主成分即可。

通常,主成分分析方法主要包括以下 5 个步骤。

步骤 1 对原始数据进行标准化处理,使之零均值、归一化以消除量纲不同的影响。

原始数据 $Y = (Y_1, Y_2, \dots, Y_n)$ 为 n 维空间的随机变量,其中, $Y_j = (y_{1j}, y_{2j}, \dots, y_{mj})^T$, m 为样本数量。 Y 经标准化处理后得到 n 维随机变量 $X = (X_1, X_2, \dots, X_n)$, $X_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$, 其中 x_{ij} 的计算式为:

$$x_{ij} = \frac{y_{ij} - \bar{Y}_j}{S_j} \quad (1)$$

其中, $\bar{Y}_j$ 、 S_j 分别为 Y_j 的样本均值和标准差,计算式为:

$$\bar{Y}_j = \frac{1}{m} \sum_{i=1}^m y_{ij}, (j=1, 2, \dots, n) \quad (2)$$

$$S_j^2 = \frac{1}{m-1} \sum_{i=1}^m (y_{ij} - \bar{Y}_j)^2, (j=1, 2, \dots, n) \quad (3)$$

步骤 2 计算指标变量的相关系数矩阵 (或协方差矩阵) 并求解特征值和特征向量。

X 的相关系数矩阵为 $R = (r_{ij})_{n \times n}$, 其中相关系数 r_{ij} 的计算式如下:

$$r_{ij} = \frac{\sum_{k=1}^n x_{ki} x_{kj}}{n-1}, (i, j=1, 2, \dots, n) \quad (4)$$

求解相关系数矩阵的特征值, 并从大到小排序 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$, 特征值所对应的特征向量为 $a_1, a_2, \dots, a_n$, 其中 $a_j = (a_{1j}, a_{2j}, \dots, a_{nj})^T$ 。由特征向量得到 n 个新的指标变量:

$$\begin{cases} Z_1 = a_{11}X_1 + a_{21}X_2 + \dots + a_{n1}X_n \\ Z_2 = a_{12}X_1 + a_{22}X_2 + \dots + a_{n2}X_n \\ \vdots \\ Z_n = a_{1n}X_1 + a_{2n}X_2 + \dots + a_{nn}X_n \end{cases} \quad (5)$$

Z_1 称为第一主成分, Z_2 称为第二主成分, ..., Z_n 称为第 n 主成分。

步骤 3 计算主成分的方差贡献率及累计贡献率, 并依据方差累计贡献率选择前 p 个 ($p < n$) 主成分。

$$\text{第 } j \text{ 主成分 } z_j \text{ 的方差贡献率 } b_j = \frac{\lambda_j}{\sum_{k=1}^n \lambda_k},$$

($j=1, 2, \dots, n$), 前 p 个主成分的方差累计贡献率为:

$$\alpha_p = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^n \lambda_k} \quad (6)$$

根据研究问题的实际情况, 通常选取方差累计贡献率大于 80% 的前 p 个主成分代替原来的 n

个指标变量,从而实现数据的降维。

步骤 4 根据主成分方差贡献率和主成分表达式计算主成分综合得分 $w = \sum_{j=1}^p b_j z_j$ 。

步骤 5 结合主成分综合得分及实际情况,做进一步的统计分析。

4 大型 IP 网络时延的 PCA 分析

4.1 大型 IP 网络时延简介

对于大型 IP 网络而言,端到端距离跨度少则数千千米、多则数万千米,此时网络时延将成为评估和考核网络性能的一个重要指标。通常运维部门会利用相应的技术手段实时监测并定期采集记录网络各节点间端到端的时延,并形成时延矩阵报表。通过对这些网络时延数据的统计和分析,可以及时评估网络的运行状况并发现异常情况,进而采取相应的处理措施。

为了适应业务的飞速发展和网络扁平化的演进,近年中国电信在原有 IP 骨干网基础上启动网络新平面的建设工作,同时原有旧平面将逐步缩容并分批退网,该过程将持续 3~5 年,最终实现业务的无缝切换和高效承载。本文将以前中国电信新、旧平面全国 28 个地市到 IP 骨干网 9 个大型节点的时延矩阵数据为基础,通过主成分分析方法对 IP 骨干网的新、旧两个平面的网络时延矩阵进行研究分析,试图找出网络时延与大型节点之间深层的相互依赖关系,并对其中相对重要的节点提出优化改造建议;同时通过主成分分析搭建更科学合理的时延评价体系,评估和分析新平面相对旧平面网络时延的改善情况,并借鉴旧平面运营经验进一步为新平面的时延优化和网络建设提供有价值的参考。

4.2 网络时延的 PCA 分析过程

4.2.1 时延数据标准化处理

大型 IP 网络的时延矩阵经标准化处理后见表 1、

表 2, 其中表 1 对应网络旧平面的时延矩阵,表 2 对应网络新平面的时延矩阵;其中 A1~A28 代表分布于全国各省份的 28 个地市, X1~X9 代表 IP 网络的 9 个大型节点;表中每行指示了某地 $A_m (m=1, 2, \dots, 28)$ 到各大节点 $X_n (n=1, 2, \dots, 9)$ 标准化后的网络时延,每列指示了全国 28 个地市到某大型节点 $X_n (n=1, 2, \dots, 9)$ 标准化后的网络时延。

4.2.2 时延相关系数矩阵

网络旧平面相关系数矩阵见表 3。

网络新平面相关系数矩阵见表 4。

两个节点之间的相关系数越接近 1 则相关性越高。由表 3 可见,网络旧平面中节点 X1、X2 间的相关度很高,节点 X3、X7、X8、X9 间的相关度很高,X5、X6 间的相关度也较高。由表 4 可见,新平面中各节点间的相关性整体较旧平面有一定增强,其中 X1、X2 间的相关度仍很高,X3、X4、X7、X8、X9 之间的相关度较高,X4 与以上其他几个节点的相关性较旧平面有增强。大型节点相互之间的相关性说明,这些节点并非相互独立,而是密切关联的,结合网络实际情况分析,节点间的关联性大体与它们之间的链路数量、带宽大小、距离远近等因素呈正相关的关系。

4.2.3 主成分个数的选取

通过求解相关系数矩阵的特征值及特征向量,得到图 2 的碎石图,图 2 (a) 对应网络旧平面的碎石图,图 2 (b) 对应网络新平面的碎石图。图 2 中每个圆点对应时延相关系数矩阵的每个特征值,这些特征值由大到小排序,形成断崖式的碎石图。由碎石图可见,可以选取前两个或前三个特征值所对应的主成分来进行后续的分析。

表 5、表 6 是网络旧平面和网络新平面的各主成分的方差贡献率,其中第一行是各主成分的标准差,第二行为各主成分的方差贡献率,第三行为主成分的方差累计贡献率。



表 1 旧平面标准化后的时延矩阵

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9
A1	-1.508 53	-1.307 03	-0.354 17	0.000 783	-0.691 53	1.156 154	-0.371 68	-0.196 41	-0.277 24
A2	-1.298 51	-1.489 23	0.017 812	0.259 75	0.173 263	0.638 81	0.010 35	0.361 49	0.104 387
A3	0.395 988	-0.003 4	-0.265 32	-2.103 84	0.189 925	-0.54925	-0.679 14	-0.034 48	-0.344 64
A4	-0.619 12	-0.384 27	-0.272 43	-0.548 64	-2.113 7	-1.218 25	-0.501 8	-0.185 44	-0.338 13
A5	1.330 479	0.517 775	0.205 578	-0.242 63	-1.268 9	-2.089 26	0.062 879	0.367 232	-0.063 88
A6	-0.703 44	-0.484 55	-0.970 18	-1.31 527	-1.007 29	-0.509 19	-1.602 12	-1.117 89	-0.903 83
A7	-0.236 73	-0.111 21	-1.553 62	-0.440 58	-0.555 73	0.160 382	-0.737 84	-1.691 46	-1.053 97
A8	-0.226 65	0.962 677	0.085 337	1.182 025	-1.441 36	-0.794 18	-0.196 16	0.133 729	0.615 7
A9	0.618 208	0.968 327	0.561 564	-0.854 65	0.481 523	-1.023 67	-1.101 59	0.160 892	0.234 075
A10	1.822 121	1.736 196	0.872 533	-0.468 4	-0.263 47	-1.237 13	0.710 479	1.005 578	1.308 761
A11	0.027 92	0.346 876	-1.539 4	-0.722 51	-0.004 2	0.318 904	-0.523 59	-1.114 23	-1.564 82
A12	-1.346 77	-1.054 68	-0.151	0.806 467	-0.518 24	0.137 491	-0.728 16	-0.312 9	0.627 321
A13	-0.489 18	-0.678 52	0.761 176	0.630 458	1.731 228	1.509 824	0.890 233	0.930 355	0.991 747
A14	-0.129 6	-0.107 44	1.180 541	1.241 943	1.779 55	1.452 024	1.471 257	1.427 137	1.595 561
A15	-0.002 31	0.419 379	-0.687 64	0.548 071	-0.279 96	-0.402 17	-0.855 86	-0.657 15	-1.083 72
A16	-0.847 17	-1.016 55	0.415 26	0.240 456	0.855 601	1.164 783	0.420 572	0.280 517	0.635 686
A17	-0.224 53	0.067 9	-0.134 41	0.083 706	-0.017 53	0.009 3	-0.988 41	-0.397 53	-0.772 28
A18	-0.314 16	-0.467 6	-1.862 81	-0.655 11	-0.265 8	-0.087 99	-0.461 86	-1.402 59	-1.460 69
A19	1.712 337	0.890 646	0.709 052	-0.641 2	-0.794 01	-1.548 45	0.621 51	0.108 132	0.479 04
A20	-0.392 12	-1.223 23	-0.769 98	-0.723 46	0.526 513	0.915 223	0.222 056	-0.406 93	-0.688 61
A21	-0.828 61	-0.011 87	-0.134 41	0.759 927	-1.362 21	-0.569 85	0.012 645	0.151 49	-0.194 5
A22	1.267 897	2.146 73	1.886 589	2.431 744	0.961 41	1.082 901	2.672 039	2.533 534	2.149 638
A23	1.079 62	1.886 85	2.425 01	2.192 072	1.898 688	1.545 878	2.308 9	2.227 943	1.699 682
A24	2.220 42	1.250 334	1.007 582	0.4779 88	0.214 086	-1.045 42	0.556 75	0.749 613	0.469 279
A25	-0.335 91	-0.760 91	-1.342 75	-0.326 63	-0.249 14	-0.266 54	-0.907 91	-1.326 32	-1.168 78
A26	0.685 033	-0.077 78	-0.524 16	-1.084 16	1.158 03	0.128 334	0.242 634	-0.491 04	-1.142 75
A27	-0.995 14	-0.835 29	0.282 58	-0.724 65	0.085 783	-0.006 72	-1.030 17	-0.637 3	-0.545 44
A28	-0.661 55	-1.044 33	0.151 677	0.215 846	0.781 452	1.128 112	0.504 7	0.257 01	0.692 397

表2 新平面标准化后时延矩阵

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9
A1	-1.425 06	-1.629 89	-0.375 77	-0.193 08	-0.836 54	-0.191 69	-0.469 21	-0.265 49	-0.054
A2	-1.282 4	-1.810 22	-0.043 51	0.029 081	-0.697 38	-0.106 13.	-0.367 67	-0.170 61	-0.071 61
A3	0.083 615	0.309 951	-0.337 13	-1.652 64	1.117 554	0.041 056	-0.419 52	-0.448 74	-0.180 78
A4	-0.599 39	-0.67 248	-0.398 52	0.133 1	-2.211 62	-0.918 31	0.578 044	-0.332 89	-0.204 55
A5	-0.115	-0.0746	0.024 315	-0.376 17	-1.241 78	-1.453 41	-0.154 33	0.106 732	-0.060 17
A6	-0.547 56	-0.639 49	-0.805 9	-0.827 54	0.225 592	-0.556 11	-1.434 92	-0.947 68	-0.908
A7	-0.248 94	-0.189 51	-1.186 67	-0.572 91	-0.421 93	0.063 222	-0.884 01	-1.362 83	-1.147 47
A8	0.491 402	0.666 631	0.587 095	0.573 65	-1.480 46	-0.801 72	0.201 593	0.430 869	0.137 926
A9	0.490 485	0.676 302	0.241 529	1.227 15	0.260 924	-0.662 07	-0.059 28	0.097 716	0.256 5
A10	-0.063 63	0.461 839	0.093 429	-0.834 6	-0.683 68	-1.089 44	-0.138 13	0.416 272	0.251 939
A11	0.031 323	-0.021 13	-1.197 4	0.363 46	-0.174 6	-0.059 14	-0.800 84	-1.073 04	-1.450 33
A12	-1.217 73	-1.452 4	-0.294 64	-0.316 86	0.406 58	0.182 478	-0.658 25	-0.432 06	-0.300 52
A13	-0.355 82	-0.377 24	0.597 827	0.864 526	0.919 981	0.896 684	0.804 347	0.708 21	1.014 813
A14	-0.263 16	-0.200 89	0.690 55	0.988 784	1.086 548	1.001 753	0.946 393	0.797 938	1.153 918
A15	0.0634 32	0.087 524	-0.959 58	-0.640 69	0.082 82	-0.337 11	-0.866 73	-1.041 27	-1.181 81
A16	-0.787 46	-1.162 8 5	0.189 157	0.495 518	0.071 283	2.251 502	0.215 096	0.169 413	0.438 585
Z17	-0.385 18	-0.411 94	-0.632 05	-0.758 39	0.420 28	-0.394 74	-0.948 82	-0.702 11	-0.644 32
A18	-0.296 65	0.152 37 5	-1.428 78	-0.472 19	-0.424 81	0.061 892	-0.637 19	-1.129 71	-1.212 18
A19	-0.240 69	0.179 112	-0.067 55	-0.538 08	-1.003 83	-1.243 27	-0.501 62	-0.701 87	-0.069 41
A20	-0.905 81	-0.498 41	-0.559 5	-0.063 64	0.500 319	0.017 559	0.068 188	-0.694 81	-0.772 42
A21	-0.100 78	-0.620 9	0.059 516	0.764 273	-1.185 54	-0.449 71	0.145 422	-0.004 46	0.236 531
A22	1.323 031	1.876 042	1.355 068	1.768 219	0.584 684	-0.045 84	1.588 574	1.325 572	1.328 239
A23	2.827 118	2.203 71	3.127 975	3.223 071	2.822 161	2.413 318	3.746 281	2.899 031	2.870 716
A24	2.501 439	0.982 921	2.003 703	0.581 181	0.454 892	-0.830 53	0.200 513	2.166 181	1.930 439
A25	-0.153 53	0.074 44	-1.076 8	-0.664 22	-0.225 8	-0.054 26	-0.965 03	-1.252 07	-1.094 65
A26	0.680 846	0.745 135	-0.720 48	-1.063 82	1.061 311	0.662 605	-0.280 72	-0.468 13	-0.725 32
A27	-0.766 36	-1.124 73	0.526 997	0.852 759	0.105 894	-0.654 97	0.500 809	0.829 278	0.412 173
A28	1.262 482	1.911 881	0.618 003	0.291 246	0.467 15	2.256 379	0.591 006	0.490 544	0.558 761

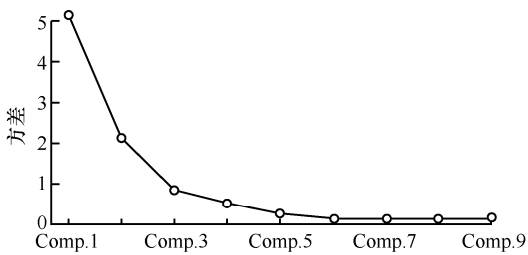


表 3 网络旧平面相关系数矩阵

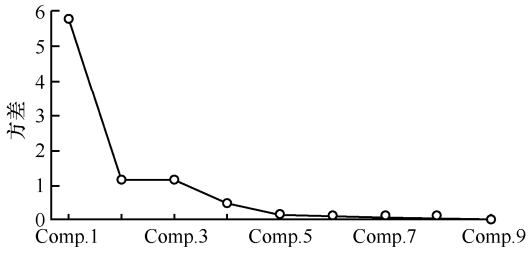
	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1.000 000 0	0.844 984 45	0.455 616 1	0.081 517 4	0.150 223 24	-0.431 343 8	0.453 992 8	0.465 268 5	0.318 890 0
X2	0.844 984 4	1.000 000 0	0.544 895 8	0.366 292 6	0.076 777 72	-0.326 413 9	0.480 077 5	0.577 083 7	0.441 787 1
X3	0.455 616 1	0.544 895 75	1.000 000 0	0.617 904 4	0.464 248 96	0.182 207 3	0.766 221 69	0.943 872 5	0.909 731 5
X4	0.081 517 4	0.366 292 60	0.617 904 4	1.000 000 0	0.314 794 02	0.441 670 4	0.694 922 9	0.687 131 2	0.679 282 8
X5	0.150 223 2	0.076 777 72	0.464 249 0	0.314 794 0	1.000 000 0	0.719 740 6	0.572 707 5	0.482 343 5	0.413 956 0
X6	-0.431 343 8	-0.326 413 92	0.182 207 3	0.441 670 4	0.719 740 56	1.000 000 0	0.428 194 2	0.257 572 2	0.269 500 4
X7	0.453 992 8	0.480 077 49	0.762 216 9	0.694 922 9	0.572 707 47	0.428 194 2	1.000 000 0	0.866 955 1	0.787 833 5
X8	0.465 268 5	0.577 083 66	0.943 872 5	0.687 131 2	0.482 343 55	0.257 572 2	0.866 955 1	1.000 000 0	0.919 055 0
X9	0.318 890 0	0.441 787 05	0.909 731 5	0.679 282 8	0.413 955 99	0.269 500 4	0.787 833 5	0.919 055 0	1.000 000

表 4 网络新平面相关系数矩阵

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1.000 000 0	0.905 770 7	0.686 749 3	0.455 820 7	0.479 148 5	0.263 908 4	0.592 238 0	0.672 006 1	0.603 464 1
X2	0.905 770 7	1.000 000 0	0.505 781 3	0.321 450 0	0.401 916 1	0.243 530 9	0.521 176 2	0.487 740 6	0.426 389 2
X3	0.686 749 3	0.505 781 3	1.000 000 0	0.791 016 7	0.444 873 4	0.321 905 0	0.860 865 1	0.986 730 0	0.973 379 9
X4	0.455 820 7	0.321 450 0	0.791 016 7	1.000 000 0	0.327 240 7	0.458 370 9	0.875 350 2	0.780 197 3	0.799 051 1
X5	0.479 148 5	0.401 916 1	0.444 873 4	0.327 240 7	1.000 000 0	0.650 277 9	0.466 149 9	0.417 715 9	0.433 087 0
X6	0.263 908 4	0.243 530 9	0.321 905 0	0.458 370 9	0.650 277 9	1.000 000 0	0.500 610 0	0.284 164 6	0.369 065 4
X7	0.592 238 0	0.521 176 2	0.860 865 1	0.875 350 2	0.466 149 9	0.500 610 0	1.000 000 0	0.851 804 8	0.858 474 1
X8	0.672 006 1	0.487 740 6	0.986 730 0	0.780 197 3	0.417 715 9	0.284 164 6	0.851 804 8	1.000 000 0	0.974 949 9
X9	0.603 464 1	0.426 389 2	0.973 379 9	0.799 051 1	0.433 087 0	0.369 065 4	0.858 474 1	0.974 949 9	1.000 000



(a) 旧平面碎石图



(b) 新平面碎石图

图 2 碎石图

由表 5 看到, 网络旧平面第 1 主成分 Comp.1 贡献率为 57%, 前两个主成分的累计贡献率为 81%, 前 3 个主成分的累计贡献率为 90%。由表 6 可知, 网络新平面第 1 主成分 Comp.1 贡献率为 65%, 前两个主成分的累计贡献率为 80%, 前 3 个主成分的累计贡献率为 91%。

结合碎石图考虑, 前两个主成分方差累计贡献率已达到 80%, 新、旧平面统一选取前两个主成分来进行分析, 图 3 是前两个主成分的载荷图。

由图 3 可见, 对于网络旧平面, 第 1 主成分 Comp.1 中, 9 个大型节点的载荷在 0.15~0.43, 说明 9 个大型节点的权重存在差异, 节点 X3、X7、

表5 网络旧平面主成分贡献率

对比项	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
标准差	2.261 982 9	1.469 033 9	0.904 578 71	0.684 403 32	0.458 152 71	0.318 612 65	0.269 466 744	0.181 465 553	0.147 453 193
方差贡献率	0.568 507 4	0.239 784 5	0.090 918 07	0.052 045 32	0.023 322 66	0.011 279 34	0.008 068 036	0.003 658 861	0.002 415 827
方差累计贡献率	0.568 5074	0.808 291 9	0.899 209 96	0.951 255 28	0.974 577 94	0.985 857 28	0.993 925 312	0.997 584 173	1.000 000 0

表6 网络新平面主成分贡献率

对比项	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
标准差	2.416 156 0	1.150 895 6	1.065 301 0	0.660 799 80	0.435 937 01	0.342 479 85	0.203 853 186	0.134 152 915	0.108 672 994
方差贡献率	0.648 645 5	0.150 663 4	0.110 096 3	0.048 517 37	0.021 115 68	0.013 032 49	0.004 617 347	0.001 999 667	0.001 312 202
方差累计贡献率	0.648 645 5	0.804 330 90	0.909 405 2	0.957 922 61	0.979 038 29	0.992 070 78	0.996 688 131	0.998 687 798	1.000 000 000

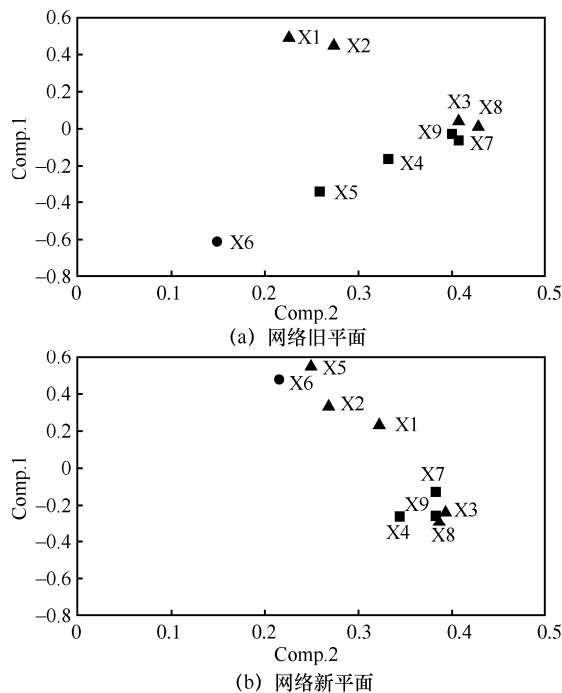


图3 前两个主成分的载荷图

X8、X9 的载荷比较高,说明这 4 个节点比较重要;第 2 组成分 Comp.2 中,节点 X1、X2 扮演主要的正向作用,X6 扮演主要的负向作用。对于网络新平面,第 1 主成分 Comp.1 中,9 个节点的载荷分布较旧平面更均衡些,说明各节点的重要性差异不大,但节点 X3、X7、X8、X9 与网络旧平面一样,载荷也比较高,说明它们在新平面中也同样比较重要;第二主成分 Comp.2 中,网络新平面与

旧平面有所不同,新平面中节点 X5、X6 扮演着主要的正向作用,没有节点扮演主要负向作用。

4.2.4 计算主成分得分

利用以上步骤得到的各主成分的载荷,可以写出各个主成分的表达式。以第 1 主成分 Comp.1 为例,写出第 1 主成分 Comp.1 的计算式。

网络旧平面第 1 主成分 Comp.1 的表达式为:

$$\text{Comp.1} = 0.23 \times X1 + 0.27 \times X2 + 0.41 \times X3 + 0.33 \times X4 + 0.2586 \times X5 + 0.15 \times X6 + 0.41 \times X7 + 0.43 \times X8 + 0.40 \times X9$$
(1)

网络新平面第 1 主成分 Comp.1 的计算式为:

$$\text{Comp.1} = 0.32 \times X1 + 0.27 \times X2 + 0.39 \times X3 + 0.35 \times X4 + 0.25 \times X5 + 0.22 \times X6 + 0.38 \times X7 + 0.39 \times X8 + 0.38 \times X9$$
(2)

为了更直观地对 28 个地市的网络时延进行综合评估和对比,定义主成分综合指数 Comp.,用每个主成分对应的方差占所提取主成分总方差的比例作为该主成分的权值系数,计算主成分综合指数 Comp.。对应于提取前两个主成分的情况,主成分综合指数的计算式如下:

$$\text{Comp.} = \text{Comp.1} \times \frac{\text{Comp.1 方差}}{\text{Comp.1 方差} + \text{Comp.2 方差}} + \text{Comp.2} \times \frac{\text{Comp.2 方差}}{(\text{Comp.1 方差} + \text{Comp.2 方差})} \quad (3)$$

利用主成分表达式及主成分综合指数表达



式,可以计算得到各地市的主成分得分及主成分综合指数。前两个主成分的得分及综合指数如图4、图5所示。

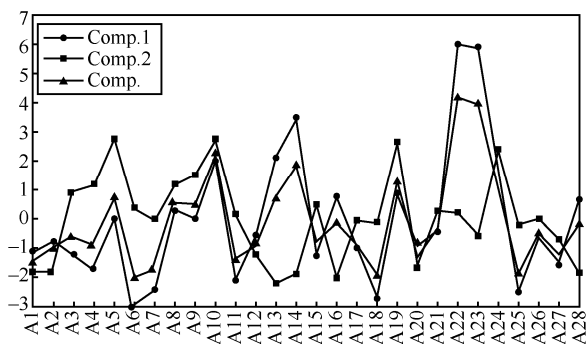


图4 网络旧平面主成分得分

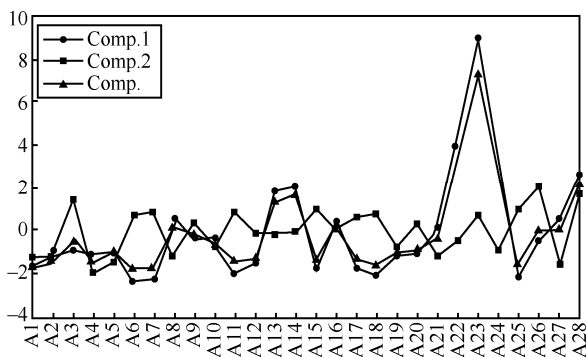


图5 网络新平面主成分得分

由图4、图5可见,新、旧两个平面中,地市A22、A23、A24的第1主成分Comp.1得分较其他地市都要大;第2主成分Comp.2旧平面各地市的差异起伏较大,而新平面各地市表现较平稳。旧平面中,由于第2主成分Comp.2的占比不算小,因此主成分综合指数Comp.介于第1主成分Comp.1与第2主成分Comp.2之间,与第1主成分Comp.1的偏离度稍大一些;而新平面中,由于第2主成分Comp.2的占比较小,因此主成分综合指数Comp.与第1主成分Comp.1的重合度很高。

两个网络平面在前两个主成分样本的分布如图6所示,两个平面的样本分布并未呈现出规律性的结构,说明很难对它们进行归类。但旧平面的A22、A23和新平面的A23远离其他地市,它们的第1主成分Comp.1特别大,说明这几个地市

的时延受第1主成分Comp.1的重要节点X3、X7、X8、X9的影响很大,在优化这几个地市的时延时建议着重优化X3、X7、X8、X9节点所对应的设备及链路。对于第2主成分Comp.2较大的地市,如旧平面的A5、A10、A19、A24,建议重点优化第2主成分Comp.2的大权重节点X1、X2和X6所对应的设备及链路。

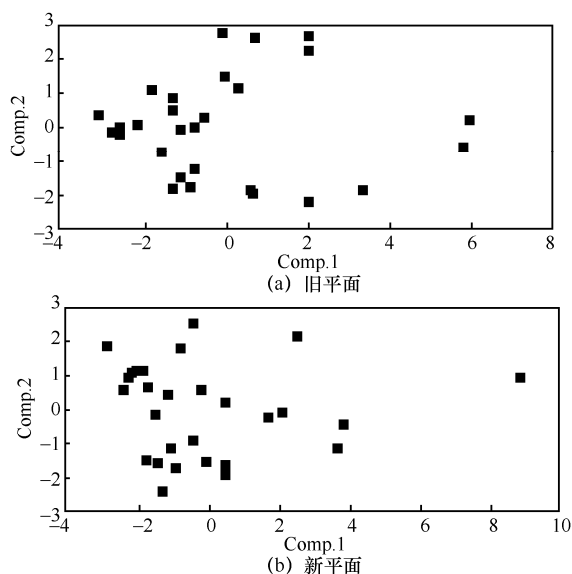


图6 各地市网络时延分布

4.3 时延评价及建议

为了更好地对各地市的IP网络时延进行分析和评估,可以基于主成分分析结果建立一个针对网络时延的评价体系。这个评价体系与通常采用的平均值、中位数等简单指标来评估的系统不同,可以更科学、更准确地揭示网络时延的真实情况,以及各地市时延与各大型节点之间的内在关联关系。

两种方法对两个平面中各地市网络时延进行排序如图7所示,由图7可知,新、旧两个平面两种方法的网络时延排序情况在头、尾两端一致性比较好,但中间区域则出现较大差异,如图7(a)旧平面的A1、A2、A4、A15、A16、A28等地市,图7(b)新平面的A1、A2、A4、A15、A17、A20、A26等地市,而采用主成分分析结果的排

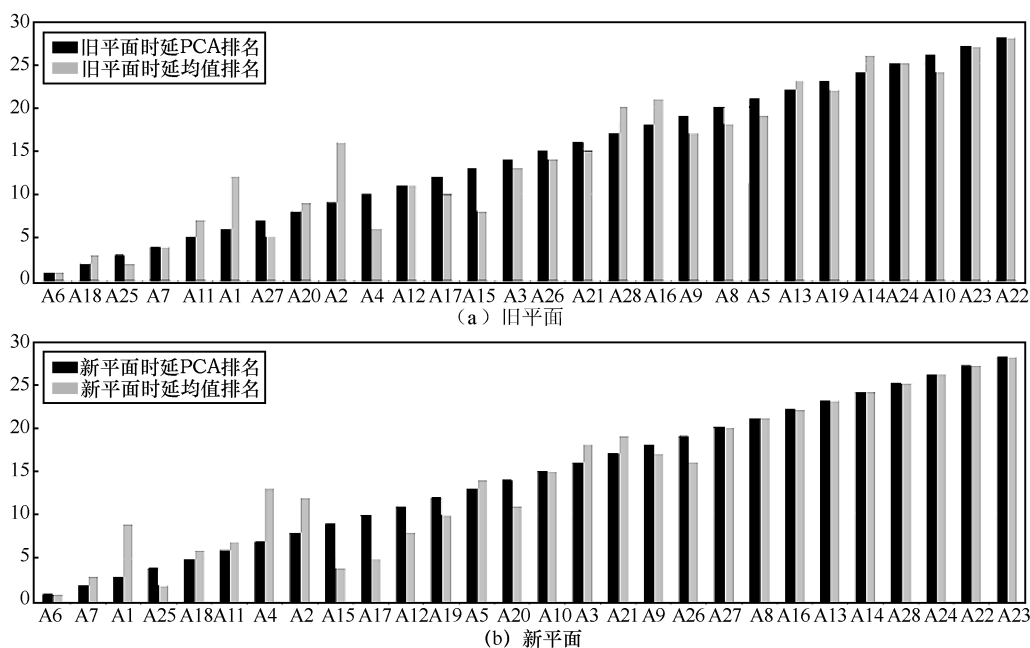


图7 各地市网络时延排名情况

序更符合网络情况的客观实际。因此,这两种方法的差异表明,通常所使用的平均值评价方法对于多指标、复杂变量系统是不适用的,需要寻求更合适的方法(如主成分方法)来进行分析。

为了研究新平面相对旧平面的网络时延的改进情况,用以上得到的新平面的主成分综合指数 $Comp.$ 减去旧平面的主成分综合指数 $Comp.$, 得到两者的差值并画图,对应图8中的深灰色柱子,而图8中浅灰色柱子则对应新平面网络时延平均值减去旧平面网络时延平均值所得到的差值。从左往右按主成分综合指数 $Comp.$ 差值从大到小排序,即新平面相对旧平面时延性能的改善程度由差到优排序,最左边最差的地市是 A23, 即新平面的网络时延相对旧平面未减反增,最右边改善最好的地市是 A22, 即新平面的网络时延相对旧平面大为减少了。主成分综合指数法评价结果显示,新平面相对旧平面网络时延得到改善的地市占了多数。

主成分综合指数法及平均值法得到的网络时延改善情况对于大多数地市是吻合的,但也有一些地市两种方法得到的结果是相反的,如 A15、A17、A1、A2 等,而事实证明主成分分析方法的

结果与网络实际情况及其他相关指标所得到的结果是一致的。为何会出现这样的差异?这说明常规利用平均值进行评估判断的方法是存在问题的,这种方法可能因为数值的平均化而掩盖了事物的主要矛盾,而主成分分析方法则通过抓住事物的主要矛盾和本质关联得以很好地避免此类问题的出现。针对多指标、复杂变量评估系统的情况,主成分分析方法可以提供更科学合理的参考依据。

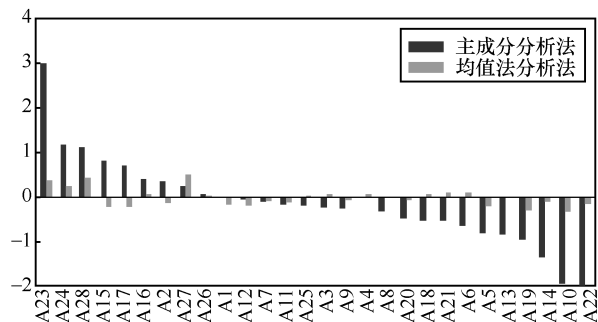


图8 两种分析方法下网络时延的改进

通过前面的分析可知,大型节点 X3、X7、X8、X9 在两个平面的第1主成分 $Comp.1$ 的权重占比较大,旧平面的 X1、X2,新平面的 X5、X6



在第2主成分 Comp.2 中占比较大,因此在实际工程建设中进行网络时延优化时,应重点对这些权值较大的节点进行调整和改造,如更新设备、增加带宽等,尤其是第1主成分 Comp.1 权值较大的节点;对于新平面时延性能大大劣于旧平面的地市 A23、A24 等,建议通过增加到大型节点 X3、X7、X8、X9 的直达方向以减少流量绕转、适当扩容带宽以降低拥塞可能性等方式进行改造;对于新平面其他地市的时延优化,建议着重优化 X5、X6 节点将更容易达到较好的结果。另外,通过对新、旧平面的 PCA 分析得知,增加大型节点之间的连接和保持节点间的均衡性有利于网络时延性能的提升,网络扁平化依然是未来网络发展的方向。

5 结束语

网络时延是衡量网络通信性能的关键指标之一,对于具有多变量、多指标的复杂的大型 IP 网络时延,简单的平均值分析方法不再适用,而主成分分析方法则为此提供了一种更科学的统计方法及分析思路,它能够挖掘网络时延的深层原因及网络各节点间的相互依赖关系,最终获取网络建设、改造优化的有效建议。目前,本文仅对历史网络时延矩阵进行离线的分析,今后可考虑将主成分分析方法应用到针对网络流量、网络时延、网络丢包等网络性能的实时在线监测中,同时可结合网络拓扑结构、现网流量流向、路由、距离等相关因素进行更广泛深入的分析,发挥数据挖掘统计方法的更大效能,进一步提升网络运营的效率和质量。

参考文献:

- [1] 陈小峰, 赵雅迪, 张利鹏, 等. 电力潜在敏感客户预测的大数据方法应用[J]. 电信科学, 2019, 35 (11): 117-124.
CHEN X F, ZHAO Y D, ZHANG L P, et al. Application of big

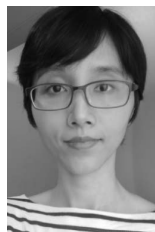
data method in forecasting potential sensitive customers of electric power[J]. Telecommunications Science, 2019, 35(11): 117-124.

- [2] 许方敏, 伍丽娇, 杨帆, 等. 时间敏感网络 (TSN) 及无线 TSN 技术[J]. 电信科学, 2020, 36 (8): 81-91.
XU F M, WU L J, YANG F, et al. Time-sensitive networking (TSN) and wireless TSN technology[J]. Telecommunications Science, 2020, 36(8): 81-91.
- [3] 李昊奇, 应娜, 郭春生, 等. 基于深度信念网络和线性单分类 SVM 的高维异常检测[J]. 电信科学, 2018, 34(1): 34-42.
LI Q H, YING N, GUO C S, et al. High-dimensional outlier detection based on deep belief network and linear one-class SVM[J]. Telecommunications Science, 2018, 34(1): 34-42.
- [4] 杨明极, 刘恺悌, 邵丹. 用于 WLAN 室内定位的 PCA 聚类算法[J]. 电信科学, 2016, 32(7): 21-26.
YANG M J, LIU K Y, SHAO D. PCA clustering algorithm for indoor positioning in WLAN[J]. Telecommunications Science, 2016, 32(7): 21-26.
- [5] 林川, 赵海, 毕远国, 等. 互联网网络时延特征研究[J]. 通信学报, 2015, 36(3): 149-160.
LIN C, ZHAO H, BI Y G, et al. Research on network delay of Internet[J]. Journal on Communications, 2015, 36(3): 149-160.
- [6] 牟澄. 互联网流量特征智能提取关键技术研究[D]. 北京: 北京邮电大学, 2014.
MU C. Research on key technologies of intelligent extraction of Internet traffic characteristics[D]. Beijing: Beijing University of Posts and Telecommunications, 2014.

[作者简介]



韦烜 (1974—), 女, 中国电信股份有限公司研究院工程师, 主要从事 IP 网络规划、新技术研究方面的工作。



黄晓莹 (1980—), 女, 中国电信股份有限公司研究院工程师, 主要从事 IP 网络架构演进及新技术研究方面的工作。