



一种基于主成分分析的综合网管 系统设计及实现

孔祥云, 初铁男

(中国联合网络通信有限公司网络分公司 北京 100033)

摘要:在大数据时代,能否有效利用数据已经成为决定运营商竞争力的重要因素。当前运营商对网络数据分析的主要方法是基于数据仓库进行联机分析处理(OLAP)。OLAP 的性能严重依赖数据模型的预先定义,其无法对业务和状态实时变化的现有网络情况进行及时有效的跟踪,针对这一问题提出了一种基于主成分分析的网络数据挖掘方法,并通过对实际网络数据的分析、比较,验证了该方法的有效性。

关键词:综合网管;大数据;主成分分析;数据挖掘

doi: 10.3969/j.issn.1000-0801.2013.12.020

Design and Implementation of an Integrated Network Management System Based on Principal Component Analysis

Kong Xiangyun, Chu Tienan

(China United Network Communications Group Co., Ltd., Beijing 100033, China)

Abstract: In the big data era, efficient using of data has become an important factor in determining the competitiveness of operators. Currently the main method for network operators in data analysis is OLAP based on data warehouse. As the performance of OLAP analysis relies heavily on pre-defined data model, it's difficult for the system track the real-time changes of wireless network, which cause the current network management system always doesn't meet the actual needs of operators. A method to solve the above mentioned problem based on principal component analysis was presented, and the implementation example and the comparison of the method were given.

Key words: integrated network management, big data, principal component analysis, data mining

1 引言

随着网络规模的不断扩展,目前网络管理系统中的数据规模正在经历着“爆炸性”增长,如何有效利用这些数据和网络并利用数据中的信息辅助决策,已经成为运营商迫切需要解决的问题;另一方面,随着 3GPP 在 LTE 中提出自组织网络(self-organized network, SON)的概念,未来的网络维护、优化、升级等工作也将越来越多地依靠网络数据自发实现,这给运营商对数据的处理和运用能力提出了更

高的要求。

在这种情况下,各大运营商纷纷建立综合网管平台,实现了对网络数据的有效采集和规范化存储,并开始针对数据中的信息进行数据挖掘。由于网络中的数据具有时变、海量、非结构化等特点,综合网络主要依靠数据仓库进行搭建。同时,基于数据仓库的特点,联机分析处理(on-line analytical processing, OLAP)成为综合网管中数据挖掘的主要手段。

联机分析处理的概念最早由关系数据库之父 Codd E F



于1993年提出。Codd认为,传统基于SQL对大数据库的简单查询已经不能满足用户分析的需求。用户的决策分析需要对关系数据库进行大量计算才能得到结果,而OLAP可以通过多维星型或雪花型模型实现OLAP“立方体”,使得庞大的记录集合可以被迅速、高效地概括、分类、查询和分析。因此,以OLAP为核心的数据存储和分析工具逐渐成为当前数据挖掘工具的主流。

然而,多维OLAP立方体的主要缺点是不够灵活。立方体是通过预先将关系型数据表连接构成固定、特定主题的结构构建的。由于实际网络的业务灵活多变,导致业务模型经常随之发生变化,同时由于不同地区的业务特点有着较大的区别,在系统实际运行时,OLAP的性能通常很难达到预期水平。

OLAP对数据准确性有着较高的要求,数据中的噪音和错误对OLAP的准确性有着较为严重的影响。因此,OLAP对源数据的准确性、数据抽取/清洗/转换处理的有效性都有较高的要求,这不仅提高了计算复杂度,也在很大程度上提高了数据分析系统的构建难度和成本。

本文针对上述问题提出了一种基于主成分分析的电信网络数据挖掘方法,通过数据压缩去除数据中的噪音,最大程度地提取数据中蕴含的各种信息,继而提高数据挖掘的效率以及算法的自适应性。最后,通过对网络中实际数据的分析,结合对比传统OLAP方法的性能,证明本文所提方法能有效提升综合网管的性能。

2 综合网管数据分析的现状和问题

OLAP可以使网络管理人员对网络建立起有效的模型,从而大大提高网络管理人员管理网络的效率。OLAP基于大量的数据分组和表间关联,使得网络管理人员可以不去探究网络事件和数据背后的原因,而把注意力放在数据之间的相关关系上,实现了运营商网络管理水平的显著提高。

OLAP通常基于数据仓库,通过对数据仓库中的数据进行多维分析获取数据中的相关信息。其分析过程如下:首先根据数据分析的要求,通过预先设定的模型将数据构造成数据立方体,继而采用上卷、下钻、旋转等数据分析方法获取数据中的相关性,最后将分析得到的结果返回给用户。

OLAP的核心是数据立方体的构建。为了准确地从数据中获取信息,需要采用多维方式组织和处理数据,而数据立方体就是这样一种数据的组织和呈现形式。数据立方

体的模型就是以数据中多维模式中的一些属性作为分析数据的维度,通过分析不同维度之间的关联获取数据中的信息。数据立方体的形式如图1所示。

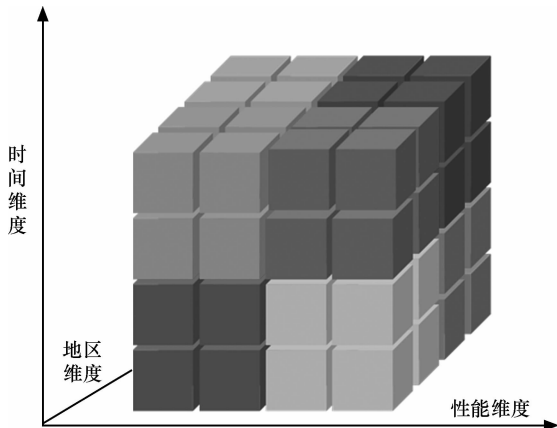


图1 OLAP中的数据立方体模型

然而,在实际应用中,由于业务和网管环境常常发生变化,分析模型也需要实时更新。而由于模型的重新定义和生成不仅需要大量的人力、物力成本,更需要花费大量的时间。因此,网络管理人员对问题分析的角度总是滞后,使得运营商花费大量成本建立的综合网管系统再次沦为死板的日常报表系统。

业务和环境的时变特性是构建综合网管系统的重要挑战,采用OLAP的最大难题就是无法自适应地调整模型以适应业务和环境的变化。因此,寻找一种具有自适应能力的数据分析方法,成为综合网管建立以及运营商数据分析水平提升亟待解决的问题。

3 基于主成分分析的数据分析方法

综合网管的一个主要目标是从采集到的数据中找到影响关键性能指标(key performance indicator,KPI)的主要因素,KPI是运营商对网络质量和网络状况进行判断的主要指标,受到业务和网络环境的影响,影响KPI的主要因素也常常发生改变,给网络管理和维护工作带来很大的困难。

在基于OLAP的综合网管系统中,由于分析模型的构建无法根据业务和网络环境的变化进行自适应调整,加之数据中的噪音也常常对分析产生很大的不良影响,因此在系统实际运行时,其性能往往无法达到预期要求。为了解决这一问题,需要自适应地去除数据中噪音的影响,并将数据中的主要信息提取出来。主成分分析(principal

component analysis, PCA)正是这样一种数学工具。

主成分分析,又称主分量分析,是一种从多个变量中挑选出少数重要变量的多元统计分析方法。在实际的网管数据中,每个变量都可能在一定程度上影响 KPI 的变化。因此,可以采用主成分分析的方法挑选出对 KPI 影响最大的相关关系(即主成分)。基于主成分分析的思想,简要介绍本系统的设计思路。

在相关关系分析的过程中,首先需要构建样本空间,通过数据分析方法分析各个参数对 KPI 所造成的影响。设 $X_i = \{x_1, x_2, \dots, x_n\}$ 为第 i 次测量得到的样本集,其中 x_1 为 KPI,而 $x_2, x_3, \dots, x_n$ 为网管系统中采集到的参数。假设采集到 n 组样本,于是有数据集:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = (X_1 \ X_2 \ \cdots \ X_n) \quad (1)$$

$$\text{其中, } X_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}, i=1, 2, 3, \dots, n。$$

各数据与 KPI 之间的相关关系可以看作这一数据中蕴含的有关 KPI 的信息,由于数据中的信息由方差进行表征,故计算协方差矩阵:

$$Q = (\sigma_{ij}) = E[(X - E(X))(X - E(X))^T] \quad (2)$$

设 Q 的特征值为 $\lambda_1, \lambda_2, \dots, \lambda_n$, 其中 $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n| \geq 0$ 对应的特征向量为 $L_1, L_2, \dots, L_n, L_i = (l_{1i}, l_{2i}, \dots, l_{ni}) (i=1, 2, \dots, n)$ 。由特征向量的特点可以知道, $L_1, L_2, \dots, L_n$ 为一组线性无关向量,因此可以以 $L_1, L_2, \dots, L_n$ 为基,得到样本空间 X 的线性变换 $Y = (Y_{11}, Y_{12}, \dots, Y_{1n})^T$, 其中:

$$\begin{aligned} Y_1 &= L_1^T X = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1n}x_n \\ Y_2 &= L_2^T X = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2n}x_n \\ &\vdots \\ Y_n &= L_n^T X = l_{n1}x_1 + l_{n2}x_2 + \cdots + l_{nn}x_n \end{aligned} \quad (3)$$

在 Y 中,非零元素的个数即 X 的秩,表征在采集到的数据中,可以找到的 KPI 与各参数可能的相关关系。其中, λ_1 对应的 Y_1 是第一主成分,由于 λ_1 为最大的特征值,其对应的线性变换 Y_1 表征 KPI 与网络参数最显著的相关关系,而其他成分 $Y_2 \cdots Y_n$ 表述 KPI 与网络参数其他可能的关系,由于其相关关系较弱,可以认为这些关系都是噪音。

下面需要对获得的相关关系的有效性进行分析,即判

断分析出来的相关关系是否具有充分的显著性,即判断第一主成分所占的比例 $\lambda_1 / \sum_{j=1}^n \lambda_j$ 。只有当第一主成分的贡献率大于某一阈值时,才能认为这一关系有效表征了 KPI 和网络性能的相关性,若第一主成分的贡献率不够大,则认为 KPI 与网络参数 $x_2, x_3, \dots, x_n$ 无明显关系。

以二维样本集 $X = (x_1, x_2)$ 为例说明这一过程。设 x_1 为基站的掉线率, x_2 为基站的 CPU 利用率,假设网管系统采集了 n 个基站的数据,且其分布情况如图 2 所示。对这一数据进行分析,可以看到数据集的第一主成分远大于第二主成分,说明基站的 CPU 利用率与掉话率之间存在明显的关系,其存在的线性相关性为第一主成分的斜率 $\tan\theta$,通过这种方法实现了从大数据中获取相关性的功能。

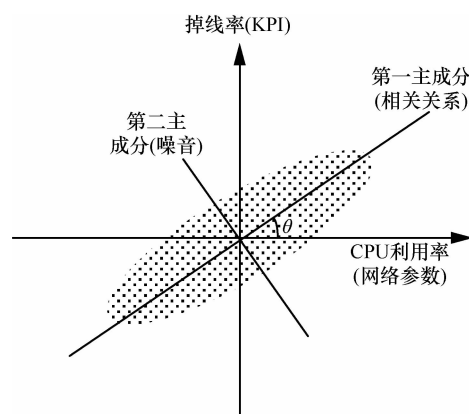


图 2 具有显著相关关系的样本集

另一方面,假设 x_1 为掉线率而 x_2 为通话时长的样本集,且样本分布如图 3 所示。在这一样本中,显然第一主成分与第二主成分的大小差别不显著,此时可以认为通话时长并不影响掉线率。

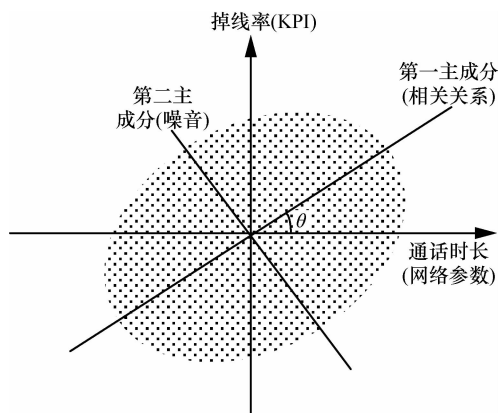


图 3 没有显著相关关系的样本集



综上所述,通过对网络中的采样数据进行数据挖掘,可以准确找到影响掉线率的参数,并定量地分析出其相关关系。同时,通过海量数据的处理与分析,保证了最大程度地利用数据中的信息,并发掘出数据内蕴含的各种相关信息。

4 本文方案的验证与比较

4.1 基于实际数据的方案验证

为了验证本文所提出的综合网管系统中的海量数据分析方法,以中国联通某省公司的实际数据为例,阐述海量数据分析的实际过程以及分析效果。

分析数据为某省会城市 10 000 个小区的统计指标,希望通过海量数据分析技术得到影响掉线率(属于 KPI)的主要因素。为了简化讨论,只针对语音业务时长和数据业务流量两个因素进行分析。根据前文所述方法,分别构建二元组样本集。

样本集 1:(各个小区相邻两个月间总掉话次数的差值,各个小区相邻两个月间总数据流量的差值);

样本集 2:(各个小区相邻两个月间总掉话次数的差值,各个小区相邻两个月间总语音业务通话时长的差值)。

对两组样本进行分析,提出各组样本的主成分,以分析各参数对掉线次数的影响,结果如图 4、图 5 所示。其中,每个点代表样本集中的一个样本,颜色深浅代表对于样本点的不同分类。可以看到,在所分析的样本中,各小区的掉线次数都会随着数据流量的增大而增大;相对的,总通话时长的变化对总掉线次数几乎没有显著影响。

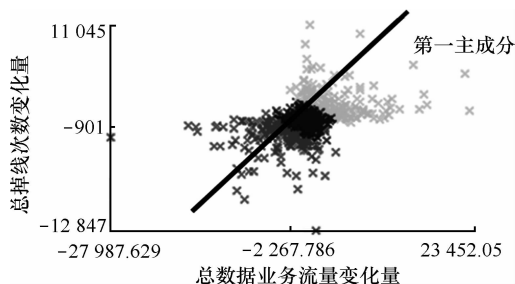


图 4 无线小区掉线次数变化与数据流量变化的样本集

传统上,语音业务和数据业务都可能对小区的掉线产生影响,且语音业务造成的影响更为显著。相关结果不仅有助于网络管理人员更好地管理和配置网络,更为今后的网络扩容、网络优化提供了有益的参考,带来了客观的经济效益。

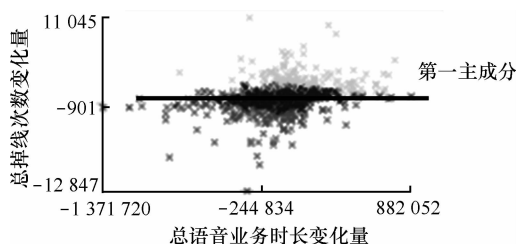


图 5 无线小区掉线次数变化与语音业务时长变化的样本集

4.2 与 OLAP 方法的性能比较

为了对比本文所提出的方法与传统 OLAP 方法的性能,选出 1 000 个无线小区样本,对比两种算法在相关挖掘方面的准确性。其中,样本小区均为城市中心的高负载小区,且样本小区均存在由于 CPU 负载过高引起掉话率高的现象。现将样本小区分为 10 组,对比两种算法判断样本小区掉话原因的准确性,结果如图 6 所示。

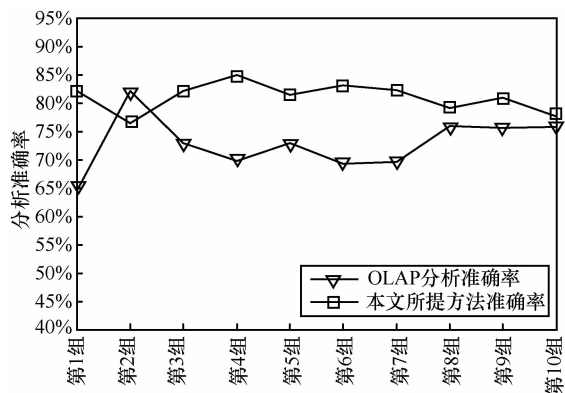


图 6 本文所提方法与 OLAP 方法的性能比较

从图 6 可以看到,采用本文所提的数据分析方法可以有效地提高数据分析的准确性。对于运营商而言,其业务质量并不完全取决于网络的性能参数,但通过不断地对网络中的数据进行分析,采用本文所提出的算法可以在很大程度上及时发现网络运行中可能存在的问题,这样的准确率已经基本可以满足运营商对网络管理的要求,同时这也说明本文所提出的分析方法具有一定的适用价值。

5 结束语

综合网管的应用将成为运营商增强竞争力、提高网络管理效率的重要手段。基于数据分析的综合网管系统未来不仅能够成为公司竞争力的来源,也将成为运营商的重要资产。本文从综合网管的应用需求出发,提出了一种有效自适应获取数据中信息的有效方法。最后通过实际数据的验证及比较分析,验证了本文所提方法的有效性。本文研

(下转第 133 页)

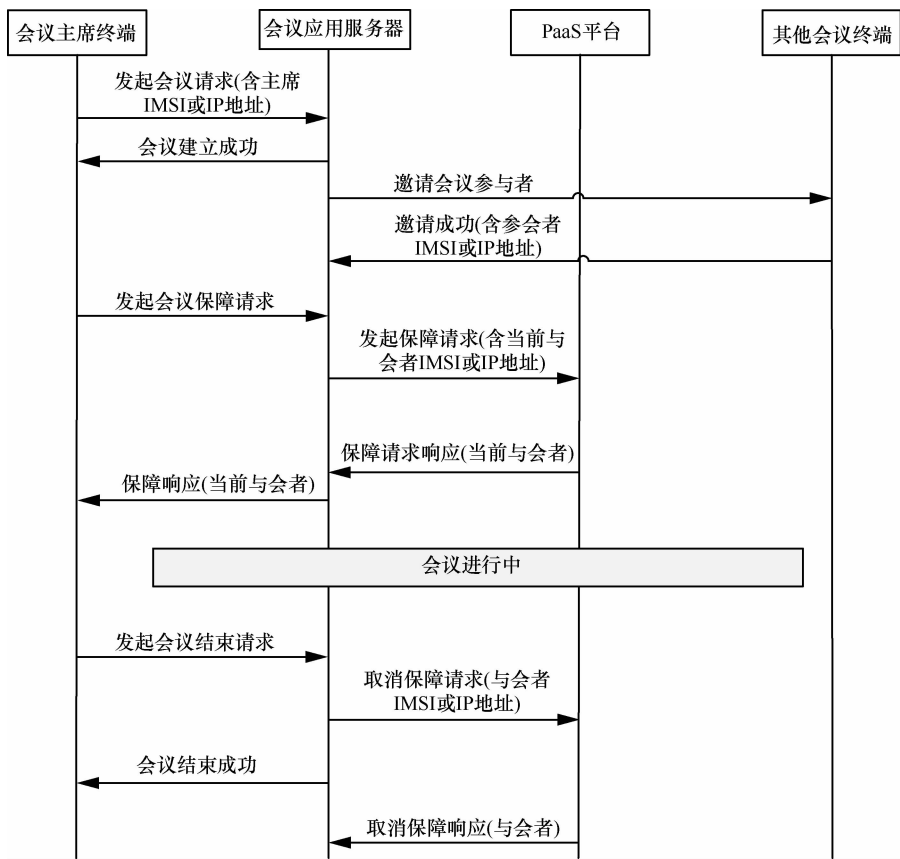


图 2 视频会议应用智能管道能力调用示意

[作者简介]

赵学军,男,中国电信集团公司高级工程师、技术部部长,主要负责电信核心网络技术、业务平台、三网融合和终端技术研究。

陆钢,男,中国电信股份有限公司广东研究院高级工程

师,主要从事业务平台、云计算技术、移动互联网研发工作。

何震苇,男,中国电信股份有限公司广东研究院工程师,主要从事云计算研发工作。

(收稿日期:2013-12-10)

(上接第 128 页)

究成果对未来综合网管的设计、数据的使用有着很大的借鉴意义。

参考文献

- 1 Eryilmaz A, Srikant R. Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. IEEE/ACM Transactions on Networking, 2007, 15 (6): 1333~1344
- 2 Shlens J. A tutorial on principal component analysis. <http://www.docin.com/p-422532195.html>,2009
- 3 Inmon W H. 数据仓库. 北京: 机械工业出版社, 2000
- 4 张维明. 数据仓库原理与应用. 北京: 电子工业出版社, 2002

[作者简介]



孔祥云,男,中国联合网络通信有限公司网络分公司运行维护部工程师、主管,主要研究方向为网络数据统计分析。

初铁男,男,中国联合网络通信有限公司网络分公司运行维护部工程师、处长,主要研究方向为网络管理与优化。

(收稿日期:2013-11-25)