DOI: 10.3969/j.issn.1009-6868.2015.05.015
网络出版地址：http://www.cnki.net/kcms/detail/34.1228.TN.20150921.1011.002.html

基于大数据的业务并发度分析

Service Concurrency Analysis Based on Big Data

中图分类号：TN929.1 文献标志码：A 文章编号：1009-6868 (2015) 05-0060-005

摘要： 指出不同业务之间的关系对于网络优化具有很重要的意义。使用大数据的分析方法处理蜂窝网络的实测数据，可以得到各种业务的并发关系，同时将并发关系通过业务关系网络的方式展现出来，具有很好的可视性。通过业务关系网络可以直接看出业务并发情况，为蜂窝网络的优化扩容等方面提供理论指导。

关键词： 大数据；蜂窝网络；业务并发度；网络优化

Abstract: The relationship between different services is important for network optimization. In this paper, we process the real data in cellular networks with a method used in big data to attain a concurrent relationship between the various services. Then the service concurrency is presented by the way of a relationship network with good visibility. Service concurrency can be seen directly through the relationship network, providing theoretical guidance for cellular network optimization and expansion.

Key words: big data; cellular networks; service concurrency; network optimization

易正磊/YI Zhenglei [1]
顾军/GU Jun [2]
张兴/ZHANG Xing [1]
（1. 北京邮电大学无线信号处理与网络实验室，北京 100876；
2. 中兴通讯股份有限公司，上海 201203）
(1.Wireless Signal Processing and Network Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. ZTE Corporation, Shanghai 201203, China)

1 大数据的意义和影响

1.1 大数据的价值和挑战

大数据是由于规模、复杂性、实时而导致的使之无法在一定时间内用常规软件工具对其进行获取、存贮、搜索、分享、分析、可视化的数据集合。由于大数据的上述特点，如何将数据进行合理应用是目前大数据领域的核心目标[1]。

大数据对于整个社会来说蕴含着巨大的潜在价值，大数据的价值并不在于数据本身，而在于如何将数据的作用反馈于社会决策。维克托·迈尔-舍恩伯格在《大数据时代》中指出，人类从依靠自身判断做决定到依靠数据做决定的转变，是大数据做出的最大贡献之一。因此，能否正确利用大数据的内在规律，是决策成功或者失败的关键因素[2]。

一般来讲，数据的生命周期包括数据采集、数据归纳、数据重构、数据挖掘、数据预测、数据可视化等6个方面，大数据亦是如此。但是由于大数据的体积庞大、结构复杂，常规的处理方法并不能挖掘出数据的内在价值，这也正是大数据时代人们面临的最大挑战。

1.2 大数据对移动互联网的影响

在移动互联网的环境中，大数据的特点并不仅仅体现在数据量的巨大，更体现在数据的实时性与关联性，这些特点让数据搜集变得容易，对数据的分析可以反过来快速影响基础网络，这也正是移动互联网的自身特点。在大数据与移动互联网高度融合的背景下，最重要的目的就是发现和挖掘真正有价值的数据，使我们能够根据这些信息精确地指导每一次的网络规划和优化，而这些有价值的信息往往来源于对用户大量网络行为数据的抽象和分析。

如图1所示，大数据应用于移动互联网的基本目的包括：负载优化、用户划分、趋势预测等。

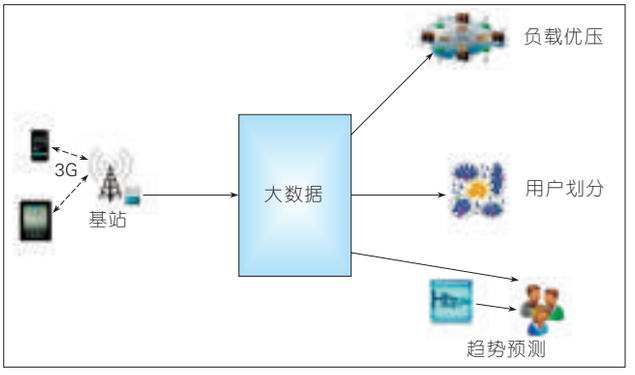
2 蜂窝网络中的大数据应用

2.1 蜂窝网络的发展现状

蜂窝网络是一种最常见的移动通信网络结构，通过把移动设备的服务区分为很多正六边形的子区域，并在每个子区域设置一个基站，形成了形状似“蜂窝”的结构，因而把这种移动通信网络称为蜂窝网络。

作为支撑信息传输的关键基础设施，蜂窝网络的性能优劣对于整个通信系统的效率起着至关重要的作用。近年来，为了应对移动网络流量爆炸性增长，欧洲、美洲、日韩等地区

基金项目：国家自然科学基金（61372114）；国家重点基础研究发展（“973”）规划（2012CB316005）
收稿日期：2015-08-17
网络出版时间：2015-09-21



◀图1
大数据在移动互联网中的应用

已经大规模建设长期演进(LTE)网络,中国也迎来了LTE网络建设高峰。在这一建设进程中,除了基于IP多媒体子系统(IMS)的Vo LTE作为最终的语言解决方案之外,Small Cell和无线局域网(WLAN)将成为提升网络容量的重要手段,在超宽带移动网络部署[3-4]中发挥越来越重要的作用。此外,行业应用正从窄带向宽带演进,除了语音通信外,数据、视频传输需求逐步增加。LTE的100 Mbit/s高速数据传送能力,可以更好地服务于政务网、公共安全和应急救灾等行业。因此,利用新一代无线技术来实现行业应用,已成为一个发展趋势[5]。

随着分布式计算和云平台的逐步实现,作为云服务的关键应用之一,大数据应用具备了大规模发展的条件[6]。运营商的大数据业务正从分散孤立系统向统一的标准化平台方向发展、从数据的采集、存储向检索和挖掘的方向发展。在网络大数据的收集与预处理方面,分布式存储方案、内存数据库技术将成为主流技术;Hadoop和一体机成为主流的数据分析平台。

根据以往的研究,蜂窝网络的承载情况随着通信协议的发展也发生着巨大变化。根据某一地区的真实历史数据分析可以得知,全市所有基站的平均吞吐量在2G、3G和4G网络的时代是有着很大区别的,具体数值如图2所示。

根据图2可以看出,发展速度最快并且承载压力最大的信道是下行数据信道,而目前国内的4G网络还处于发展初期,吞吐量的增长了已达到将近100%。由此可知,数据业务的有效传输是蜂窝网络的核心难题。

此外,随着移动互联网、物联网的发展以及4G技术的逐渐普及,无线通信网络的各种功能需求也日益扩大,用户的网络行为也随之发生改变,所产生的数据体积也达到了新的数量级。因此,如何利用有限的蜂窝网络资源,应对大数据时代的挑战,是当今通信领域的一个重要课题。

2.2 有效的网络优化

当今的蜂窝网络的发展速度明显慢于基站吞吐量的增长,而用户日益丰富的行为对蜂窝网络的承载能力提出了更高的要求。可见,优化网络,提升网络负载能力已经成为一个亟待解决的问题。

我们已经提到,大数据时代的到来使得人们的思维方式发生了巨大的变革,数据驱动了决策制定,因此对蜂窝网络的优化策略正是通过对蜂窝网络中产生的数据分析处理来制订。同时,蜂窝网络系统本身就是一个巨大的数据仓库,我们可以从中采集到丰富的数据,通过对这些数据的分析,可以识别用户的地理位置,洞察客户接触不同信息的渠道,了解用户的各种网络行为。常用的分析方法包括:预测业务流量、探寻不同业务之间的关联、探寻不同业务模型下的资源瓶颈等等。

接下来以业务并发度探寻实例来阐述在蜂窝网络系统中如何应用大数据来解决问题。在该实例中,我们分析各种常见业务在不同场景不同时间的并发度,最终得出不同区域的业务并发以及对网络资源的消耗情况,并据此提出网络优化策略。

3 大数据时代的业务并发度分析

3.1 应用场景

随着通信技术的飞速发展和移动终端的功能逐渐丰富,蜂窝网络中的数据业务流量迅猛增长,移动互联网下终端用户需求更加多样化和复杂化,这也促使移动互联网由传统的单业务向着多业务平台发展。多业务的出现使得网络数据更趋向于复杂多样与结构各异,这给用户的行为分析带来了很大的困难和挑战。与此同时,由于业务种类的繁多,分析某一种业务对于整体蜂窝网络影响甚至微乎其微,因此为了提高网络的承载能力,需要科学准确地分析各数

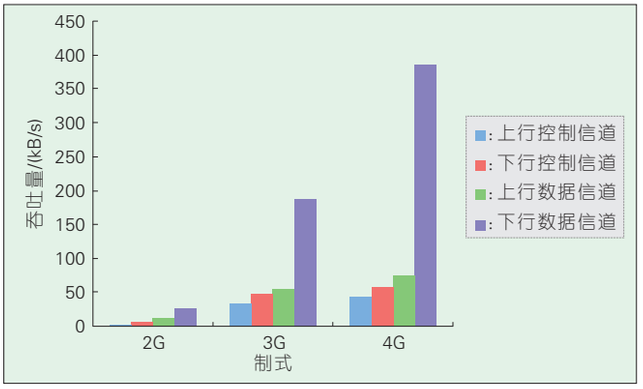


图2▶
不同制式下基站平均吞吐量

据业务之间的并发性。

由图3可见，由于目前蜂窝网络用户数和终端数快速增长，以及业务场景多种多样，数据种类也趋于全面，包括用户使用各业务的时间信息、位置信息、链接次数、业务量大小等，因此，用户——业务网络资源之间的映射十分复杂，海量的数据具有极低的价值密度，如果仅仅对某一部分数据进行分析，不能反映整体网络情况，所具有的价值意义也就很小。

此外，无线侧采集到的数据格式并不统一，应用传统数据库无法建立统一的数据表结构。根据数据的以上特点，我们采用Hadoop平台进行数据的预处理与所有算法的执行，实现了高效处理非结构化数据。

3.2 核心技术

为了分析不同场景下的海量数据，首先平台应该具有较高的数据处理能力。以无线侧的呼叫细节记录（CDR）数据为例，某地区全市全天CDR数据条数达到千亿数量级，数据体积约为70 GB，连续一个月的数据量将达到2 TB，如果空间维度扩展到全国，时间维度扩展到几年，这样的数据量是传统数据库和单机环境的处理能力远远不能达到的，而以Hadoop为代表的大数据处理平台则可以轻松应对[7-8]。

其次，平台能够处理并管理非结构化数据。非结构化数据相对于结构化数据而言，不方便用数据库二维逻辑表来表现的数据即称为非结构化数据，包括所有格式的办公文档、文本、图片、标准通用标记语言下的子集XML、HTML、各类报表、图像和音频、视频信息等等。这些数据需要特殊的筛选方法进行预处理和归一化，之后才可以应用到各类数据挖掘算法中去。

再次，算法需要分析复杂场景并对场景进行区分。如果将全市或者全国所有的数据不加以区分而进行统一处理，那么所有的特殊区域都将被平均化，然而某些区域的特殊情况的内在原因正是数据分析者们更感兴趣的地方，因此将数据进行合理的场景划分，是数据预处理的中重要一步。

最后，平台必须能够分布式结构并行处理，提高海量数据的处理速度。由于大数据更新速度快的特点，能否快速实时的对海量数据进行处理是整个数据分析的基础。

3.3 处理流程

为了探究未来一段时间某地区的业务并发关系，我们需要对该地区采集到的海量数据做预处理。原始数据包括了许多字段，如基站信息（经纬度）、网络类型、上下行流量、数据包数、业务持续时长等。我们主要研究各业务之间的联系，因此只需提取出与业务量有关的字段。首先，从所有数据中出剔除了许多不常用的或是数据不全的业务，最终得到了60种有效业务，这些业务对象包括了除语音、短信外所有数据类业务，如即时消息、社交、流媒体、邮件等；然后再从这60种需要分析的业务中选取可能会使用到的各业务的上下行流量、用户数、数据包个数等；最终我们选取了各业务流量来探究业务并发关系。我们将相关的数据整理为表1格式。

现在各种网络业务越来越多，但是各种业务之间并不是孤立存在的，用户使用习惯、业务本身属性等都会使得各业务是息息相关的。为了衡量业务之间关系的大小，我们定义了各业务之间的距离。

对于采集到的N种业务，要得到第i（i=1,2,3……N）种业务与第j（i=1,2,3……N，$j\neq i$）种业务之间的距离，首先需要计算出第i种业务与第j种业务的皮尔逊相关系数，计算公式为：

$$\rho_{ij}=\frac{\mathrm{Cov}(x^{(i)},x^{(j)})}{\sqrt{\mathrm{Var}(x^{(i)})}\sqrt{\mathrm{Var}(x^{(j)})}} \quad (1)$$

其中$x^{(i)}$，$x^{(j)}$分别为第i，j两种业务流量的时间序列。在得到业务相关系数的基础上，我们就可以计算第i种业务与第j种业务之间的距离d_{ij}[9]，计算公式为：

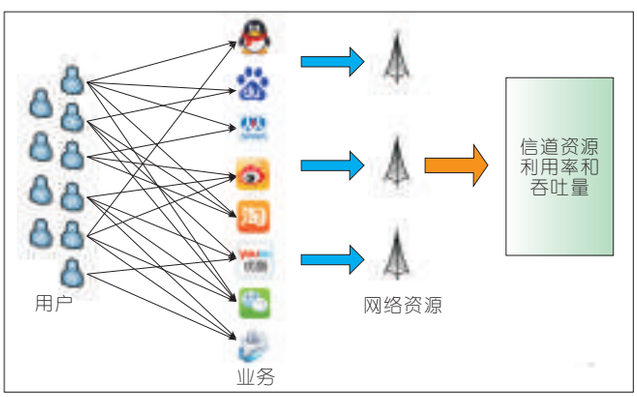
$$d_{ij}=\sqrt{2\left(1-\rho_{ij}\right)} \quad (2)$$

为了直观地看出各业务之间的并发关系，我们使用kruskal算法构建最小生成树网络[10-11]，对于由N种业务之间的C_N^2个距离构建的集合U，首先找出U中的最小值，即距离最小的两种业务，在这两种业务之间添加一条无向边，连接这两种业务，接着在剩下的C_N^2-1条边中继续寻找最小值连接业务，同时保证业务之间不连成环，直至遍历所有距离值。其流程如图4所示。

3.4 结果展示

图5所示为某使用地区，从2014年1月5日开始连续15天全网60种业务构建的最小生成树业务网络。

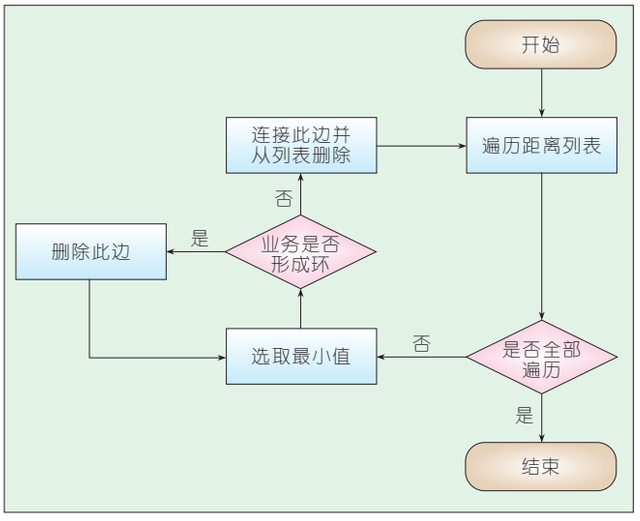
在该业务网络中，各节点代表了不同业务，如快播、优酷、微信、淘



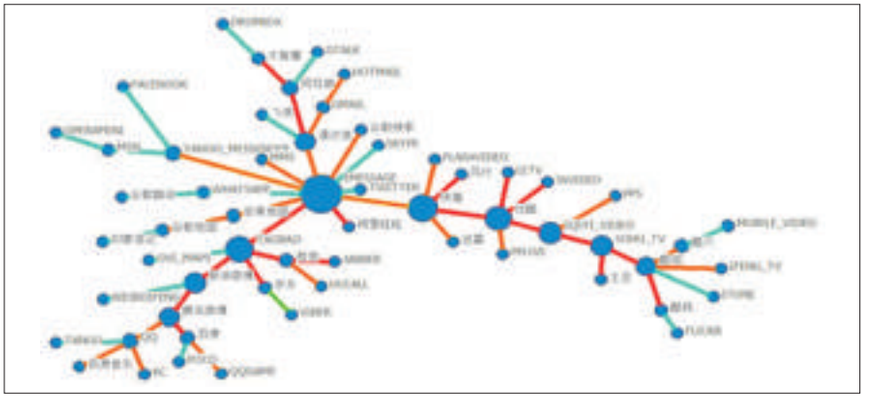
◀图3
用户—业务—网络资源之间的映射

▼表1 业务流量数据示例

时间	JABBER的流量/M	GTALK的流量/M	56VIDEO的流量/M	优酷的流量/M	土豆的流量/M	风行的流量/M	阿里旺旺的流量/M
2014/1/5 8:00	13 22	0	2 858	714 084	32 236	27 502	595
2014/1/5 9:00	23 740	194	577 924	163 836	615 363	70 646	2 997
2014/1/5 10:00	20 451	0	529 369	174 232	335 228	90 245	3 293
2014/1/5 11:00	19 979	0	155 743	137 776	56 886	11 620	3 085
2014/1/5 12:00	17 595	0	151 262	953 784	118 481	98 255	3 155
2014/1/5 13:00	16 151	0	210 168	829 077	189 708	36 301	2 594
2014/1/5 14:00	17 448	175	198 973	679 424	93 845	84 362	2 685
2014/1/5 15:00	22 507	0	69 295	546 109	40 653	60 004	2 778
2014/1/5 16:00	31 648	0	130 530	780 224	108 563	66 752	3 788
2014/1/5 17:00	31 085	0	151 783	716 352	75 607	58 511	4 587
2014/1/5 18:00	27 859	1 067	96 042	195 300	44 656	8 781	6 342
2014/1/5 19:00	33 841	710	32 882	139 539	15 262	13 070	6 676
2014/1/5 20:00	26 495	475	98 747	115 989	30 487	15 968	5 618
2014/1/5 21:00	23 522	965	89 497	103 930	46 340	28 453	4 424
2014/1/5 22:00	31 415	200	189 714	105 917	97 643	2 898	5 971



◀图4 业务网络构建流程



▲图5 最小生成树业务网络

宝、新浪微博等。节点大小是由该业务在网络中介数的大小决定的，节点越大代表了该业务介数越大。而对于这样一个由业务构成的网络来说，业务介数越大，其他的业务就越容易通过它关联在一起，例如在图5中，淘宝与优酷这两种业务并未直接连接在一起，这说明它们之间的相关性不是最大的，但是它们还是可以通过一定的路径连接起来，从图中具体来看就是：淘宝——IMESSAGE——快播——优酷，而且从我们构建该网络的方法来看，这样连接起来的路径一定是相关系数权重最大的。对于其他业务也可以此类推，可以看出任意两种业务相关的路径通过IMESSAGE的次数是最多的。

该网络中的边选取了不同颜色来标注，不同颜色代表不同的业务距离范围。在阐述该网络的构建方法时我们就已经说明，边连接的节点是两种距离最小的业务，因此从结合点与边即可看出各业务之间的关系：距离越小就说明两种业务越容易并发。从图5中可以发现任意一种业务与其他业务的并发情况。

对于不同地区或不同时间段的业务数据，我们都可以构建出最小生成树业务网络，从该网络中找到各业务之间的关联关系，通过业务关联关系可以预测出任意一种业务在未来一段时间与其他业务的并发情况。例如采用某地区一定时间内的数据可以得到图5所示的业务网络，从中可以预测出，在出现优酷这种业务时，很可能会同时出现快播、LETV（乐视）、56VIDEO、PPLIVE这4类相关的业务。

该业务关系网络图是根据历史数据得出。为了预测未来一段时间的业务并发关系，需根据预测需要，不断使用新数据来更新业务关系网络图，从而保证预测的准确性。

得到不同业务之间的并发关系后，我们就可以结合不同业务对网络资源的消耗情况对网络的调控与优

化提供一定的理论指导。例如，如果某地区的业务呈现出图5所示并发情况，则可以知道IMESSAGE业务会与多种业务并发，那么在做网络调控时，需要优先满足该业务消耗的信道资源。同时对于容易并发的业务，在做优化时可以当做同一类业务来处理，因为它们会同时消耗不同的网络资源。

4 结束语

提出了一种大数据背景下基于业务并发度来分析用户网络行为的方法，该方法分析所得到的结果可以对网络规划和优化进行理论指导。我们需要进一步分析不同种类的业务对于蜂窝网络资源消耗的映射关系，从而精确预测整体网络的负载情况，并据此提出更准确、更全面的网络优化指导。

参考文献
[1] Howe D, Costanzo M, Fey P, et al. Big data: The Future of Bio Curation [J]. Nature, 2008, 455(7209): 47–50
[2] Viktor Mayer Schonberg. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M].UK: Hodder & Stoughtonr, 2013
[3] Hwang I, Song B, Soliman S S. A Holistic View on Hyper–Dense Heterogeneous and Small Cell Networks [J]. Communications Magazine, IEEE, 2013, 51(6), pp.20–27
[4] Hoadley J, Maveddat P. Enabling Small Cell Deployment with HetNet [J]. Wireless Communications, IEEE, 2012, 19(2): 4–5.doi: 10.1109/MWC.2012.6189405
[5] LTE: the UMTS Long Term Evolution [M]. New York: John Wiley & Sons, 2009
[6] 张建勋, 古志民, 郑超. 云计算研究进展综述 [J]. 计算机应用研究, 2010,27(2): 429–433
[7] 吴吉义,平玲娣,潘雪增. 云计算: 从概念到平台 [J]. 电信科学, 2009 (12): 23–30
[8] Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data[M]. McGraw–Hill Osborne Media, 2011
[9] Zheng Z, Yamasaki K, Tenenbaum J N, et al. Carbon–Dioxide Emissions Trading and Hierarchical Structure in Worldwide Finance and Commodities markets [J]. Physical Review E, 2013, 87(1): 012814
[10] Gallager R G, Humblet P A, Spira P M. A Distributed Algorithm for Minimum–Weight Spanning Trees[J]. ACM Transactions on Programming Languages and systems (TOPLAS), 1983, 5(1): 66–77
[11] Cheriton D, Tarjan R E. Finding Minimum Spanning Trees[J]. SIAM Journal on Computing, 1976, 5(4): 724–742

作者简介

易正磊，北京邮电大学无线信号处理与网络实验室在读研究生；目前主要研究领域为大数据与社会网络分析。

顾军，中兴通讯FDD LTE无线网络规划总工；长期从事3/4G无线组网方案研究与应用；申请国家专利10项，国际专利2项，发表技术论文5篇。

张兴，北京邮电大学副教授、博士生导师；主要研究领域为异构无线网络融合、物联网与蜂窝网络融合等；近期出版专著3本，发表SCI检索论文7篇，其他国内期刊会议论文等30余篇。

综合信息

Gartner调查报告称：信息安全管理实践正日趋完善

全球领先的信息技术研究和顾问公司Gartner近日表示，越来越多除IT以外的其他部门也在设置安全保障，这一趋势反映了有效安全管理的必要性。

根据Gartner公司有关终端用户隐私、IT风险管理、信息安全、业务连续性和合规管理的年度调查结果显示：信息安全管理实践正在日趋成熟。2015年2—4月，Gartner在全球7个国家开展调查，964名在大型机构工作的受访者参与了调查，这些大型机构在2014财政年度的总收入不低于5 000万美元，拥有不少于100名员工。

Gartner的副总裁兼院士级分析师Tom Scholtz表示："人们对数字业务的风险意识日益增强，再加上有关网络安全事件的深度宣传，促使IT风险列为董事会级别的讨论事宜，71%的受访者表示，IT风险管理数据影响董事会层面的决策。这也反映出，越来越多的企业开始重视把应对IT风险作为公司管理的一部分。"

Scholtz指出，在IT以外的部门建立汇报程序的主要原因在于促进执行与监督相分离，提升企业信息安全的形象，打破员工和利益相关者认为"安全只是IT问题"的思维定势。企业日益认识到，必须将安全性作为企业风险问题进行管理，而不仅仅是一个IT运营问题。

支持安全项目的高管级别也在提高。63%的受访者表示，他们的信息安全项目获得的资助和支持来自IT以外部门的领导层，这一调查数据相较于2014年的54%有了大幅上升。企业首席执行官和董事会的支持率保持不变，为30%（2014年是29%），而指导委员会的支持率从7%升至12%。地区差异引人注目，57%的北美受访者表示信息安全项目获得的支持来自IT以外部门，明显低于西欧的63%和亚太地区的67%。

Scholtz认为，企业高管重视安全项目至关重要，否则安全项目将难以得到企业其他部门的必要支持。在安全策略的有效性方面，虽然有一半的受访者表示管理层参与评估和审批这些策略，但仅30%的受访者表示，业务部门会积极参与到这些会影响其业务的策略制订中来。（转载自《中国信息产业网》）