基于网络编码的文件备份方案

林国庆[1,2] 陈汝伟[3] 李 颖[2] 王新梅[2]
(长安大学汽车学院 西安 710064)[1]
(西安电子科技大学综合业务网理论及关键技术国家重点实验室 西安 710071)[2]
(桂林电子科技大学数学与计算科学学院 桂林 541004)[3]

摘 要 针对数据备份需占用大量空间和需额外手段对其安全保密性进行保护的问题,提出了一种基于网络编码技术的文件备份方案,其核心是对多个备份文件进行网络编码操作,生成编码备份文件后存储于备份服务器中。该方案包括以下部分:备份的基本原理、备份过程、文件更新对备份恢复的影响和处理、备份的更新和全局备份控制系统。实验与理论分析表明,该方案较传统备份方案大幅节省了存储空间,提高了备份文件的安全保密性,但文件的可恢复性微降并增加了系统的复杂度。

关键词 网络编码,数据备份,文件备份,存储空间,安全保密性
中图法分类号 TP391.41 **文献标识码** A

Data Backup Scheme Based on Network Coding

LIN Guo-qing[1,2] CHEN Ru-wei[3] LI Ying[2] WANG Xin-mei[2]
(School of Automobile, Chang'an University, Xi'an 710064, China)[1]
(State Key Lab of Integrated Services Networks, Xidian University, Xi'an 710071, China)[2]
(School of Mathematics and Computational Science, Guilin University of Electronic Technology, Guilin 541004, China)[3]

Abstract According to the problems that data backup needs a lot of storage space and its safe privacy needs to be protected, a file backup scheme based on network coding was proposed. Specifically, adopting network coding, many backup files are encoded into one encoded file, and then saved in the backup server. This scheme includes the following aspects: the principle of backup, the process of backup, influence of file update on recovery and how to handle it, backup file update and the global control system. Experiment and theoretical analysis show, compared with traditional method, this scheme can save storage space greatly and enhance the safe privacy of the backup data, but decrease appreciably the recoverability of it and increase the complexity of the system.

Keywords Network coding, Data backup, File backup, Storage space, Safe privacy

1 引言

数据备份,就是在现时使用的数据之外,实现或设置另外一份不同物理体现的、内容相同的有效数据拷贝。文件备份是指针对文件的备份,是数据备份的一种。传统数据备份方法是将多个主机通过局域网(Local Area Network, LAN)或存储网络(Storage Area Network, SAN)与备份服务器相连,备份时各主机将需备份的数据传输到备份服务器上独立存储。节省磁盘存储空间是众多备份软件所追求的共同目标之一。现有的备份软件多是通过各种方式的压缩来减少备份所需空间[1-3]。而这种方式只是对文件个体的缩小,对空间的节省是很有限的。并且,要实现备份数据存储的安全保密性,必须采取额外的手段,如加密[4,5]等方式实现。为此,本文提出了一种新的文件备份方案,利用网络编码技术,首先对多个主机传输的需备份的文件进行编码,生成编码备份文件后存储于备份服务器中,从而达到节省存储空间的目的,且自动实现备份文件存储的安全保密性。本文首先介绍网络编码的原理,在其基础上给出具体的基于网络编码的文件备份方案,对方案涉及到的一些具体问题进行了分析并提出了相应的处理方法。

2 网络编码

传统的通信网络中,中间节点仅进行数据包的存储、转发,不对数据包进行任何操作。网络编码[6]的具体思想是具有编码功能的节点对其多条输入链路上的数据进行编码,然后发送到输出链路上。网络编码的本质是利用节点的计算能力提高链路带宽的利用率。图1(b)阐述了网络编码的基本原理,图中 s 是信源,x,y 是信宿,各条链路的带宽均为1比特/单位时间,现要将2比特数据 a,b 同时从 s 传到 x,y。易知 s 与 x,y 之间均分别存在两条独立路径,若采用传统路由

到稿日期:2009-03-31 返修日期:2009-06-11 本文受国家自然科学基金(U0635003),863基金(2007AA01Z215),桂林电子科技大学科学研究基金(UF09006Y)资助。

林国庆(1978—),男,助理工程师,研究方向为网络编码,E-mail:linguoqinghao@163.com;陈汝伟(1973—),男,副教授,研究方向为信道编码;李 颖(1973—),女,教授,研究方向为信道编码;王新梅(1937—),男,教授,研究方向为信道编码。

方法，如图 1(a)所示，由于两组路径间存在共有链路 wz，a，b 不能同时在边 wz 上传输，则 s 到 x，y 的最大信息流速率为 1.5 比特/单位时间。若采用网络编码方法，在节点 w 上对 a，b 执行异或操作并转发，则节点 x 可以通过 $a \oplus b \oplus a$ 的计算解出 b，同理 y 也可以解出 a，从而使 s 到 x，y 的信息流速率达到 2 比特/单位时间，带宽利用率提高 33%。

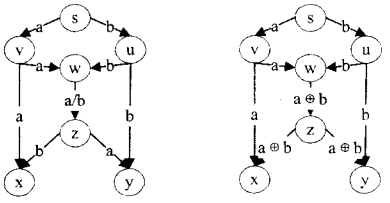

图 1　用网络编码提高网络性能示意图

随机线性网络编码[7]是一种实用的分布式网络编码方式，不需要知道整个网络的拓扑结构，就可以进行编码。在随机线性编码中，编码后的数据包是输入数据包的线性组合，其系数随机地取自某个有限域。编码符号域足够大时，可保证接收节点以接近 1 的概率恢复出信源信息[8]。现已证明，随机线性编码可以达到网络编码理论上的最优[9]，并且随机线性编码容易实现，计算开销也较小。

下面介绍随机线性网络编码，如图 2 所示。服务器 Server 中存放多个等长的数据块 $B_1 \cdots B_n$，客户端为 A 和 B，客户端可以从服务器或者邻近节点那里得到数据块。

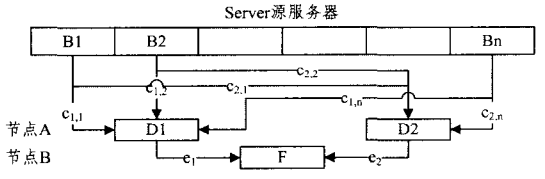

图 2　随机线性网络编码过程

数据块在节点间的编码及传输过程如图 2 所示。首先对源服务器 S 而言，如果节点 A 向其请求传输一个数据块，源服务器 S 随机生成向量 $C_1=(C_{1,1},C_{1,2},\cdots,C_{1,n})$ 作为系数，由 $\sum_{i=1}^{n}(C_{1,i}\times B_i)$ 得到一个编码后的块 D_1，将系数 C_1 和编码块 D_1 作为回复发给 A。同理，如果节点 A 再向 S 请求传输一个数据块，则 s 按类似的操作生成一新的数据块 $D_2=\sum_{i=1}^{n}(C_{2,i}\times B_i)$ 及与 D_2 相对应的系数向量 $C_2=(C_{2,1},C_{2,2},\cdots,C_{2,n})$，并把它们传递给节点 A。对于中间节点，比如 A，如果节点 B 向其请求传输一个数据块，且 A 已收到 m 个编码块 $D_1,\cdots,D_m$ 和对应的系数向量 $C_1,\cdots,C_m$，可类似地随机生成向量 $E=(e_1,e_2,\cdots,e_m)$，并由 $F=\sum_{i=1}^{m}(e_i\times D_i)$ 得到编码块 F，

$$F=(e_1,e_2,\cdots,e_m)\times\begin{bmatrix}D_1\\D_2\\\cdots\\D_m\end{bmatrix}$$

$$=(e_1,e_2,\cdots,e_m)\times\begin{bmatrix}C_{1,1}&C_{1,2}&\cdots&C_{1,n}\\C_{2,1}&C_{2,2}&\cdots&C_{2,n}\\\cdots&\cdots&\cdots&\cdots\\C_{m,1}&C_{m,2}&\cdots&C_{m,n}\end{bmatrix}\times\begin{bmatrix}B_1\\B_2\\\cdots\\B_n\end{bmatrix}$$

$$=\left(\sum_{i=1}^{m}(e_i\times C_{i,1}),\sum_{i=1}^{m}(e_i\times C_{i,2}),\cdots,\sum_{i=1}^{m}(e_i\times C_{i,n})\right)\times\begin{bmatrix}B_1\\B_2\\\cdots\\B_n\end{bmatrix}=G\times\begin{bmatrix}B_1\\B_2\\\cdots\\B_n\end{bmatrix}$$

其中，$G=\left(\sum_{i=1}^{m}(e_i\times C_{i,1}),\sum_{i=1}^{m}(e_i\times C_{i,2}),\cdots,\sum_{i=1}^{m}(e_i\times C_{i,n})\right)$ 为对应原始数据块 B_i 的系数向量。A 将新的编码块 F 和系数向量 G 发送给节点 B。同理，B 还会向 A 以及其他对等节点请求传输数据块，等到它获得足够多的数据块使得系数向量矩阵满秩，即其中 n 个数据块线性无关，B 通过高斯消元法就能恢复出原始数据块 $B_1 \cdots B_n$。

3　基于网络编码的文件备份方案

网络编码的实质就是将多个数据块进行合成，使编码后的数据块带有更多的信息。本文旨在将网络编码技术用于文件备份中，在合理的可恢复性前提下达到节约存储空间的目的。下面介绍基于网络编码的数据备份方案，并分析其相对于传统备份方案的性能优势。

如图 3 所示，主机 A，B，C 和 D 分别有大小为 1 的文件 a，b，c，d，S 为备份服务器。

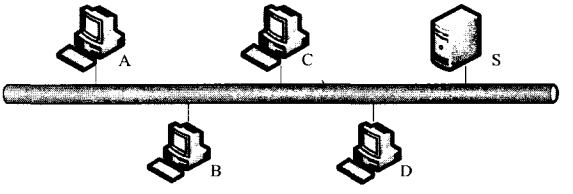

图 3　局域网示意图

传统备份方案：

$A:a,B:b,C:c,D:d,S:a,b,c,d$。占用总存储空间为 8，备份时的总数据传输量为 4。只要两个相同文件不同时丢失，就可得到恢复。

基于网络编码的备份方案一：

$A:a,B:b,C:c,D:d,S:e=k_1a+k_2b+k_3c+k_4d$，系数 k_i $(i=1,\cdots,4)$ 从有限域 $GF(q)$ 中随机选取且不为 0。占用总存储空间为 5，备份时的总数据传输量为 4。任何一个文件丢失都可以得到恢复。如文件 a 丢失，可通过将文件 b,c,d 重新传输到 S 上，并对 b,c,d,e 应用高斯消元法就可恢复出 a，然后将 a 传给主机 A；但如果两个文件同时丢失，则导致系数向量矩阵不满秩，就不能完成恢复。

基于网络编码的备份方案二：

$A:a,B:b,C:c,D:d,S:e1=k_1a+k_2b+k_3c+k_4d,e_2=l_1a+l_2b+l_3c+l_4d$。系数 $k_i,l_i(i=1,\cdots,4)$ 从有限域 $GF(q)$ 中随机选取且不为 0。占用总存储空间为 6，备份时的总数据传输量为 4。任何两个文件同时丢失都可恢复。超过 3 个文件同时丢失，则不能完成恢复。

对 3 种方案进行比较发现，采用网络编码原理的方案在总存储空间方面大幅减少，分别减少了 3/8 和 2/8 的存储空间。即在备份服务器 S 上对数据进行了网络编码操作，用一个文件的存储空间存放了四个文件的信息量。备份时的数据传输量相同。

在可恢复性方面，基于网络编码的方案较传统方案略有

所降低。传统方案中文件是独立备份，各个文件不相关联，即使所有文件丢失也可以从 S 中恢复，可恢复性好。在基于网络编码的方案一和方案二中，由于文件是经编码操作后存储于备份服务器中的，其在恢复过程中互相依赖，一旦丢失文件数超过要求，便会造成系数向量矩阵不满秩，便恢复不出原始数据。可通过增加经过编码的副本来提高在可恢复性方面的健壮性，如基于网络编码的方案二比方案一增加了一个备份副本，则对其他原文件的依赖性降低，可恢复性增强。

同时，基于网络编码的方案在数据恢复时的传输量较传统方案有所增加，传统方案只要将副本从服务器 S 传回到对应主机即可，而基于网络编码的方案需要从其他主机上重新获取相应文件并进行解码，恢复出所需文件后再传回到对应主机。尽管可恢复性有所降低，数据恢复时的数据传输量增加，利用网络编码原理的备份方案仍不失为特殊情况下节省存储空间的合理方法。原因有以下几点：

1) 硬盘数据恢复技术一定程度上可以做到自身数据恢复。少量不能恢复的数据才有必要调用备份恢复。即使多个参加共同编码备份的文件同时被破坏，用硬盘恢复技术可恢复 80%的文件[10]，可以极高的概率恢复出足以解码备份文件的文件数，用以恢复少量不能用硬盘恢复技术恢复的文件。

2)具有安全保护体系，局域网中的所有主机的安全性是有保证的，外来的破坏在一定程度上被遏制。尤其在存在多个编码备份(如方案二)的前提下，参加编码的多个文件被同时破坏以至于难以完成备份解码的可能性并不大。

3)参加备份的多台主机由于自身原因，如硬盘毁坏，而同时丢失数据的可能性也非常小。若一台主机出现此情况，可以及时解码备份数据恢复。只要做到及时恢复，就可以保证不出现无法解码恢复的情况。另外，通过编码后的备份数据，安全保密性增强了，因为编码后的备份文件在偷窃者看来只是一个乱码文件，只获得一个编码副本，无法复原出原始文件。

4 基于网络编码原理的文件备份方案具体实现中的几个关键问题

本小节介绍基于网络编码的文件备份方案在实现过程中的几个关键问题。

4.1 备份的过程

各个需要备份的文件到达服务器的时间是不一致的，有先有后，可采用先来先编码的方式，比如若 a，b 先到，则先对其进行编码，生成编码文件 r；而后 c 到达，c 和 r 编码生成 t；最后 d 到达，d 与 t 编码最终生成备份文件 e。在现实中很少有文件大小是完全相同的，在编码时可用添位的方式补齐。还可以对编码文件添加校验位以防止编码存储过程中的差错。

4.2 文件更新对备份恢复的影响和处理

基于网络编码的文件备份方案一中，在备份后，经过一段时间更新后，各文件发生了变化，文件 a,b,c,d 分别变为 a_1，b_1,c_1,d_1，此时若主机 A 崩溃，文件 a_1 丢失，可采用如下方法恢复文件 a：

(1) 对每个文件建立日志，记录对该文件的每次操作并保证对其可以进行逆操作；

(2) 各主机在每次备份时设置一个恢复点，必要时可还原出一个备份时的副本。若 a_1 丢失，可将 b_1,c_1,d_1 结合日志进行逆操作，还原出文件 b,c,d，结合 S 上的文件 e 便可还原出 a。

4.3 备份的更新

上面提到文件经过一段时间后会有更新，此时根据需要重新进行备份。可设定一个共同的备份周期，如每日一次或每周一次。更新后的 a_1,b_1,c_1,d_1 被重新上传到服务器上，经过编码生成新的备份 e_1，此时可删除原备份 e。某文件若额外单独进行了一次备份，也可在下次周期备份后删除。

4.4 全局备份控制系统

全局备份控制系统运行在备份服务器上，对整个备份和恢复过程起到管理和协调的作用，其工作主要包括需要备份文件的传输和保存管理、备份文件的生成、备份文件更新管理、恢复时各文件的再次上传与丢失文件的具体恢复。

5 验证实验

实验方案：对本文第 3 节所描述的 3 种方案进行实验验证。备份过程中均不采用压缩，文件大小取为 10M。

(1)对比两者占用的存储空间的大小

各文件大小及 3 个方案中服务器的存储空间使用量如表 1 所列。

表 1 验证实验数据(kB)

a	b	c	d
10240	10240	10240	10240

e	e_1	e_2	传统方案	新方案一	新方案二
10241	10241	10241	40960	10241	20482

(2)对 3 个方案进行数据恢复验证

取需恢复的文件为 a。实验中 3 个方案均顺利将 a 从服务器恢复到主机 A 上。3 个方案在恢复 a 时的数据传输量如表 2 所列。

表 2 恢复 a 时的数据传输量(kB)

传统方案	新方案一	新方案二
10240	40960	30720

验证实验所得数据与理论推测基本吻合，即新方案节省了备份空间，但增加了数据恢复时的数据传输量。

结束语 本文利用网络编码的原理，设计了一个旨在节省存储空间的文件备份方案。其主要实现原理是将多个需备份文件利用随机网络编码原理编码成一个编码备份文件后存储。该方案主要包括以下 4 个部分：备份的基本原理和过程、文件更新对备份恢复的影响和处理、备份的更新和全局备份控制系统。理论分析和实验证明了该方案较传统备份方案大幅节省了存储空间，同时也自动实现了对备份数据的加密。网络编码是有代价的，具体到本方案就是以数据可恢复性的微降和增加系统的计算量和复杂度，来换取存储空间的节省和备份数据安全保密性的自身实现。基于其自身特点，该备份方案可作为特殊情况下如磁盘空间紧张时节省存储空间的合理方法，也适用于一些长期不更新的存档数据的备份。下一步的工作将集中于备份系统的具体实现和进一步优化。同时，该方案是基于文件的，下一步将考虑如何将该方案的原理用于数据库系统中，以节约空间和提高其安全保密性。

参 考 文 献

[1] 陈飞翔,周治武,张建兵. 基于动态规划算法的矢量数据压缩改进算法[J]. 计算机应用,2008,28(2):168-170

[2] 赵锴,李建中,骆吉洲. 基于谓词索引的海量数据压缩存储及数据操作算法[J]. 计算机科学,2005,32(9):86-90

[3] 陈艳拢. 等值线数据压缩算法研究[J]. 大连理工大学学报,2005(10):18-19

[4] 关宗安,仲丛久. 数据加密标准算法的 DSP 实现[J]. 数据采集与处理,2006,21(12):266-268

[5] Schneier B. 应用密码学[M]. 吴世忠,祝世雄,译. 北京:机械工业出版社,2001

[6] Ahlswede R,Cai N,Li S R,et al. Network information flow[J]. IEEE Trans. On Information Theory,2000,46(4):1204-1216

[7] Ho T,Karger D R,Medard M,et al. The benefits of coding over routing in a randomized setting[A] // Hideki I,eds. The 2003 IEEE International Symposium on Information Theory[C]. San Jose:IEEE Press,2003:442-447

[8] Jaggi S,Sanders P,Chou P A,et al. Polynomial time algorithms for multicast network code construction[J]. IEEE Transactions on Information Theory,2003,49:831- 836

[9] Chou P A,Wu Y,Jain K. Practical network coding[A] // Proceedings of 41st Allerton Conference on Communication,Control and Computing[C]. Allerton:Monticello House,2003:473-482

[10] 吴景裕. 浅谈图书馆数字化建设中数据恢复[J]. 情报探索,2007(12):53-55

(上接第 98 页)

串。因此,虚拟卫星广播信道模型和 Maurer 的卫星广播信道模型在功能上是等价的。

2.3 模型比较

对虚拟卫星广播信道模型与卫星广播信道模型做如下方面的比较。

接收同步问题:虚拟卫星广播信道模型不存在接收同步问题。而卫星广播信道模型中通信双方为实现同步接收,须在初始化阶段之前至少通信一次,以协调同步接收的问题。

误比特率比较:一般来说,在卫星广播信道模型中,合法的通信双方对窃听方知之甚少。在密钥协商阶段,通信双方需要对窃听方接收信道的误比特率有一个保守的估计。而在虚拟卫星广播信道模型中,通信双方不仅可以计算出窃听方接收信道的误比特率 $p_E=\varepsilon+\delta-2\varepsilon\delta$,而且完全掌握了窃听方接收到的比特串 $Z=A^*\oplus B^*$。如果再注意到,

$$p_E=\delta+\varepsilon(1-2\delta)>\delta=p_A,\ p_E=\varepsilon+\delta(1-2\varepsilon)>\varepsilon=p_B$$

则可以断定,在虚拟卫星广播信道模型中,窃听方的接收信道的误比特率 p_E 要大于合法通信双方的接收信道的误比特率 p_A,p_B。这样,Alice 和 Bob 就赢得了优势。

通信成本:较之虚拟卫星广播信道模型,在卫星广播信道模型中,通信双方除了要为协调同步接收而多进行至少一次通信外,还需要新信号的接收设备。在虚拟卫星广播信道模型中,虽然也使用了虚拟二元对称信道,但实际上通信双方的虚拟二元对称信道完全可以通过软件模拟来实现。换言之,通信双方分别随机选取比特串 A 和 B,然后根据选定的误比特率 ε 和 δ 来随机地改变 A 和 B 中某些比特的值,所得到的比特串 A^* 和 B^* 就是 A 和 B 通过误比特率为 ε 和 δ 的虚拟二元对称信道的输出结果。因此,本文给出的虚拟卫星广播信道模型易于软件实现。

综上所述,本文的模型较之卫星广播信道模型具有无需同步、易于控制窃听方接收信道的误比特率、通信成本低、易于软件实现等优点。

结束语 针对信息理论安全卫星广播信道模型中存在的一些缺陷,给出了虚拟卫星广播信道模型。模型无需同步,且通信成本低,易于软件实现,可以人为地对窃听方的初始信息进行控制等。一方面,在卫星广播信道模型中的优先提取、信息协调和保密增强协议可以原封不动地移植到虚拟卫星广播信道模型中;另一方面,虚拟卫星广播信道模型比卫星广播信道模型具有诸多优点。因此,如何根据这些优点设计在虚拟卫星广播信道模型中更为高效的优先提取、信息协调和保密增强协议,则属于将来需要进一步研究的工作。

参 考 文 献

[1] Maurer U. Information-theoretic cryptography[C] // Advances in Cryptology-CRYPTO'99. LNCS 1666. Berlin:Springer-Verlag,1999:47-64

[2] Shannon C E. Communication theory of secrecy systems[J]. Bell System Technical Journal,1949,28:656-715

[3] Zhao Yi-bo,Gui You-zhen,Chen Jin-jian,et al. Computational complexity of continuous variable quantum key distribution[J]. IEEE Trans. on Information Theory,2008,54(6):2803-2807

[4] Maurer U. Secret key agreement by public discussion from common information[J]. IEEE Trans. on Information Theory,1993,39(3):733-742

[5] 王保仓. 基于有扰认证信道的信息理论安全密钥协商[D]. 西安:西安电子科技大学,2004

[6] Maurer U. Information - theoretically secure secret - key agreement by not authenticated public discussion[C] // Advances in Cryptology-EUROCRYPT'97. LNCS 1233. Berlin: Springer-Verlag,1997:209-225

[7] Jones N S,Masanes L. Key distillation and the secret-bit fraction [J]. IEEE Trans. on Information Theory,2008,54(2):680-691

[8] Maurer U. Protocols for secret key agreement based on common information[C] // Advances in Cryptology-CRYPTO'92. LNCS 740. Berlin:Springer-Verlag,1993:461-470

[9] Yan Hao,Ren Tienan,Peng Xiang,et al. Information reconciliation protocol in quantum key distribution system[C] // The Fourth International Conference on Natural Computation-ICNC'08. Jinan,China,Oct. 2008,3:637-641

[10] 王保仓,杨波,胡予濮. 一种新的信息协调协议[J]. 西安电子科技大学学报,2006,33(3):486-490

[11] Watanabe Y. Privacy amplification for quantum key distribution [J]. Journal of Physics A:Mathematical and Theoretical,2007,40(3):99-104

[12] Liu Shengli,Wang Yumin. Privacy amplification against active attacks with strong robustness[J]. Electron. Lett.,1999,35(9):712-713