第48卷 第6期 世界地震译丛 Vol. 48, No. 6
2017年11月(496－516) TRANSLATED WORLD SEISMOLOGY Nov., 2017

C. E. Yoon, O. O'Reilly, K. J. Bergen, G. C. Beroza. 2015. Earthquake detection through computationally efficient similarity search. *Sci. Adv.* **1**(11): e1501057
李万金 译. 2017. **通过计算高效的相似性搜索进行地震检测. 世界地震译丛. 48**(6): 496－516. doi: 10.16738/j.cnki.issn.1003－3238.201706003

通过计算高效的相似性搜索进行地震检测

C. E. Yoon　O. O'Reilly　K. J. Bergen　G. C. Beroza

摘要　地震学正经历着数据量的快速增长，它已超过处理算法发展的速度。地震检测——连续数据中地震事件的识别——是观测地震学的一项基本操作。使用波形相似性克服了现有检测方法的缺点，从而开发了一种有效的方法来检测地震。该方法称为指纹和相似性阈值法(FAST)，分析1个星期的连续地震波形数据用时不到2小时，或者比自相关快140倍。指纹和相似性阈值法采用数据挖掘算法，最初设计用于大数据库中识别相似的音频剪辑。它首先通过提取关键的判别特征来创建波形的紧凑“指纹”；然后把相似的指纹组合在一个数据库中，以利于相似指纹对快速、可扩展的搜索；最后生成地震检测的一个列表。指纹和相似性阈值法从位于美国加利福尼亚中部卡拉韦拉斯断层附近的一个台站1个星期的连续数据中检测到了大部分编目地震(24个中的21个)和68个非编目地震，实现了能与自相关相媲美的检测性能，尽管有一些额外的虚假检测。当应用于地震台站的分布式台网、非常长的持续时间数据集上时，指纹和相似性阈值法可望发挥其充分的潜力。指纹和相似性阈值法的广泛应用，可能有助于发现意想不到的地震信号，改善地震监测，促进对各种地震过程的更深的了解。

0　引言

地震学是一门数据驱动的科学，其突破性进展往往来自观测能力的进展(1)。现在有巨大的数据集：拥有高达数千个传感器的台网已经记录了几年的连续地震数据流，并且数据采集的速度持续加快。地震学可以从处理和分析这些海量数据的新的可扩展算法中获益，从它们中提取尽可能多的有用信息。我们工作的重点是使用最初设计用于音频识别、图像检索和网络搜索引擎开发的数据挖掘技术来改善地震检测。

0.1　背景

地震台网包括位置分散的多个台站(接收器)，其中每个台站都有一台连续记录地面运动的地震仪。传统上，地震被每个台站使用如短时平均/长时平均(STA/LTA)的能量检测器检测到一次。当这些窗口沿着连续数据滑动时，STA/LTA就计算短时窗口内的短时平均能量与长时窗口内的长时平均能量的比值。当STA/LTA比值超过某些阈值时，就宣布一次检测(2，3)；然后算法

云南省地震局个旧地震台　李万金　译
中国地震局地球物理研究所　朱玉萍　校

确定该台网的多个台站是否检测到一致的地震源。如果一个地震事件被至少 4 个台站检测到，它就会被编入地震目录中。地震目录是一个包含已知地震位置、发震时间和震级的数据库。

STA/LTA 能成功识别具有脉冲、高信噪比(SNR)P 波和 S 波到达的地震。STA/LTA 普遍适用性高(图 1)，我们将普遍适用性定义为不用地震波形或震源信息的先验知识来检测各种各样地震的能力。但在诸如低信噪比、具有突至震相的波形、重叠地震事件、背景噪声以及稀疏的台站间距这些更具挑战性的情况下，STA/LTA 不能检测到地震或者可能产生错误检测。因此，STA/LTA 具有低检测灵敏度(图 1)。所以，地震目录中低震级地震不完整。

我们可以将整个地震波形的信息用于检测，而不仅仅是脉动的体波震相，以克服 STA/LTA 的局限性。在几个星期、几个月甚至几年的时间内的重复地震震源，在同一台站记录到的波形高度相似(4，5)。路径效应几乎相同：对大地震前(6)后(5)地震波走时随时间变化的搜索表明，地球速度结构的时间变化非常微小，所以地球结构在地震时间尺度上基本是恒定的。波形互相关利用所产生的波形相似性来实现灵敏的地震检波器。

波形互相关，又称为匹配滤波或模板匹配，已被证明是在噪声数据中查找已知地震信号的一种灵敏的、判别式的方法。它具有高检测灵敏度(图 1)。它是“一对多”搜索

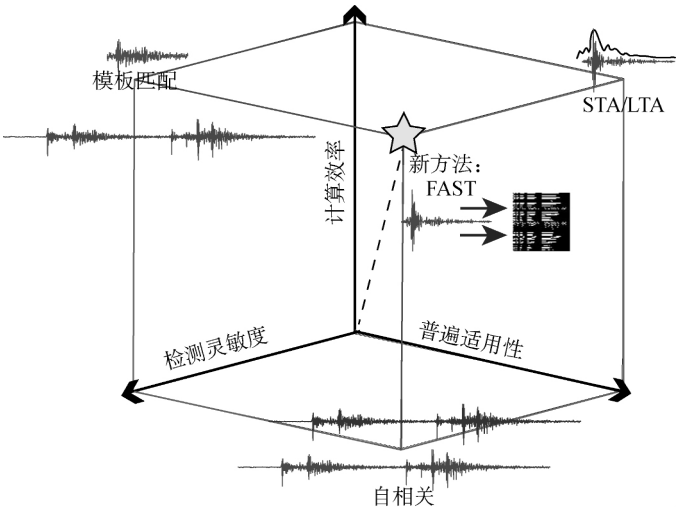

图 1　地震检测方法三个定性指标的比较：检测灵敏度、普遍适用性和计算效率(原图为彩色图——译注)。STA/LTA 的普遍适用性高，因为它能发现未知来源；计算效率高，因为它是实时检测地震；但检测灵敏度低，因为它会漏掉低信噪比的地震事件。模板匹配的检测灵敏度高，因为互相关可以找到低信噪比事件；计算效率高，因为我们只需要将连续数据和一小组模板波形作互相关；但普遍适用性低，因为模板波形需要事先确定。自相关的检测灵敏度高，因为它对波形作互相关；普遍适用性高，因为它可以找到未知的类似来源；但计算效率非常低，因为它与连续数据集的大小极不相符。指纹和相似性阈值法(FAST)对所有三个指标都表现良好，将检测灵敏度与基于相关性检测的普遍适用性相结合，具有较高的计算效率和可扩展性

方法，它计算具有连续波形数据的连续候选时间窗口的模板波形的归一化相关系数(CC)，且具有相关系数超过特定阈值的任一候选窗口就被认为是检测到地震(7)。两段时域波形 $\vec{a}$ 和 $\vec{b}$ 之间的归一化相关系数定义为：

$$\mathrm{CC}(\vec{a},\vec{b}) = \frac{\vec{a}^{T}\vec{a}}{\|\vec{a}\|_2 \|\vec{b}\|_2} = \frac{\sum_{i=1}^{M} a_i b_i}{\sqrt{\sum_{i=1}^{M} a_i a_i}\sqrt{\sum_{i=1}^{M} b_i b_i}} \quad (1)$$

式中，M 是每个波形的样点数。模板匹配能检测非常低信噪比的事件，很少有误报。当模板包含的波形来自多个台站和多个通道时，检测是基于综合的台网互相关系数(CC)(7，8)。模板匹配是一种通用的功能强大的技术，已在各种地震活动性研究中发现未检测到的地震：未编入目录的低震级地震(9)、前震(10)、余震(11)、触发地震(12)、震群(13)、构造颤动中的低频地震(LFE)(8)和触发颤动(14)、地震台网稀疏的潜在诱发地震地区的小震级事件(15)、核监测与鉴别(7，16)，以及地热(17)与油气藏(18)中的微震监测。

然而匹配的一个主要局限是它需要一个先验波形模板，因此，它的普遍适用性低(图 1)。模板通常通过提取目录地震的波形或通过人工检测从连续波形中提取脉冲地震波形来选定。这不是一种查找低信噪比重复信号未知来源的有效、全面的方法。已开发了子空间检测(19)和经验子空间检测(20)方法，以将模板匹配推广到波形变化更大的、相似的、非重复震源；然而，我们感兴趣的是最一般的情况——系统地对连续数据中具有相似波形的信号进行盲搜索，而无需信号先验知识。

当所需信号波形未知时，自相关是一种搜索相似波形的完备的“多对多”方法。我们知道，感兴趣的地震信号有短暂的持续时间(每个通道一般几秒钟)，所以我们将连续数据分成 N 个短的重叠窗口，并对所有可能的窗口对作互相关。当窗口对的互相关系数超过检测阀值时，就被标记为候选事件，它可以用额外的互相关进行后续处理，或被组合进“家族”，并被叠加来形成低噪声模板波形。自相关已在构造地震中成功地找到已知和以前未知的低频地震(21，22)。自相关提供了超过 STA/LTA 的改进的波形互相关灵敏度，也能检测具有相似波形的未知源(图 1)。

自相关有一个主要的缺点，因为它属于计算密集型(图 1)，所以对于在大量连续数据集中检测地震来说，最终是不可行的。对于 N 个窗口，我们必须计算 $N(N-1)/2$ 次互相关系数来考虑所有可能的窗口对；因此，自相关运算时间与数据长度呈 2 次方关系，算法复杂度为 $O(N^2)$。自相关执行了大量的冗余工作，因为大多数窗口对是不相关的，且对检测不感兴趣(图 S1A，见补充材料，下同——译注)；自相关检测到的高度相似地震只占窗口对总数的一小部分。自相关非常适合用于检测几个小时连续数据中发生的重复地震(21)，其中 N 值小。但自相关的 $O(N^2)$ 运行时间，对于要使用它来从具有数百个通道和地震台站的台网几天、几个星期、几个月、甚至几年的连续地震数据中，且不使用大规模计算资源来查找不经常重复的事件是不切实际的。我们已经开发出一种结合了自相关的优势(查找未知源的检测灵敏度和综合能力)和可扩展的运行时间的新方法用于大的 N 值(图 1)。我们的技术有可能改善地震监测并对地震过程给出新见解。

0.2 地震检测的新方法

人们已经开发了许多算法来高效地搜索

大数据集中相似的条目(23)；应用包括识别大文件系统中相似的文件(24)、查找近乎重复的网页(25)、检测文档抄袭(26)和作为音乐鉴定来识别相似的音频剪辑(27)，如 Shazam 移动应用程序(28)。我们可以通过利用计算机科学界广泛使用的可扩展算法，来对大量连续数据中的相似地震波形进行快速、高效、自动化的盲检测来达到我们的目标。地震学家刚开始利用数据密集型搜索技术来分析地震记录；最新的一个应用是用于快速震源机制识别的地震搜索引擎，它从大数据库中检索最佳吻合的合成地震记录(29)，而另一项研究开发了一种快速近似算法，从大量目录中查找相似的事件波形(30)。

局部敏感散列(LSH)，是一种广泛用于高维近似近邻搜索的方法，可以使我们避免比较构成数据中大多数波形的不相似对；局部敏感散列法是返回可能具有高概率相似性的“候选对”的一个较短列表(23，31)。在计算机科学中，散列法经常用于数据库中条目的高效插入、搜索和删除，有恒定的 $O(1)$ 运行时间；每个条目被插入到根据散列函数的输出选择的一个散列桶中(32)。散列表包含许多散列桶，而散列函数确定条目如何分配到不同的散列桶中(32)。使用局部敏感散列法(图 S1B)，我们只需要搜索同一散列桶中相似的条目对(地震信号)——这些对成为候选对，我们可以忽略没有出现在同一个散列桶中的条目对，它们包括大多数对。因此，局部敏感散列法允许用与连续数据中的窗口数成近似线性关系的运行时间来搜索相似条目，这比自相关的二次关系要好得多。

并非直接比较波形，我们首先进行特征提取，将每个波形浓缩成一个只保留其主要鉴别特征的紧凑的“指纹”。指纹用作波形的替代品；因此，两个相似的波形应该有相似的指纹，而两个不相似的波形应该有不同的指纹。我们将指纹(而不是波形)分配到局部敏感散列的散列桶中。

我们的方法是一种称为指纹和相似性阈值的算法，基于 Waveprint 音频指纹算法(33)，它结合了计算机视觉技术和大规模数据处理方法来匹配相似音频剪辑。我们根据从连续地震数据中检测相似地震的特殊应用的性质和要求来修改 Waveprint 算法。我们选择 Waveprint，是因为它的音频识别方面展示的性能以及该技术容易映射到我们应用中。首先，音频信号与地震记录在几个方面相似：它们都是连续时间序列波形数据，而且感兴趣的信号往往是非脉冲；其次，Waveprint 使用如自相关中的短叠加音频剪辑来计算指纹；再次，Waveprint 利用局部敏感散列法仅搜索指纹中的一小部分。Waveprint 还能报告高精度的快速检索结果，且其特征提取步骤很容易并行化。指纹和相似性阈值法用作地震检测方法在 3 个定性理想化指标(检测灵敏度、普遍适用性和计算效率)得分都高(图 1)，而其他地震检测算法(STA/LTA，模板匹配和自相关)在 3 个指标中仅有 2 个指标较好。

1　结果

1.1　数据集

我们在一个包含可能有相似波形的未编入目录的地震的连续数据集上测试了指纹和相似性阈值法的检测能力。众所周知，美国加利福尼亚州中部的卡拉韦拉斯断层(图 2)有重复地震(34)。我们检索了来自北加州地震台网(NCSN)的台站 CCOB. EHN(水平北—南分量)以速度测量的从 2011 年 1 月 8 日(00：00：00)至 2011 年 1 月 15 日(00：00：00)期间 1 个星期(168 小时)的连续波形数据。根据北加州地震台网目录，2011 年 1 月 8 日该断层上发生了一次 M_W4.1 地震，

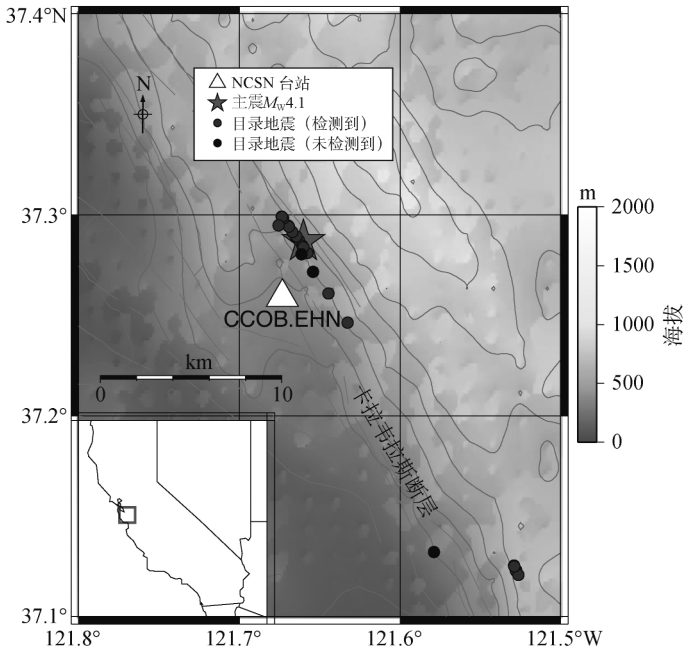

图 2 卡拉韦拉斯断层上目录地震和有数据的地震台站的位置图（原图为彩色图——译注）。2011 年 1 月 8 日 M_W4.1 地震的双差目录位置（红五角星）和 2011 年 1 月 8 日至 15 日期间卡拉韦拉斯断层上发生的北加州地震台网（NCSN）目录事件（点），以及我们处理了从 2011 年 1 月 8 日至 15 日 1 个星期数据的台站 CCOB. EHN（白色三角形）。蓝点表示指纹和相似性阈值法检测到的 21 个目录事件，而黑点表示指纹和相似性阈值法漏检的 3 个目录事件。（插图）地图在美国加利福尼亚州中的位置（红框）

之后有几次余震。这些编目事件中的大多数事件位于距离台站 3km 的范围内。

在运行指纹和相似性阈值算法前，我们对连续时间序列数据进行了预处理。我们对数据进行 4～10Hz 的带通滤波，因为低频段的相关噪声会干扰我们检测未编目地震的能力。这种相关噪声，似乎是此台站特有的，包括出现在数据中不同时间的相似的非地震信号。然后我们将滤波后的数据从原来每秒 100 个样本的采样率抽样成每秒 20 个样本，所以奈奎斯特频率为 10Hz。

1.2 指纹和相似性阈值法检测结果

我们证实指纹和相似性阈值法在 1 个星期的连续时间序列数据中能成功检测到未编目地震，我们将它的检测性能和运行时间与自相关进行了比较。表 1 包含了我们用于指纹和相似性阈值法的参数，表 S1 显示了自相关的参数；虽然这些参数没有调整到最佳值，但它们表现已相当好。一般情况下，我们不能期望来自指纹和相似性阈值法、自相关和目录的事件时间能准确匹配，它们都有自己的事件检测时间列表。因此，为方便比较，我们将匹配事件定义为发生在互相的 19s 范围内（表 1），这是具有 1s 滞后的 10s 长的指纹（表 1）和 10s 长的自相关窗口（表 S1）之间重叠的最大时间。

表 2 通过几个指标总结了自相关与指纹和相似性阈值法的性能：检测事件、错误检测、目录检测，新（未编目的）检测、漏检测

以及运行时间的数值。指纹和相似性阈值法从这些数据中总共检测到 89 个地震(图 3)，而自相关检测到 86 个事件，因此它们在检测事件总数上性能相当。指纹和相似性阈值法比自相关有更多的虚假检测，但运行得更快。大部分事件都被自相关(86 个中的 64 个)与指纹和相似性阈值法(89 个中的 64 个)检测到。但新事件中相当一部分或被自相关(22 个事件)或被指纹和相似性阈值法(25 个事件)检测到，但没有被两者同时检测到。

指纹和相似性阈值法检测到图 2 中位于感兴趣区域内(37.1°～37.4°N，121.8°～121.5°W)的 24 个目录事件中的 21 个(图 3)，而自相关发现了全部 24 个。只使用来自 CCOB.EHN 台站的数据，自相关与指纹和相似性阈值法都没有检测此区域外的目录地震。图 S2 显示了被指纹和相似性阈值法

表 1　指纹和相似性阈值法输入参数。这些参数被用于合成数据检测(除事件检测阈值)和 CCOB.EHN 台站的 1 个星期的数据

指纹和相似性阈值法的参数	值
用于频谱图生成的时间序列窗口长度	200 个采样点(10s)
用于频谱图生成的时间序列窗口滞后	2 个采样点(0.1s)
频谱图像窗口长度	100 个采样点(10s)
频谱图像窗口滞后＝指纹采样周期	10 个采样点(1s)
顶层 k 个幅度标准化哈尔系数的数目	800
局部敏感散列法：每个散列表中散列函数的数目 r	5
局部敏感散列法：散列表的数目 b	100
初始对阈值：表中数目 v(分数)，在同一桶中的配对	4(4/100＝0.04)
事件检测阀值：表中数目 v(分数)，在同一桶中的对	19(19/100＝0.19)
相似性搜索：近重复排除参数	5 个采样点(5s)
近重复对和事件消除时间窗	21s
自相关和目录比较时间窗	19s

表 2　自相关与指纹和相似性阈值法之间几个指标性能比较总结。指标 3～5 的数目总和应该等于指标 1 的数目

指标	自相关	指纹和相似性阈值法
1. 检测到的事件的总数	86	89
2. 虚假检测数量(误报)	0	12
3. 目录检测的数量和百分比	24/24＝100%	21/24＝87.5%
4. 这两种算法都有的新检测数	43	43
5. 一个检测到而另一个错过的新检测数	19	25
6. 错过检测的数目(漏报)	25	22
7. 运行时间	9 天 13 小时	1 小时 36 分钟

检测到的 21 个目录地震按目录时间排序的 20s 归一化波形(图 S2A)，震级范围从主震 M_W4.10 到最小事件 M_d0.84(表 S2)，指纹和相似性阈值法未检测到的 3 个目录事件属于漏检(图 S2B)。指纹和相似性阈值法没有检测到这 3 个目录事件，是因为它们在这 1 个星期的连续数据中没有重复(图 2)。位于 361 736s 处检测到的那个地震的位置是(37.132 08°N 和－121.578 79°W)，不同于其他目录事件。在 314 077s 和 336 727s 处的另外两个事件更靠近主震附近的大多数目录事件，但与大多数目录事件的 6～7km 深度(表 S2)相比它们的深度更浅(分别为 3.50km 和 3.53km)。自相关检测到这 3 个目录事件，因为它们的初始震相到时与另一个地震的初始震相到时的匹配具有高相关系数；然而对 5s 后地震对的检查显示它们波形的其余部分不相似(图 S3)，因此指纹和相似性阈值法没有检测到它们就不足为奇了。

除了这 21 个目录事件，指纹和相似性阈值法还检测到了目录中没有的 68 个新事件(图 3)。这些额外的事件提供了对卡拉韦拉斯断层上地震活动更完整的描述；此余震序列更高的时间分辨率可以用来更准确地预测传染型余震序列模型的余震速率。图 S4 显示了 CCOB.EHN 台站一个星期之内的数据按事件检测时间排序的这些新事件的 20s 归一化波形，指纹和相似性阈值法检测到自相关也检测到的 43 个新事件(图 S4A)，还检测到自相关漏检的 25 个新事件(图 S4B)。这些事件比图 S2 中的目录地震波形的噪声大。

图 S4 中的波形在时间上没有完全对齐有两个原因：第一，指纹和相似性阈值法事件时间准确性最高只有 1s，等于相邻指纹之间的时间滞后(表 1)；第二，同一事件有多个检测时间，而我们只考虑具有指纹和相似性阈值法最高相似度的那个时间(见补充材料)。指纹和相似性阈值法相似度被定义为在同一个桶中指纹对的散列表的分数(见材料与方法一节)。指纹和相似性阈值法不能估计精确的到达时间，但这可以由检测过程的后续步骤中的互相关很容易地计算出来。

假定以表 1 中的参数为选项，我们还估计了指纹和相似性阈值法所产生的误报和漏检的数量。此估计是基于对波形仔细的目视检查：虽然指纹和相似性阈值法检测时只使用 EHN 通道，但是 CCOB 台站数据的所有三个分向波形必须看起来像一个脉冲地震信号才被归类为“真实的检测”。在我们的应用中，我们只想检测地震，所以我们没有将具有非脉冲波形的相似信号归类为真实的检测。指纹和相似性阈值法返回超过地震检测阈值，但根据 20s 归一化波形视觉识别为低幅度噪声的 12 个误报检测(图 S5A)。自相关没有任何误报，因为我们故意设置了一个高检测阈值(CC＝0.818)；我们可以为自相关设置低检测阈值以检测更多事件，但这也将会引起误报，使指纹和相似性阈值法与自相关检测之间的自动比较变得复杂。指纹和相似性阈值法未能检测到但自相关检测到的 19 个未编目事件(图 S5B)，所以这些都是漏报。这 19 个检测中的 10 个与那 3 个目录事件(图 S3)的漏检原因是一样的：自相关匹配了初始 P 波到时，但整个波形不相似。指纹和相似性阈值法总共漏检了，但被自相关检测到的 22 个事件(包括 3 个目录事件)。但指纹和相似性阈值法检测到，而自相关没有检测到的 25 个新事件，可以被解释为自相关漏检；它们的相关系数介于 0.672～0.807 之间，所以它们低于 CC＝0.818 的阈值。这 25 个事件波形对的整体形状是相似的，但时间上没有精确对齐(图 S6)。

最后，我们比较了指纹和相似性阈值法和自相关的串行运行性能来检测 CCOB.EHN 台站一个星期数据中的事件。当在英特尔至强处理器 E5-2620(2.1GHz 中央处理单元)上处理时，自相关用了 9 天 13

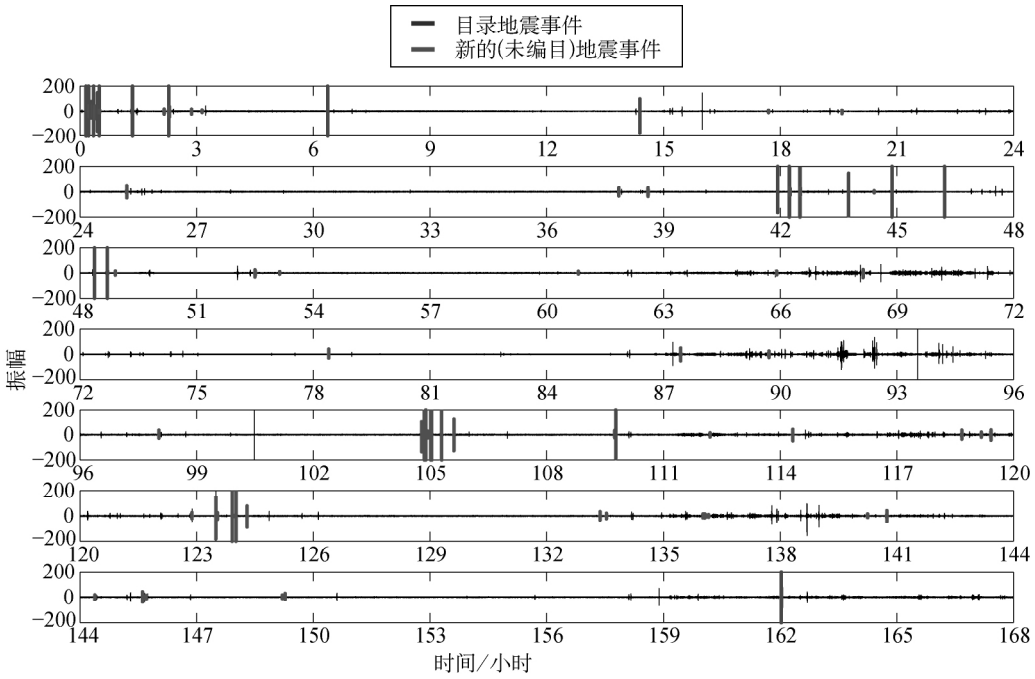

图 3　将指纹和相似性阈值法事件检测绘制在 1 个星期的连续数据上(原图为彩色图——译注)。数据来自台站 CCOB. EHN(带通滤波，4～10Hz)，起始于 2011 年 1 月 8 日(00：00：00)。指纹和相似性阈值法共检测到 89 次地震，其中包括 24 个目录事件中的 21 个(蓝色)和 68 个新事件(红色)

小时生成了地震检测的一个列表，而指纹和相似性阈值法仅用了 1 小时 36 分钟，加速了 143 倍。加速因子的估计有些不确定性，因为自相关与指纹和相似性阈值法的实现都没优化为最快的运行时间。指纹和相似性阈值法将 38％的时间用于特征提取，11％的时间用于生成数据库，以及 51％的时间用于相似性搜索。指纹和相似性阈值法在运行时间上比自相关有巨大优势，而且根据这两种算法的可扩展性，我们预计对于更长的连续数据集，这一优势将会增加。

图 S7 解释了来自指纹和相似性阈值法的候选对输出的小数目，这有助于其计算效率。它显示了在对数刻度上按指纹和相似性阈值法相似度分级的相似指纹对的直方图(包括近似重复对)。有 $N_{fp}(N_{fp}-1)/2\sim 1.8\times10^{11}$ 个可能的指纹对，但指纹和相似性阈值法输出初始阈值至少为 0.04 相似度的有 978 904 对(表 1)，这仅占总对数的 0.000 5％。在应用 0.19 的事件检测相似性阈值(表 1)后，我们仅保留 918 对。进一步后续处理(见补充材料)返回了一个 101 个检测的列表，其中包括 89 个真实事件和 12 个误检：去除近似重复对会将指纹对数减少到 105 对，而去除近似重复事件将检测数从 2×105＝210 个减少到 101 个。虽然指纹和相似性阈值法计算指纹和特征提取时要花费一些运行时间，这与从避免不必要的比较获得的加速相比是很小的。

2　讨论

2.1　扩展到大数据集

为了量化对于大数据集指纹和相似性阈值法运行时间和内存使用的可扩展性，我们

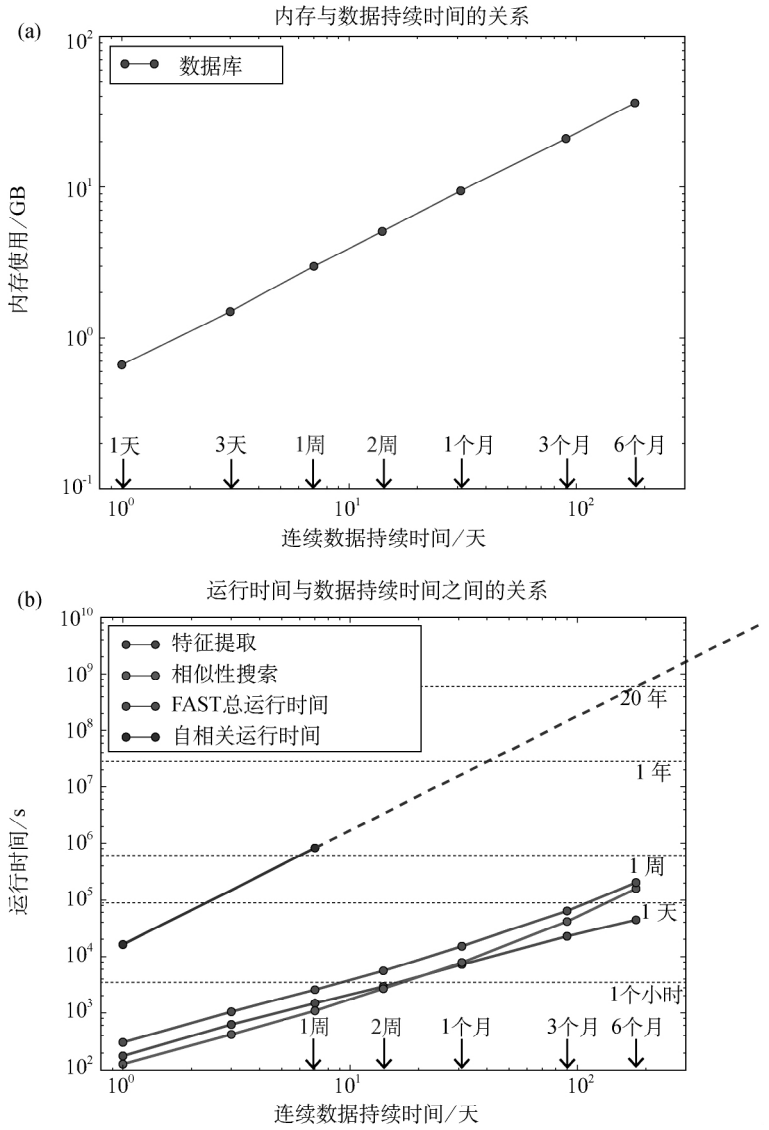

图 4 **以持续长达 6 个月的连续数据的函数作为指纹和相似性阈值法的比例属性(原图为彩色图——译注)。(a)由局部敏感散列法产生的数据库的内存使用。(b)指纹和相似性阈值法的总运行时间(红色),细分为特征提取的(蓝色)和相似性搜索的运行时间(绿色)。长度超过 1 周的连续数据的自相关运行时间(紫色)是根据二次比例关系外推得来的(虚线)。这些结果是用表 1 中的参数运行指纹和相似性阈值法得来的,同时将散列函数的数目 r 从 5 增加到 7,这样做可将 1 周的连续数据的总运行时间减少到一个小时之内**

从台站 CCOB. EHN 下载了 6 个月(181 天)的连续数据(2011 年 1 月 1 日至 2011 年 6 月 30 日),并在范围从 1 天至 6 个月(包括 1 个星期)的七个不同数据持续时间(表 S3)上运行指纹和相似性阈值法。对于此扩展测试,我们使用了表 1 中的参数,但我们将散

列函数的数量 r 从 5 个增加到 7 个。此参数的变化降低了检测性能，但提高了计算效率。

局部敏感散列法生成的数据库的内存使用取决于散列表的数量、指纹的数量和针对散列表实现的额外开销。我们使用 Linux 的 Top 命令来估计长持续时间数据的内存使用情况。我们发现 6 个月的连续数据大约需要 36GB 内存(图 4a)。

我们分别通过测量特征提取和相似性搜索步骤的时钟时间来以连续数据持续时间函数的形式调查了指纹和相似性阈值法的运行时间(图 4b)。特征提取与数据持续时间成线性比例，而相似性搜索与 $O(N^{1.36})$ 呈近线性比例。为了比较，我们记录了长达 1 个星期的数据的自相关运行时间，然后通过假设二次方比例关系来外推到更长的时间。指纹和相似性阈值法可以仅用 2 天 8.5 小时检测 6 个月连续数据内的相似地震——比我们的自相关实现至少快三个数量级，预计自相关需要约 20 年才能完成同样的任务。

2.2　局限性

指纹和相似性阈值法以更高内存要求来换取更快的运行时间和降低算法的复杂度。与自相关不同，指纹和相似性阈值法需要大量的内存，因为局部敏感散列法生成的数据库要存储散列表，每个散列表包含对分布在其散列桶集合中所有 $N_{fp}=604\ 781$ 个指纹的引用。因为我们预期搜索几个月到几年连续数据中的事件，因此这些内存需求会增加。对于几年的连续数据，内存可能成为瓶颈，而且在分布式计算环境中数据库的并行实现将是必要的。

我们可以通过几种方法来改善指纹和相似性阈值法的检测灵敏度和阈值算法。我们目前的实现中使用了值为 0.19 的事件检测阈值(表 1)作为指纹和相似性阈值法的相似性指标，这是通过视觉检查波形后设定的：高于此阈值的大多数事件看起来像地震，而低于此阈值的大多数事件看起来像噪声。由于我们处理持续时间更长的连续数据，我们将需要一个在特定时间段内能随噪声水平变化的自动和自适应的检测阈值。我们不想因为一个短时不常见的噪声时段而通过使用一个提高的恒定检测阈值来降低连续数据整年的检测灵敏度。另外，指纹和相似性阈值法输出的相似指纹对(图 S8)是真正的候选对(23)，它们需要经过额外的后处理来归类为事件检测。例如，没有使用 0.19 的指纹和相似性阈值法的相似性事件检测阈值(表 1)，我们可以取超过 0.04 初始阈值的所有对(表 1)，并基于侯选对的波形直接计算的相关系数来设置事件检测阈值。

从指纹和相似性阈值法输出的相似指纹对(图 8)仅识别波形之间的两两相似性；然而，我们希望找到三个或更多个彼此相似的波形组。丛集相似性具有有效的地震学应用，从识别重复地震序列到查找构造颤动中的低频地震，它在一次处理期间可包含成千上万个相似事件(8，21，22)。将来可以开发后处理步骤来确定相似波形对之间的“链接”以创建相似波形组。进一步研究识别多次重复的指纹对之间的联系(22)或应用经常用于分析社交网络(23)的聚类和图形算法(30)的组合，可以帮助解决这个问题。

指纹和相似性阈值法设计用于在连续数据中查找相似信号，但这些信号不一定是地震。指纹和相似性阈值法也可用于相关噪声的检测，尤其是如果数据具有重复的噪声信号，如图 S5A 中的 12 个虚假检测。由于该台站特有的低频相关噪声降低了我们的检测性能，因此我们对 CCOB. EHN 台站数据使用一个 4～10Hz 的带通滤波器。相关噪声的其他例子包括短时、大振幅的假信号、尖峰以及其他扰动。相关噪声也对自相关事件的检测产生了负面影响：当我们不对连续数据应用 4～10Hz 带通滤波器时，自相关也检测到许多比地震波形还具有高相似性的非

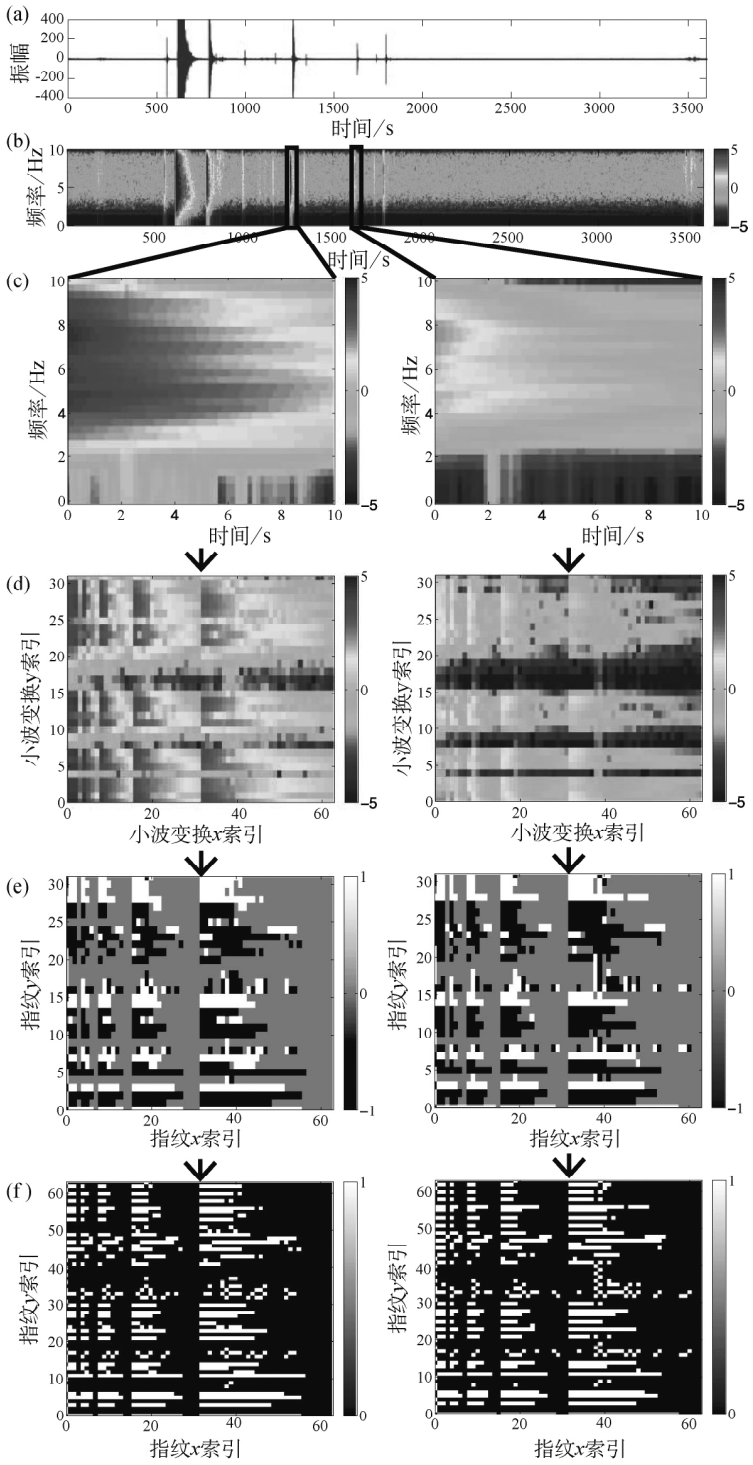

图 5 指纹和相似性阈值法中的特征提取步骤(原图为彩色图——译注)。(a)连续时间序列数据。(b)频谱图：振幅为对数刻度。(c)1 267s 和 1 629s 处两个相似地震的频谱图像。(d)哈尔小波系数：振幅为对数尺度。(e)经过数据压缩后最高标准哈尔小波系数的符号。(f)二进制指纹：特征提取的输出。注意：相似的频谱图像会产生相似的指纹

(a)

指纹 A	指纹 B
155	155
64	64
231	207
35	35
110	110
21	21

(b)

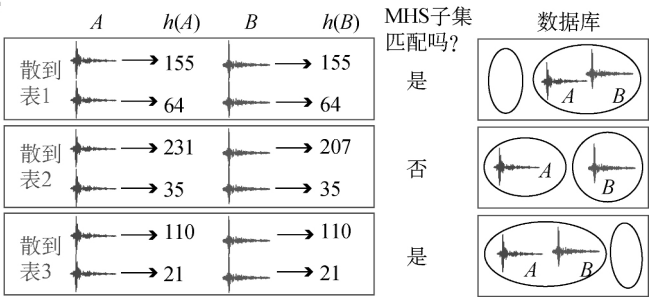

图 6　局部敏感散列法如何将指纹分组进数据库的示例(原图为彩色图——译注)。(a)两个相似指纹 A 和 B 的最小散列签名(MHS)的例子，$p=6$。(b)局部敏感散列法决定如何将两个相似指纹 A(蓝色)和 B(绿色)放入每个散列表(红框)的散列桶(椭圆形)中；为方便可视化显示了波形。最小散列签名长度为 $p=6$，且有 $b=3$ 个散列表，所以每个散列表得到每个指纹最小散列签名的一个不同的子集，它是 $6/3=2$ 个整数长；$r=2$ 最小散列函数的输出。分别处理每个散列表：如果 A 和 B 的最小散列签名子集都相等，那么 A 和 B 进入到数据库的同一散列桶中；在散列表 1 和 3 中就是如此，其中分别为 $h(A)=h(B)=[155, 64]$ 以及 $h(A)=h(B)=[110, 21]$。然而在散列表 2 中，A 和 B 的最小散列签名子集不相等，因为 $h(A)=[231, 35]$ 而 $h(B)=[207, 35]$，所以 A 和 B 进入不同的散列桶中

地震信号。减轻相关噪声影响的可能策略包括：应用一个高于噪声段的自适应检测阈值用分离相似地震信号和相似噪声信号的方式来分组检测，以及开发如从噪声中区分地震的特征分类器(35)这样的后处理算法。

由于指纹和相似性阈值法被设计来检测相似信号，我们不要指望它查找到与指纹和相似性阈值法处理过的连续数据中其他信号不相似的独特地震信号。例如，如果数据包含 100 个事件信号，但其中只有两个有相似的波形，指纹和相似性阈值法将只返回两个检测。持续时间越长的数据越有可能包含相似的地震信号，因此指纹和相似性阈值法将可能检测到更多地震事件。如果数据还包含独特的、不重复的地震信号，只要它有一个具有足够能量的脉冲到达，就可以用 STA/LTA 来检测它。此外，指纹和相似性阈值法可应用于本研究不探究的一个另外的“模板匹配模式”，其中连续数据部分的指纹是从其他数据集抽取的模板信号的指纹数据库中查询的，这使得类似于已知波形的检测而无需去匹配连续数据滞后期间出现的信号。

2.3 结论和未来影响

地震学是一门数据驱动的科学，其认识的新进展往往来自观测(1)，且地震台网采集的地震数据量从未像今天这么大。计算机科学家们开创了用于相似性搜索的数据挖掘算法，应用范围从音频剪辑到大型数据库中的图像，再到互联网网页。指纹和相似性阈值法证实我们可以利用这些算法来解决地震学中的一个基本问题：识别未知地震。

指纹和相似性阈值法超过其他竞争性方法的最重要的优势在于它的快速运行和可扩展性。对于 1 个星期的连续数据，指纹和相似性阈值法运行比自相关快约 140 倍，同时检测到相同总数的事件。然而，对于更长的连续数据流，我们预测串行指纹和相似性阈值法的运行会比自相关快几个数量级，纯粹基于这些算法的运行时间复杂度：自相关是二次方而指纹和相似性阈值法是近似线性(图 4b)。

地震学家以前应用并行处理来加速图形处理单元上(36)和分布式集群上(37)的模板匹配。我们也使用并行自相关实现(见补充材料)作为参考来与指纹和相似性阈值法检测结果进行比较。用并行实现可使指纹和相似性阈值法运行时间进一步减少，虽然只有特征提取这步难以实现并行；要跨多个节点分配局部敏感散列生成的指纹数据库需要重新设计一个非常重要的算法。

为了能够检测低信噪比环境中的地震，需要将指纹和相似性阈值法应用到分布的地震台网。现有的指纹和相似性阈值算法从单台站(CCOB)的单通道连续数据中检测地震事件；我们正在研究指纹和相似性阈值法的扩展应用，以便能够使用一个台站的所有三个分量来检测地震而且可以组合多个台站。许多模板匹配研究(7，8，11)已经表明，多个台站的组合通道提高了检测灵敏度，揭示出埋在噪声中的低信噪比信号。此外，多个台站以不同距离和方位角记录到震源的相干信号，更可能是地震而不是台站的局部噪声。因此，我们期望用一种多台站检测方法来减少虚假检测的数量，并减轻相关噪声对指纹和相似性阈值法检测性能的负面影响，假设不同台站之间的相关噪声在时间上是独立的。任何检测方法需要对网络结构的变化具有稳定性，比如长持续数据中新台站增加或台站减少。

指纹和相似性阈值法的检测能力有待通过对各种数据集的测试进一步探索，这些数据集由于低信噪比、具有非脉冲到时波形、重叠波形和相关噪声而带来检测挑战。今后的工作也应该开发更多区别性的指纹，并探索不同的方法来将指纹散列到数据库中。

由于指纹和相似性阈值法可以在近似恒定的时间内识别与给定查询事件相似的地震事件，该技术也可以适用于实时地震监测。如果在地震台网上大规模实施，提高的检测灵敏度可以降低目录完整性震级。实时指纹和相似性阈值法实现可以存储来自连续数据库的指纹数据库；并随着新数据的流入，新指纹将被创建和添加到数据库中，检查它们与其他指纹的相似性，并归类为检测到或没有检测到。指纹和相似性阈值法也可以进行大规模的模板匹配：成千上万的模板指纹可以被用作对庞大指纹数据库的搜索查询。指纹和相似性阈值法会在不同地震序列范围内发现 STA/LTA 或模板漏检的相似地震：前震、余震、触发地震、震群、低频地震、火山活动和诱发地震活动。指纹和相似性阈值法也可以识别不经常重复的可能几个月一次的低震级地震信号。

3 材料与方法

指纹和相似性阈值算法在连续地震时间序列数据的单通道内检测到相似的信号。它有两大部分：(Ⅰ)特征提取，(Ⅱ)相似性

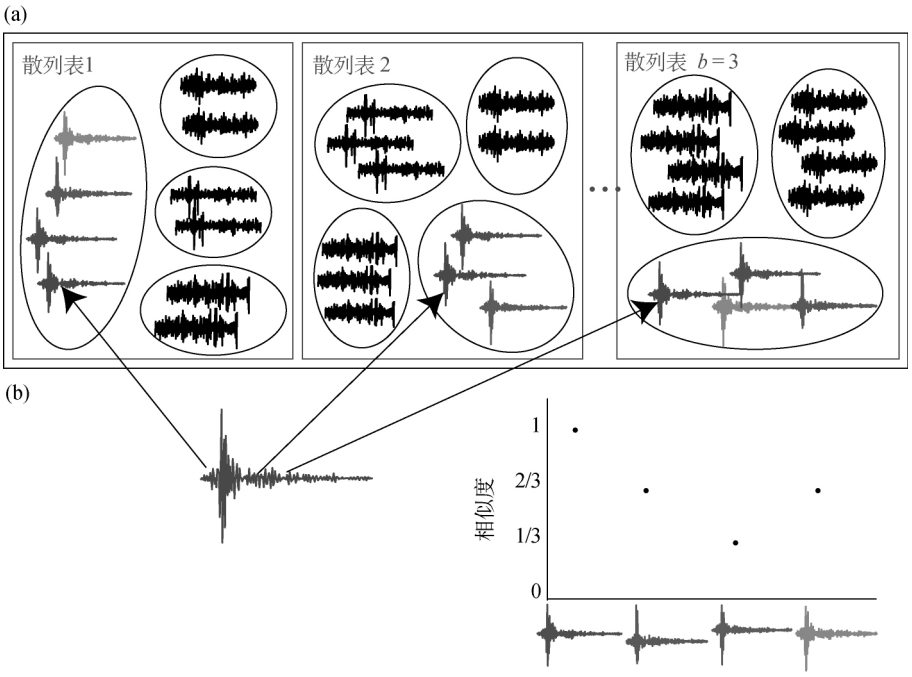

图 7　局部敏感散列数据库和相似性搜索的示例(原图为彩色图——译注)。(a)使用局部敏感散列生成的数据库，有 $b=3$ 个散列表(红框)；每个散列表有许多散列桶(椭圆形)。局部敏感散列以高概率将相似的指纹分组进相同的散列桶中；地震信号(彩色)有可能进入同一个桶中，而噪声(黑色)被分组进不同的桶中。(b)在数据库中搜索类似于查询波形(蓝色)的波形。首先，局部敏感散列确定每个散列表中的哪个桶中有波形与查询匹配。然后，我们将每个散列表中的所有其他波形放入同一个桶中并计算每个(查询，数据库)波形对之间的指纹和相似性阈值法的相似性：散列表中在同一个桶中包含对的分数。红色波形与所有 3 个散列表中的蓝色查询波形在同一个桶中，所以它们的相似性为 1，绿色波形放在 3 个散列表中的 2 个表的同一个桶中等等。这个图形为方便可视化显示了波形，但数据库将对指纹的引用存储在散列桶中，而且搜索查询需要将波形转换成它的指纹

搜索。特征提取通过将每个波形转换成一个稀疏二进制指纹来压缩时间序列数据。使用局部敏感散列函数将所有指纹插入一个数据库。给定所期望的“搜索查询”指纹，该数据库以几乎恒定的时间以高概率返回最相似的匹配指纹(23)。在我们目前的多对多搜索应用中，我们使用数据库中的每个指纹作为一次搜索查询，以便我们能找到此数据集内的所有相似指纹对。然而，我们也可以选择指纹的一个子集或使用其他数据源作为搜索查询。最后，删除从搜索查询返回的最不相似对，并且附加的后处理和阈值(见补充材料)产生了地震检测次数列表。

3.1　指纹：特征提取

图 5a～f 包含了指纹和相似性阈值法中特征提取步骤的概述，它遵循 Baluja 等(33)中的大部分工作流程：连续时间序列数据(a)、频谱图(b)、频谱图像(c)、哈尔小波变换(d)、数据压缩(e)和二进制指纹(f)。

频谱图 我们利用短时傅里叶变换(STFT)来计算时间序列数据(图 5a)的频谱图(图 5b)。我们在时间序列中取重叠的 10s 窗口(由 0.1s 的时间滞后隔开)(表 1),然后对每个窗口应用汉明锥形函数,并计算出每个锥形窗口的傅里叶变换。我们计算生成的复数短时傅里叶变换的功率(振幅的平方),然后对频谱图降低采样成 32 个频点,这样做可以消除一些噪声。地震在频谱图中以短暂的高能量事件出现(图 5b)。

频谱图像 我们要比较和检测短持续时间信号形式的相似地震,所以我们在时间维度里将频谱图划分成多个重叠窗口,并将每个窗口称为一个“频谱图像”。频谱图像之间的匹配模式以前已经作为一种地震检测方法被提出来(38)。与频谱图的其余相比(图 5b),地震信号的频谱图像具有较高的能量(图 5c)。频谱图像也有窗口长度和滞后参数;我们选择 $L_{fp}=100$ 个样点作为频谱图像长度和 $\tau_{fp}=10$ 个样点作为相邻频谱图像之间的滞后(表 1),这对应于 10s 的频谱图像长度和 1s 的频谱图像滞后,更短的频谱图像滞后以更多的运行时间为代价来增加检测灵敏度和定时精度。频谱图像窗口的总数,以及最终指纹的数目 N_{fp} 为:

$$N_{fp}=\left\lfloor\frac{N_t-(L_{fp}-\tau_{fp})}{\tau_{fp}}\right\rfloor \qquad (2)$$

式中 N_t 是频谱图像中时间样点的数量。对于来自 CCOB 台站的 1 个星期的连续数据来说,$N_{fp}=604\ 781$。

由于频谱图像内容随时间缓慢变化(33),与时间序列自相关中使用的 0.1s 滞后相比,我们可以找到具有更长的频谱图像滞后为 1s 的相似地震信号,这有助于指纹和相似性阈值法的快速运行。我们从相同持续时间的连续数据中获得很少的频谱图像(与自相关时间窗的数量相比),因此我们有更少的指纹来首先计算,随后比较相似性。虽然频谱图像长度是 10s,但它包含了 20s 的波形数据。频谱图像中每个 $L_{fp}=100$ 个时间样本都包含 10s 数据,各个样本之间有 0.1s 的偏移。

下一步(哈尔小波变换)要求每个频谱图像的维数是 2 的幂。因此我们要将时间维度从 $L_{fp}=100$ 降低采样到 $2^6=64$ 个样本。我们之前在频率维度上降低采样到 $2^5=32$ 个样本,所以每个频谱图像的最终维度是 32×64 个样本。

哈尔小波变换 接下来我们计算每个频谱图像的二维哈尔小波变换以得到它的小波表示,这有利于使用快速算法来进行图像数据有损压缩,同时保留对小噪声扰动的稳定性(33,39)。图 5d 显示了图 5c 中频谱图像哈尔小波系数的幅度;地震信号在全分辨率小波系数中具有高能量,以不同的模式出现。

小波是多分辨率分析的数学工具:它们按层次关系将数据分解成它们的整体平均形状,并详细描述偏离平均形状连续的水平,从最粗糙到最精细的分辨率(40)。在图 5d 中,最精细的分辨率细节系数在右上角象限内,当我们沿着左下对角方向移动时,它们变得更粗糙,直到我们到达左下角整个频谱图像的平均系数。傅立叶变换的基函数为正弦和余弦,只能在频率域内定位,仅使用几个系数来描述周期信号。类似的方式,离散小波变换(DWT)具有不同的小波基函数,能同时在时间域和频率域中定位,只使用几个小波系数能表示非平稳、突发的信号(如地震)(41)。离散小波变换已被用来改善 STA/LTA 地震检测以及准确估计震相到时(42)。可以使用快速小波变换来递归计算离散小波变换,但要求输入数据的维数是 2 的幂。也可以使用其他小波函数来计算离散小波变换,比如不同阶数的 Daubechies 小波基函数(41),但这比哈尔小波需要更多的计

算量。

数据压缩：小波系数选择　我们现在通过为每个频谱图像选取一小部分哈尔小波系数来对数据进行压缩，丢弃剩余的小波系数。由于大部分连续信号是噪声，我们期望地震的诊断小波系数与噪声的那些诊断小波系数脱离开来。因此，我们保持与其平均值偏离最大的 k 个哈尔小波系数，通过标准化将每个哈尔小波系数进行偏差量化。我们使用基于标准化系数的 z 分量，而不是简单的哈尔小波系数的幅值，因为它们有更大的区别值，并且经验性地导致了地震检测性能的提高。

我们现在描述如何获得标准化的哈尔小波系数。将 $N=N_{\mathrm{fp}}$ 个频谱图像的 $M=32\times 64=2\ 048$ 个哈尔系数放置进 $R^{M\times N}$ 阶 H 矩阵中。令 $\hat{H}$ 为 $\hat{h}_{\mathrm{j}}=h_{\mathrm{j}}/\parallel h_{\mathrm{j}}\parallel_2$ 列的矩阵，通过归一化 H 的每列 h_{j} 来获得；然后，对于矩阵的每个 i 行，我们在所有频谱图像 j 上计算哈尔小波系数 i 的样本均值 μ_{i} 和校正的样本标准差 σ_{i}：

$$\mu_i=\frac{1}{N}\left(\sum_{j=1}^{N}\hat{H}_{ij}\right),$$
$$\sigma_i=\sqrt{\frac{1}{N-1}\left(\sum_{j=1}^{N}(\hat{H}_{ij}-\mu_i)^2\right)} \qquad (3)$$

标准化的哈尔小波系数 $\hat{Z}_{ij}$，根据每个哈尔系数 i 和频谱图像 j 的 z 分值来计算，然后给出那个系数值距离整个数据集的平均值的标准偏差数：

$$\hat{Z}_{ij}=\frac{\hat{H}_{ij}-\mu_i}{\sigma_i} \qquad (4)$$

对于每个频谱图像，我们只选择前 $k=800$ 个标准化的具有最大幅度（保留具有最大幅度的负数 z 分值）的哈尔系数（表 1，800/2 048=39%），并将其余的系数置为 0。只保留前 k 个系数的符号（图 5e）：+1 代表正数（白色），−1 代表负数（黑色），0 代表被丢弃的系数（灰色）。存储符号而不是振幅提供了额外的数据压缩，同时保持了对噪声污染的较强稳定性（33，39）。

二进制指纹　我们生成一个二进制（只包含 0 和 1）且稀疏（大多数为 0）的指纹，以便我们可以使用在下节中描述的局部敏感散列算法来有效地搜索相似的指纹，并最大限度地减少存储所需的比特数。我们用 2 个比特来表示每个标准哈尔系数的符号：−1→01，0→00，1→10。因此，每个指纹使用的比特数是哈尔系数的 2 倍。因为每个频谱图像窗口有 2 048 个哈尔系数，每个指纹有 2×2 048=4 096 比特。图 5f 显示从地震频谱图像衍生出来的二进制指纹，其中 1 是白色，0 是黑色。

3.2　相似性搜索

特征提取后，我们得到了 N_{fp} 个指纹的一个集合，每个指纹对应着一个频谱图像（乃至波形）。我们的目标是识别相似的指纹对来检测地震。指纹和相似性阈值法首先生成一个数据库，其中相似指纹高概率地被分组到相同的散列桶中。然后，在相似性搜索中，数据库返回与通过杰卡德相似度测量的所有与给定的搜索查询指纹相似的指纹。此搜索对日益增大的数据库来说是快速和可扩展的，单次搜索查询有近乎恒定的运行时间。指纹和相似性阈值法使用数据库中的每个指纹作为一个搜索查询，所以总运行时间几乎是线性的。

杰卡德相似度　在模板匹配和自相关中，我们使用归一化的相关系数[公式(1)]来测量两个时域波形之间的相似性。本文中，我们使用杰卡德相似度作为在局部敏感散列执行中比较指纹的相似性度量标准。两个二进制指纹 A 和 B 的杰卡德相似度定义为(23)：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \quad (5)$$

在式(5)中，分子包含 A 和 B 的比特数都等于 1，而分母是 A，B 或者 A 和 B 的比特数等于 1。图 S9A 显示了两个非常相似的归一化地震波形，而图 S9B 显示了它们对应指纹的杰卡德相似度。

数据库生成 使用一种被称为最小合理的独立排列(最小散列)来将每个指纹的维数从一个 4096 个元素的位向量缩减到一个较短的整数数组(43)。最小散列使用多个随机函数 $h_i(x)$ 置换 i，其中每个散列函数将任一稀疏的二进制高维度指纹 x 映射成一个整数 $h_i(x)$。最小散列有一个重要的局部敏感散列属性——两个指纹 A 和 B 映射成同一整数的概率等于它们的杰卡德相似度：

$$\Pr[h(A)=h(B)]=J(A,B) \quad (6)$$

因此，最小散列降低了维数，同时以概率的方式保留了 A 和 B 之间的相似性(23，33)。

若给定一个稀疏二进制指纹作为输入(23)，则最小散列输出的是一个被称为最小散列签名(MHS)的无符号整数数组 p。最小散列签名可用于估计指纹 A 和 B 之间的杰卡德相似度，只要通过从 A 和 B 的最小散列签名计算出匹配整数的数值，然后除以 p；杰卡德相似度估计会随着 p 的增加而提高(33)。使用应用于同一指纹的不同的随机散列函数 h_i 来计算 p 的每个整数。这些 p 最小散列函数通过绘制来自均匀分布的 $p\times 4\,096$(其中 4 096 是指纹的比特数)个独立同分布的随机样本来创建，通过调用均匀随机散列函数返回，以得到一个数组 $r(i,j)$，其中 $i=1,\cdots,p$ 和 $j=1,\cdots,4\,096$。然后，为了获得一个给定指纹 x 的最小散列函数 $h_i(x)$ 的输出，我们使用指纹中 k 个非零位的索引来从 $r(i,j)$ 数组中选择 k 个值。例如，如果我们考虑所有 p 散列函数中的第一个散列函数 $h_1(x)$，并且如果指纹 x 中非零位的索引是 $j=4$，则选择 $r(1,4)$。在所有 k 个选择值 $r(i,j)$ 中，我们选择最小值，并分配获得最小值的索引 j 为最小散列函数 $h_i(x)$ 的输出(23)。我们通过只保持 8 比特数来进一步降低输出大小，以便最小散列签名总共有 $8p$ 比特数；最小散列签名中的每个整数的值在 0～255 之间(33)。图 6a 显示两个相似指纹 A 和 B 的样本最小散列签名数组，其中 $p=6$。

局部敏感散列使用最小散列签名来将每个指纹插入数据库中。图 6b 演示了最小散列签名如何将两个相似的指纹 A 和 B 放进数据库每个散列表的散列桶中，给定它们的最小散列签名数组(图 6a)。每个最小散列签名的 $8p$ 比特数被划分为 b 个子集，每个子集($p=rb$)有 $8r$ 位。这些 $8r$ 比特数被连接起来生成散列键，它只属于 b 个散列表中的一个。每个散列键是用于检索散列桶的 64 位整数索引，它可以包含多个值(指纹的参考)。对于一个给定的散列表，如果 A 和 B 共享同一个散列键(等价地说，如果图 6 中它们的最小散列签名相同)，那么它们被插入相同的散列桶中；否则，它们被插入不同的散列桶中(23，33)。

图 7a 展示了由局部敏感散列创建的一个数据库的示意图；高度相似的指纹有可能被组合到同一个散列桶中。该数据库将指纹的 32 位整数索引存储到散列桶中，而不是指纹(或波形)本身。我们从所有 N_{fp} 个指纹来为每个最小散列签名生成 bN_{fp} 个散列键和值，并将所有值插入到 b 个散列表内的散列桶中，只要给出它们相应的散列键。$N_{\mathrm{fp}}=604\,781$ 个指纹中每个指纹代表一些散列桶中的每个散列表，所以局部敏感散列产生多个指纹分组到散列桶中。

数据库内的相似性搜索 局部敏感散列生成的数据库为搜索相似指纹对以及相似波

形提供了一种快速、有效的方式。图 7b 展示了如何在数据库内搜索与查询指纹(蓝色)相似的指纹。对于每个搜索查询指纹，我们使用图 6 所示的步骤来确定每个散列表中它的散列桶，并使用相应的散列键来检索包含在这些选定的散列桶中的所有指纹参考，形成指纹对，其中第一个条目是查询指纹，而第二个条目是来自从数据库同一个散列桶的相似指纹(23，33)。因此，对于每一个查询，我们执行与散列表数量一样多的查找。查找的平摊时间是 $O(1)$。检索时间取决于每个桶中条目的数量，每个桶都需要包含数据库中指纹参考的一小部分指纹引用子集，因此，我们忽略了所有其他散列桶中的指纹参考，使搜索随着数据库大小的增加而可升级。对于所有检索对中一个给定的查询，我们只保留 $b=100$ 个散列表中至少出现的 $v=4$ 个，对于一个初始的指纹和相似性阈值法的相似性阈值 $4/100=0.04$(表 1)而言，这些成为我们的候选对(33)。我们后来设置 $b=100$ 个散列表中的 $v=19$，在对这些指纹对对应的波形进行视觉检查后，选择 0.19 的指纹和相似性阈值法的相似度作为事件检测阈值(表 1)。我们将指纹和相似性阈值法相似度定义为同一散列桶中包含指纹对的散列表的分数。

一次成功搜索的理论概率——两个指纹在至少 v 个 b 散列表中的同一个数据桶之间的概率(存在散列碰撞)，每个表有 r 个散列函数，作为杰卡德相似度 s 的函数，由下式给出(23)：

$$\mathrm{Pr}=1-\sum_{i=0}^{v-1}\left[\binom{b}{i}(1-s^r)^{b-i}(s^r)^i\right] \quad (7)$$

$$\binom{b}{i}=\frac{b!}{i!(b-i)!} \quad (8)$$

图 S10 中的红色曲线(与所有子图中的相同)绘制了公式(7)随杰卡德相似度 s 的变化，给定我们特定的输入参数(表 1)：每个表的 $r=5$ 个散列函数，$b=100$ 个散列表(所以每个指纹的最小散列签名有 $p=rb=500$ 个整数)，且 $v=19$ 作为同一个散列桶中包含一个指纹对的散列表数量的阈值。此概率随着相似度单调增加。

我们通过改变 r，b 和 v 参数来调整公式(7)中曲线的位置和斜率，因此我们可以将杰卡德相似度修改成具有 50% 的成功搜索概率。图 S10A 修改 r，同时保持 b 和 v 的恒定；曲线随着 r 的增大右移，对于一次成功的搜索需要更高的杰卡德相似性。对于一个大的 r 值，有大量的散列桶，因此，在这些桶内所产生的低密度指纹可能会导致漏检，因为相似的指纹更可能出现在不同的桶内。但如果 r 太小，就有太少的散列桶；每个桶中可能有太多指纹，这将增加搜索相似指纹的运行时间和虚假检测的可能性。图 S10B 修改 b，同时保持 r 和 v 不变；随着 b 的增大曲线左移，由于有更多的散列表增加了同一个桶中搜索到两个相似指纹的概率，即使它们有中等的杰卡德相似度。但是，这是以增加内存需求、搜索运行时间和虚假检测为代价(23)。图 S10C 修改 v，同时保持 r 和 b 恒定；随着 v 的增加，曲线以更陡的斜率移动到右边，一次成功的搜索要求指纹对之间具有更高的杰卡德相似度以及检测和非检测之间具有更鲜明的区隔。

为了检测连续数据中的相似地震，我们指纹和相似性阈值法的多对多搜索使用数据库中的每个指纹作为一个搜索查询，这样我们就可以找到数据库中与每个查询指纹相似的所有其他指纹，具有近似线性的时间复杂度：$O[(N_{\mathrm{fp}})^{1+\rho}]$，其中 $0<\rho<1$。对于这个数据集，使用表 1 中的参数，且散列函数的数目增加到 $r=7$，我们估计 $\rho=0.36$，假定相似性搜索的运行时间 t 是图 4b 中连续数据持续时间 d 的函数；我们假设一个幂律

比例关系 $t=Cd^{(1+\rho)}$，其中我们用最小二乘线性拟合在对数空间中求解因子 C 和 ρ：$\log t=\log C+(1+\rho)\log d$。由于 d 与 N_{fp} 之间存在线性关系，$O(d^{1+\rho})\sim O[(N_{fp})^{1+\rho}]$。这比自相关的二次方运行时间更快和更具可扩展性：当 $N>N_{fp}$，$O(N^2)$。相似性搜索的输出是相似指纹索引对的一个列表，我们将其转换成连续数据中的时间，具有相关的指纹和相似性阈值法的相似性值。我们可以将此列表可视化为一个稀疏对称的 $N_{fp}\times N_{fp}$ 阶相似度矩阵（图 S8）。这个矩阵是稀疏的，因为局部敏感散列算法避免搜索构成大部分可能对的不同指纹对。

补充材料

本文的补充材料可从 http://advances.sciencemag.org/cgi/content/full/1/11/e1501057/DC1 上获取。

参考文献

1. P. M. Shearer, *Introduction to Seismology* (Cambridge Univ. Press, New York, ed. 2, 2009).
2. R. Allen, Automatic phase pickers: Their present use and future prospects. *Bull. Seismol. Soc. Am.* **72**, S225－S242 (1982).
3. M. Withers, R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, J. Trujillo, A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bull. Seismol. Soc. Am.* **88**, 95－106 (1998).
4. R. J. Geller, C. S. Mueller, Four similar earthquakes in central California. *Geophys. Res. Lett.* **7**, 821－824 (1980).
5. D. P. Schaff, G. C. Beroza, Coseismic and postseismic velocity changes measured by repeating earthquakes. *J. Geophys. Res.* **109**, B10302 (2004).
6. G. Poupinet, W. L. Ellsworth, J. Frechet, Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. *J. Geophys. Res.* **89**, 5719－5731 (1984).
7. S. J. Gibbons, F. Ringdal, The detection of low magnitude seismic events using array-based waveform correlation. *Geophys. J. Int.* **165**, 149－166 (2006).
8. D. R. Shelly, G. C. Beroza, S. Ide, Non-volcanic tremor and low-frequency earthquake swarms. *Nature* **446**, 305－307 (2007).
9. D. P. Schaff, F. Waldhauser, One magnitude unit reduction in detection threshold by cross correlation applied to Parkfield (California) and China seismicity. *Bull. Seismol. Soc. Am.* **100**, 3224－3238 (2010).
10. A. Kato, S. Nakagawa, Multiple slow-slip events during a foreshock sequence of the 2014 Iquique, Chile M_W 8.1 earthquake. *Geophys. Res. Lett.* **41**, 5420－5427 (2014).
11. Z. Peng, P. Zhao, Migration of early aftershocks following the 2004 Parkfield earthquake. *Nat. Geosci.* **2**, 877－881 (2009).
12. X. Meng, Z. Peng, J. L. Hardebeck, Seismicity around Parkfield correlates with static shear stress changes following the 2003 M_W 6.5 San Simeon earthquake. *J. Geophys. Res.* **118**, 3576－3591 (2013).
13. D. R. Shelly, D. P. Hill, F. Massin, J. Farrell, R. B. Smith, T. Taira, A fluid-driven earthquake swarm on the margin of the Yellowstone caldera. *J. Geophys. Res.* **118**, 4872－4886 (2013).
14. C.-C. Tang, Z. Peng, K. Chao, C.-H. Chen, C.-H. Lin, Detecting low-frequency earthquakes within non-volcanic tremor in southern Taiwan triggered by the 2005 M_W 8.6 Nias earthquake. *Geophys. Res. Lett.* **37**, L16307 (2010).
15. R. J. Skoumal, M. R. Brudzinski, B. S. Currie, J. Levy, Optimizing multi-station earthquake template matching through re-examination of the Youngstown, Ohio, sequence. *Earth Planet. Sci. Lett.* **405**, 274－280 (2014).
16. D. Bobrov, I. Kitov, L. Zerbo, Perspectives of cross-correlation in seismic monitoring at the international data centre. *Pure Appl. Geophys.* **171**, 439－468 (2014).

17. K. Plenkers, J. R. R. Ritter, M. Schindler, Low signal-to-noise event detection based on waveform stacking and cross-correlation: Application to a stimulation experiment. *J. Seismol.* **17**, 27－49(2013).

18. F. Song, H. S. Kuleli, M. N. Toksöz, E. Ay, H. Zhang, An improved method for hydrofracture-induced microseismic event detection and phase picking. *Geophysics* **75**, A47－A52(2010).

19. D. Harris, Subspace Detectors: Theory(*Lawrence Livermore National Laboratory Report UCRL-TR-222758*)(Lawrence Livermore National Laboratory, Livermore, CA, 2006), p. 46.

20. S. A. Barrett, G. C. Beroza, An empirical approach to subspace detection. *Seismol. Res. Lett.* **85**, 594－600(2014).

21. J. R. Brown, G. C. Beroza, D. R. Shelly, An autocorrelation method to detect low frequency earthquakes within tremor. *Geophys. Res. Lett.* **35**, L16305(2008).

22. A. C. Aguiar, G. C. Beroza, PageRank for earthquakes. *Seismol. Res. Lett.* **85**, 344－350(2014).

23. J. Leskovec, A. Rajaraman, J. D. Ullman, Finding similar items, in *Mining of Massive Datasets* (Cambridge Univ. Press, New York, ed. 2, 2014), pp. 73－130; http://www.mmds.org.

24. U. Manber, Finding similar files in a large file system, *Proceedings of the USENIX Conference*, San Francisco, CA, 17 to 21 January 1994, pp. 1－10.

25. M. Henzinger, Finding near-duplicate Web pages: A large-scale evaluation of algorithms, *Proceedings of the 29th SIGIR Conference*, Seattle, WA, 06 to 10 August 2006(ACM).

26. B. Stein, S. M. zu Eissen, Near-similarity search and plagiarism analysis, *Proceedings of the 29th Annual Conference German Classification Society*, Magdeburg, Germany, 09 to 11 March 2005, pp. 430－437.

27. J. Haitsma, T. Kalker, A highly robust audio fingerprinting system, *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, 13 to 17 October 2002, pp. 144－148.

28. A. Wang, An industrial-strength audio search algorithm, *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, MD, 27 to 30 October 2003, pp. 713－718.

29. J. Zhang, H. Zhang, E. Chen, Y. Zheng, W. Kuang, X. Zhang, Real-time earthquake monitoring using a search engine method. *Nat. Commun.* **5**, 5664(2014).

30. M. Rodgers, S. Rodgers, D. C. Roman, Peakmatch: A Java program for multiplet analysis of large seismic datasets. *Seismol. Res. Lett.* **86**, 1208－1218(2015).

31. A. Andoni, P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* **51**, 117 － 122 (2008).

32. A. Levitin, *Introduction to the Design and Analysis of Algorithms*(Pearson Education, Addison-Wesley, Upper Saddle River, NJ, ed. 3, 2012).

33. S. Baluja, M. Covell, Waveprint: Efficient wavelet-based audio fingerprinting. *Pattern Recognit.* **41**, 3467－3480(2008).

34. D. P. Schaff, G. H. R. Bokelmann, G. C. Beroza, F. Waldhauser, W. L. Ellsworth, High-resolution image of Calaveras Fault seismicity. *J. Geophys. Res.* **107**, ESE 5-1－ESE 5-16(2002).

35. D. A. Dodge, W. R. Walter, Initial global seismic cross-correlation results: Implications for empirical signal detectors. *Bull. Seismol. Soc. Am.* **105**, 240－256(2015).

36. X. Meng, X. Yu, Z. Peng, B. Hong, Detecting earthquakes around Salton Sea following the 2010 M_W7.2 El Mayor-Cucapah earthquake using GPU parallel computing. *Proc. Comput. Sci.* **9**, 937－946(2012).

37. T. G. Addair, D. A. Dodge, W. R. Walter, S. D. Ruppert, Large-scale seismic signal analysis with Hadoop. *Comput. Geosci.* **66**, 145－154(2014).

38. M. Joswig, Pattern recognition for earthquake detection. *Bull. Seismol. Soc. Am.* **80**, 170 － 186 (1990).

39. C. E. Jacobs, A. Finkelstein, D. H. Salesin, Fast

multiresolution image querying, *Proceedings of SIGGRAPH* 95, Los Angeles, CA, 06 to 11 August 1995.

40. E. J. Stollnitz, T. D. Derose, D. H. Salesin, Wavelets for computer graphics: A primer. 1. *IEEE Comput. Graphics Appl.* **15**, 76－84(1995).

41. D. Shasha, Y. Zhu, *High Performance Discovery in Time Series: Techniques and Case Studies* (Springer, Berlin, 2004).

42. H. Zhang, C. Thurber, C. Rowe, Automatic P-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings. *Bull. Seismol. Soc. Am.* **93**, 1904－1912 (2003).

43. A. Z. Broder, M. Charikar, A. M. Frieze, M. Mitzenmacher, Min-wise independent permutations. *J. Comput. Syst. Sci.* **60**, 630－659(2000).

44. J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 25 to 29 June 2006, pp. 233－240.

译 者 简 介

李万金(1977—)，**男，中国科学技术大学地质工程专业硕士研究生毕业，云南省地震局个旧地震台高级工程师，主要从事地震观测及相关研究工作**，E-mail: ynlwj@sina.com。