一种新的不确定性时间序列概率相似查找方法①

廖建平

(衢州职业技术学院 信息工程学院, 衢州 324000)

摘 要: 针对传统的数据管理中的数据表示、存储与索引、查询与挖掘等所有技术, 不能直接应用于不确定性时间序列数据的相似性查找的不足. 研究了可用于不确定性时间序列数据的降维表示、索引与剪枝、查找等理论与技术, 针对不确定性时间序列数据结构的复杂性, 首次给出了不确定性时间序列上的概率最近邻的定义; 将不确定性时间序列进行了 PLA 降维, 转换到 PLA 空间, 并提出了三个引理, 用以加速查找效率; 基于该三个引理, 提出了概率 K 最近邻查找算法 PKNNS. 通过实验, 验证了 PKNNS 算法的有效性和效率.

关键词: 不确定性时间序列; 分段线性逼近; 相似性查找; 最近邻查找

A Novel Probabilistic Similarity Search Method for Uncertain Time-Series

LIAO Jian-Ping

(College of Information Engineering, Quzhou College of Technology, Quzhou 324000, China)

Abstract: The traditional data management, data representation, storage and indexing, querying, mining and all other technical can not be directly applied to the similarity search of time series data with uncertainty. Our work study for other theories and technologies, the uncertainty of time-series data for the complexity of the structure, and for the first time we give the formal definition of probability nearest neighbor search over uncertain time series database; the PLA dimensionality reduction over time series of uncertainty. After conversion to the PLA space, we propose three lemmas to accelerate the search efficiency; based on three lemmas the appropriate searching algorithm PKNNS is also given. A serials of experiments are also made to test the effectiveness and efficiency of algorithm PKNNS.

Key words: uncertain time series; Piecewise Linear Approximation; similarity search; nearest neighbor search

从二十世纪九十年代初以来, 时间序列的最近邻查找已经得到广泛的研究. 迄今为止, 因为在传感器网络监控[3]、移动对象跟踪[2]和股票数据分析等领域的广泛应用, 该课题仍是一个热门的研究方向. 例如, 在煤矿中[1,4], 应用传感器采集诸如温度和氧气密度等时间序列数据, 因为紧急事件通常和一些特殊的模式相关, 因此事件检测考虑时间序列上这些特殊模式的查找, 为了保护矿工的生命, 要求能够实现快速的查找. 不同于传统的确定性时间序列数据库中的相似性查找, 数据的不确定性使得从查询结果中直接检索准确数据变得没有意义[5,6]. 也就是说, 我们不能简单地在查询结果中用是或不是来标记一个时间序列, 而是返回这些查询结果的最近邻概率.

近年来, 时间序列的相似性查找研究主要集中在两个方面: (1)新的降维技术; (2)新的度量两个时间序列相似性的方法. 其中代表性的降维技术包括 DFT 方法[7]、PLA 方法[8]和 CP 方法[9]. 这些方法首先将对高维的时间序列进行降维处理, 然后使用新的度量距离函数计算两个转换后的时间序列的相似性. 两个转换后的时间序列数据在降维空间中的距离应该是在原始空间中欧氏距离的降维下界. 以上这些关于时间序列相似性查找方法仅能处理确定性时间序列, 还不能直接用于不确定性时间序列. 因此本文提出的新技术能够解决不确定性时间序列的相似性查找问题.

① 收稿时间:2012-09-28;收到修改稿时间:2012-11-09

本文的主要贡献包括: 1)给出了在不确定时间序列上的概率最近邻查找的定义; 2)基于在 PLA 空间的转换后的不确定时间序列, 提出了用于提高查找效率的三个引理; 3)通过大量的实验验证了新方法的有效性.

1 问题定义

定义 1(不确定性时间序列). 不确定性时间序列数据库中的一个元组, 是指该数据库中的一个不确定性时间序列, 记为 TS_u=<S_{1l}, S_{1u}; S_{2l}, S_{2u}; …; S_{nl}, S_{nu} >. 其中, S_{il}, S_{iu}(1≤i≤n)分别是元组 Si 的上界和下界; n 是时间序列 TS_u 的长度; 元组的每一个数据项 Si 取值是在[S_{il}, S_{iu}]域内服从某种非零的概率函数分布, 比如均匀分布. 即 Si 的取值是不确定的.

定义 2(概率 K 最近邻查找). 不确定性时间序列数据库中的概率 K 最近邻查找, 是指对于一个给定的查询时间序列 Q, 在不确定性时间序列数据库中查找与 Q 距离最小的概率不为零的 K 个不确定性时间序列 TS_u, 查找结果返回 K 个时间序列以及相应的概率. 查询时间序列 Q 或者是确定的, 或者是不确定的, 即 Q=<q_1, q_2, …, q_n>或者 Q_u=<q_{1l}, q_{1u}; q_{2l}, q_{2u}; …; q_{nl}, q_{nu}>. 概率 K 最近邻查找, 简记为 PKNN 查找(Probabilistic K Nearest Neighbors Search), 当 K=1 时, 简记为 PNN 查找, 称为概率最近邻查找, 返回 1 个最近邻时间序列及其概率.

定义 3(不确定时间序列距离). 一个不确定性时间序列 TS_u 与 Q 的距离, 定义为 TS_u 与 Q 的方差, 即

$$dist^2(TS_u,Q)=\sum_{i=1}^{n}(s_i-q_i)^2, s_i\in[s_{il},s_{iu}] \quad (1)$$

2 不确定性时间序列概率相似查找算法

2.1 降维和表示

由于不确定性时间序列的长度很长, 不能直接使用空间索引的方法, 需要采用降维技术将原始的不确定性时间序列降维到一个低维空间. 因为 PLA (Piecewise Linear Approximation)降维方法具有较高的重建精度和较强的剪枝力度, 所以采用 PLA 方法进行降维和表示.

不确定时间序列 TS_u=<S_{1l},S_{1u};S_{2l},S_{2u};…;S_{nl},S_{nu} >, 可以被分为两个确定的时间序列 TS_{lu}= <S_{1l};S_{2l};S_{3l};S_{4l};…;S_{nl}>和 TS_{uu}=<S_{1u};S_{2u};S_{3u};S_{4u};…;S_{nu} >, TS_{lu} 和 TS_{uu} 分别是 TS_u 的下界和上界. 这样, 可以将长度为 n 的时间序列 TS_{lu}、TS_{uu} 和 Q 均划分为长度为 l 的 m 个彼此不重叠的分段. 然后, 根据 Qiuxia 等人[1]提出的 PLA 方法, 得到降维后的 PLA 空间中的序列数据为

$$TS_{lu\text{-}PLA}=<a_{11},b_{11};a_{12},b_{12};\ldots;a_{1m},b_{1m}> \quad (2)$$

$$TS_{uu\text{-}PLA}=<a_{21},b_{21};a_{22},b_{22};\ldots;a_{2m},b_{2m}> \quad (3)$$

$$Q_{PLA}=<a_{31},b_{31};a_{32},b_{32};\ldots;a_{3m},b_{3m}> \quad (4)$$

2.2 相似性度量计算

TSlu-PLA 和 QPLA 的下界方差距离为

$$dist^2_{PLA}(TS_{lu},Q)=\sum_{i=1}^{m}\sum_{j=1}^{l}[(a_{1i}-a_{3i})j+(b_{1i}-b_{3i})]^2$$
$$=\sum_{i=1}^{m}\left(\frac{l(l+1)(2l+1)}{6}(a_{1i}-a_{3i})^2+l(l+1)(a_{1i}-a_{3i})\times(b_{1i}-b_{3i})+l(b_{1i}-b_{3i})^2\right) \quad (5)$$

$TS_{lu\text{-}PLA}$ 和 Q_{PLA} 的上界方差距离为

$$dist^2_{PLA}(TS_{uu},Q)=\sum_{i=1}^{m}\sum_{j=1}^{l}[(a_{2i}-a_{3i})j+(b_{2i}-b_{3i})]^2$$
$$=\sum_{i=1}^{m}\left(\frac{l(l+1)(2l+1)}{6}(a_{2i}-a_{3i})^2+l(l+1)(a_{2i}-a_{3i})\times(b_{2i}-b_{3i})+l(b_{2i}-b_{3i})^2\right) \quad (6)$$

在原始的没有降维的高维空间中, TS_{lu} 和 Q 的下界欧几里德方差距离为

$$dist^2(TS_{lu},Q)=\sum_{i=1}^{n}[(a_{il}-q_i)]^2=\sum_{i=1}^{m}\sum_{j=(i-1)+1}^{i-l}[(a_{jl}-q_j)]^2 \quad (7)$$

在原始的没有降维的高维空间中, TS_{uu} 和 Q 的上界欧几里德方差距离为

$$dist^2(TS_{uu},Q)=\sum_{i=1}^{n}[(a_{iu}-q_i)]^2=\sum_{i=1}^{m}\sum_{j=(i-1)+1}^{i-l}[(a_{ju}-q_j)]^2 \quad (8)$$

因为 PLA 降维能保持降维下界定理, 所以

$$dist^2_{PLA}(TS_{lu},q)\le dist^2(TS_{lu},Q) \quad (9)$$

$$dist^2_{PLA}(TS_{uu},q)\le dist^2(TS_{uu},Q) \quad (10)$$

PLA 的这个性质, 能保证在降维后的 PLA 空间中剪枝时不会导致漏报.

2.3 索引和剪枝

形式化地, 对于一个在 PLA 空间的不确定性时间序列 TS_u 和一个查询点 Q 的最小(最大)方差距离, 要么是 $dist^2_{PLA}(TS_{lu},Q)$, 要么是 $dist^2_{PLA}(TS_{uu},Q)$. 这两个值均来自于 m 个离散的分段的和, 而对于第一个分段内, 如第 i-th 个分段内, 最小(或最大)方差距离为

$$dist^2_{PLA}(S_{i-lu},Q_i)=\sum_{j=1}^{l}[(a_{1i}-a_{3i})j+(b_{1i}-b_{3i})]^2 \quad (11)$$

$$dist^2_{PLA}(S_{i-uu}, Q_i) = \sum_{j=1}^{l} [(a_{2i} - a_{3i})\, j + (b_{2i} - b_{3i})]^2 \quad (12)$$

引理 1. 对于 PLA 空间中的两个不确定性时间序列 S_u 和 T_u, 对于查找点 Q, 如果 S_u 和 Q 在第 i-th 个分段的最小距离 $\min dist^2_{PLA}(S_{i-u}, Q_i)$ 大于 T_u 和 Q 在第 i-th 个分段的最大距离 $\max dist^2_{PLA}(T_{i-u}, Q_i)$, 那么在第 i-th 个分段的索引树中可以安全地剪去 S_u. 这里和下文中, 我们称不会导致漏报的剪枝为安全的剪枝.

证明: 显然, 由于 S_u 和 Q 的最小距离大于 T_u 和 Q 的最大距离, 所以, S_u 没有机会成为 Q 的 NN, 即 S_u 成为 Q 的 NN 的概率为 0, 因此, 可以安全地剪去 S_u.

基于引理 1, 在不确定性时间序列数据库中的 NN 查询的关键之一, 是在其对应的 PLA 空间中, 在第 i-th 个分段, 怎样找到剪枝所需要的最小最大 minmaxdist^2(), 简记 mM, 例如, 如果 minmax dist$^2_{PLA}$(S_{i-u}, Q_i) 是这个最小最大距离, 那么 mM=minmaxdist$^2_{PLA}$(S_{i-u}, Q_i). 为此, PKNNS 算法提出对每一个 PLA 分段, 逐个各自构建两个 2 维的 R 树 R_{il} 和 R_{iu}, 这两个树的节点中分别包含着第 i-th 个分段内的 PLA 降维所对应的上界系数 2 维点和下界系数 2 维点. 例如, <a_{1l}, b_{1l}>作为一个 2 维的节点插入到 R_{1l} 树中, <a_{2l}, b_{2l}>作为一个 2 维节点插入到 R_{1u} 树中. 与此相同地, 将不确定性时间序列数据库中的每一个时间序列的 PLA 降维所对应的系数 2 维点分别插入到 R_{1l} 树中, 假设数据库中有 S 个时间序列, 那么, 最后这 S 个时间序列的系数 2 维点<a_{il}, b_{il}>均成为 R_{1l} 树中的叶子节点.

至此, 可以通过搜索 R_{1u}, 找到最小最大距离 mM, 然后用该 mM 对 R_{1l} 剪枝, 剪去 R_{1l} 中大于 mM 的分支, 直到搜索到小于 mM 的分支节点停止, 那么, 该分支下的所有节点的索引项所对应的不确定性时间序列均作为第 1-th 个分段上的 Q 的 NN 选项. 相同地, 对 2、3、…、m 个分段, 逐个各自构建 R_{iu} 和 R_{il}, 同时进行同样的剪枝和搜索, 得到每一个分段上的 Q 的 NN 候选. 接下来, 按引理 2 计算这些候选作为 Q 的 NN 的概率.

引理 2. 对于 PLA 空间中的不确定性时间序列 S_u, 对于查找点 Q, 假设 S_u 在第 i-th 分段上作为 Q 的 NN 候选的概率为 p_i, 那么最终, S_u 在整个时间序列上作为 Q 的 NN 候选概率, 也即 S_u 在全部所有的 m 个分段上作为 Q 的 NN 候选的概率的和, 记为 P_s, 则

$$P_s = \sum_{i=1}^{m} p_i / m \quad .$$

证明: 因为不确定性时间序列 S_u 是独立划分成 m 个分段的, 所以, 在每一个分段均有可以作为 Q 的 NN 候选的相等概率 1/m, 则

$$P_s = \sum_{i=1}^{m} p_i \times (1/m) = \sum_{i=1}^{m} p_i / m \quad .$$

实际上, 大多数查询用户并不关心具体的概率值是多少, 而是需要对于返回的 NN, 有一个大于给定的阀值的置信度, 即返回的最终结果 NN 集合中的每一项的概率大于给定的阀值概率. 引理 3 可以进一步改进查找效率.

引理 3. 如果 S_u 在 u (0≤u≤m)个分段上是 Q 的 NN 候选, 那么, 最终 S_u 作为 Q 的 NN 候选的概率 P_s 的上限为 u/m.

证明: 如果 S_u 是 u 个分段内的每一个分段上的 Q 的唯一的 NN 候选, 那么 S_u 在该 u 个分段内每一个分段上的 Q 的 NN 候选概率为 1. 这样, 根据引理 2, 则 Su 最终作为 Q 的 NN 候选的概率 P_s 的上限为(1+1+…+1)/m=1×u/m=u/m.

具体地, 如果 S_u 在第 i-th 个分段上作为 Q 的 NN 候选概率为 p_i, 那么 S_u 最终作为 Q 的 NN 候选的概率 P_s 的上限下降为 p_i/m+(u-1)/m=(p_i+u-1)/m. 因此, 不需要等到计算出全部每个分段上的 p_i, 而是可以每计算出一个分段上的 p_i, 就可以立即更新和测算 P_s 的上限, 从而可以不再计算 P_s 的上限小于给定的概率阀值的 NN 候选的 p_i 了, 同时也可以立即放弃将 S_u 作为最终的 NN. 一旦 P_s 的上限小于给定的概率阀值的 NN 候选, 那么 P_s 不能作为最终要返回给查询用户的 NN 集合中的一项, 因此, 提高了在 NN 候选集合中计算最终可以作为 NN 的概率的计算效率.

2.4 算法过程

结合上述三个引理, 不确定时间序列概率相似查找算法 PKNNS 描述如下:

假设我们已经为所有不确定性时间序列在每个分段上构建了 R_{il} 和 R_{iu} 树, 那么在 R_{il} 和 R_{iu} 树上, 基于给定的查询点 Q, 以最佳方式执行两个标准的最近邻查询. 当算法返回头两个 1NN 点时, 最大的一个将作为不确定性时间序列第 i-th 分段最小最大距离值. 例如, minmaxdist^2(R_i, Q_i)=dist$^2_{PLA}$(S_{i-u}, Q_i). 接着, 输出较小 1NN 点的在 R 树上的最近邻查询过程将继续搜索直到返回值大于 minmaxdist^2(R_i, Q_i). 按照上面的示例, 将继续遍历 R_{il} 树, 直到返回点 C_u, 使得 dist$^2_{PLA}$(C_{i-lu}, Q_i)

$\geqslant \text{minmaxdist}^2(R_i, Q_i)$.

3 实验结果和分析

3.1 数据集

通过实验验证 PKNNS 算法的有效性和查找效率. 效率从剪枝力度和 WCT 时间(Wall Clock Time, 时钟时间)两个方面进行与 LS(Linear Scan)算法的对比实验. 剪枝力度用在 PLA 降维空间中, 能够被剪去的时间序列的比例来表明. WCT 时间包括 CPU 时间和 I/O 时间, 实验中每一次 I/O 的时间为 10ms.

数据集: 为了验证 PKNNS 的有效性, 本文通过一定量的拟合数据进行了实验, 每一个拟合数据集包含了 50K 左右的不确定性时间序列, 每个时间序列的长度为 128-1024 个数据点.

3.2 PKNNS 算法与 LS 算法的性能比较

3.2.1 不同的降维维度对剪枝力度和 WCT 时间的影响的实验

如图 1(a)(b)所示, 对不同的降维维度, PKNNS 算法与 LS 算法的剪枝力度和 CPU 的对比, 时间序列数据长度取默认值 256. 剪枝力度方面, 图 1(a)表明, 算法中被剪去元组的比率, 即剪枝力度, LS 算法的每一组的结果几乎为 0, 而 PKNNS 算法的每一组的结果均大于 0.3; 另外, 随着降维度的增加, PKNNS 算法的剪枝力度也相应的增加; 但是, 图 1(b)也表明, 随着降维维度的增加, PKNNS 的 WCT 时间也相应稍有增加, 表明 PKNNS 算法依然轻微受到“维灾”的影响, 而 LS 算法的 WCT 时间每一组都较长, 每一组都有严重的“维灾”, 即使是维度较低时间也有“维灾”.

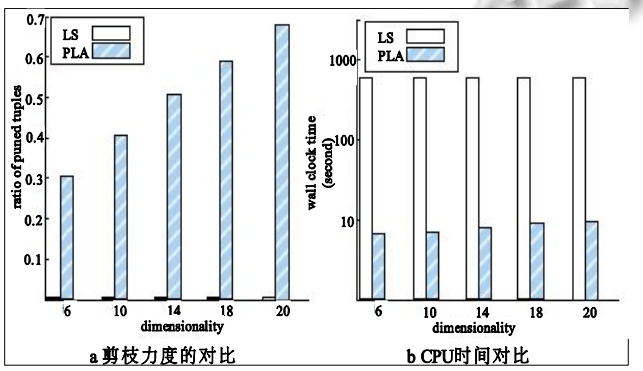

图 1 对不同的降维维度，PLA(PKNNS)与 LS 的剪枝力度和 CPU 时间的对比

3.2.2 不同的时间序列数据长度对剪枝力度和 WCT 时间的影响的实验

如图 2(a)(b)所示, 对不同的时间序列长度, PKNNS 算法与 LS 算法的剪枝力度和 CPU 时间的对比, 时间序列的长度取默认值 14. 图 2(a)表明, PKNNS 算法的前枝力度比 LS 大很多, 图 4(b)表明 PKNNS 算法的 WCT 时间比 LS 算法的 WCT 时间少很多. 进一步地, 图 3(a)表明, 随着时间序列长度的增大, PKNNS 算法的剪枝力度有所下降, 但仍然在有效的剪枝力度内. 这是合理的, 因为随着时间序列长度的增大, 相应的要增加 PLS 降维后的维度, 从而会增多不能被剪去的时间序列. 再进一步地, 图 2(b)表明, 随着时间序列长度的增大, PKNNS 算法的 WCT 时间有所增多, 但仍然比 LS 算法小很多, 而 LS 算法的 WCT 时间在每一组均比 PKNNS 算法的 WCT 时间多很多. 总体上, 这些实验验证了 PKNNS 算法的有效性和效率.

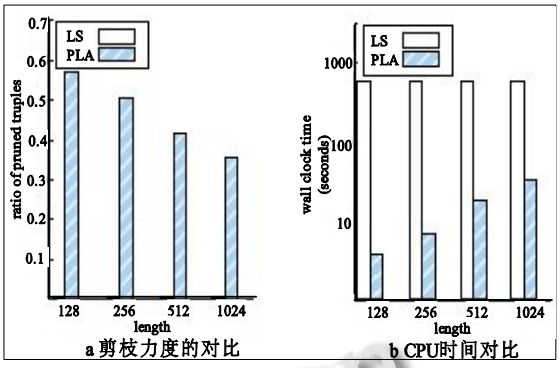

图 2 对不同的时间序列长度，PLA(PKNNS)与 LS 的剪枝力度和 CPU 时间的对比

4 结语

本文集中研究了不确定性时间序列的概率相似性查找, 提出了新颖的不确定性时间序列的概率相似性查找算法 PKNNS. 集中研究了在不确定性时间序列数据库中, 对确定性时间序列 Q 的 PKNN 查找方法, 当 K=1 时, 即降低为 PNN 查找. 而对于在不确定性时间序列数据库中, 对不确定性时间序列 Q 的 PNN 查询以及 PKNN 查找的问题, 作为后续的进一步的研究.

参考文献

1 Chen Q, Chen L, Lian X, Liu Y, Yu JX. Indexable PLA for efficient similarity search. Proc. of the 33st International Conference on Very Large Data Bases(VLDB). Vienna, Austria. 2007

(下转第 124 页)

根据测试结果我们对事务数为 1000 的 Apriori 算法和改进后的 Apriori 算法数据进行分析，如图 3、图 4 所示.

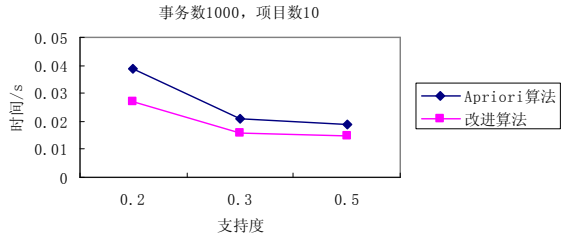

图 3　不同支持度下两种算法执行时间

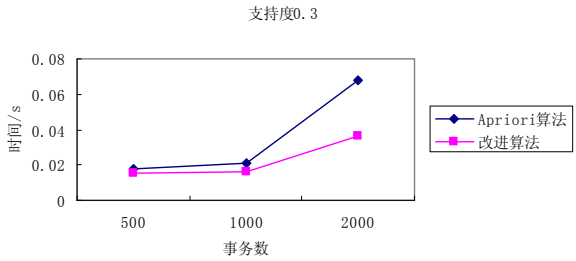

图 4　不同事务数下两种算法执行时间

从上图可以看出，当最小支持度比较小时，改进的算法所需的时间比 Apriori 算法少很多，当最小支持度逐渐增大时，两种算法执行时间比较接近；当事务数量不断增加时，改进的算法所需的时间比 Apriori 算法少很多，但在事务数量比较少时，改进的算法不具备明显优势. 这是因为改进的算法主要是针对减少候选项的数量和跳过对应频繁项集产生无贡献的记录的考虑而获得性能改进的.

4　结语

本文在深入研究关联规则算法的基础上，针对 Apriori 算法可能产生大量的候选集和可能需要重复扫描数据库的问题，提出了一种基于布尔矩阵的改进算法，实验表明，改进算法适合在大规模的事务数据库中挖掘关联规则，对于产生大量候选项集的情况下具有较高的挖掘效率.

参考文献

1 郭云峰,张集祥.对关联规则挖掘中 Apriori 算法的一种改进.杭州电子科技大学学报,2009,4:60–63.

2 Han JW, Kamber Mi. 范明,孟小峰译.数据挖掘概念与技术. 北京:机械工业出版社,2006.

3 周怡,王世伟.医学数据挖掘－SQL Server 2005 案例分析. 北京:中国铁道出版社,2008.

4 袁剑,王文海.挖掘关联规则 Apriori 算法的一种改进.青岛科技大学学报,2008,10:448–451.

(上接第 141 页)

2 Chen L, Ozsu MT, Oria V. Robust and fast similarity search for moving object trajectories. Proc. of ACM SIGMOD Int. Conf. on Management of Data. 2005.

3 Cranor CD, Johnson T, spatscheck O, Gigascope: A stream database for network applications. Proc. of ACM SIGMOD Int. Conf. on Management of Data. 2003.

4 Xue W, Luo Q, Chen L, Liu Y. Contour map matching for event detection in sensor networks. Proc. of ACM SIGMOD Int. Conf. on Management of Data. 2006.

5 Srikant R, Agrawal R. Mining generalized association rules. Proc. of the 21st VLDB Conf. 1995: 409-419.

6 Lian X, Chen L. Monochromatic and Bichromatic Reverse Skyline Search over Uncertain Database. Proc. ACM SIGMOD Int. Conf. on Management of Data. Vancouver, Canada. 2008. 213–226.

7 Agrawal R, Faloutsos C, Swami AN. Efficient similarity search in sequence databases. FODO, 1993.

8 Morinaka Y, Yoshikawa M, Amagasa T, Uemura S. The L-index: An indexing structure for efficient subsequence matching in time sequence databases. PAKDD. 2001.

9 Cai Y, Ng R. Indexing spatio-temporal trajectories with Chebyshev polynomials. SIGMOD. 2004.