一种时间序列相似搜索中提前终止效率的估算方法

李俊奎 王元珍 李海波 左 琼
(华中科技大学数据库与多媒体研究所 武汉 430074)

摘 要 提前终止(Early Abandon)是在受限的相似搜索中的一项技术,在提高时间序列相似搜索的效率,减少冗余计算中取得成功应用。但是以往的工作中提前终止的效率往往都只是通过大量的实验测试来体现,而缺少一种理论化的方法。从理论上提出了一种对提前终止技术的实际效率的估算方法,采用统计概率的方式分析了提前终止技术在时间序列相似搜索中的效率,同时对理论结果进行了实验验证。实验结果表明,理论上的估计方法在一定程度上可以估算出提前终止的效率,为时间序列相似搜索的实际效率计算提供了理论工具。

关键词 时间序列,相似搜索,提前终止,概率

Estimate on the Effects of Early Abandon Technique in Time Series Similarity Search

LI Jun-kui WANG Yuan-zhen LI Hai-bo ZUO Qiong
(Research Institute of Database & Multimedia, Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract Early abandon is one of the techniques in the constrained similarity search, and has found great success in accelerating time series similarity search, as well as reducing the redundant computations. However, previous works on early abandon were focused on the empirical experimental demonstrations on the effects of the technique, while no theoretical analysis is available. A theoretical estimate method on the effects of early abandon was proposed, which adopts the statistical analysis in the process. Substantial experiments were performed to evaluate the results of the estimate. The experimental results show that the estimate can get the value of the effects in most cases, and can be applied in the real efficiency calculation of time series similarity search.

Keywords Time series, Similarity search, Early abandon, Probability

1 引言

提前终止(Early Abandon)是在受限的相似性搜索中采用的一项技术,其递进地计算距离,同时不断与给定差异阈值进行比较。一旦发现计算的距离已经超过差异阈值,则断定最终的距离将超过差异阈值,此时停止其余的计算,从而节省计算资源,提高计算效率。

提前终止技术在时间序列相似搜索中被用于缩减比对次数,尽快终止在不符合条件的搜索序列上的计算,从而在计算层面上提高搜索效率。

文献[1]在讨论顺序扫描算法时,指出顺序扫描时在每个迭代后都将得到的部分距离与阈值进行提前终止的比较,与普通的完全计算距离后与阈值进行比较的方法相比,可以减少计算的时间。

文献[5]采用提前终止技术来消除在计算序列的欧拉距离(Euclidean Distance)时的冗余计算,效率有明显增长。

文献[2]从欧拉距离的提前终止过渡到动态时间弯曲距离(DTW,Dynamic Time Warping)的提前终止。文献[4]中从DTW弯曲矩阵的性质出发,提出了精确DTW下的提前终止方法EA_DTW。

但是以前有关提前终止的工作对于提前终止的效率都只是从实验角度进行分析,为测试实际的提前终止技术在时间序列搜索中的效率需要大量的实验测试和实验分析,而缺乏一种理论化的分析工具。本文则提出运用统计的方法对提前终止技术进行分析,主要解决提前终止技术能够节约多少计算量的问题,从而为评估提前终止的效率提供参考。

本文的其余部分如下组织:第2节给出相关背景介绍;第3节给出提前终止效率估算;第4节对理论分析结果进行实验研究,比较理论分析和实际的结果;最后总结全文。

2 相关背景

为说明严谨,我们首先给出本文的相关定义。

2.1 相关定义

定义1(时间序列) 一组按照时间顺序观测得到的值,记为 $T=t_1,t_2,\cdots,t_n$,其中 $|T|=n$ 为其长度。

定义2(受限的相似性搜索) 给定搜索时间序列 $Q=q_1,q_2,\cdots,q_m$,距离度量 D,距离差异阈值 $\varepsilon(\varepsilon>0)$,若序列 $C=c_1,c_2,\cdots,c_v$ 满足 $D(Q,C)\leqslant\varepsilon$,则称序列 C 为结果序列。由于需

到稿日期:2008-01-29 本文受国家发展与改革委员会"安全智能数据整合平台开发及产业化"项目(项目编号[2005]538号)资助。
李俊奎 博士生,CCF会员,主要研究方向为数据挖掘、机器学习,E-mail:jkltk2000@126.com;**王元珍** 教授,博士生导师,主要研究方向为现代数据库理论与实现技术、数据挖掘中间件技术;**李海波** 讲师,博士生,主要研究方向为数据挖掘;**左 琼** 讲师,博士生,主要研究方向为多媒体数据库建模分析。

要给定阈值 ε，寻找结果序列的过程称为受限的相似性搜索。

提前终止是在受限的相似性搜索中使用的一项技术，它主要作用在 $D(Q,C)\leqslant\varepsilon$ 的计算过程中。

定义 3(有效计算路径) 在计算 $D(Q,C)$ 过程中，若点对距离 $d(q_i,c_j)$ 包含在最终的 $D(Q,C)$ 中，则点对 (q_i,c_j) 对最终的 $D(Q,C)$ 大小有贡献。两序列间所有点对按时间顺序排列形成的一条路径称为有效计算路径，记为 $s_1,s_2,\cdots,s_n$，有效计算路径上的点对距离记为 $d_1,d_2,\cdots,d_n$。

由定义 3，可得

$$D(Q,C)=\sqrt{\sum_{i=1}^{n}d_i^2} \tag{1}$$

在欧拉距离计算中，由于 Q,C 之间的所有点对都对 $D(Q,C)$ 的大小有贡献，故欧拉距离的有效计算路径即为两序列之间的点对路径。但是在 DTW 距离中，只有在最小弯曲路径上的点对才对最终的 $DTW(Q,C)$ 有贡献，DTW 计算中的有效计算路径仅是两序列间的最小弯曲路径。关于 DTW 的计算和弯曲路径请参见文献[3,4]。

定义 4(有效计算路径溢出) 在求解有效计算路径的过程中，如果 $\sum_{i=1}^{k}d_i^2>\varepsilon^2(1\leqslant k\leqslant n)$，则称有效计算路径在 k 处发生溢出。

定义 5(提前终止的效率) 若进行提前终止技术前后的计算方格数分别为 c_n 和 c_r，则提前终止省却了 $c_\Delta=c_n-c_r$ 的计算量，效率为 $eff=c_\Delta/c_n$。

eff 越大，表明省却的计算量相对于总的计算量越大，提前终止越有效。

2.2 提前终止原理

提前终止在有效计算路径的计算过程中，都会检查有效计算路径是否溢出，一旦溢出，则说明有效计算路径上存在 k $(k<n)\sum_{i=1}^{k}d_i^2>\varepsilon^2$。由于距离的非负性，因此由式(1)可得：

$$D(Q,C)^2=\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{k}d_i^2+\sum_{i=k+1}^{n}d_i^2>\sum_{i=1}^{k}d_i^2>\varepsilon^2 \tag{2}$$

因此由受限的相似性搜索定义可得，序列 C 不可能是最终的结果序列，此时可以停止计算。

提前终止的原理说明如图 1 所示。在有效计算路径上发现溢出时(行 4)，则停止计算(行 5－6)。

```
1.  sum←0;overflow←0; // 初始化
2. for i=1 to n do // 有效计算路径
3.      sum+=d_i^2 // 计算部分和
4.      if sum>ε^2 then
5.          overflow←i; // 在 i 处溢出
6.          break;
7.      endif
8. endfor
9. if overflow>0 then // 发生溢出
10.     return false; // 返回不匹配
11. else
12.     return sum≤ε^2;
13. endif
```

图 1 提前终止的原理

提前终止的示意图如图 2 所示。从图上可以看出，在检测到发现溢出后，虽然得不到最终 Q,C 之间距离 $D(Q,C)$ 的具体数值，但是由于在受限的相似性搜索中，已经由式(2)确定后续的计算不会改变序列 C 不是结果序列的事实，因此后续的有效计算路径都不必计算，从而节省了计算资源，提高了处理效率。

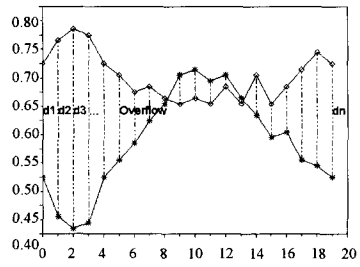

图 2 在有效计算路径上的提前终止

3 提前终止效率估算

在前面对提前终止的原理进行了说明。提前终止可以提高计算效率，同时这项技术也在时间序列搜索中得到了成功应用。但是以往的工作对其效率分析都是采用大量的实验测试得出，而缺乏一种理论的工具。下面我们对提前终止效率进行理论估算。

3.1 独立于距离的估算

令 Δ_i 为两序列的对应点对距离的平方，即

$$\Delta_i=d_i^2(i=1,2,\cdots,n) \tag{3}$$

由于序列点对随机分布，则 Δ_i i. i. d.(独立同分布)，并令

$$E(\Delta_i)=\mu \tag{4}$$

$$D(\Delta_i)=\delta^2\neq 0 \tag{5}$$

$$\text{设 } C_{\Delta k}=\sum_{i=1}^{k}\Delta_i(1\leqslant k\leqslant n) \tag{6}$$

定理 1 有效计算路径在 $k(1\leqslant k\leqslant n)$ 处溢出，当且仅当 $C_{\Delta k}>\varepsilon^2$。

证明：由 $C_{\Delta k}$ 的定义及溢出定义易得，证略。

推论 1 若有效计算路径在 $k(1\leqslant k\leqslant n)$ 处溢出，有 $C_{\Delta i}>\varepsilon^2(k\leqslant i\leqslant n)$。

证明：有效计算路径在 k 处溢出，则由定理 1，有 $C_{\Delta k}>\varepsilon^2$，同时由式(4)，有

$$C_{\Delta i}=C_{\Delta k}+\sum_{j=k+1}^{i}\Delta_j\geqslant C_{\Delta k}>\varepsilon^2(k\leqslant i\leqslant n)$$ 证毕。

推论 1 说明，若有效计算路径在某处溢出，则路径上后续段都将溢出。

推论 2 若有效计算路径在 k 处未溢出，则有 $C_{\Delta i}\leqslant\varepsilon^2(1\leqslant i\leqslant k)$。

证明：有效计算路径在 k 处未溢出，则由定理 1，有 $C_{\Delta k}\leqslant\varepsilon^2$，同时由式(4)，有

$$C_{\Delta i}=C_{\Delta k}-\sum_{j=i}^{k}\Delta_j\leqslant C_{\Delta k}\leqslant\varepsilon^2(1\leqslant i\leqslant k)$$ 证毕。

推论 2 说明，若有效计算路径在某处未溢出，则路径上前段都未溢出。

推论 3 若有效计算路径上存在溢出，则可以提前终止。

证明：设在 $k(1\leqslant k\leqslant n)$ 处溢出，由推论 1 可得 $C_{\Delta n}>\varepsilon^2$，由前面的提前终止原理部分式(2)，可提前终止。 证毕。

令 p_{ok} 为有效计算路径上在 k 处溢出的概率，则有

$$p_{ok}=\Pr\{C_{\Delta k}>\varepsilon^2\} \tag{7}$$

定理 2 $p_{ok}\approx 1-\phi\left(\frac{\varepsilon^2-k\mu}{\sqrt{k}\sigma}\right)$，其中 $\phi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}e^{\frac{-x^2}{2}}dx$。

证明：令 $f_k=\frac{C_{\Delta k}-E(C_{\Delta k})}{\sqrt{D(C_{\Delta k})}}$，则

$$p_{ok}=\Pr\{C_{\Delta k}>\varepsilon^2\}=\Pr\{\frac{C_{\Delta k}-E(C_{\Delta k})}{\sqrt{D(C_{\Delta k})}}>\frac{\varepsilon^2-E(C_{\Delta k})}{\sqrt{D(C_{\Delta k})}}\}=\Pr\{f_k>\frac{\varepsilon^2-E(C_{\Delta k})}{\sqrt{D(C_{\Delta k})}}\}$$

由式(4)，(5)，(6)可得

$E(C_{\Delta k})=k\mu, D(C_{\Delta k})=k\sigma^2$

又根据中心极限定理[6]，有 $f_k \doteq N(0,1)$，故

$$p_{ok}=1-\Pr\{f_k\leqslant\frac{\varepsilon^2-E(C_{\Delta k})}{\sqrt{D(C_{\Delta k})}}\}=1-\Pr\{f_k\leqslant\frac{\varepsilon^2-k\mu}{\sqrt{k}\sigma}\}\approx 1-\phi(\frac{\varepsilon^2-k\mu}{\sqrt{k}\sigma}) \quad (8)$$

定理 2 给出了有效计算路径上在 k 处溢出的概率的估算方法。根据推论 3，有效计算路径上有溢出，则可提前终止。又根据提前终止原理，由于是递进式计算有效计算路径，因此一旦发现溢出则立刻停止。

提前终止获益为 k(省略计算 k 步)的概率为

$$p_{savek}=p_{o(n-k)}\approx 1-\phi(\frac{\varepsilon^2-(n-k)\mu}{\sqrt{n-k}\delta}) \quad (9)$$

由于已经得到了提前终止获益概率，因此可以计算提前终止获益的期望：

$$E(x)=\sum_{k=1}^{n} k p_{savek} \quad (10)$$

于是提前终止的效率可以估算为

$$eff=\frac{\sum_{k=1}^{n} k p_{savek}}{n} \quad (11)$$

其中 p_{savek} 用式(9)进行估算。

3.2 DTW 距离下的估算

由于欧拉距离的点对之间的路径即为有效计算路径，因此欧拉距离下的估算可以使用前面的估算结论，所作的修改仅是在式(3)中对应点对的距离是欧拉距离计算中逐一点对的距离。

DTW 距离下有效计算路径(最小弯曲路径)包含在弯曲路径的计算中。虽然前面的结论适用在最小弯曲路径上的估算，但是如图 3 所示，最小弯曲路径中仅占弯曲矩阵中的一部分，如果仅对有效计算路径上的效率进行估算，则似乎会漏掉许多的未估算模块(实际并不是如此，后面会解释)。下面对所有计算路径部分进行估算。

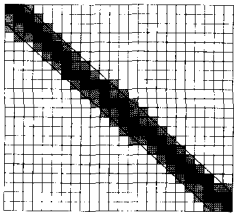

图 3 动态弯曲距离中有效计算路径(Sakoe-Chiba Band 限制)

设动态弯曲距离计算的两序列分别为 $U=u_1,u_2,\cdots,u_m$ 和 $V=v_1,v_2,\cdots,v_v$，最小弯曲路径的长度 n 满足：

$$\max(m,v)\leqslant n\leqslant m+v \quad (12)$$

我们分如下情形来进行估算：

· 不施加限制的动态弯曲距离效率估算

在原始的动态时间弯曲距离计算中，有两种实现方式：一种是构建大小为 $m\times v$ 的弯曲矩阵 M，按顺序计算矩阵中的累计距离和，最后 $M[m][v]$即为两序列间的动态时间弯曲距离；另一种是不构造弯曲矩阵，而是采用两个与查询序列相同大小的向量，分别记录当前时间点和前一时间点的累计距离和，计算时循环更新两向量。这两种方式的区别在于构建弯曲矩阵的方式在得到距离的同时还可以重建出最小弯曲路径，而向量的方式仅可以得到距离。但是它们的计算量相同，都需要计算 $m\times v$ 的方格。我们假设每次从 U 方向进行填充，计算方格占整个填充矩阵的面积，因此根据提前终止效率的定义，有

$$eff=\frac{m\times v-v\times\sum_{i=1}^{n} m\times k p_{ok}/n}{m\times v}=1-\frac{\sum_{i=1}^{n} k p_{ok}}{n} \quad (13)$$

其中 n 为最小弯曲路径长度，p_{ok} 用式(8)进行估算。

· 施加限制的动态弯曲距离效率估算

动态弯曲距离的计算一般会对弯曲路径施加全局限制，最为常用的限制是 Sakoe-Chiba Band 和 Itakura Parallelogram 限制。本文主要考虑 Sakoe-Chiba Band，如图 3 所示。它将弯曲路径的弯曲程度限制在大小为 $w(w>0)$的窗口内，因此需要计算的方格数约为 $\max(m,v)\times w$，同样计算提前终止计算方块占总计算方块(限制范围内的方块)的面积，按照提前终止的效率估算为：

$$eff=\frac{\max(m,v)\times w-\max(m,v)\times\sum_{i=0}^{n} w k p_{ok}/n}{\max(m,v)\times w}=1-\frac{\sum_{i=0}^{n} k p_{ok}}{n} \quad (14)$$

其中 n 为最小弯曲路径长度，p_{ok} 用式(8)进行估算。

从式(13)和(14)可以看出，对于提前终止的效率估算并不会随着 DTW 的实现不同而有改变，因此我们有定理 3。

定理 3 提前终止的效率估算与 DTW 距离的实现方式相互独立。

这是因为在不同的 DTW 实现方式下都需要计算最小弯曲路径，即使添加限制也是对最小弯曲路径的弯曲程度的限制，而提前终止作用在有效计算路径上(DTW 即为最小弯曲路径上)，因此不同的实现方式对提前终止的效率并没有影响。值得指出的是，这对于实际 DTW 的计算却有十分明显的影响。从图 3 中可以看出，施加限制后所计算的方格数有明显减少。另外，注意到

$$\sum_{k=1}^{n} k p_{savek}=\sum_{k=1}^{n} k p_{o(n-k)}=\sum_{k=1}^{n}(n-k)p_{ok}=n^2-\sum_{k=1}^{n} k p_{ok}$$

因此式(13)和式(11)等价，于是有定理 4。

定理 4 DTW 下的提前终止效率估算也可以采用前面在有效计算路径上的估计结论。

4 实验研究

4.1 实验准备

这一节对理论分析结果进行实验研究，通过在实际操作上的结果与理论分析的结果进行对比来确定理论分析结果的有效性。

使用 C++实现了欧拉距离和动态弯曲距离下的提前终止算法，实验环境如表 1 所示。

表 1 实验环境配置

配置项目	项目值
CPU	Intel Pentium 3-866

操作系统	Ubuntu Linux 4.1.1.13(内核版本:2.6.2)
RAM	256M
磁盘	40G
编译环境	GNU g++ 4.1.2.20060928
数据处理	GNU R2.6.1

我们使用 eff_{est} 和 eff_{real} 分别表示提前终止理论估算的效率和实际的效率,并规定 eff_{est} 落入区间$[(1-\alpha)eff_{real},(1+\alpha)eff_{real}]$为一次有效估算(实验中 $\alpha=0.2$),估算准确率如下计算:

$$\text{估算准确率}=\frac{\text{有效估算数}}{\text{总估算数}}\times 100\% \qquad (15)$$

实验数据来自程序产生的随机数发生器,它能够产生满足正态、普通和指数分布的随机序列。在这 3 种情形下分别产生了 50 条长度为 300 的随机序列。随机选取其中 1 条序列作为查询序列,其余序列作为候选序列,估算在候选序列上进行查询时的提前终止效率。

估算时式(4),(5)中点对距离平方的期望 μ 和方差 δ^2 的值通过计算前面 20 个点对的距离平方的期望和方差近似得出。动态弯曲距离的最小弯曲路径长度 n 由式(15)给出范围,估算时采用均值,即 $n=\frac{(m+v+\max(m,v))}{2}$。由于 $\phi(.)$ 函数的值已经制成表格[6],按式(9)计算出参数后,查 $\phi(.)$ 函数表得出具体数值。

4.2 实验结果及分析

(1)欧拉距离下的提前终止估算

首先是在欧拉距离下的提前终止估算。估算式采用式(11),其中 $n=300,\varepsilon=85$,实验结果分别如图 4,5 和 6 所示。

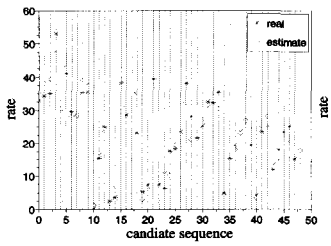

图 4 满足正态分布的随机序列集上的欧拉距离估算结果对比(估算准确率 71.4%)

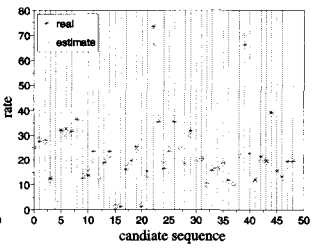

图 5 满足普通分布的随机序列集上的欧拉距离估算结果对比(估算准确率 91.8%)

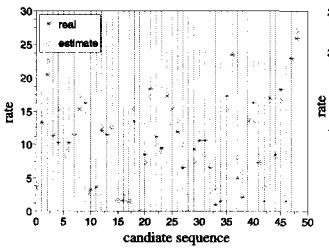

图 6 满足指数分布的随机序列集上的欧拉距离估算结果对比(估算准确率 77.6%)

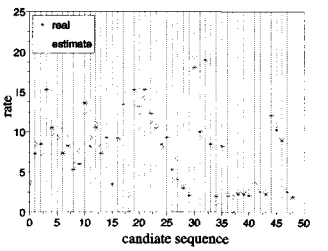

图 7 满足正态分布的随机序列集上的 DTW 距离估算结果对比(估算准确率 77.6%)

(2)动态弯曲距离下的提前终止估算

下面是对动态弯曲距离下的提前终止估算。估算式采用式(13),其中 $n=450,\varepsilon=60$。实验结果分别如图 7,8 和 9 所示。

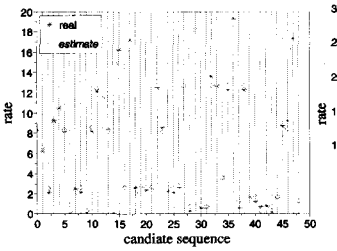

图 8 满足普通分布的随机序列集上的 DTW 距离估算结果对比(估算准确率 73.5%)

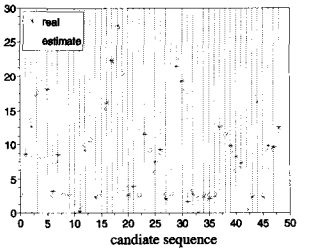

图 9 满足指数分布的随机序列集上的 DTW 距离估算结果对比(估算准确率 81.6%)

(3)实验结果分析

从实验结果可以看出,与实际的计算相比,理论估算可以取得 70%以上的准确率,在图 5 中还取得了 45 次有效计算、91.8%的准确率。说明理论估算在一定程度上是有效的。在实验中期望和方差是通过计算前 20 个点的期望和方差得到,这样即使不完全计算,依靠理论估算也可以辅助评估效率,节省了计算。

对于在不同距离上的估计没有明显的差异性,说明理论估算是可以独立于距离的。这是由于只要限定了有效计算路径的范围,理论估算可以根据范围确定一个大致的值,帮助用户确定计算中的有效值。

实验中另外一个值得关注的地方是在实际计算中提前终止效率为 0 的情形。比如在图 5 中的对 11 号序列的计算,此时没有提前终止,而理论估算值为 1.2,基本接近实际值,说明理论估算对于没有提前终止的情形也能够给出一定的指示。

结束语 本文根据以前工作中对时间序列搜索中使用的提前终止技术缺乏理论工具的事实,提出了一种以概率统计为分析手段的估算方法,给出了估算计算式及推导,并在实验中验证了其实际效果。这种理论估算的意义在于通过少部分的计算,就可以估算出实际的效果,从而减少盲目计算量。

参 考 文 献

[1] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration [C]// Proc. of SIGKDD. Edmonton, Alberta, Canada, 2002: 102-111

[2] Keogh E, Li Wei, Xi Xiaopeng, et al. LB_Keogh Supports Exact Indexing of Shapes Under Rotation Invariance with Arbitrary Representations and Distance Measures[C]// Proc. of 32nd Very Large Databases Conf. (VLDB). Seoul, Korea, 2006

[3] Keogh E. Exact Indexing of dynamic time warping // Proc. of 28th International Conference on Very Large Data Bases (VLDB). Hong Kong, China, 2002: 406-417

[4] Li Junkui, Wang Yuanzhen. EA_DTW: Early Abandon to Accelerate Exactly Warping Matching of Time Series[C]// Proc. of Int'l Conf. on Intelligent Systems and Knowledge Engineering (ISKE). 2007

[5] Li Wei, Keogh E, Van H H, et al. Atomic Wedgie: Efficient Query Filtering for Streaming Time Series[C]// Proc. of 5th IEEE Int'l Conf. on Data Mining (ICDM). Houston, Texax, 2005: 490-497

[6] 盛骤,谢式千,潘承毅. 概率论与数理统计. 第 2 版. 北京:高等教育出版社,1997