# 动态规划字符串匹配算法在曲线对比中的应用

魏 莲,吴信才
WEI Lian,WU Xin-cai

中国地质大学 信息工程学院,武汉 430074
Faculty of Information Engineering,China University of Geosciences,Wuhan 430074,China
E-mail:wl9926@sohu.com

**WEI Lian,WU Xin-cai.Application of string matching in curve comparison by dynamic programming algorithm.Computer Engineering and Applications,2007,43(8):8-9.**

**Abstract**: Curve Comparison is a foundation correlation method in formation correlation.Proposes a comparison method based on string match,it symbols the log curve by segregating the geology occurrence,and calculates the longest common sequence by dynamic programming,and applies the string similarity searching in curve Comparison.Algorithm allows skipping over characters when matching,thus permits a lack and dissimilar between the matching sequences.It realizes a disconnected match in curve comparison,and applies to strata repeat or lack when formation correlation.

**Key words**: curve comparison;string match;formation correlation;dynamic programming

**摘 要**:曲线对比是地层对比的基础手段。提出了一种基于字符串的曲线对比方法,通过对地质事件的识别来符号化测井曲线,采用动态规划方法计算二个序列的最长公共子序列,将字符序列的相似性计算用于曲线对比中;算法允许在匹配过程跳过一定的字符数,实现了曲线的不连续对比。算法对于地层重复与缺失状况下的地层对比具有很好的适用性。

**关键词**:曲线对比;字符串;地层对比;动态规划
**文章编号**:1002-8331(2007)08-0008-02 **文献标识码**:A **中图分类号**:TP631.8

曲线对比是地层对比常用的方法。曲线是在井中连续采样获得的,采样值是地层物理性质的反映,曲线变化反映出地层变化情况。

在曲线相似性对比研究中,有基于距离表示的算法,有基于形态表示的算法。基于距离的度量适于井间曲线幅值差别小的地层,曲线测量由于仪器刻度以及噪声等的影响,会存在刻度不一致、曲线波动、数值偏移等情况,对匹配结果造成一定的影响;基于形态的方法通过从曲线中提取一系列的特征参数作为对比因子,这种方法有较强的专业性,特征参数要选择适当,主要的影响因素有采样间距、地层重复与缺失等。

本文提出的符号化的曲线对比算法是一种基于形态表示的方法,将连续的数值曲线转换为字符序列,简化了特征参数的提取过程,通过动态规划法计算曲线的相似程度,能有效消除采样间距、地层重复等因素的影响。

## 1 曲线分段与符号化

数值形式表示的曲线不便于对比的描述。为此,本文提出将数值曲线序列转换成离散的、相对抽象的符号序列,每一种符号代表一种基本的、相对独立的变化趋势,这些符号构成了曲线相似性计算的基本元素。

### 1.1 基于重要点的分段

对曲线分段的方法有多种,有基于距离的分段方法,有移动窗口法,基于重要点的分段方法认为曲线变化是由一系列的事件引起的,如在地质过程中,水进水退会引起曲线幅度的变化,按事件发生的重要程度对曲线进行划分。

重要极小/(大)点定义:给定常量 $R$ 和测井序列$\{x_1,\cdots,x_n\}$,对于 $1<m<n$,存在下标 $i$ 和 $j$,且 $1\leqslant i<m<j\leqslant n$,有 $x_m$ 是 $x_i,\cdots,x_j$ 中的最小/(大)值;$x_i/x_m\geqslant R$ 且 $x_j/x_m\geqslant R$ 成立。其中:$R$ 是可控选取的参数,$R$ 值大则被选中的相对重要点少,曲线分段就越粗;反之,分段就越细。因此,通过选择 $R$,可以在不同的精细程度上进行曲线分段。基于重要点的分段方法在一定程度消除了仪器、背景等噪声的影响,保持了序列变化的主要特征模式[1]。如图 1 所示。

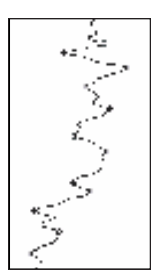

图 1 基于重要点的曲线分段

**基金项目**:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2003AA133010)。
**作者简介**:魏莲(1970-),女,博士研究生,讲师,主要研究方向:GIS 空间分析与应用;吴信才(1952-),男,教授,博士生导师,教育部 GIS 工程研究中心主任,长江学者首批特聘教授,目前主要从事地理信息系统软件的研究和开发工作。

### 1.2　曲线序列符号化

采用重要点分段方法，得到一个由子段组成的序列集合，该序列代表了曲线变化的主要趋势。分段曲线的形态特征主要由斜率 $K_i$ 及在垂向上的厚度决定。本文将斜率作为符号化的主要参数，将厚度作为序列相似性计算的影响因素。在利用斜率对曲线符号化的过程中，有必要对曲线采样值进行预处理，以消除由于仪器刻度差异造成的对计算结果的影响。

$$K_i=\frac{v_{i+1}-v_i}{H_{i+1}-H_i}$$

其中：$K_i$ 为第 $i$ 个分段曲线的斜率；为分段序列第 $i$ 个特征点的采样值；$H_i$ 为分段序列第 $i$ 个特征点的深度值。

对分段序列分别计算斜率，设立阈值将其转换成字符序列，这样，连续的、以数值形式表示的时间序列转换成离散的符号序列，每个符号代表一种基本的、相对独立的变化模式。这个符号序列是定义曲线相似性搜索的模式空间。

## 2　基于字符串的匹配算法

经过计算后，连续曲线转换为离散的、以字符表示的序列，该序列中蕴含了由于地质事件引起的地层变化。这里通过分析二个字符序列的相似度来判断二条曲线的是否匹配。

### 2.1　字符串完全匹配法

在计算字符匹配时，通常的算法是采用一种完全匹配法，若 $S$='**ABCD**'，$T$='**XYABCDZ**'，匹配的结果就是找出了 $T$ 中的一个 **ABCD**，使得 $S$ 中的'**ABCD**'能与 $T$ 中的'**ABCD**'完全匹配，也就是说在 $S$ 中有 100%的字符与 $T$ 中的匹配，而在 $T$ 中有 57%的字符与 $S$ 中的匹配。

但如果字符串中掺杂了一些其它的字符，如：$S$='**ABCDE**'，$T$='**ATXBYCDZ**'，则不存在这样的字串既在 $S$ 中存在，又在 $T$ 中存在，其结果是不匹配。而实际上，在这两个序列中，会发现都存在'**A**'，'**B**'，'**C**'，'**D**'字符，且具有顺序性，二者也是具有一定的相似性的，如图 2。本文采用动态规划方法计算二者的相似程度。

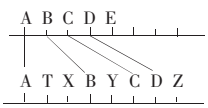


图 2　字符不连续匹配情况

### 2.2　动态规划法求最长公共子序列

子序列概念：一个给定序列的子序列是在该序列中删去若干元素后得到的序列。对于序列 $X$ 和 $Y$，当另一序列 $Z$ 既是 $X$ 的子序列又是 $Y$ 的子序列时，称 $Z$ 是序列 $X$ 和 $Y$ 的公共子序列[2]。

计算两个序列的最长公共子序列，根据子序列的长度来判断二者是否相似。

$X=\{x_1,x_2,\cdots,x_m\}$ 和 $Y=\{y_1,y_2,\cdots,y_n\}$ 的最长公共子序列，可按以下方式递归地计算：当 $x_m=y_n$ 时，二者的最长公共子序列为 $x_{m-1}$ 和 $y_{n-1}$ 的最长公共子序列，在序列尾部加上 $x_m$ 即可；当 $x_m\neq y_n$ 时，需要找出 $M_{m-1}$ 和 $Y$ 的一个最长公共子序列及 $X$ 和 $Y_{n-1}$ 的一个最长公共子序列，二者中长度大者即为 $X$ 和 $Y$ 的最长公共子序列。

$$L[i][j]=\begin{cases}0 & i=0,j=0\\ L\ [i-1][j-1]+1 & i,j>0;x_1=y_1\\ Max\{L[i][j-1],L[i-1][j]\} & i,j>0;x_1\neq y_1\end{cases}$$

定义 $X$ 与 $Y$ 的匹配度为 $\xi=\frac{L[i][j]}{m}$。其中：$L[i][j]$ 为 $X$ 与 $Y$ 的最长公共子序列长度；$m$ 为序列 $X$ 的字符个数。

为了提高字符匹配的精度，还可以设置下标 $i$ 与 $j$ 允许的距离，匹配时跳过不需匹配的字符[3]。它限定了两个匹配字符之间允许出现其它字符的数量，称之为模糊因子。当模糊因子等于 0 时，也就是前面所说的完全匹配。

该算法充分考虑到字符不连续匹配的情况，当两个序列具有足够长的相似子序列时，则认为此两序列是相似的[4]。这些子序列可以是不连续的、残缺的，但它们出现的顺序应保持一致，反映了地质事件的作用过程。

可以利用最长公共子序列来计算二个字符序列的相似度，在利用字符序列计算曲线是否相似时，还要考虑到厚度因素的影响。

$$\xi'=A*\xi$$

$$A=\frac{L_1+\cdots+L_i}{L}$$

其中：$A$ 为厚度权系数；$L_1,\cdots,L_i$ 分别为被匹配曲线中匹配字符所代表的地层厚度；$L$ 为被匹配曲线总地层厚度。

利用字符串相似来计算曲线是否匹配，该方法对于地层重复与缺失情况具有很好的适用性。

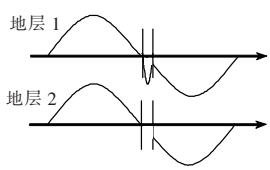


图 3　地层缺失示意图

## 3　曲线对比与分析

曲线对比是地层对比的基础手段。采用基于字符串的对比方法，曲线对比步骤如图 4 所示。

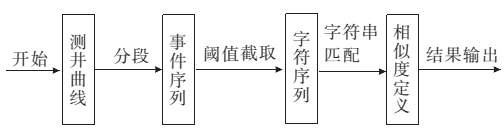


图 4　基于字符串的测井曲线对比流程图

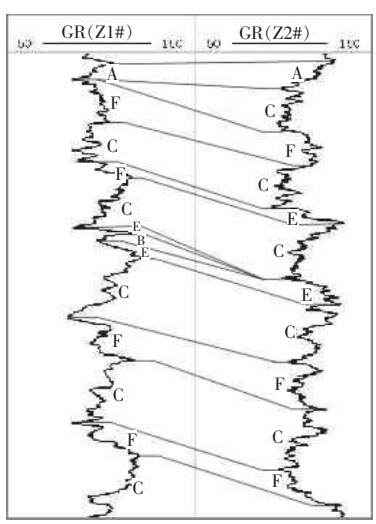


图 5　基于字符串的曲线匹配

（下转 17 页）