水文时间序列相似性挖掘系统

杨 敏 王志坚 尹燕敏 (南京河海大学计算机及信息工程学院 210098)

摘要：时间序列分析正成为数据挖掘研究的热点，本文讨论了时间序列相似性研究的现状和典型方法，介绍了水文时间序列相似性挖掘系统的设计与实现，详细分析了系统采用的相似性度量方法。

关键词：相似性 时间序列 多维时间序列 数据挖掘

1 引言

时间序列(time series)是指按时间顺序排列的观测值集合。按照研究的现象和问题的不同，可以得到各种时间序列。例如产品销售记录、股票价格数据、地区降雨量数据等，针对时间序列的数据挖掘研究从大量时间序列历史数据中发掘有价值信息的算法及实现技术，是一个新的、极具挑战性的研究领域。时间序列相似性挖掘的目的是在时间序列数据库中发现与给定序列模式相似的序列，在这方面前人已开展了大量工作。

时间序列相似性挖掘的典型算法是基于离散傅立叶变换(DFT)的相似性比较方法[1]。该方法通过DFT将时间序列从时域空间映射到频域空间，用R*-树作为其索引结构。Faloutsos等人又在此基础上提出了快速子序列匹配技术[2]，采用滑动窗口和MBR方法，ST-index作为索引和存储。这些方法在判断序列相似时采用了欧氏距离作为序列间的相似性评价函数，即两序列间欧氏距离小于给定阈值时，就认为序列相似。但实际中，由于序列长度不等或取样率不同等问题，使得欧氏距离难以直接应用。因而R.Agrawal[3]和Davood Rafilei[4]在各自工作中提出了序列规范变换和仿射安全变换的技术，允许序列比较前首先进行适当变换，并对各种变换实施顺序及其对效率的影响等问题进行了讨论。

上述基于DFT、欧氏距离和各种序列变换的相似挖掘方法具有较好的适应性和灵活性。但其局限性时间排序的数据元素组成，每个数据元素又是一个n维向量。目前关于多维时间序列相似性研究尚不多见。但在实际中，多维时间序列具有重要的意义，例如为了根据水雨情数据的历史观测值作出正确的预报，需要将水位、流量、降雨、气温等多个指标综合考虑，这就需要考虑多维时间序列。目前，多维时间序列研究，特别是基于多维序列数据的相似性挖掘研究是具有重要应用前景的研究方向。

水文数据库相似性挖掘系统是我们正在从事的水文数据挖掘HYDM(Hydrological Data Mining)研究工作的一部分。水文数据库中存储着大量时间序列数据(水位、流量、雨量等)，而相似性挖掘对洪水预报、防洪调度具有重要的现实意义。我们在分析现有的相似挖掘模型和算法的基础上，设计了针对水文时间序列数据的特点及要求的相似性挖掘系统。

2 水文相似性挖掘系统

经过对现有多种时间序列相似性挖掘算法的比较，依据水文时间序列的特点和要求，我们借鉴前人的研究成果并进行了改进，设计出水文相似性挖掘系统。系统主要基于欧氏距离度量，根据需要允许平移并引入移动平均值比较；处理较长序列时，采用子序列MBR方法，降低复杂度；特别提出了对多维时间序列进行相似性挖掘的方法。

2.1 水文时间序列

水文数据库中存储着大量时间序列数据，例如各类测站(河道站、闸坝站等)水文数据观测值，数据包含实时数据和历史数据。

以江苏省水文数据库为例。如河道水情数据：存储

各河道站记录的水文数据，主要信息包括站号、时间、水位、流量等。降雨量数据：存储各雨量站的降雨数据，主要信息包括站号、时间、降雨量等。

由于拍报级别不同，不同地区不同类型测站的报汛时间可能不同。但同一地区内各类测站均按照级别规定，有严格统一的时间间隔。如雨量站基本报汛时间为每日上午8时，加报除外。因而水文时间序列可以看作是离散、规则的时间序列。

2.2 系统功能与结构

水文时序数据相似性挖掘系统主要目的是指定一时间序列(称为查询序列)，在水文时间序列数据库的相应序列集上自动挖掘出与之具有相似模式的一个或多个时间序列。例如可以发现今年汛期某河道站的水情与历史上某年相似。从而辅助防汛部门做出相应的防洪决策。

按挖掘对象信息类别分，系统包括雨情、河道水情、闸坝水情、水库水情等多个功能模块；按时序数据的维数分，包括一维时间序列挖掘和多维时间序列挖掘。

例如，单站降雨量是一维时间序列，而同时考虑某时间段内多个站的降雨量就可看作是多维序列。仅考虑单站流量或单站水位是一维时间序列，但某些站既有流量又有水位，就是二维序列，当同时比较某时间段内n个站的流量时，就成了n维序列。在实际应用中，某地区内多站降雨、某河道上的多站流量或水位通常比单站数据更有意义，从而也更有相似性挖掘的必要。

水文时序数据相似性挖掘系统层次结构如图1所示。

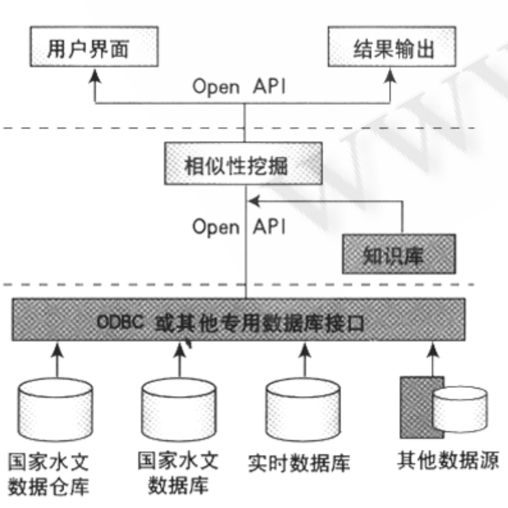

图 1

相似性挖掘主要步骤包括：

· 数据选择：相似性挖掘的对象可以来自水文数据库(包括实时数据库)，也可以来自水文数据仓库及其他数据源。需对选择的时间序列进行数据预处理，进行一致性检查、消除噪声干扰等。

· 相似挖掘：按照预先设计的相似性度量方法，进行挖掘。包括参数定义、序列分割、子序列MBR距离计算、相似程度判断等，是系统的核心。

· 结果表达：将挖掘得到的结果以可视化的形式呈现给用户，如果不能另用户满意，则需重复上面的步骤。

3 结束语

我们利用江苏省水文数据库的部分数据对挖掘系统进行了测试，不仅验证了许多已知的相似序列模式，更发现了许多新的相似性。

水文时间序列相似性挖掘系统中提出了较为可行的多维时间序列的相似性挖掘方法，挖掘算法基于MBR，并不需对序列上的每个点进行比较，降低了时间和空间复杂度，提高了效率。尚须努力的方面包括：不规则时间序列上的相似性挖掘；对噪声数据的有效处理；当多维时间序列维度较高时，需考虑采取降维技术。■

参考文献

1 R.Agrawal,C.Faloutsos,and A.Swami.Efficient similarity search in sequence databases,Proc.4[th] Int'1.Conf.Foundations of Data Organization and Algorithms.Oct.1993.69～84。

2 C.Faloutsos,M.Ranganathan,and Y.Manolopoulos.Fast subsequence matching in time-series databases.In Proc.1994 ACM-SIGMOD Int.Conf.Management of Data,May 1994.419～429。

3 R.Agrwal,K.-I.Lin,H.S.Sawhney,and K.Shim.Fast Similarity search in the presence of noise,casling,and translation in time-series databases.Proc.21[st]Int'1.Conf.Very Large Data Bases,Sept.1995.490～501。

4 Davood Rafiei,and Alberto O.Mendelzon.Querying Time Series Data Based on Similarity,IEEE Transactions on Knowledge and Data Engineering.VOL.12,No.5,Sept/Oct.2000.675～693。