多度量水文时间序列相似性分析

王继民[1,2],朱跃龙[1],李 薇[3],万定生[1,2],李士进[1]

(1.河海大学计算机与信息学院,江苏 南京 210098; 2.河海大学水资源高效利用与工程安全国家工程研究中心,江苏 南京 210098;3.水利部水文局,北京 100053)

摘 要:多度量组合可以提高相似性分析的准确性,基于该思想,提出多度量水文时间序列相似性分析方法,首先,使用多个单一相似度量分别计算相似时间子序列,然后,采用改进 BORDA 投票法对各度量分析得到的相似子序列进行组合和排序,得到最终的相似时间子序列。为了证明提出方法的可行性和有效性,以淮河流域王家坝水闸洪水过程相似性分析为例进行了验证。分析结果表明,基于改进 BORDA 投票法的多度量水文时间序列相似性度量方法可以提高相似查询的准确性。

关键词:时间序列;相似性分析;水文;BORDA 投票法;多度量

中图分类号:P333 **文献标识码**:A **文章编号**:1000-0852(2014)04-0015-06

1 研究背景

长期的水文观测和实践积累了大量水文数据,这些数据中蕴涵着自然界长期的演变规律和人类活动影响的信息。将信息技术和水文学有机结合,利用人工智能和数据挖掘技术从这些大量数据中发现隐含的未知水文知识已经成为一个重要研究领域。水文过程由于受相似的季节性气候因素以及其他随机因素影响而呈现相似性,水文过程相似性分析可以回答防汛指挥中经常会问到的“当前水文过程相当于历史上哪一时期的同类过程”等问题,因而在洪水预报、防洪调度等方面有着重要的现实意义。本文主要研究水文时间序列相似性,其目标是 k 个最近邻水文过程发现,特别是洪水过程的相似性。

时间序列数据挖掘研究主要包括时间序列特征提取、相似性搜索、预测、分类、聚类以及序列模式挖掘等,其中,相似性搜索是其他挖掘的基础,相似性搜索在 1993 年由 R.Agrawal[1]首次提出,在时间序列相似性搜索中,需解决的问题包括时间序列特征提取、时间序列索引以及相似度量等。针对相似度量,研究人员提出了各种度量方法,如欧氏距离及其基于 L_P 准则的变种[2]、动态时间弯曲(Dynamic Time Warping,DTW)[3]距离、编辑距离[4-5]以及模式距离[6]等。

在已有相似度量的基础上,众多学者结合水文时间序列的特点,进行了水文时间序列相似性的研究,李薇等[7]抽取时间序列的模式特征(包括长度和斜率),然后借鉴动态时间弯曲的思想定义序列之间的动态模式匹配(Dynamic Pattern Matching,DPM)距离。欧阳如琳等[8]采用 DTW 距离计算流域内多水文站之间的相似洪水过程,发现流域的洪水过程形态。朱跃龙等[9]提出基于语义的水文时间序列相似性度量,定义水文时间序列的上升、保持和下降等语义模式,在此基础上,定义序列的语义距离描述序列的相似度。针对多维水文时间序列相似性,李士进等[10]首先逐维进行一元时间序列相似性分析,然后采用 BORDA 投票法对各维的相似子序列进行组合和排序,获取多维相似子序列。

目前水文时间序列相似性搜索大多采用单一相似

收稿日期:2013-06-25
基金项目:国家自然科学基金项目(51079040);河海大学中央高校基本科研业务费(2009B22014);水利部 948项目(201016)
作者简介:王继民 (1976-),男,安徽全椒人,讲师,主要研究方向为智能信息处理和数据挖掘。E-mail: wangjimin@hhu.edu.cn

度量来评价序列之间的相似性，文献[11-12]提出多度量相似性分析，采用启发式搜索确定各度量的权重，相似距离为各度量距离的加权和，实验表明多度量可以提高相似性查询的准确率。本文基于多度量组合的思想，提出基于改进 BORDA 投票法的多度量水文时间序列 k 近邻查询方法，并以淮河流域王家坝水闸洪水过程相似性为例进行分析，验证了方法的可行性和有效性。

2 基于改进 BORDA 投票法的多度量水文时间序列相似性分析

2.1 多度量水文时间序列相似性查询模型

多度量水文时间序列相似性分析可以采取两种处理策略，一是采用串行处理方式，即前一个相似度量算法的处理结果作为后一个相似度量算法的输入，并进行相应处理，最后一个相似度量算法的输出作为最终的相似子序列，如图 1(a)所示。二是采用并行处理方式，即多个相似度量算法同时对时间序列进行相似性分析，并得到各自的相似子序列，然后采用组合方法将多个相似度量的结果进行组合，得到最终的相似分析结果，如图 1(b)所示。相对于寻找单个复杂相似性度量的方法，多度量组合有着很好的灵活性，可以充分利用每个单一相似性度量的特长，而且能根据特定应用需求进行灵活调整。

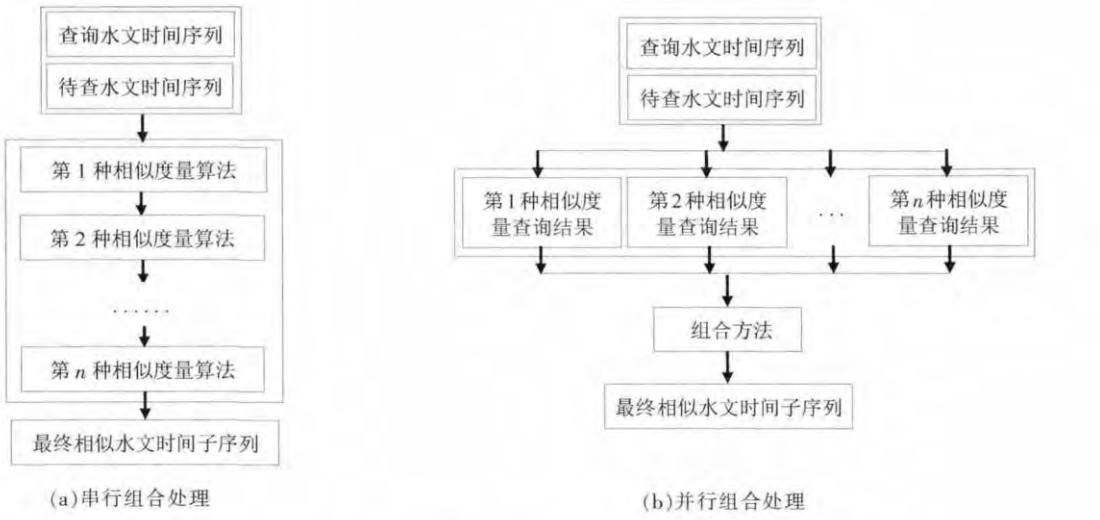


图 1 多度量水文时间序列相似模型

Fig.1 The model of multi-measure similarity searching of hydorlogic time series

本文关注 k 近邻搜索问题，即查询与指定序列最相似的前 k 个子序列。从分类角度来看，k 近邻相似搜索可以视为采用相似度量将时间子序列划分为第 1 相似子序列、第 2 相似子序列，…，第 k 相似子序列以及不相似子序列。采用多个单一相似度量进行相似搜索相当于采用多个分类器对时间序列进行分类，重点是如何将单一相似度量的相似结果组合得到最终的相似序列。本文采用改进的 BORDA 投票法[13]对多个单一相似度量算法计算得到的相似子序列进行并行组合，得到最终的 k 近邻子序列。

2.2 BORDA 投票法的改进

2.2.1 BORDA 投票法的缺点

假设 m 为胜者数，p 为候选人数，n 个投票人每人都通过由高到低的顺序对全体候选人进行排序来表示他的偏好。针对每个投票人的排序，给每个候选人设定一个排序分数，规定排在最后的候选人的排序分数为 1 分，倒数第 2 位的候选人为 2 分，依次类推，排在第 1 位为 p 分，候选人的排序分数的累积称为 BORDA 分数，BORDA 分数进入前 m 名的候选称作 BORDA 胜者。BORDA 投票法只考虑所有候选人排序先后，没有考虑前后候选人差距的大小，这样可能造成无法计算出真实的排序。如假设，有 A、B、C、D 四匹马进行四场比赛，四场比赛名次排序分别为：A、B、C、D，B、A、D、C，D、C、A、B 和 C、D、B、A。四匹赛马 BORDA 分数都为 10 分，出现了四匹赛马并列第一名的情况，因为 BORDA 投票法仅考虑四匹马的名次排序，而没有考虑各匹马在比赛时成绩的具体差异。针对该缺点，本文提出了改进的 BORDA 投票法，在计算 BORDA 分数时，利

用前后候选人之间的差距对最终的分数进行加权,从而较好的避免上面出现的问题。

2.2.2　改进 BORDA 投票法

假设 n 个投票人每人都通过由高到低的顺序对 p 个候选人进行排序来表示他的偏好,并且给出相邻两个候选人之间量化的差异大小。设某投票人按照由高到低对候选人的排序为 $c_1,c_2,\cdots,c_p$,相邻候选人之间差距记作为 $d_2,\cdots,d_i,\cdots,d_p$,其中,$d_i$ 表示 p_{i-1},p_i 之间的差距。规定排在第一位的候选人的改进排序分数为 p 分,排在最后一位的候选人的改进排序分数为 1 分,排在第i 位的候选人的改进排序分数采用公式(1)计算。

$$ss_i=ss_{i-1}-(p-1)\times\frac{d_i}{\sum_{j=2}^{p}d_j}\quad(i=2,\cdots,p-1)\qquad(1)$$

对候选人的改进排序分数进行累加得到该候选人的改进 BORDA 分数,对所有候选人按照改进 BORDA 分数进行排序,得到最终的排名,排名前 k 的序列为查询序列的 k 近邻。当 $d_2=d_3=\cdots=d_p$ 时,$ss_i=p-(i-1),(i=2,\cdots,p-1)$,即为传统的 BORDA 投票法。改进 BORDA 投票法只适合于前后候选人之间的差距可以量化的情况下使用。

2.3　基于改进 BORDA 投票法的多度量组合

在如图 1(b)所示的多度量水文时间序列相似性模型中,采用改进 BORDA 投票法作为多度量的组合方法,单一相似度量看作是投票人,其计算得到的前 k 个相似子序列作为投票人的一次投票结果,相似子序列与查询序列之间距离的差距作为各相似子序列之间的排序差距。水文时间序列经过特征提取后,采用 n 种单一相似度量方法分别计算相似子序列,然后采用改进 BORDA 投票法对相似子序列计算改进的 BORDA 分数,最后对相似子序列按照改进的 BORDA 分数排序,获得最终的相似水文时间子序列。

在各单一相似度量的查询结果中,出现次数越多的子序列,说明其被越多的相似度量认可为相似子序列,其改进 BORDA 分数将可能越高;在单一相似度量的查询结果中,某子序列排序靠前,但若其和前面子序列的排序差距较大,则其改进 BORDA 分数仍然可能会较低,从而可以过滤排序靠前或者出现次数较多但真实相似程度较差的子序列。相对于单一相似度量,多度量组合可以在结果中包含参与组合的多个单一相似度量结果中的优秀结果,从而提高相似搜索的准确性。

3　实验验证与分析

3.1　实验数据与方法

王家坝水闸是淮河流域重要的水利枢纽工程,王家坝水闸洪水预报对整个淮河流域的防洪调度起着举足轻重的作用,王家坝水闸洪水过程相似性分析是整个王家坝水闸洪水预报的基础核心部分,因此,对王家坝水闸洪水过程的相似性研究非常有实际意义。

实验数据为王家坝水闸 1998 年 6 月 1 日到 2009 年 7 月 12 日期间每年 6 月 1 日~9 月 30 日记录的流量数据,每天有 2:00、8:00、14:00、20:00 4 个监测时间点。基于特征点提取洪水时间序列的特征。分别选取“单洪峰倒 V 型”和“双洪峰 M 型”两种形态的洪水过程作为查询序列,剩下的时间序列作为被查序列,检索查询序列的 5 个最近邻序列。选择欧式距离、斜率距离以及 DTW 距离作为单一相似度量,分别计算相似时间子序列,利用传统 BORDA 投票法和改进 BORDA 投票法分别进行多度量组合,完成相似性分析,并对结果进行比较分析。

3.2　实验结果分析

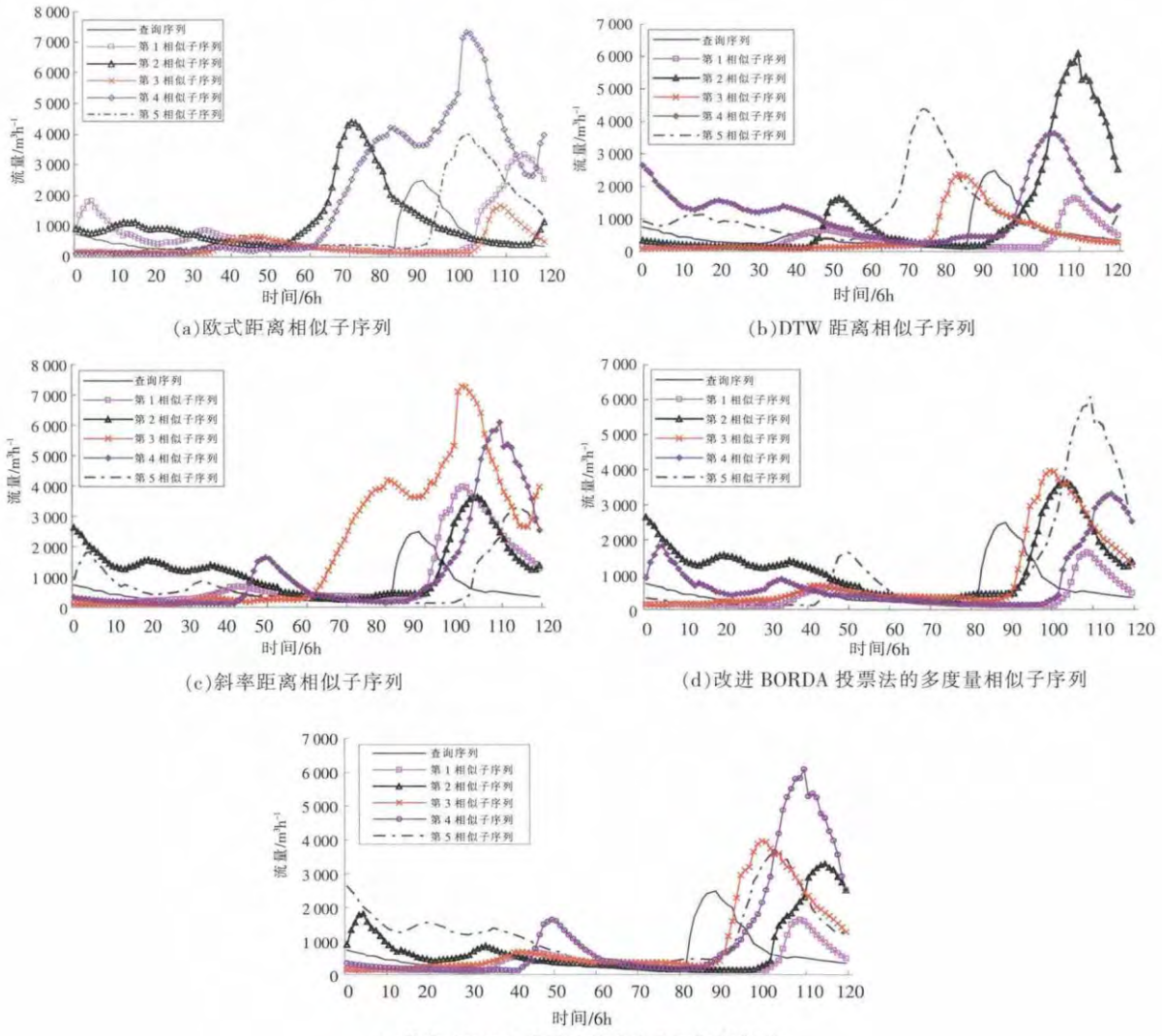
3.2.1　“单洪峰倒 V 型”洪水过程相似性分析

选取 2000.7.31 2:00~2000.8.29 20:00 期间的“单洪峰倒 V 型”洪水过程时间序列作为查询序列进行相似性分析,各单一相似度量以及多度量组合的结果见表 1,图 2 给出了相似子序列和查询序列的比较。

表1单洪峰洪水过程相似子序列

Table1　Similar subsequences of single-peak flood

改进 BORDA 投票法的多度量组合		传统 BORDA 投票法的多度量组合		欧式距离		DTW 距离		斜率距离	
序列起点	BORDA 分数	序列起点	BORDA 分数	序列起点	相似距离	序列起点	相似距离	序列起点	相似距离
2005.6.1 8:00	7.45	2005.6.1 8:00	8	2000.8.31 2:00	799.86	2005.6.1 8:00	465	2008.7.1 8:00	0.11
2005.7.31 2:00	6.01	2000.8.31 2:00	6	2008.8.1 8:00	848.37	2005.6.16 2:00	1830	2005.7.31 2:00	0.16
2008.7.1 8:00	6	2008.7.1 8:00	6	2005.6.1 8:00	944.75	2004.7.1 2:00	3730	2007.6.16 2:00	0.20
2000.8.31 2:00	6	2005.6.16 2:00	6	2007.6.16 2:00	971.13	2005.7.31 2:00	5230	2005.6.16 2:00	0.25
2005.6.16 2:00	5.90	2005.7.31 2:00	6	2008.7.1 8:00	1027.45	2008.8.1 8:00	7230	2000.8.31 2:00	0.28



(a)欧式距离相似子序列

(b)DTW 距离相似子序列

(c)斜率距离相似子序列

(d)改进 BORDA 投票法的多度量相似子序列

(e)传统 BORDA 投票法的多度量相似子序列

图 2 单洪峰洪水过程相似子序列

Fig.2 Similar subsequences of single-peak flood

表 1 中，改进 BORDA 投票法的多度量组合查询结果中的相似子序列是在各单一相似度量结果中多次出现而且排序靠前的子序列。起点为 2004.7.1 2:00 的子序列由于只在 DTW 距离度量结果中出现且排序差距大，造成改进 BORDA 分数低被淘汰，起点为 2007.6.16 2:00 和 2008.8.1 8:00 的子序列虽然分别在两种单一相似度量结果中出现，但其他多次出现的子序列在各自的排序中差距较小，改进 BORDA 分数提高，因此造成这两个序列在最终的相似排序中位置靠后而被淘汰。从图 2(d)看出，改进 BORDA 投票法的多度量组合集成了三种单一相似度量的优点，查询出的相似子序列的洪峰变换过程与查询序列几乎完全一致，被过滤的三个子序列和图 2(d)结果中的其他相似子序列相比，相似程度较弱。

基于传统 BORDA 投票法多度量组合和改进 BORDA 投票法多度量组合的查询结果中包含相同的子序列，但改进 BORDA 投票法多度量组合对部分序列给出了排序。

3.2.2 “双洪峰 M 型”洪水相似性分析

选取 2000.8.15 2:00~2000.9.13 20:00 期间的“双洪峰 M 型”洪水流量过程时间序列作为查询序列进行相似性分析，各单一相似度量以及多度量组合的结果见表 2，图 3 给出了相似子序列和查询序列的比较。

表2　双洪峰洪水过程相似子序列

Table2　Similar subsequences of double-peak flood

改进 BORDA 投票法的多度量组合		传统 BORDA 投票法的多度量组合		欧式距离列		DTW 距离		斜率距离	
序列起点	BORDA 得分	序列起点	BORDA 得分	序列起点	相似距离	序列起点	相似距离	序列起点	相似距离
2004.7.31 2:00	8.19	2004.7.31 2:00	8	2004.8.15 2:00	675.72	2004.7.16 2:00	830	2004.7.31 2:00	0.52
2004.7.16 2:00	6.36	2004.7.16 2:00	7	2007.7.16 2:00	943.29	2008.8.17 2:00	2260	2007.8.15 2:00	0.57
2008.8.17 2:00	6.18	2008.8.17 2:00	6	2003.7.1 2:00	997.07	2004.7.31 2:00	4460	2003.7.1 2:00	0.75
2004.8.15 2:00	6	2004.8.15 2:00	6	2004.7.16 2:00	1037.63	2005.7.1 2:00	7660	2008.8.17 2:00	0.80
2007.8.15 2:00	4.45	2003.7.1 2:00	6	2005.7.1 2:00	1073.35	2007.7.16 2:00	8860	2004.8.15 2:00	0.88

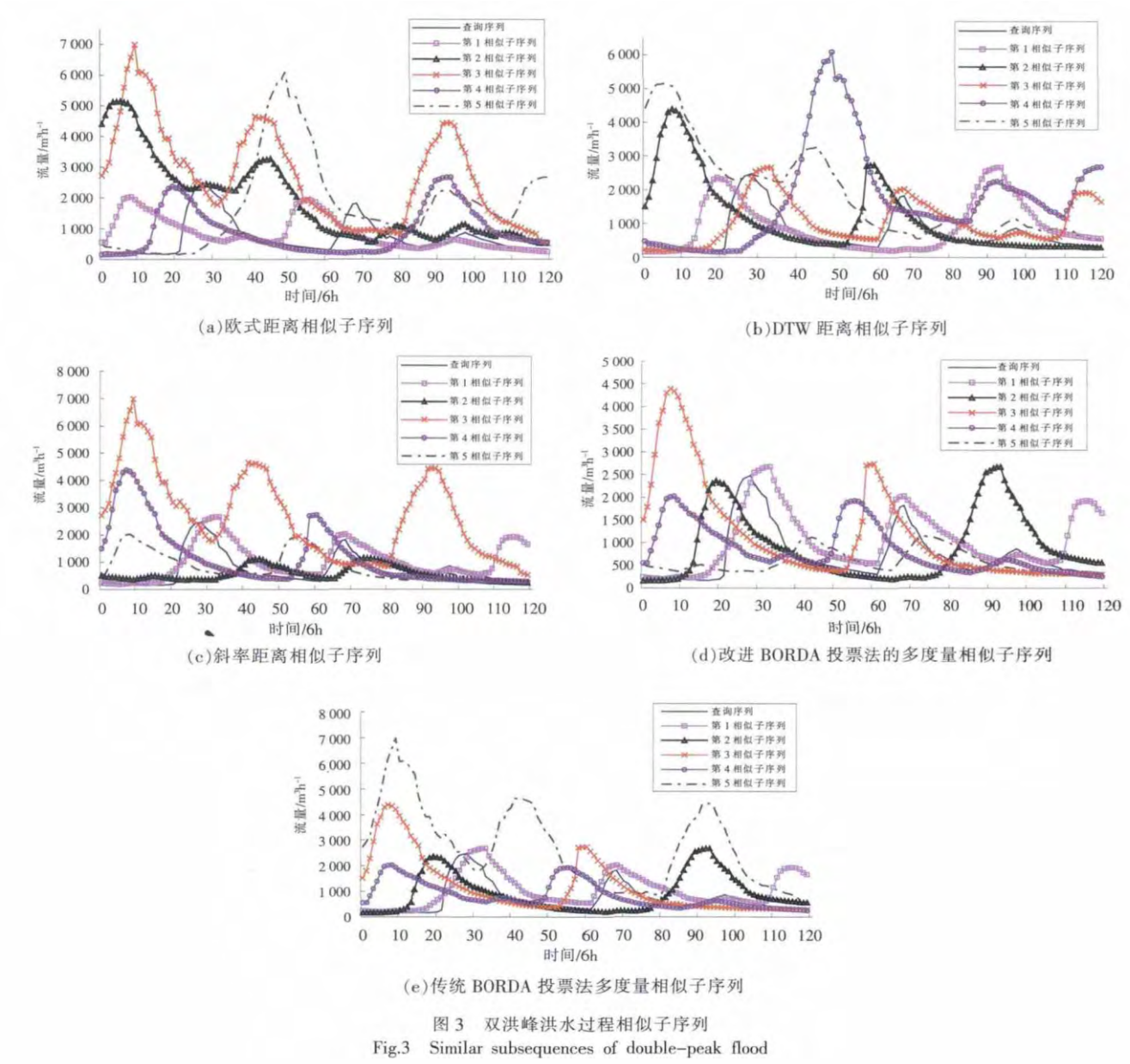


图3　双洪峰洪水过程相似子序列

Fig.3　Similar subsequences of double-peak flood

表2中，改进BORDA投票法的多度量组合查询结果中，除了起点为2007.8.15 2:00的子序列，都是在各单一相似度量结果中多次出现的子序列。起点为2007.7.16 2:00、2003.7.1 2:00和2005.7.1 2:00的子序

列虽然在不同的单一相似度量中多次出现，但是各自的排序差距较大，造成改进 BORDA 得分较低，因此被淘汰。从图 3(d)看出，改进 BORDA 投票法的多度量组合查询出的相似子序列的洪峰变换过程与查询序列几乎完全一致，都是双洪峰的洪水过程。被淘汰的三个子序列和图 3(d)结果中的其他相似子序列相比，相似程度较弱。

相对传统 BORDA 投票法的多度量组合，改进 BORDA 投票法的多度量组合淘汰了起点为 2003.7.1 2:00 的三洪峰洪水流量过程子序列，保留起点为 2007.8.15 2:00 的双洪峰洪水流量过程子序列。起点 2003.7.1 2:00 的子序列出现两次，但是其排序差距大，因此改进 BORDA 得分比传统 BORDA 得分大大降低，起点 2007.8.15 2:00 的子序列虽然只出现一次，但是其排序差距小，反而改进 BORDA 得分比传统 BORDA 得分高，从图 3(d)和图 3(e)中，起点为 2007.8.15 2:00 的子序列比 2003.7.1 2:00 的子序列相似程度大。同时，传统 BORDA 投票法的多度量组合没有很好的对最终的相似子序列进行排序。

4 结语

本文提出多度量组合的水文时间序列相似度量方法，首先使用多个单一相似度量分别计算相似时间子序列，然后采用改进的 BORDA 投票法对各单一相似度量查询结果进行组合，以得到最终的相似子序列。通过分析淮河流域王家坝水闸相似洪水过程可以看出，针对水文时间序列的相似性分析，基于改进 BORDA 投票法的多度量组合比单一相似度量以及基于传统 BORDA 投票法的多度量组合够获得更加准确的相似子序列。

改进的 BORDA 投票法中，排序最前和最后的相似子序列排序得分都是固定的，不能灵活反映其和查询序列的相似程度，今后还需要进一步研究符合水文领域知识的多度量组合方法问题。

参考文献：

[1] R.Agrawal, C.Faloutsos, A.Swami. Efficient similarity search in sequence databases [A]. Proceedings of the 4th International Conference on Foundations of Data Organizations and Algrithms(FODO'93) [C]. 1993:69–84.

[2] Yi, B, K., Faloutsos ,C. Fast time sequence indexing for arbitrary Lp norms [A]. Proceedings of the 26th Int'l Conference on Very Large Databases [C]. 2000:385–394.

[3] Bemdt Donald J, Clifford James. Using dynamic time warping to find patterns in time series[A]. Proceedings of the KDD Workshop, Seattle,WA [C]. 1994:359–370.

[4] André–Jönsson, H. and Badal. D. Using signature files for querying time–series data [A]. Proceedings of Principles of Data Mining and Knowledge Discovery, 1st European Symposium [C]. 1997:211–220.

[5] L.Chen, M.T.Ozsu. V.Oria. Robust and fast similarity search for moving object trajectories[A]. In SIGMOD Conference, 2005.

[6] 张建业,潘泉,张鹏,等. 基于斜率表示的时间序列相似性度量方法[J]. 模式识别与人工智能, 2007, 20(2): 271–274. (ZHANG Jianye, PAN Quan, ZHANG Peng, et al. Similarity measuring method in time series based on slope [J]. Pattern Recognition and Artificial Intelligence, 2007, 20(2): 271–274. (in Chinese))

[7] 李薇,孙洪林. 水文时间序列相似性查询的分析与研究—以漯河站、何口站汛期降雨量相似性查询为例[J]. 水文, 2009,29(6):76–80. (LI Wei, SUN Honglin. Analysis and study on hydrological time series similarity search [J]. Journal of China Hydrology, 2009,29(6):76–80. (in Chinese))

[8] 欧阳如琳,任立良,周成虎. 水文时间序列的相似性搜索研究[J]. 河海大学学报(自然科学版), 2010,38(3):241–245. (OUYANG Rulin, REN Liliang, ZHOU Chenghu. Similarity search in hydrological time series [J]. Journal of Hohai University (Natural Sciences), 2010,38(3):241–245. (in Chinese))

[9] 朱跃龙,王咏梅,万定生,等. 基于语义相似的水文时间序列相似性挖掘—以太湖流域大浦口站水位数据为例[J]. 水文, 2011,31(1):35–40. (ZHU Yuelong, WANG Yongmei, WAN Dingsheng, et al. Similarity mining of hydrological time series based on semantic similarity measures [J]. Journal of China Hydrology, 2011,31(1):35–40. (in Chinese))

[10] 李士进,朱跃龙,张晓花,等. 基于 BORDA 计数法的多元水文时间序列相似性分析[J]. 水利学报, 2009,40(3):378–384. (LI Shijin, ZHU Yuelong, ZHANG Xiaohua, et al. BORDA counting method based similarity analysis of multivariate hydrological time series [J]. Journal of Hydraulic Engineering, 2009,40(3):378–384. (in Chinese))

[11] F.Fabris,I.Drago, F.M.Varejao. A multi–measure nearest neighbor algorithm for time series classication [A]. Proceedings of the 11th Ibero–American Conference on AI, Springer–Verlag, Berlin, Heidelberg [C]. 2008: 153–162.

[12] Barrios, J.M., Bustos, B. Automatic weight selection for multi–metric distances [A]. Proceedings of the 4th International Conference on Similarity Search and Applications [C]. 2011: 61–68.

[13] Black D. The Theory of Committees and Elections (2nd edition) [M]. London: Cambridge University Press, 1963.

(下转第 73 页)

随着各行业的快速发展,产生大量污染源聚集入河,对局部水域水环境有一定污染影响。依靠现状引江调水增加内河水源,乘潮排泄弃水入江,该调度水源方式制约因素较多,水生态用水的补给水量严重不足。将火电厂的温排水源调度入内河,作为河道水生态用水水源,能有效地增加补水量,增加水生态调水力度,对实现内河水环境质量的稳定提高、水生态的健康发展,能创出较好的环境效益[2]。

4.3 利用温排水为内河补充水源的经济效益

火电厂温排水源是一种已经过电厂抽提长江水体至发电机组循环冷却使用后的水源,然后由高处向低处自流排放。调度温排水入内河,不需要再用动力抽提,可免去建设运行同等规模提水站的投资建设及运行成本。按照现有从长江提水入内河的抽水站运行成本参考推算,约需 350 元/万 m^3 的抽水运行费。若将两座电厂现年产生 $11.87\times10^8 m^3$ 温水量全部调度入内河,每年可节省提水运行费约 4154.5 万元。利用火电厂温排水源入河可创出较为显著的经济效益。

5 结语

沿江火电厂的温排水源取自长江,水量充沛,水源稳定,水质良好。利用火电厂温排水源入内河作为区域水资源的补充配置,提增供水能力;增加区域河道水环境容量,增强河道纳污能力;与城市污水处理厂达标尾水混合后排放,减轻尾水污染混合带的影响程度,均具有十分重要的意义。

参考文献:

[1] 黄向阳,谢磊. 江水源热泵系统温排水对江水水温及水质的影响[J]. 水电能源科学, 2010,28 (7):34-36. (HUANG Xiangyang, XIE Lei. Effect of warm water discharge of river water resources heat pump system on water temperature and water quality [J]. Water Resources and Power, 2010,28(7):34-36. (in Chinese))

[2] 翟水晶. 电厂温排水对湿地生态系统的影响研究—以江苏射阳港电厂为例[D]. 南京: 南京师范大学, 2006. (ZHAI Shuijing. Effects of Thermal Discharge from Power Plant on Wetland Ecology: A Case of Sheyang Port Power Plant in Jiangsu [D]. Nanjing: Nanjing Normal University, 2006. (in Chinese))

How to Use Thermal Discharge from Power Plant along Yangtze River

CHEN Yan, LIU Jiansheng

(*Nantong Hydrology and Water Resources Survey Bureau of Jiangsu Province*, *Nantong* 226006, *China*)

Abstract: This paper analyzed and evaluated the water stability, water quality and temperature effect scope of the thermal discharge from the power plant along the Yangtze River. The results show that the thermal discharge can increase the water supply capacity in the river of the enhancing area, increase the local water environmental capacity and pollutant carrying capacity of river water.

Key words: power plant; thermal discharge; use

(上接第 20 页)

Multi-measure Similarity Analysis of Hydrological Time Series

WANG Jimin[1,2], ZHU Yuelong[1], LI Wei[3], WAN Dingsheng[1,2], LI Shijin[1]

(1. *College of Computer & Information, Hohai University*, *Nanjing* 210098, *China*; 2. *National Engineering Research Center of Water Resources Efficient Utilization Engineering Safety*, *Hohai University*, *Nanjing* 210098, *China*; 3. *Bureau of Hydrology*, *MWR*, *Beijing* 100053, *China*)

Abstract: Based on the idea that multi-measure combination can improve the accuracy of similarity analysis, multi-measure similarity analysis method was proposed for hydrological time series. Firstly, the similarity of time series was computed by several similarity measures, respectively. Then, the modified BORDA voting method was used to synthesize the similar time series of each similarity measure to obtain the final similar time series. The proposed method was validated by the analysis results of the flood data obtained from the Wangjiaba Station in the Huaihe River Basin.

Key words: time series; similarity analysis; hydrology; BORDA voting method; multi-measure