基于嵌入式索引的水文时间序列相似性搜索模型

沈 强，万定生，王亚明

（河海大学计算机与信息学院，江苏 南京 210098）

摘 要：相似模式挖掘已成为水文领域一个重要研究方向。对水文数据的相似性挖掘，有利于掌握水文数据变化规律和趋势，为洪水预报、防洪调度提供支持，是具有重要意义的工作。为此，在引入时间序列嵌入索引的基础上，结合水文时间序列的特点提出水文时间序列的快速搜索方法。该方法通过序列分割、聚类和参考集训练从原始序列中获取参考序列集，在此基础上通过索引计算方法，将相似性搜索过程映射到欧氏向量空间的搜索，从而提高了搜索效率。

关键词：相似性分析；时间序列分割；聚类；嵌入索引

中图分类号：TP391 **文献标识码**：A **文章编号**：1000-0852(2016)06-0064-06

1 引言

在水文领域，水文时间数据挖掘(Hydrological Time Series Data Mining)是指是从海量的，多噪声、随机的水文时序数据库中，提取隐含其中有用的水文信息和知识的过程[1]。

基于嵌入式索引的子序列相似性搜索[2](EBSM)是一种子序列匹配的近似算法，核心思想是利用嵌入索引[3]把子序列匹配过程转换为欧氏向量空间中的匹配。针对水文时间序列的特征，在EBSM方法的基础上提出了基于嵌入式索引的相似性搜索模型HTS-EBSM(Hydrologic Time Series Embedding-Based Subsequence Matching)，主要包括：提出了洪峰分割和聚类算法；重定义索引计算方法，降低计算的复杂度和减少索引空间的冗余；优化在线搜索中候选点选择方法，避免了候选点的冗余搜索。

2 模型框架

时间序列相似性查询中过滤-精炼框架（Filter-and-refine framework）[4]常用于对于给定查询序列从原始序列中寻找前个匹配序列中。

HTS-EBSM是基于上述框架的，包括离线和在线两个过程：(1) 离线部分实现了水文时间序列分割，聚类，参考序列选择训练和时间序列嵌入索引过程。(2)在线阶段包含对查询序列的索引，候选集的选择和最终相似性序列的获取过程。HTS_EBSM具体流程见图1。

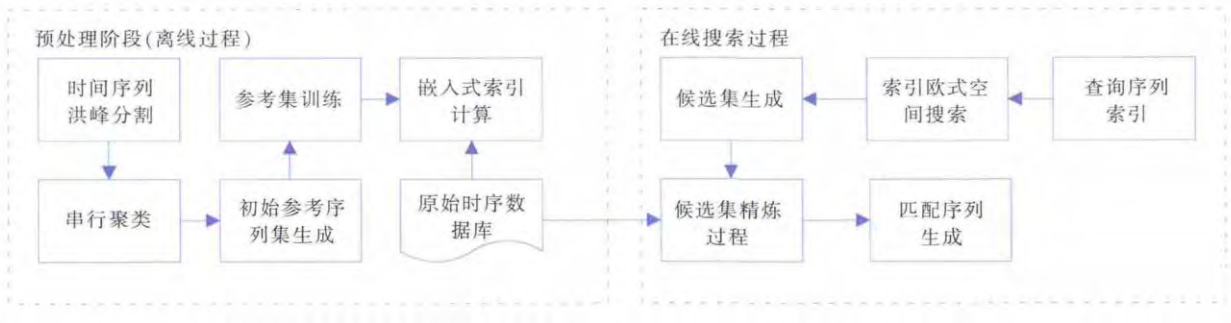

图1 水文时间序列嵌入式索引搜索流程

Fig.1 The flow of hydrologic time series embedding-based index search

收稿日期：2015-06-03
基金项目：水利部公益性行业科研专项经费项目(201501022)
作者简介：沈强(1991-)，男，江苏泗阳人，硕士研究生，主要研究方向为数据挖掘与信息系统。E-mail:shenqiangz@126.com

3　基于嵌入式索引的相似性搜索

3.1　洪峰模式提取和串行聚类

文献[5]定义了极值点提取过程，但实验发现原始序列过多的噪声会导致局部极值点过多，结合小波变换[6]和极值点方式提出了洪峰分割算法，其流程如下：

(1)小波平滑去噪。

(2)根据水位阈值进行粗分割。

(3)改进的模式提取方法进行细分割。

分割方法见算法 1。

算法 1：HTS_RefineSegment(F)

输入：洪峰粗分割集合 F

输出：洪峰模式提取结果 RF

1. 对 F 中每个子序列 X，进行以下操作；
2. 提取 X 中的所有极大值构成新的序列 M；
3. 提取 X 中的所有极小值构成新的序列 N；
4. 对于 M，找到第一个极大值点 (M_i,P_i)，在 N 中寻找具有最小水位值的点 (N_i,P_i) 作为洪峰模式的起始位置，其中 $N_i<M_i$，再从中寻找处于 M_i 之后而在 M_{i+1} 之前具有最小水位值的点 (N_j,P_j) 作为洪峰模式的终止位置；
5. 寻找 M 中下一个极大值点，重复 4 直到 M 遍历结束，得到所有的分割点集，生成洪峰模式提取结果 RF。

聚类算法进行相似性序列段的剔除，假设 $R_i\in RF$，而 $R_j\in RF(j\neq i)$，若 $D(R_i,R_j)<\varepsilon$，即距离比较小，相似性显著，只需其中的一个为候选序列。给定相似性的阈值 ε，各聚类中，所有序列与质心序列的距离都小于 ε，从而每个聚类中选取单个序列组成参考序列训练的初始集。

3.2　嵌入式索引生成方法

利用 EBSM 方法训练得到的参考序列集，进行嵌入索引计算，把原始序列映射到欧氏向量空间中。EBSM 方法对原始序列每个位置都生成对应的欧氏向量。索引向量相邻位置对应向量非常相似，导致索引结构冗余，欧氏空间搜索时间消耗过多。

HTS-EBSM 提出基于窗口的欧氏索引生成算法，该算法定义了原始序列中的一个窗口，索引计算通过该窗口进行，从而降低数据的冗余，提高了索引结构的精简性，进而提高查询的效率。

定义 1：时间序列 X 中第 i 位置的索引结构为 $(F_i^{R^1}(X),F_i^{R^2}(X),\cdots,F_i^{R^d}(X))$，其中 $F_i^{R^k}(X)=D(R^k,X_{i-\omega+1}\cdots X_i)$ $(1\leq k\leq d)$，ω 为窗口大小，$F_i^{R^k}(X)$ 为对应维度的参考序列与序列 X 中以 i 为终点、长度为 ω 的子序列之间的动态弯曲距离 (DTW)[7]。

定义 1 为对应参考序列集和时间序列的索引向量计算公式。在实际应用中，我们定义了两个参数 ω 和 δ，ω 表示窗口的长度，δ 表示相邻两个窗口之间的重叠部分，具体索引过程见算法 2。

算法 2：HTS_EmbeddingWithWindow (*R*,*X*, Width, Overlap)

输入：*R*:参考序列集；*X*：原始时间序列；

Width：索引窗口长度；Overlap：相邻窗口的重叠长度

输出：VectorSpace：欧氏向量索引空间

BEGIN:

1.endPoint=Width;step=Width−Overlap;index=1; //初始化对应的初始索引位置

2.While (endpoint<length(*X*))

3.Vector 初始化为空

4.For *i*=1:length(*R*)

5.dist=$DTW(R_i,X$[(endPoint −Width+1):endPoint]); //计算第 i 个参考序列与原始序列 X 中 endPoint−Width+1 至 endPoint 的序列段之间的 DTW 距离

6.Vector←dist; //对应距离放入向量中

7.End

8.VectorSpace←Width,endPoint,Vector;//存储对应位置的索引向量和标记到欧氏向量索引空间中

9.End

END

算法 2 中每个窗口计算过程中时间复杂度为 $O(\sum_{i=1}^{a}\omega|R_i|)$，其中 ω 和 $|R_i|$ 分别为窗口的长度和参考序列的长度。窗口每次移动的步长都为 $(\omega-\delta)$，主循环次数为 $\eta=\frac{|X|-\omega}{\omega-\delta}$，所以全部时间复杂度为 $O(\eta\sum_{i=1}^{a}\omega|R_i|)$。

3.3　时间序列在线搜索方法

在线搜索的任务为：给定一个查询序列从原始序

列中寻找相似性匹配对象。对于查询序列的在线搜索过程需要通过过滤–精炼过程从嵌入式索引向量空间中过滤生成对应的候选集然后通过精炼候选集寻找匹配结果。

对于给定的查询序列 Q,其在线搜索过程如下:

(1)对查询序列 Q 进行索引计算:

$$F(Q)=(D(R_1,Q),D(R_2,Q),\cdots,D(R_d,Q))$$

(2)对于每一个 $F(Q)$,与时间序列嵌入索引 VectorSpace 中的每一个向量 Vector 之间进行欧氏距离的计算从而得到前 k 个候选点,候选点为 Cand$P$=$\{P_1,P_2,\cdots,P_k\}$。

(3)在候选集 Cand$P$ 附近范围内进行原始 DTW 的匹配从而得到相似性搜索结果。

候选集中可能出现距离相近的点,所以对候选集进行优化处理,以避免精炼过程中重复计算,处理流程如下:

对候选集中的点进行排序,设置精炼过程中搜索范围为 δ,从最后一个点开始进行以下操作:

(1)设置当前点为 $P_i(i>1)$,若 P_{i-1} 搜索范围($[P_{i-1}-\delta,P_{i-1}+\delta]$)与 P_i 搜索范围($[P_i-\delta,P_i+\delta]$)有重叠部分,则进行两点范围的合并。

(2)迭代 a 直到 P_2 或者 $P_i-\delta<0$。

HTS_EBSM 时间序列搜索方法,时间花费主要是在线搜索部分,其余部分都在离线状态下完成。在线搜索时间复杂度主要分为三部分:①查询序列的嵌入索引,具体为 $O(|Q|\sum_{i=1}^{a}|R_i|)$,表示为查询序列与参考序列集中所有序列长度的乘积;②欧氏空间搜索部分复杂度为 $O(n)$,具体为索引维度与索引向量个数的乘积;③精炼过程中查询序列与原始序列中进行 DTW 匹配的复杂度,为 $O(k|Q||X|^2)$,k 为候选点个数,$|X|$为各候选范围的原始序列子段的长度。

4 实验分析

4.1 有效性测试

长江流域特别是长江中下游平原地区是我国防汛的重点地区[8],选择长江中下游南京段下关站水位数据作为实验数据集。下关站记录了 1912~2007 年共 31 777 条长江水位数据。设置控制水位阈值 6m,分割出洪峰模式共 144 个,时间跨度最小为 15d,最大为 168d。对 144 个洪峰模式聚类计算,得到 50 个聚类结果,获取了容量为 50 的初始参考序列集。对应的洪峰模式的聚类结果具体请见图 2。

容量为 50 的初始参考集再通过 EBSM 中的参考集训练方法,最终得到对训练查询序列效果最好的容量为 30 的参考序列集。下关站参考序列集容量为 30,索引向量维度为 30,窗口每次往后移动 10 个时间,对于长度为 31 777 的历史水位数据库一共生成了 3 172 个索引向量。从历史序列中选取了 3 条查询序列,查询序列详见表 1。

表1 查询序列数据详细

Table1 The query sequence data query

编号	起始时间	终止时间	时间间隔 / d
1	2000-06-01	2000-09-15	107
2	2005-07-15	2005-09-15	63
3	1990-07-01	1990-08-15	46

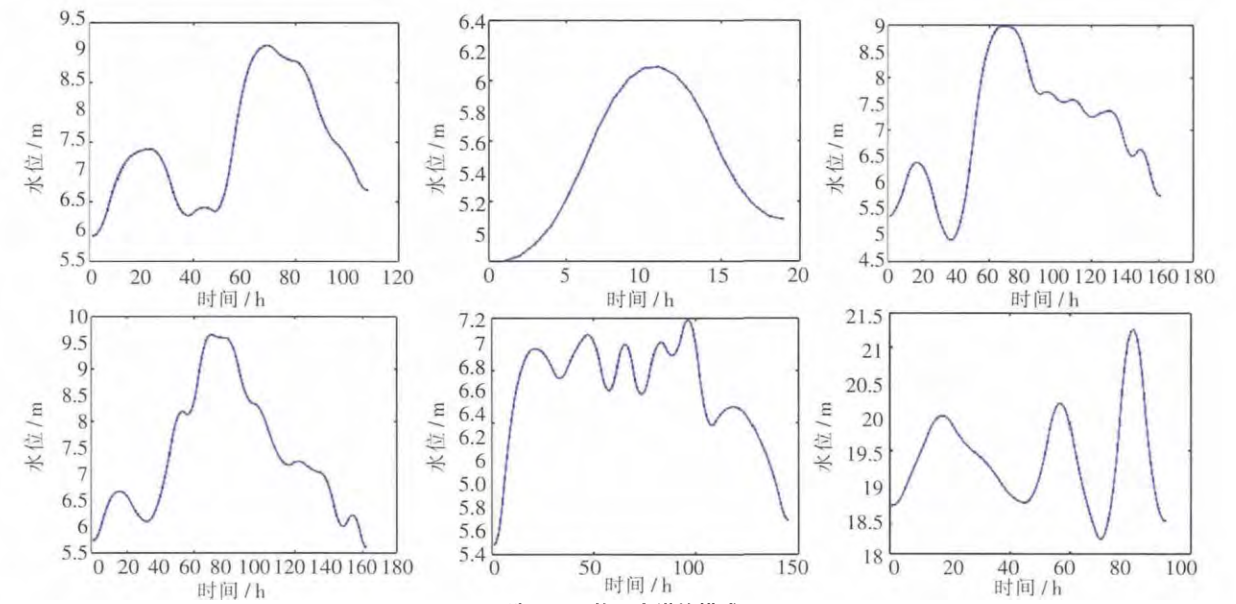

图 2 选取展示的 6 个洪峰模式

Fig.2 The 6 selected flood peak modes

表2 下关站HTS_EBSM相似性搜索结果

Table2 The results of HTS_EBSM similarity searching

编号	HTS_EBSM 模型			
	匹配序列段		DTW	时间
	起始时间	终止时间	距离	/s
1	1981-05-31	1981-09-01	1.7947	60.04
	1927-06-02	1927-09-18	2.6258	
	2004-05-10	2004-09-20	4.6551	
	1970-04-14	1970-07-16	4.7908	
2	1989-05-27	1989-07-12	0.9982	21.44
	1948-05-12	1948-07-27	1.2476	
	1976-06-09	1976-07-30	1.4359	
	1999-05-30	1999-06-30	1.4532	
3	1975-06-29	1975-08-15	1.3144	8.67
	2007-07-31	2007-09-26	1.3747	
	1968-07-15	1968-09-10	1.3906	
	1989-07-05	1989-08-31	1.3990	

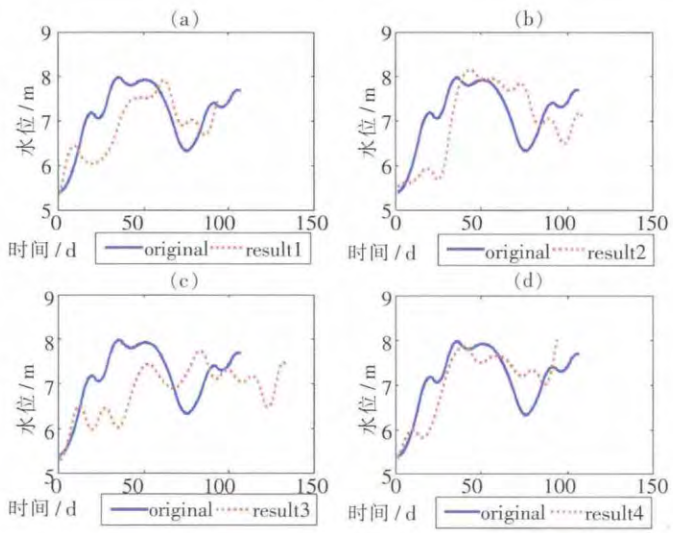

图3　下关站2000年6月1日~9月15日的水位相似性查找结果

Fig.3 The results of searching the water level similarity of the Xiaguan station from June 1 to September 15, 2000

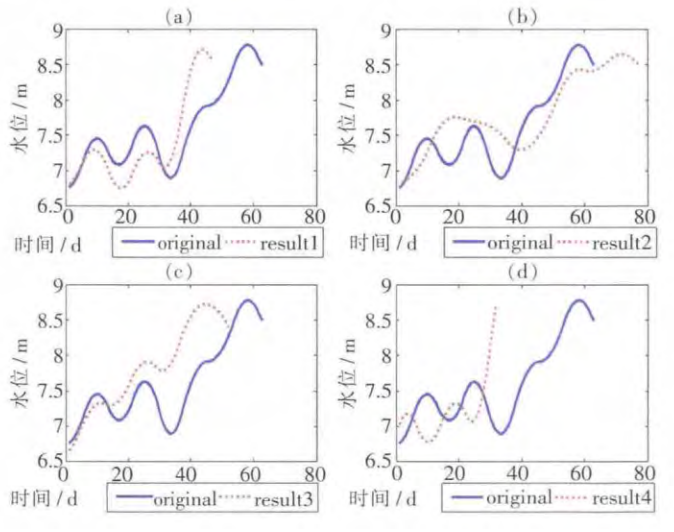

图4　下关站2005年7月15日~9月15日的水位相似性查找结果

Fig.4 The results of searching the water level similarity of the Xiaguan station from July 15 to September 15, 2005

从HTS_EBSM搜索得到的匹配序列集中选取前4个相似性度最高的序列作为搜索结果，对应的匹配序列详细见表2。

图3~5分别对应了下关站3个查询序列经过HTS_EBSM方法得到的前4个匹配结果(分别为图a~d),从图中可以直观的发现所获的序列与查询序列有着比较相似的走势和相同的水位波动，可知相似性搜索结果是有效的。

4.2　性能分析

本节讨论HTS_EBSM和同样基于过滤-精炼框架的基于语义相似的水文序列相似性搜索方法(DTW_SS)[9]的对比。对下关站的三个查询序列实现了DTW_SS搜索。同样选择前4个相似性最高的序列,得到的匹配结果详细如表3。

对比表2和表3可以发现,对于编号为1的查询序列,两种搜索方法得到的匹配结果中存在3个相同的匹配序列；编号为2的查询序列得到了1989年匹配段;对于编号为3的查询序列,一样的得到了1968年、1975年类似的匹配序列段。以上相同的匹配结果,表3都作了标注。因此,对于编号为1~3的查询序列,HTS_EBSM方法和DTW_SS方法都能得到同样的匹配序列段,证明HTS_EBSM方法能有效地获得相似性

表3 DTW_SS相似性搜索结果

Table3 The results of DTW_SS similarity searching

编号	HTS_EBSM 模型			
	匹配序列段		DTW	时间
	起始时间	终止时间	距离	/s
1	1981-05-21	1981-09-04	1.9531	470.11
	1927-06-02	1927-09-16	2.5466	
	1965-06-06	1965-09-20	4.4252	
	2004-06-09	2004-09-23	4.9039	
2	2005-07-13	2005-09-13	0.9863	148.03
	1989-05-16	1989-07-17	1.1061	
	1983-05-04	1983-07-05	1.5705	
	1935-06-04	1935-08-05	1.6910	
3	1964-06-28	1964-08-12	0.7007	78.53
	1962-07-01	1962-08-15	1.2159	
	1968-07-20	1968-09-03	1.2664	
	1975-06-30	1975-08-14	1.3039	

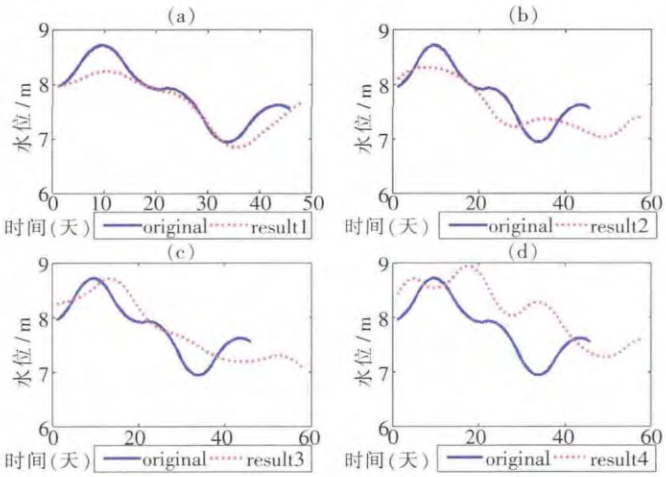

图5 下关站1990年7月1日~8月15日的水位相似性查找结果

Fig.5 The results of searching the water level similarity of the Xiaguan station from July 1 to August 15, 1990

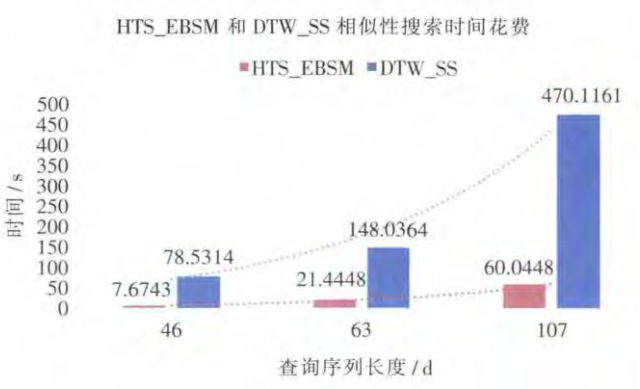

图6 HTS_EBSM 和 DTW_SS 查询时间对比柱状图

Fig.6 The comparison between HTS_EBSM and DTW_SS in searching time

匹配结果。

同时,从图6可以发现,在相似性匹配效果相似的前提下,HTS_EBSM 相似性搜索方法时间耗费只相当于 DTW_SS 方法时间耗费的 1/7 到 1/10 之间。

上述测试结果和性能实验证明,HTS_EBSM 方法能有效的得到相似性匹配结果,时间耗费也比较低,具有比较大的实际应用性。

5 结语

基于水文数据的短期波动特点和周期性特征,本文引入嵌入式索引思想提出了水文数据领域的时间序列快速搜索模型 HTS_EBSM, 并证明其确有实际的应用性。但是还可以对 HTS_EBSM 算法继续研究分析,如原始时间序列长度对在线搜索时间复杂度的影响;参考序列维度所产生的作用等。

参考文献

[1] 艾萍,倪伟新. 我国水文数据挖掘技术研究的回顾与展望[J]. 计算机工程与应用, 2003,39(28):13-17. (AI Ping, NI Weixin. Review and preview of the research on hydrological data mining technology in China [J]. Computer Engineering and Applications, 2003,39(28): 13-17. (in Chinese))

[2] Athitsos V, Papapetrou P, Potamias M, et al. Approximate embedding-based subsequence matching of time series [A]. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data [C]. 2008:365-378.

[3] V. Athitsos, M. Hadjieleftheriou, G. Kollios, et al. Query-sensitive embeddings [A]. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data [C]. 2005:706-717.

[4] Hjaltason G R, Samet H. Properties of embedding methods for similarity searching in metric spaces [J]. Pattern Analysis and Machine Intelligence, 2003,25(5):530-549.

[5] 朱跃龙, 彭力, 李士进, 等. 水文时间序列模体挖掘 [J]. 水利学报, 2012,43(12):1422-1430. (ZHU Yuelong, PENG Li, LI Shijin, et al. Research on hydrological time series motifs mining [J]. Journal of Hydraulic Engineering, 2012,43(12):1422-1430. (in Chinese))

[6] Popivanov I, Miller R J. Similarity search over time-series data using wavelets [A]. Proceedings of 18th International Conference on Data Engineering [C]. 2002:212-221.

[7] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series [J]. KDD Workshop, 1994,10(16):359-370.

[8] 黄荣辉,陈栋,刘永,等. 中国长江流域洪涝灾害和持续性暴雨的发生特征及成因 [J]. 成都信息工程学院学报, 2012,(1):1-19. (HUANG Ronghui, CHEN Dong, LIU Yong. Characteristics and causes of the occurrence of flooding disaster and persistent heavy rainfall in the Yangtze River valley of China [J]. Journal of Chengdu University of Information Technology, 2012,(1):1-19. (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))

Embedding-based Index Model for Hydrological Time Series Similarity Searching

SHEN Qiang, WAN Dingsheng, WANG Yaming

(*College of Computer and Information, Hohai University, Nanjing* 210098, *China*)

Abstract: Similar pattern mining has become an important research direction in the field of Hydrology. It is a significant work to process similarity mining in historical data that can be conducive to recognize the trend pattern of hydrological data and provide technical support for the flood forecasting and flood control. Thus, this paper proposed a quick similarity search model according to hydrological sequence features. This model employed series segment, serial cluster and reference training method to generate reference set, and transferred similarity search to European vector space search with indexed by reference set so as to improve the searching efficiency.

Key words: similarity analysis; time series segmentation; clustering; embedded index

(上接第 59 页)

Junhua, LUO Longfu, ZHANG Zhiwen, et al. Comprehensive evaluation of power quality based on fuzzy set pair analysis [J]. Power System Technology, 2012,36(7):81-85. (in Chinese))

[12] 徐健,吴玮,黄天寅,等. 改进的模糊综合评价法在同里古镇水质评价中的应用 [J]. 河海大学学报 (自然科学版), 2014,(2):143-149. (XU Jiang, WU Wei, HUANG Tianyin, et al. Application of improved fuzzy comprehensive evaluation towater quality evaluation in Tongli town [J]. Journal of Hohai University (Natural Sciences), 2014,(2):143-149. (in Chinese))

[13] 童英伟,刘志斌,常欢.集对分析法在河流水质评价中的应用[J]. 安全与环境学报, 2008,(6):84-86. (TONG Yingwei, LIU Zhibin, CHANG Huan. Application of set pair analysis in appraising the river water quality of Fuxin[J]. Journal of Safety and Environment, 2008,(6):84-86. (in Chinese))

Water Resources Monitoring Ability Evaluation for Yunnan Province Based on Fuzzy Set Pair Analysis Assessment Model

WANG Jing[1], HE Xing[2], LIU Yanhui[3], ZHANG Yunying[4], ZHANG Liangen[1], BAI Yong[2]

(1. *College of Water Conservancy, Yunnan Agricultural University, Kunming* 650201,*China*; 2. *College of Resources and Environment, Yunnan Agricultural University, Kunming* 650201,*China*; 3. *College of Architecture and Engineering, Yunnan Agricultural University, Kunming* 650201, *China*; 4. *Hydrology and Water Resources of Yunnan Province, Kunming* 650106, *China*)

Abstract: In order to provide scientific basis for sustainable utilization of water resources and water resources management in Yunnan Province, this paper used fuzzy set pair analysis assessment model to select 11 indices1including the water quantity monitoring coverage rate of water users, water quality monitoring coverage rate of water function area and so on for constructing a water resources monitoring ability evaluation index system, and determining the evaluation grading standard. AHP was used to get weight value of each index, and the water resources monitoring ability for 16 regions were evaluated. The results show that the water resources monitoring ability is at the middle level. The fuzzy set pair analysis method is not only simple for calculation, but also reasonable and reliable.

Key words: fuzzy set pair analysis method; water resources monitoring; Yunnan Province; analytic hierarchy process