Multi-measure Similarity Searching for Time Series

Jimin Wang, Yuelong Zhu, Dingsheng Wan, Pengcheng Zhang, Jun Feng
College of Computer & Information, HoHai University, Nanjing 211100, China
Email: {wangjimin, ylzhu, dshwan, pchzhang, fengjun}@hhu.edu.cn

***Abstract*—In this paper, we evaluate some techniques for the time series similarity searching. Many distance measures have been proposed as alternatives to the Euclidean distance in the similarity searching. To verify the assumption that the combination of various similarity measures may produce more accurate similarity searching results, we propose an multi-measure algorithm to combine several measures based on weighted BORDA voting method. The proposed method is validated by the analysis results of the flood data obtained from Wangjiaba in the Huaihe basin of China.**

***Index Terms*—multi-measure, similarity searching, hydrological, BORDA voting, time series**

I. INTRODUCTION

With the development of the information technology and the sensor technique, there are more and more datasets stored in the form of time series, including financial, stock prices, climatic, biological, hydrological and other fields. A time series is a collection of observations obtained sequentially through time. It's one of the hottest problems to discover knowledge from those data. Data mining on time series is mainly about similarity searching, classification, clustering, sequential patterns mining and prediction. As the important basis of other tasks, similarity searching has been paid more attention.

Similarity searching is firstly presented in [1] and mainly focuses on representation, indexing and similarity measure. A univariate time series is often regarded as a point in multidimensional space, so one of the major reasons for time series representation is to reduce the dimension (i.e. the number of data point) because of the curse of dimensionality. Many approaches are used to extract the pattern, which contains the main information of original time series, to reduce the dimension. Piecewise linear representation (PLA) [2, 3], Piecewise Aggregate Approximation (PAA) [4], etc. use k adjacent segments to represent the time series with length n(n>>k). Furthermore, perceptually important points (PIP) [5], critical point model (CMP) [6], etc. reduce the dimension by preserving the salient points. Another common family of time series representation approaches transform time series into discrete symbols, so that string operation could be performed on time series, e.g. Symbolic Aggregate Approximation (SAX) [7], Shape description Alphabet (SDA) [8], and other symbol-generation method based on clustering [9, 10]. Representing time series in the transformation domain is another large family of approaches, e.g. Discrete Fourier Transform (DFT) [11], Discrete Wavelet Transform (DWT) [12] ,FD [13] transform the original time series into frequency domain. After transformation, only the first few or the best few coefficients are chosen to represent the original time series [14]. Recently, the extraction of semantic characteristics for semantic similarity are paied more and more attentaion in different domain[15]. Many of the representation schemes are incorporated with different multi-dimensional spatial indexing techniques, e.g. the k_d tree [16], b_tree [17], r_tree and its variants [18, 19], are used to indexing sequences to improve the the query efficiency during similarity searching. Given two time series S, Q and their representation PS, PQ, a similarity measure function D calculates the distance between the two time series, denoted by D(PQ, PS) to describe the similarity/dissimilarity between Q and S, such as Euclidean distance (ED) [1] and the other Lp Norms, dynamic time warping (DTW) [20], longest common subsequence (LCS) [21], slope distance [22] and pattern distance [23].

During similarity searching, the traditional approach is to select one similarity distance to measure the similarity. For k nearest neighbor (kNN) searching, the similarity measure could be considered as classifier, and classifies the time sequences into the 1st similar sequence, the 2nd similar sequence, ..., the kth similar sequence and the not similar sequence categories. Inspired by the concept, that multi-classifier could improve the accuracy of classification [24], multi-measure method is used for kNN searching. Reference [25] proposed a new heuristic method based on weights to combine measures in the nearest neighbor decision rule, and find in some cases, combining metrics brought a good accuracy gain. Reference [26] used a multi-metric function that combines distances from many descriptors (colors, edges, textures, etc.) to retrieve content-based multimedia information, and presented three novel techniques to find a different weight to each descriptor representing its relative importance in the combination.

In this work, a multi-measure method based on the weighted BORDA voting method is proposed for univariate kNN similarity searching. Several similarity

doi:10.4304/jcp.9.10.2266-2273

measures are used to search similar sequences respectively, then, the weighted BORDA voting method is used to synthesize the similar sequences and obtain the final *k*NN sequences.

In the next section, we briefly describe the BORDA voting method and some similarity measures widely used. Next, Section 3 presents the proposed algorithm to search the kNN sequences. Data sets and experimental results are shown in section 4. Finally, Section 5 concludes the paper.

II. Related Work

A. BORDA Voting Method

BORDA voting, a classical voting method in group decision theory, is proposed by Jena-Charles de BORDA. Supposing *k* is the number of winners, *m* is the number of candidates; *n* electors express their preferences from high to low in the sort of candidates. To every elector's vote, provide the No.1 candidate m points (called voting score), the second candidate *m*-1 points, followed by analogy, the last one put 1 points. The accumulated voting score of the candidate is BORDA score. The candidates, BORDA scores in the top *k*, are called BORDA winners.

B. Similarity Measures

In many fields, the similarity between two sequences is usually measured by a distance function. Similarity is inversely proportional to distance, the smaller distance the more similar of two sequences.

Minkowski Distance. Minkowski distance is a commonly adopted similarity measure. Manhattan distance and Euclidean distance are both special cases of Minkowski distance. This measurement method has the advantage of easily calculating, indexing and clustering. However, it is sensitive to noise and small variations in time axis. So Minkowski distance is not well suitable for similarity comparison of two sequences directly.

Dynamic Time Warping Distance. Dynamic programming is the theoretical basis for dynamic time warping (DTW). DTW is a non-linear planning technique combining time and distance measure, which was firstly introduced to time series mining areas by Berndt and Clifford [19] to measure the similarity of two univariate time series. According to the minimum cost of time warping path, the DTW distance supports time axis stretching, but does not meet the requirement of triangle inequality, and with high computing cost.

Pattern Distance. Pattern distance [23] is an effective similarity measure too. It remedies the defect of time series match in point distance, and reflects the dynamic trends of the time series. It is closer to natural language description, with clear physical meaning of pattern definition and rapid calculation speed. To overcome the inaccuracy caused by the pattern distance, Reference [22] proposed a slope distance to measure the similarity of time series, which is claimed with more clear physical meaning, more intuitive and simple calculation process. The slope distance meets the basic criteria of similarity measure such as symmetry, self-similarity, non-negative and triangle inequality.

III. The Proposed Method

In the previous section, we have reviewed the BORDA voting method and several similarity measures, In this section, we propose a multi-measure similarity measure on weighted BORDA voting method, denoted by S_{WBORDA}, for univariate *k*NN searching. The proposed method is suitable for both the whole and subsequence matching similarity searching.

A. Multi-measure Weighted BORDA Voting: S_{WBORDA}

Traditional BORDA voting method takes just the order into consideration, without the actual gap between two adjacent candidates, that may lead to rank failure for the candidates. For example, assuming four candidate r_1, r_2, r_3, r_4 take part in race, the first round position is r_1, r_2, r_3, r_4, the second is r_2, r_1, r_4, r_3, the third is r_4, r_3, r_1, r_2, and the last is r_3, r_4, r_2, r_1. The four runners are all ranked no.1 with traditional BORDA score (10 points), because of considering only the rank order, but not the speed gap of runners in the race. In our proposed approach, we use the complete information of candidate, including the order and the actual gap to neighbor, to generate the BORDA score.

Given the query sequence *Q*, to perform *k*NN searching in time series database, several similarity measures are used to calculate the similar sequences respectively. For one similarity measure, the *m* nearest neighbor sequences are s_1, s_2, ..., s_m, where *m* is equal or greater than the *k*, and the similarity distance to query sequence is d_1, d_2 ,..., d_m, respectively. The similarity distance d_{i-1} is less than or equal to d_i, and the distance gap, d_i-d_{i-1}, describes the similarity gap to the query sequence of s_i and s_{i-1} to query sequence. Let the weighted voting score of s_1 p points, and s_m 1 point, the weighted voting score of the sequence s_i , vs_i, is defined by

$$vs_i = m - (m\text{-}1) \times \frac{d_i - d_1}{d_m - d_1} \qquad (i = 1, ..., m) \tag{1}$$

vs_i is inversely proportional to d_i-d_1, s_1 is the baseline, the higher similarity gap between s_i and s_1, the lower weighted BORDA score s_1 will get. Traditional BRODA voting is a special case of weighted BORDA voting when the similarity gaps between adjacent candidate are all equal i.e. d_2- d_1 = d_3- d_2 = ... = d_m- d_{m-1}.

We accumulate the weighted voting scores of a sequence and obtain its weighted BORDA score. The sequences are ranked on their weighted BORDA scores, and the top *k* are the final similar sequences to *Q*. The model of multi-measure similarity searching based on weighted BORDA voting is shown in Fig. 1.

In the model of Fig. 1, several similarity measures, called single-measure, are selected to search the *m*NN sequences one by one, where *m* is greater than the final *k*, then the *m*NNs are truncated to generate candidate similar sequences. At last, the weighted BORDA voting method

is performed on candidate similar sequences to obtain the *k*NN sequences. Intuitively, multi-measure measures the similarity from different aspects(measures), and synthetizes them. The following sections describe the similarity searching in detail.

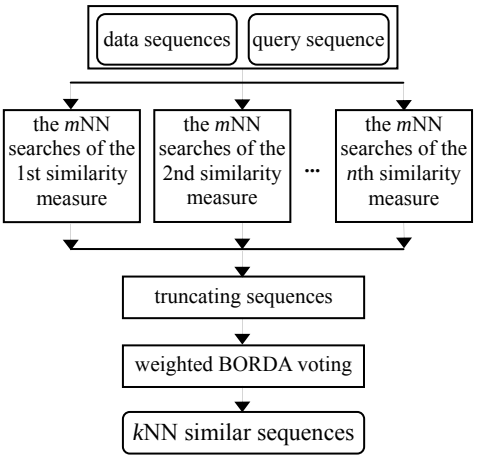


Figure 1. The model of multi-measure similarity searching

B. The Selection of Single-measures

The single-measure, included in our multi-measure, should be selected according to the analytic request, e.g. pattern distances or slope distances are suitable for shape similarity, and DTW is suitable for sequences with different lengths. Besides similarity measure, the representation, indexing, and searching method, etc. should be considered. When performing *m*NN by single-measure, the m should be greater than the final *k*, so that *k* candidate similar series could be obtained after truncation.

C. Truncating the Similar Sequences

The similar sequences of single-measures may not start from the same time, but the similar sequences with close start time could be considered as in the same candidate similar sequence, so the similar sequences should be truncated to obtain the candidate similar sequences. The truncation includes three cases: grouping the original similar sequences, deleting the isolated sequences, aligning the overlapping sequences and reordering the sequences.

The truncation for whole sequence matching is just a special case for subsequence matching, so we introduce the truncation in subsequence matching similarity searching. In Fig. 2, three single-measures have been used to search 3NN sequences for univariate query sequence with length *l*. The original 3NN sequences of the first measure are s_{11} (the subsequence from t_{11} to $t_{11}+l$), s_{12} (from t_{12} to $t_{12}+l$) and s_{13} (from t_{13} to $t_{13}+l$). The similar sequences are presented according to their occurrence time, and the present order don't reflect the similarity order to the query sequence. The original 3NN sequences of the second measure are s_{21} (from t_{21} to $t_{21}+l$), s_{22} (from t_{22} to $t_{22}+l$) and s_{23} (from t_{23} to $t_{23}+l$), and these of the third measure are s_{31} (from t_{31} to $t_{31}+l$), s_{32} (from t_{32} to $t_{32}+l$) and s_{33} (from t_{33} to $t_{33}+l$).

1) Grouping the original similar sequences

The original similar sequences of single-measures are divided into several groups, so that in each group, for any sequence *s*, at least one sequence *w*, which is overlapping with *s* more than, e.g. ten percent *l*, could be found. The original similar sequence, not overlapping with any others, will be put into a single group just including itself. In Fig. 2, all the similar sequences are divided into five groups. The group g1 includes s_{11}, s_{21}, s_{31}. s_{11} and s_{21} overlapp with s_{21} and s_{31} respectively, and the overlapping lengths are all over ten percent *l*. g2 includes s_{32}, g3 includes s_{12}, s_{22}, g4 includes s_{13}, s_{33}, and g5 includes s_{23}.

2) Deleting the isolated sequences

The group, in which the number of similar sequences included is less than half number of the single-measures, is called isolated group, and the similar sequences in isolated group are called isolated similar sequences. Isolated sequences and groups are deleted and ignored in the subsequent processing. In Fig. 2, groups g2 and g3 are both isolated groups, because the number of included sequences is less than half the number of single-measures i.e. three, and will be deleted.

3) Aligning the overlapping sequences

The original similar sequences in the same group should be set the same start time and length. For one group, the average start time t of all the included sequences is calculated, then the subsequence from t to t+*l*, denoted by *cs*, is candidate similar sequence. The *cs* is regarded as the similar sequences of all the single-measures, and the similarity distance between *cs* and the query sequence is recalculated by the single-measures one by one. If the group contains the similar sequence of the *i*th single-measure, then the corresponding similarity distance is set to *cs* for the *i*th single-measure to reduce computation. In Fig. 2, For group g1, the average of t_{11}, t_{21}, t_{31} t_{c1} is computed, then the subsequence s_{tc1}, from t_{c1} to $t_{c1}+l$, is the candidate similar sequence. for group g3, the similarity distance between s_{tc2} and the query sequence should be recalculated by the third single-measure. The same alignment operation is performed on group g4 to obtain the candidate sequence s_{tc3}.

4) Reordering the candidate similar sequences

For each single-measure, the candidate similar sequences are reordered by the similarity distance calculated in step (3), and the weighted BORDA voting method is used to synthesized the candidate similar sequences and generate the *k*NN sequences.

In whole matching *k*NN searching, the original similar sequences are either whole overlapping or not overlapping each other, and the truncation steps are the same to that of the subsequence matching.

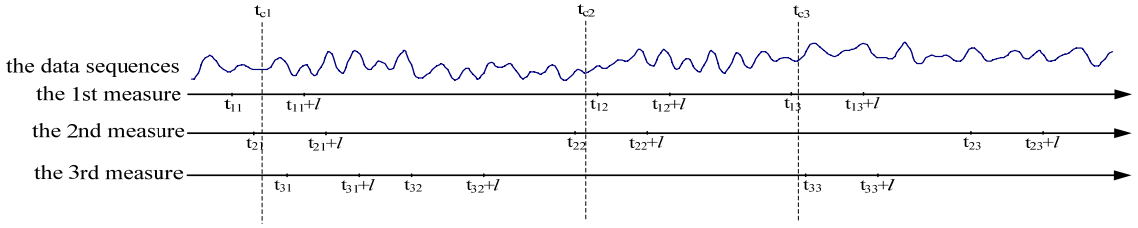


Figure 2. Truncating similar sequences in subsequence matching

IV. Experiments and Analysis

In order to evaluate the performance of our proposed techniques, we performed experiments on real-world datasets. In this section, we first describe the data sets used in the experiments, and the experiments methods followed by the results.

A. Datasets

Great deal of hydrological data, obtained by long-term observation, contains important information. Data mining plays an increasingly significant role in dealing with massive and complex hydrological data. Similarity analysis in hydrological time series is one of the most important basic technologies, which is directly applied to answer questions in flood control such as "which historical period corresponding to the current situation". The experiments have been conducted on the flood data from Wangjiaba in the Huaihe basin of China. The data was recorded from June 1st to September 30th in every year during 1998 to 2009, four observation values were obtained at 2:00, 8:00, 14:00 and 20:00 every day.

B. Methods

Our goal is to determine if multi-measure in the similarity searching could perform better than using always the same one. The Euclidean distance (ED) is the most straightforward similarity measure for time series. Normally the ED is used as the baseline when someone wants to advocate the utility of a novel measure. dynamic time warping (DTW) and slope distance (SD) is proved to be able to produce good results in hydrological [27, 28], so in the experiments, ED, DTW and SD are selected as the single-measures. The multi-measure on traditional BORDA voting method, denoted by S_{TBORDA}, is also included as a compared measure.

Time series piecewise linear representation based on features points [7] is performed to extract the pattern representation of the flood data. Two case studies, single-peak and double-peak flood process 5NN searching, was conducted to verify the feasibility and effectiveness of the proposed multi-measure similarity measure.

C. Experimental Results

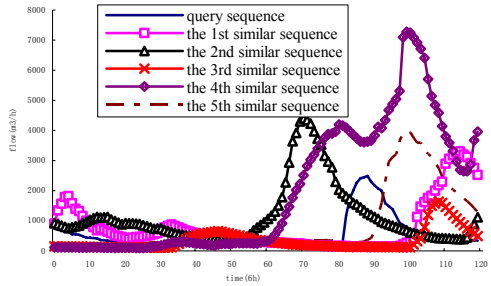
Table I illustrates the top 5 similar subsequences of single-peak flood process with query sequence from 2:00 on July 31, 2000 to 20:00 on August 29, 2000, and Fig. 3 illustrates the trend of similar subsequences. In Fig. 3, horizontal axis stands for time, vertical axis stands for flow.

In table I, the similar subsequences of multi-measure all appear in more than one single-measure results. The subsequence from 2:00 on 1 July, 2004 only appears in the result of DTW, so gets the low weighted BORDA score and is discarded. To the subsequences from 2:00 on June 16, 2007 and 8:00 on 1 August, 2008, although appear in more than one results of single-measures, but there is big similarity gap between them and their neighbor, so get low weighted voting score and are discarded. In Fig. 3(e), the similar subsequences of SWBORDA show almost the same trend with query sequence, compared to the three discarded subsequences, the subsequences in Fig. 2(e) are more similar to the query sequence.

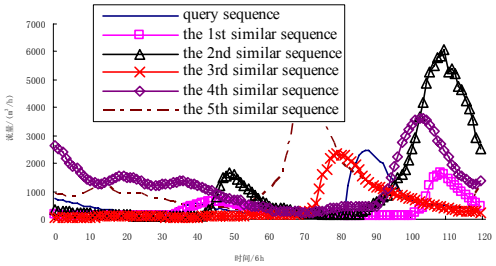
Table I shows that, S_{WBORDA} and S_{TBORDA} find the same 5 similar subsequences, but S_{TBORDA} provides 4 similar subsequences with the same BORDA scores.

TABLE I.
SIMILAR SUBSEQUENCES OF SINGLE-PEAK FLOOD PROCESS

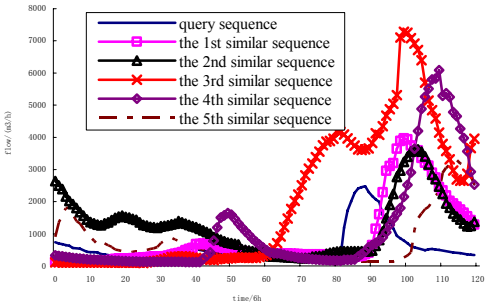
ED		DTW		SD		S_{TBORDA}		S_{WBORDA}	
start time	distance	start time	distance	start time	distance	start time	BORDA score	start time	weighted BORDA score
2000.8.31 2:00	799.86	2005.6.1 8:00	465	2008.7.1 8:00	0.11	20005.6.1 8:00	8	2005.6.1 8:00	7.45
2008.8.1 8:00	848.37	2005.6.16 2:00	1830	2005.7.31 2:00	0.16	2000.8.31 2:00	6	2005.7.31 2:00	6.01
2005.6.1 8:00	944.75	2004.7.1 2:00	3730	2007.6.16 2:00	0.20	2008.7.1 8:00	6	2008.7.1 8:00	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	6	2000.8.31 2:00	5.96
2008.7.1 8:00	1027.45	2008.8.1 8:00	7230	2000.8.31 2:00	0.29	2005.7.31 2:00	6	2005.6.16 2:00	5.90



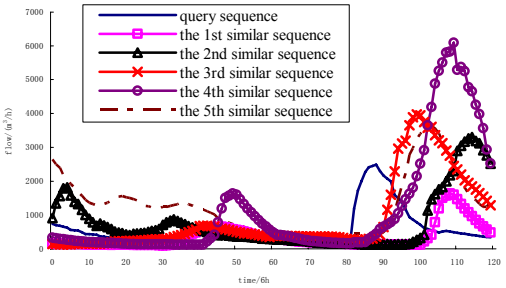
(a) 5NN subsequences of ED



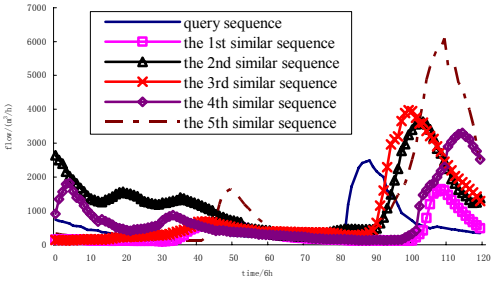
(b) 5NN subsequences of DTW



(c) 5NN subsequences of SD



(d) 5NN subsequences of S_{TBORDA}



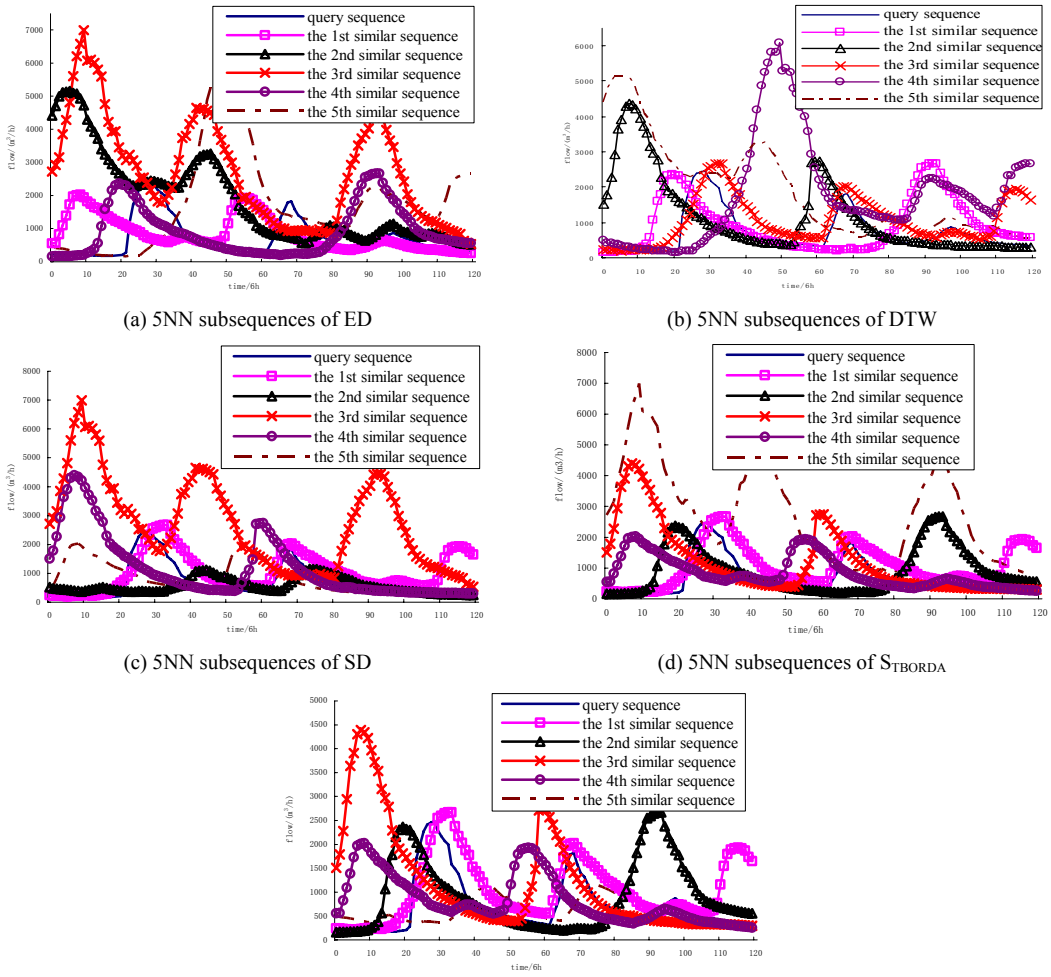
(e) 5NN subsequences of S_{WBORDA}

Figure 3. 5NN of single-peak flood process

Table II and Fig. 4 illustrates the top 5 similar subsequences of double-peak flood process with query sequence from 2:00 on August 15, 2000 to 20:00 on September 13, 2000. In Fig. 4, horizontal axis stands for time, vertical axis stands for flow. In table II, the similar subsequences of multi-measure all appear in more than one result of single-measures besides the subsequence from 2:00 on August 15, 2007. The subsequences from 2:00 on July 16, 2007, 2:00 on July 1, 2003 and 2:00 on July 1, 2005 are discarded. Although appearing in more than one results, but they get low weighted BORDA score because of big gap between them and their ahead neighbor. In Fig. 4(e), the similar subsequences of multi-measure have almost the same double-peak with query sequence.

TABLE II.
SIMILAR SUBSEQUENCES OF DOUBLE-PEAK FLOOD PROCESS

ED		DTW		SD		S_{TBORDA}		S_{WBORDA}	
start time	distance	start time	distance	start time	distance	start time	BORDA score	start time	weighted BORDA score
2004.8.15 2:00	675.72	2004.7.16 2:00	830	2004.7.31 2:00	0.52	2004.7.31 2:00	8	2004.7.31 2:00	8.19
2007.7.16 2:00	943.29	2008.8.17 2:00	2260	2007.8.15 2:00	0.57	2004.7.16 2:00	7	2004.7.16 2:00	6.36
2003.7.1 2:00	997.07	2004.7.31 2:00	4460	2003.7.1 2:00	0.75	2008.8.17 2:00	6	2008.8.17 2:00	6.18
2004.7.16 2:00	1037.63	2005.7.1 2:00	7660	2008.8.17 2:00	0.80	2004.8.15 2:00	6	2004.8.15 2:00	6
2005.7.1 2:00	1073.35	2007.7.16 2:00	8860	2004.8.15 2:00	0.88	2003.7.1 2:00	6	2007.8.15 2:00	4.45



(a) 5NN subsequences of ED

(b) 5NN subsequences of DTW

(c) 5NN subsequences of SD

(d) 5NN subsequences of S_{TBORDA}

(e) 5NN subsequences of S_{WBORDA}

Figure 4. 5NN of double-peak flood process

Compared to the S_{TBORDA}, the S_{WBORDA} discards the three-peak flood process from 2:00 on July 1, 2003, denoted by s_{2003}, and retains the double-peak flood process from 2:00 on August 15, 2007, denoted by s_{2007}. The subsequence s_{2003} appears in two results of single-measure, but big gap between it and its neighbor results to the low weighted BORDA score. The subsequence s_{2007} appears only in one result of single-measure, but it is close to the neighbor, so get high weighted BORDA score. The Fig. 4(e) and Fig. 4(d) show, compared to s_{2003}, s_{2007} is more similar to the query sequence. Table II shows that, S_{TBORDA} give the last three similar subsequences the same BORDA score.

V. Conclusions and Future Work

To verify if using different measures together improves the similarity searching accuracy, we present a multi-measure similarity analysis method based on weighted BORDA voting. We conducted two experiments and compare the accuracy of the multi-measure against three single-measures (the Euclidean distance, dynamic time warping and slope distance) on the flood records of WangIiaba in Huaihe basin of China. The experimental results show that multi-measure on weighted BORDA voting could produce more accurate similar sequences than single measures and multi-measure on traditional BORDA voting.

In the similarity result of single-measure, the voting scores of the 1st and the *k*th similar sequences are fixed to *k* points and 1 point, respectively, even if they are not so similar to query sequence. This may affect the final result, and the problem will be tackled in the future. In the literature, there are still not so many studies in multi-measure similarity analysis for time series, the new integration method for multi-measure similarity will also be further explored.

Acknowledgment

This research was partially supported by the Fundamental Research Funds for the Central Universities (No. 2009B22014). the National Natural Science Foundation of China (No. 61170200, No. 61370091, NO. 61202097).

References

[1] R.Agrawal, C.Faloutsos, and A.Swami, “Efficient similarity search in sequence databases,” In Proc. of the 4th International Conference on Foundations of Data

Organizations and Algrithoms(FODO'93), pp. 69-84, 1993.
[2] Keogh E, Pazzani M, "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback," In Proc. of the 4th International Conference of Knowledge Discovery and Data Mining, pp. 239-243, 1998.
[3] Keogh E, Smyth P, "A probabilistic approach to fast pattern matching in time series databases," In Proc. of the 3rd International Conference of Knowledge Discovery and Data Mining, pp. 24-30, 1997.
[4] Keogh, E., Pazzani, M, "A simple dimensionality reduction technique for fast similarity search in large time series databases," In Proc. of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 122–133, 2000.
[5] Fu T, Chung F, Luk R, "Representing financial time series based on data point importance," Engineering Applications of Artificial Intelligence, Vol.21, No.2, pp. 277-300, 2008.
[6] [6] Bao, D.A, "Generalized model for financial time series representation and prediction," Applied Intelligence, Vol.29, No.1, pp. 1–11, 2008.
[7] Lin J, Keogh E, Lonardi S, "A Symbolic representation of time series, with implications for streaming algorithms," In Proc. of the Eighth ACM SIGMOD International Conference on Management of Data Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2-11, 2003.
[8] André-Jönsson H, Badal D Z, "Using signature files for querying time-series data," In Proc. of the First European Symposium on Principles and Practice of Knowledge Discovery in Databases, pp. 211–220, 1997.
[9] [9] Hebrail G, Hugueney B, "Symbolic representation of long time-series," In Proc. of the Conference on Applied Statistical Models and Data Analysis, pp. 537-542, 2001.
[10] Hugueney B, Bouchon-Meunier B, "Time-series segmentation and symbolic representation, from process-monitoring to data-mining," In Proc. of the 7th International Conference on Computational Intelligence, Theory and Applications, pp. 118–123, 2001.
[11] R.Agrawal, C.Faloutsos, A.Swami, "Efficient similarity search in sequence databases," In Proc. of the 4th International Conference on Foundations of Data Organizations and Algrithoms, pp. 69-84, 1993.
[12] Chan, K.P., Fu, A.C, "Efficient time series matching by wavelets," In Proc. of the 15th IEEE International Conference on Data Engineering, pp. 126–133, 1999.
[13] Hu Y, Li Z, "An improved shape signature for shape representation and image retrieval," Journal of Software (1796217X), Vol.8, No.11, pp. 2925-2929, 2013.
[14] Kiyoung Yang and Cyrus Shahabi, "A PCA-based similarity measure for multivariate time series," In Proc. of the 2nd ACM international workshop on Multimedia databases, pp. 65-74, 2004.
[15] Wang Y, Chen S, Qiu Y, "Ontology-based semantic similarity transfer algorithm," Journal of Software (1796217X), Vol.8, No.5, pp. 1268-1274, 2013.
[16] Ooi B C, McDonell K J, Sacks-Davis R, "Spatial kd-tree: An indexing mechanism for spatial databases," In Proc. of IEEE COMPSAC Conference, pp.433-438, 1987.
[17] Ni Z, Guo J, Wang L, "An efficient method for improving query efficiency in data warehouse", Journal of Software (1796217X), Vol.6, No.5, pp.857-865, 2011.
[18] Guttman A, R-trees: A dynamic index structure for spatial searching. ACM, 1984.
[19] Beckmann N, Kriegel H P, Schneider R, The R*-tree: an efficient and robust access method for points and rectangles. ACM, 1990.
[20] Berndt D J, Clifford J, "Using dynamic time warping to find patterns in time series," In KDD workshop,Vol.10, No.16, pp. 359-370, 1994.
[21] Paterson M, Dančík V, "Longest common subsequences," Springer Berlin Heidelberg, 1994.
[22] Zhang Jian-Ye, Pan Quan, Zhang Peng, "Similarity measuring method in time series based on slope," Pattern Recognition and Artificial Intelligence, Vol. 20, No. 2, pp. 271-274, 2007.
[23] Wang Da, Rong Gang, "Pattern distance of time series," Journal of Zhejiang University (Engineering Science), Vol. 38, No. 7, pp. 795-798, 2004.
[24] Kittler J, "Combining classifiers: A theoretical framework," Pattern analysis and Applications, Vol.1, No.1, pp.18-27, 1998.
[25] F.Fábris, I.Drago, and F.M.Varejão, "A multi-measure nearest neighbor algorithm for time series classification," In Proc. of the 11th Ibero-American Conference on AI(IBERAMIA '08), pp.153-162, 2008.
[26] Barrios J M, Bustos B, "Automatic weight selection for multi-metric distances," In Proc. of the 4th International Conference on SImilarity Search and Applications, pp. 61-68, 2011.
[27] [27]Shi-jin LI, Yue-long ZHU, Xiao-hua ZHANG, "BORDA count method based similarity analysis of multivariate hydrological time series," SHUILI XUEBAO, Vol. 40, No. 3, pp. 378-384, 2009.
[28] Rulin Quyang, Liliang Ren, and Chenghu Zhou, "Similarity search in hydrological time series," Journal of Hohai University(Natural Sciences), Vol.38, No.3, pp. 241-245, 2010.

Jimin Wang is a Ph.D. candidate at HoHai University. He received his M. Sc. degree in Computer Science from the HoHai University, China in 2003. In the past, until he started his Ph.D. study in 2009, he worked as a lecturer and researcher at the College of Computer & Information, Hohai University, China. His research interests include intelligent data processing and data mining. His current research work focuses on time series data mining and it's application in hydrological field.

Yuelong Zhu (Ph.D.) is professor at HoHai University since 2001. His research interests include data mamangment, data minging and multimedia mining. His current work concerns the hydrological data mining.

Dingsheng Wan (Ph.D.) is professor at HoHai University since 2008. His research interests include data quality mamangment, data minging. His current work concerns the hydrological data mining.

Pengcheng Zhang (Ph.D.) is associate professor at HoHai University since 2013. His research interests include software service and data minging. His current work concerns the cloud-storage services.

Jun Feng (Ph.D.) is professor at HoHai University since 2008. She received her Ph.D. in computer science in 2004 at Nagoya University Nagoya, Japan. Her research interests include

temporal-spatial data storage, retrieval, and data minging. Her current work concerns the hydrological spatial-temporal data retrievl.