Time-Varying Water Quality Analysis with Semantical Mining Technology

Jun Feng, Qinghan Yu(✉), and Yirui Wu

College of Computer and Information, Hohai University, Nanjing, China
{fengjun,qinghanyu,wuyirui}@hhu.edu.cn

Abstract. Water resources is one of the most important natural resources. With the development of industry, water resource is harmed by various types of pollution. However, water pollution process is affected by many factors with high complexity and uncertainty. How to accurately predict water quality and generate scheduling plan in time is an urgent problem to be solved. In this paper, we propose a novel method with semantical mining technology to discover knowledge contained in historical water quality data, which can be further used to improve forecast accuracy and achieve early pollution warning, thus effectively avoiding unnecessary economic losses. Specifically, the proposed semantical mining method consists of two stages, namely frequent sequence extraction and association rule mining. During the first stage, we propose FOFM (Fast One-Off Mining) mining algorithm to extract frequently occurred sequences from quantity of water quality data, which can be further considered as input of the second stage. During the process of association rule mining, we propose PB-ITM (Prefix-projected Based-InterTransaction Mining) algorithm to find relationship between frequently occurred water pollution events, which can be regarded as knowledge to explain water pollution process. Through experimental comparisons, we can conclude the proposed method can result in flexible, accurate and diverse patterns of water quality events.

Keywords: Pattern mining · Sequence patterns · Association rules · Water quality forecasting

1 Introduction

Single-sequence pattern mining can be defined as technology of discovering patterns that occurs frequently in a single sequence. Considering that there is a fixed time interval between each event of time-varying water quality sequence, we can conclude a significant relationship between water pollution events. With discovering knowledge of relationship among water pollution events, researchers can be aware of changes of water quality in a few weeks. Therefore, it is necessary to adopt a semantical mining method to mine knowledge inside the time-varying water quality sequence.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2020
Published by Springer Nature Switzerland AG 2020. All Rights Reserved
X. Zhang et al. (Eds.): CloudComp 2019/SmartGift 2019, LNICST 322, pp. 349–362, 2020.
https://doi.org/10.1007/978-3-030-48513-9_29

Since we can gain quantity of water quality data with help of sensor technology, it's necessary to first extract frequent sequence of water quality events as input to be analyzed. We thus develop the semantical mining method with two stages, where the first stage is used to extract frequent sequence and the second stage is applied to discover association rules among frequent sequences of water quality events.

During the **first stage**, we adopt interval constraint to help discover more useful patters from time-varying water quality sequence, where interval constraint refers to that each event in the mode needs to meet the interval condition. To better explain such property, we take sequence $S = abcabcdefgabbbadacbabc$ and pattern $ababc$ as an example. If we examine the exact pattern inside the sequence, we can't find such pattern. Once we define a constraint as the minimum interval is 0 and the maximum interval is 2, we can find that the pattern $ababc$ appears twice, where the positions of such pattern sequences are $\{0, 1, 3, 4, 5\}$ and $\{16, 18, 19, 20, 21\}$. From such example, we can see that interval constraint makes the single-sequence pattern mining more complicated and more flexible to dig useful patterns.

Based on the property of time-varying water quality sequence, we further define One-Off condition [4] to calculate uncertainty, where the position sequence of any two same patterns shouldn't share the same position under One-Off condition. Considering sequence as $S = abacc$, pattern as $P = abac$ and interval constraint as $[0, 1]$, the positions of such pattern P can be solves as $\{0, 1, 2, 3\}$, $\{0, 1, 2, 4\}$ without One-Off condition. Once we adopt One-Off Condition, pattern P can be only solved with positions $\{0, 1, 2, 3\}$, since the three identical positions $0, 1, 2$ are shared in above two occurrences of pattern P. In the water quality time sequence, it is obviously unreasonable to think that this pattern P occurs twice.

Although quantity of fast single-sequence pattern extraction algorithms have been proposed, they own high potential to lose patterns and may decrease efficiency under interval constraint and One-Off condition. In order to solve the problem of extracting frequent sequence of water quality events, we thus propose FOFM (Fast One-Off Mining) algorithm with interval constraints and One-Off condition, which can improve efficiency and accuracy during the process of mining water quality time sequence.

During the **second stage**, traditional association rule mining algorithms can only be applied on transaction database, which are not suitable to mine association rules from water quality sequences due to their high complexity and diversity. Moreover, water quality events are often occurred with time delay, since water pollution event happened in area A cannot immediately affect the area B. Therefore, mined association rules should own the property of temporal characteristics, where we use "Water quality deteriorates in area A this week, water quality of area B deteriorates in the second week $\Longrightarrow$ Water quality of area C deteriorates in the third week" as an example. In order to mine such rules with temporal characteristics, we proposes the PB-ITM algorithm (Prefix-projected Based-InterTransaction Mining), which not only considers association items, but also pay special attention on time relationship between transactions items.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the details of the proposed FOFM algorithm to extract frequent sequences. The proposed PB-ITM algorithm to discover traditional association rules is discussed in Sect. 4. Section 5 presents the experimental results and discussions. Finally, Sect. 6 concludes the paper.

2 Related Work

With the rapid development of network and information technology, how to effectively use the gradually accumulated data is a hot research topic. As one of the core contents of data mining, pattern mining can obtain the implicit association between things, thus helping people to achieve prediction and recommendation. Pattern mining is a process of discovering high-frequency item sets, sub-sequences or sub-structures from a large amount of data, and is divided into three types of classical mining algorithms.

Association rule mining is one of the most basic methods, search for frequent item sets in the form of traversal trees. The most popular method of this type is the breadth-first search Apriori algorithm [1] and depth-first search FP-growth algorithm. Agrawal et al. proposed three test mining algorithms based on Apriori properties: AprioriAll, AprioriSome, DynamicSome [2], and then GSP [21]. The above algorithms are based on horizontal format algorithms. Zaki [23] proposed a sequence pattern algorithm SPADE based on the vertical format, which converts the sequence database into a vertical format database that records the location of the item set, and then dynamically joins the mining frequent sequence pattern. The algorithm only needs to scan the database three times, reducing the I/O overhead.

The pattern mining problem of the sequence was first proposed by Agrawal and Srikant in 1995, and introduced many related algorithms, including FreeSpan [10], PrefixSpan [20], and some improved algorithms have been proposed [3,5], Zou Xiang et al. studied the sequential pattern mining algorithm in distributed environment [26]. The traditional sequential pattern mining algorithm is to mine frequent occurrence patterns from sequence databases, without defining wildcard constraints. Ji et al. [14] and Li et al. [17] studied the problem of pattern mining with wildcards in the sequence database. The concept of minimum distinction mode with wildcards is proposed in article [13]. Another focus of sequential pattern mining research is to mine frequent patterns from a single sequence, which is usually quite long, for example. NA sequences and protein sequences, etc. Zhang et al. studied the pattern mining problem with wildcards in a single sequence, and proposed the MPP algorithm [24]. He et al. [11] studied the problem of sequential pattern mining in a single sequence, and the pattern satisfies the One-Off condition. Zhu et al. [25] proposed the MCPaS algorithm to mine frequent patterns from multiple sequences.

Graph pattern mining is the process of identifying high-frequency substructures from a set of graphs or a single large graph. Most of the current algorithms are for atlas, including SUBDUE based on greedy strategy [6], gSpan

[22] using depth-first search, GASTON [19], and AGM using breadth-first search [12], PATH [13], FGS [15]. The SUBDUE algorithm is also applicable to single pictures.

Mining interesting patterns from different types of data is quite important in many real-life applications [7,16]. Sequential pattern mining (SPM) [9] has been extensively studied a novel utility mining framework, called utility-oriented pattern mining (UPM) or high-utility pattern mining, which considers the relative importance of items, has become an emerging research topic in recent years. Affinitive utility [18] is proposed to address the special task of correlated UPM, but not used for the general task of UPM. The utility occupancy [8] is more suitable than the utility concept and average utility for discovering the high-utility patterns which have high utility contribution.

3 Fast One-Off Mining Algorithm

The proposed FOFM algorithm is used to extract frequent sequences from quantity of water quality data. We present steps of FOFM by first defining related conceptions and then presenting its detail.

3.1 Related Definition

This section mainly describes the related definitions and theorems involved in the single-sequence pattern of water quality time sequence.

Given a sequence $S = \{s_1, s_2, s_3, \ldots s_{n-1}, s_n\}$ with n characters, its length can be defined as n, where the set of all the different characters in the sequence can be denoted as $\sum$. The interval constraint consists of a minimum interval and a maximum interval, which can be denoted as N and M. Moreover, size $(M - N)$ is defined as the interval constraint length.

Given patterns $P = \{p_1, p_2, p_3, \ldots p_{m-1}, p_m\}$ and $Q = \{q_1, q_2, q_3, \ldots q_{t-1}, q_t\}$, for any k with $1 \leq k \leq t$, if existing a sequence with positions $1 \leq i_1 \leq i_2 \leq \cdots \leq i_t \leq m$ which satisfy $p_{i_k} = q_k$, Q can be regarded as a sub-pattern of P. Meanwhile, P can be regarded as the parent pattern of Q. For any j with $2 \leq j \leq t$, if existing a sequence with positions which satisfy $i_j - i_{j-1} = 1$, Q can be regarded as a continuous sub-pattern of P. Moreover, Q can be defined as the prefix pattern of P, only if Q is a continuous sub-pattern of P and $i_1 = 1$. For any k with $2 \leq k \leq m$, if existing $p_k = q_{k-1}$, pattern P and Q can be connected to generate a new pattern, which can be represented as $p_1 q_1 q_2 q_3 \ldots q_{m-1} q_m$.

Once the appearing frequency of the pattern P in the sequence S satisfies the given interval constraint and One-Off condition, i.e., P's support is bigger than minimum support as settled, pattern P can be regarded as the frequent pattern. Under One-Off conditions, single-sequence pattern mining satisfies Apriori properties: all non-empty sub-pattern of frequent patterns must be frequent, while the parent patterns of infrequent patterns must be infrequent.

3.2 Algorithm Steps

Before describing steps of the proposed FOFM algorithm, we describe the how to calculate support value of pattern in Algorithm 1. Afterwards, the proposed FOFM algorithm can be described as four steps. During the first step, the proposed method scans the original sequence to obtain all patterns with length one and records their positions, which forms the set of to-be-connected patterns. In the second step, the proposed method tries to connect two patterns in the set of to-be-connected patterns by firstly traversing the pre-sequence, and then judging whether the position in the pre-sequence and post-sequence satisfy the interval constraint. If satisfied, position of the current post-sequence is saved as the new pattern. If not satisfied, the proposed method would continue to perform traversing and judging steps. After generating patterns, the proposed method would clear the set of to-be-connected pattern and calculate support values of generated patterns with steps described in Algorithm 1. The proposed method would save pattern as frequent pattern by judging whether its support value is larger than minimum support value requirement. If not, The proposed method would save these patters into the set of to-be-connected patterns and repeat step 2 until all sequences are used.

Algorithm 1. Calculating supporting value.

```
sup = 0 //Initialize the supporting value;
for each i = vec.size do
  if The position vec[i] is already used, the next position is matched. then
    continue;
  end if
  initialize list //Initialize a list of locations to record the occurrence of patterns
  that satisfy the condition
  list.add(vec[i])
  prevval = vec[i]
  TP = P.max_p refix // Get the pattern's maximum prefix pattern
  while TP! = "" do
    pvec = TP.list // Get the tail sequence of the largest prefix pattern
    Check if there is a position that meets the interval constraint requirements;
    if meets,add this location to your location list TP.list
    TP = TP.max_p refix //Use TP as the reference pattern to obtain the maxi-
    mum prefix pattern of TP
  end while
end for
```

4 Prefix-Projected Based-InterTransaction Mining Algorithm

The PB-ITM algorithm first perform preprocessing based on the characteristics of frequent water quality sequence, which is the output of the proposed FOFM

algorithm. After preprocessing, the proposed algorithm generates frequent item set by considering time characteristics of water quality data. Finally, the proposed algorithm generate association rules, which can be regarded as knowledge on relationship between water quality events.

4.1 Data Preprocessing

Data preprocessing consists of three steps, namely frequency sequence transactionalization, sliding window for processing, and data reduction.

Frequency Sequence Transactionalization. The main purpose of frequency sequence transactionization is to transform multiple water quality time sequence into transactional data, which can be easily processed by latter steps.

Water quality can be classified into six categories based on surface water quality standards of China, i.e., I, II, III, IV, V and bad V class. In order to mine association rules of water quality sequence, it is necessary to use characters to represent water quality sequence, where we show the transformation rule in Table 1.

Table 1. Water quality characterization based on categories

Water quality categories	The corresponding representing character
I class	1
II class	2
III class	3
IV class	4
V class	5
Bad V class	6

Given a frequency water quality sequence $S = \{s_1, s_2, s_3, \ldots s_{n-1}, s_n\}$, we define its corresponding variation sequence as $S' = \{e_1 = s_2 - s_1,\ e_2 = s_3 - s_2, \ldots, e_{n-2} = s_{n-1} - s_{n-2},\ e_{n-1} = s_n - s_{n-1}\}$. In sequence S', $e > 0$, $e < 0$ and $e = 0$ implies water quality is decreasing, improving and remains constant, respectively. With multiple variant values to represent water quality, we can represent $S' = \{e_{i,j} | i = 1, ..., n; j = 1, ..., m\}$, where i implies time and j refers to the category of multiple variants. In fact, multiple dimensions variants describe attributes associated with water quality events, such as when the event occurred or where the event occurred. Above all, frequency water quality sequence transactionization converts time-varying sequence into multiple variation sequence.

Sliding Window for Pre-processing. In reality, relevant researchers generally only care about the changing association in water quality over time. When mining water quality time sequence data with time attributes, it is necessary to consider the time interval between events. In order to avoid mining unnecessary patterns that do not meet the time requirements, the proposed method uses a sliding window method to process the transaction database. The size of the sliding window W is w, W is composed of multiple transactions in the transaction database, the interval of the dimension attributes of these transactions is less than w from the beginning to the end. Where $W[i](0 \leq i \leq w-1)$ is called the child window of W.

After drawing multiple windows by the method of sliding the window, each of them can generate a new set. The number of events in the transaction database is u, then merge set $M = \{e_i(j)|e_i \in W[j], 1 \leq i \leq u, 0 \leq j \leq w-1\}$. The Expanded event, also known as expanded item, is shaped as $e_i(j)$, which is an event in the merge set that has expanded information, and call the set of all expanded events as $\sum'$. A transaction database consisting of multiple merge sets is called a cross-transaction database.

The original transaction database is transformed into multiple merge sets by sliding window method. These merge sets are separated from each other, avoiding mining the patterns that are not interested. The expanded events in the merged set retain the relative dimensional information, so that the mining process retains the relevant characteristics which the researcher is interested in. Figure 1 illustrates the processing of the transaction database sliding window, where the size of sliding window is 4.

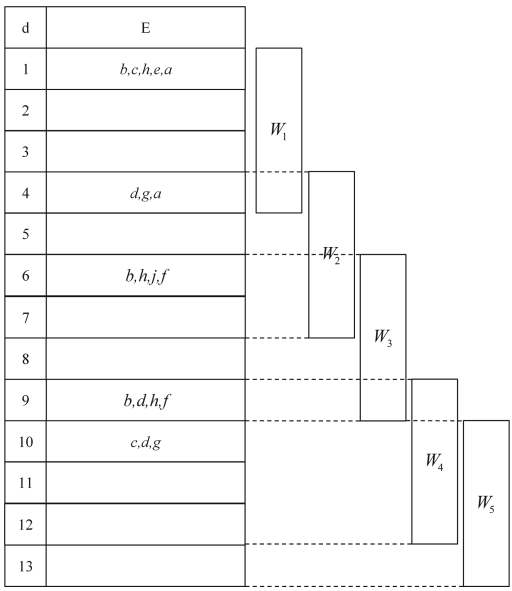

Fig. 1. Sliding window of the transaction database.

Data Reduction. The data reduction method is used to deal with the cross-transaction database formed by the water quality time sequence. The specific strategy is: in the process of generating the expansion item in the sliding window, determining whether the current event is a constant event, and if the event is a constant event, the event is not added to the appropriate window.

4.2 The Generation of Frequent Item Set

The part is to generate frequent pattern from the cross-transaction database that get by data preprocessing. It accelerates the mining efficiency of frequent patterns by expanding only from the item, whose sliding window index is 0. In the process of obtaining the suffix from the prefix projection database, the PB-ITM algorithm uses pseudo-projection technology, it only save the items that can be expanded and their position information. The specific steps of generating frequent pattern are as follows:

4.3 The Generation of Association Rules

The generation of association rules is the last step of mining association rules. The goal is to generate corresponding association rules based on the frequent patterns.

The cross-transaction association rule is shaped like $X \Longrightarrow Y$, where X and Y are both subsets of the expansion item set $\sum'$, and $X \cap Y = \emptyset$. $T(X)$ is the supporting value of the item set X in the cross-transaction database, and $T(XY)$ is the supporting value of the item set $X \cup Y$, and the confidence level for the association rule pattern of cross-transactional databases is $T(X)/T(XY)$.

The main steps of generating the association rule are as follows:

Algorithm 2. Generating frequent item set.

$freitem1 = getfreitem1(db)$; //Initialize frequent item sets
$time0item = gettime0item(freitem1)$; //Obtain an extension with time information of 0
for each s in db **do**
  if s is not frequent **then**
    Delete s from db;
  end if
end for
for each s in $time0item$ **do**
  Create a prefix for the projection database;
  Build an initial list of extended location information for each item with a sliding window value of 0;
  Start mode expansion Recursively;
end for

Algorithm 3. Generating association rules.

for each lk in $itemsets$ **do**
  initialize list //Initialize the list of items that meet the minimum confidence;
  for each j in lk **do**
    Stop association rule generation if time information is unreasonable
    Calculate the confidence of the association rule $(lk - item) \Longrightarrow item$
    Store the item if the rule meets the minimum support requirement
  end for
end for

5 Experiments

5.1 Dataset

This paper experiments on water quality time single sequence mode and water quality time sequence association rules. For the experiment of water quality time single sequence mode, we choose the water quality time sequence data of Nanjing Tuqiao 2007–2016. The water quality time sequence of the three water quality sites, Tuqiao, Liangyi and Xinyanggang, were selected. The length of the water quality time sequence of the Tuqiao in Nanjing is 521, the length of the water

Table 2. Partial mining results of Tuqiao water quality time sequence from 2007 to 2016.

Serial number	Pattern	Support	Weeks of occurrence
1	443333	10	81∼88, 130∼141, 146∼171, 188∼196, 252∼261, 392∼399, 465∼475, 503∼510
2	466664	10	15∼30, 40∼46, 95∼132, 176∼184, 227∼237, 330∼336, 365∼374
3	666666	10	21∼27, 45∼51, 97∼104, 108∼115, 116∼123, 124∼129, 177∼182, 218∼224, 225∼234, 367∼372
4	22433	10	248∼258, 304∼317, 407∼415, 424∼430, 436∼445, 459∼470, 500∼509, 515∼521
5	6634	10	4∼11, 77∼83, 103∼107, 181∼184, 233∼237, 272∼279, 284∼290, 334∼339, 360∼365, 478∼484
6	5434	10	7∼15, 36∼40, 78∼83, 90∼95, 102∼107, 131∼135, 151∼158, 259∼264, 375∼389
7	6434	12	27∼35, 80∼89, 104∼107, 129∼135, 182∼189, 233∼242, 253∼260, 335∼339, 372∼377, 390∼398, 478∼484
8	3464	10	14∼18, 39∼46, 94∼100, 106∼113, 258∼264, 287∼290, 314∼319, 328∼336, 387∼393, 475∼479

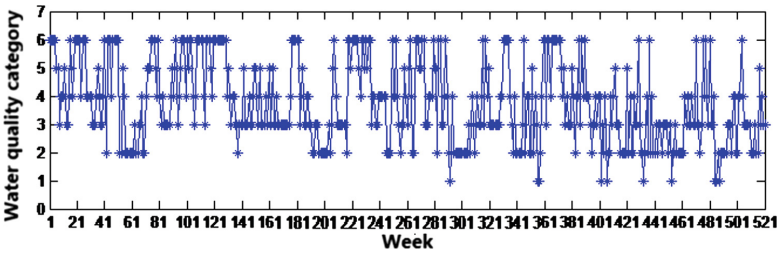

Fig. 2. Changes in water quality of Nanjing Tuqiao from 2007 to 2016.

quality time sequence of Liangyi in Dongtai is 511, and the length of the water quality time sequence of Xinyang Port in Yancheng is 509. The water quality time sequences are arranged weekly. The weekly water quality category is based on the relevant surface water environmental quality standards. The results are comprehensively evaluated by pH, dissolved oxygen, permanganate index and ammonia nitrogen. The experimental data of the water quality time sequence association rules are from 8 water quality monitoring sites in the Taihu Basin. The eight water quality monitoring stations are Lanshanzui, Dapu Port, Baiyu Port, Wujin Port, Zhihu Port, Shazhu, Wangting and Niaozuiqiao.

5.2 Result

The water quality are divided into six categories. In order to better mine the water quality time sequence, different characters are used here to represent different water quality categories, and the corresponding relationship is shown in Table 1:

This section selects the water quality time sequence data of Nanjing Tuqiao from 2007 to 2016. The minimum interval is 0, the maximum interval is 2, and the minimum support is 10. The experiment uses the FOFM algorithm for mining. The results of some mining are shown in Table 2:

The water quality change of Nanjing Tuqiao from 2007 to 2016 is shown in Fig. 2.

In order to compare the relevant algorithms, the water quality time sequence of the three water quality sites, Tuqiao, Liangyi and Xinyanggang, in 2007–2016 were selected. In the experiment, the control variable method was used to test the running time of the FOFM algorithm, OFMI algorithm, I-OFMI algorithm and the number of mining patterns under different interval constraint lengths or different minimum support. The minimum support is set to 6, 8, 10, 12, 14, 16, 18, and the interval constraint is set to $[0, 2]$.

By mining the pattern of the water quality time sequence from the three locations, it can be seen from the Fig. 3 that as the minimum support set increases, the fewer the number of models mined, the less time it takes to run. For the mining of water quality time sequence in the same place, if given the same minimum supporting value, the number of patterns mined by the method proposed

Table 3. Partial mining results of Tuqiao water quality time series from 2007 to 2016.

Serial number	Association rule	Support	Confidence level
1	Baidu Port's water quality deterioration(0), Wujin Port's water quality deterioration(1) $\Longrightarrow$ Wangting's water quality deterioration(3)	8	0.57
2	Lanshanzui's water quality improvement(0), Wujin Port's water quality improvement(0), Wangting's water quality improvement(2) $\Longrightarrow$ Niaozuiqiao's water quality improvement(3)	6	0.75
3	Baidu Port's water quality improvement(0), Lanshanzui's water quality improvement(1) $\Longrightarrow$ Niaozuiqiao's water quality improvement(3)	8	0.57
4	Shazhu's water quality improvement(0), Lanshanzui's water quality improvement(1) $\Longrightarrow$ Niaozuiqiao's water quality improvement(3)	7	0.5
5	Baidu Port's water quality improvement(0), Zhihu Port's water quality improvement(0) $\Longrightarrow$ Niaozuiqiao's water quality improvement(3)	7	0.64
6	Zhihu Port's water quality deterioration(0), Wujin Port's water quality deterioration(1) $\Longrightarrow$ Wangting's water quality deterioration(3)	10	0.59
7	Shazhu's water quality deterioration(0), Dapu Port's water quality deterioration(2) $\Longrightarrow$ Lanshanzui's water quality deterioration(3)	7	0.54

in this paper is basically the same as the I-OFMI algorithm, but far more than the OFMI. In terms of running time, the time used in the proposed method is the least among the three methods. In summary, the water quality sequence pattern mining method proposed in this paper has higher efficiency than other algorithms, while ensuring the high completeness of the patterns.

In the association rule mining experiment, the parameter sliding window size is 4, the minimum support is 5, and the minimum confidence is 0.5. The experimental results are shown in Table 3.

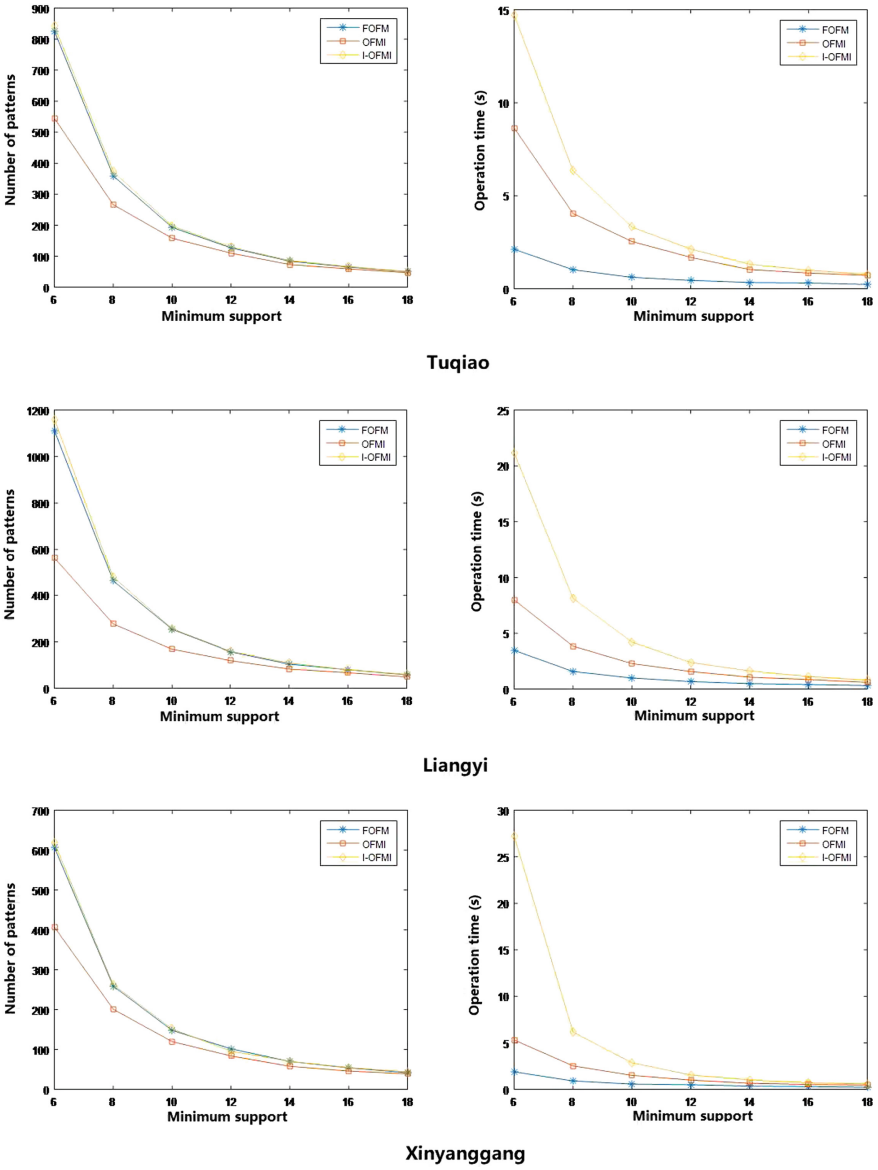

Fig. 3. Comparison of different methods.

Through the mining of the association rules between the water quality time sequence of the eight sites, we can see from the above table, in the requirements of higher minimum support and minimum confidence, the method proposed in this paper can accurately mine the association rules with time characteristics in the water quality time sequence, it fully consider the relationship between events

and events, and solve the main problem: the water quality time series has high complexity and the association rules are difficult to mine.

6 Conclusion

This paper proposes two time-varying water quality analysis methods based on pattern mining, sequence pattern mining and association rule mining for water quality data. Through theoretical and experimental analysis, first, the sequence pattern mining method FOFM has high efficiency while ensuring the completeness of pattern mining. Then, the association rule mining method PB-ITM considers the relationship and time characteristics of the transaction, so that the reliability of the prediction result is higher. In the next work, we will increase the experiment. Based on, to further improve the model and improve model performance.

References

1. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proceedings of 20th International Conference Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
2. Agrawal, R., Srikant, R., et al.: Mining sequential patterns. In: ICDE, vol. 95, pp. 3–14 (1995)
3. Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: Sequential pattern mining using a bitmap representation. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429–435. ACM (2002)
4. Chen, G., Wu, X., Zhu, X., Arslan, A.N., He, Y.: Efficient string matching with wildcards and length constraints. Knowl. Inf. Syst. **10**(4), 399–419 (2006)
5. Chiu, D.Y., Wu, Y.H., Chen, A.L.: An efficient algorithm for mining frequent sequences by a new strategy without support counting. In: Proceedings of 20th International Conference on Data Engineering, pp. 375–386. IEEE (2004)
6. Cook, D.J., Holder, L.B.: Substructure discovery using minimum description length and background knowledge. J. Artif. Intell. Res. **1**, 231–255 (1993)
7. Fournier-Viger, P., Lin, J.C.W., Kiran, R.U., Koh, Y.S., Thomas, R.: A survey of sequential pattern mining. Data Sci. Pattern Recogn. **1**(1), 54–77 (2017)
8. Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C., Philip, S.Y.: HUOPM: high-utility occupancy pattern mining. IEEE Trans. Cybern. **99**, 1–14 (2019)
9. Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C., Yu, P.S.: A survey of parallel sequential pattern mining. arXiv preprint arXiv:1805.10515 (2018)
10. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. ACM SIGMOD Rec. **29**, 1–12 (2000)
11. He, Y., Wu, X., Zhu, X., Arslan, A.N.: Mining frequent patterns with wildcards from biological sequences. In: 2007 IEEE International Conference on Information Reuse and Integration, pp. 329–334. IEEE (2007)
12. Inokuchi, A., Washio, T., Motoda, H.: An Apriori-based algorithm for mining frequent substructures from graph data. In: Zighed, D.A., Komorowski, J., Żytkow, J. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 13–23. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45372-5_2

13. Inokuchi, A., Washio, T., Motoda, H.: Complete mining of frequent patterns from graphs: mining graph data. Mach. Learn. **50**(3), 321–354 (2003)
14. Ji, X., Bailey, J., Dong, G.: Mining minimal distinguishing subsequence patterns with gap constraints. Knowl. Inf. Syst. **11**(3), 259–286 (2007)
15. Kuramochi, M., Karypis, G.: Frequent subgraph discovery. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 313–320. IEEE (2001)
16. Lepping, J.: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2018)
17. Li, C., Wang, J.: Efficiently mining closed subsequences with gap constraints. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 313–322. SIAM (2008)
18. Lin, J.C.W., Gan, W., Fournier-Viger, P., Hong, T.P., Chao, H.C.: FDHUP: fast algorithm for mining discriminative high utility patterns. Knowl. Inf. Syst. **51**(3), 873–909 (2017)
19. Nijssen, S., Kok, J.N.: A quickstart in frequent structure mining can make a difference. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 647–652. ACM (2004)
20. Pei, J., et al.: PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings 17th International Conference on Data Engineering, pp. 215–224. IEEE (2001)
21. Srikant, R., Agrawal, R.: Mining sequential patterns: generalizations and performance improvements. In: Apers, P., Bouzeghoub, M., Gardarin, G. (eds.) EDBT 1996. LNCS, vol. 1057, pp. 1–17. Springer, Heidelberg (1996). https://doi.org/10.1007/BFb0014140
22. Yan, X., Gspan, J.: Graph-based substructure pattern mining. In: Proceedings of 2002 International Conference Data Mining (ICDM 2002), pp. 721–724 (2001)
23. Zaki, M.J.: SPADE: An efficient algorithm for mining frequent sequences. Mach. Learn. **42**(1–2), 31–60 (2001)
24. Zhang, M., Kao, B., Cheung, D.W., Yip, K.Y.: Mining periodic patterns with gap requirement from sequences. ACM Trans. Knowl. Discovery Data (TKDD) **1**(2), 7 (2007)
25. Zhu, X., Wu, X.: Mining complex patterns across sequences with gap requirements. A... A 1(S2), S3 (2007)
26. Zou, X., Zhang, W., Liu, Y., Cai, Q.: Study on distributed sequential pattern discovery algorithm. J. Softw. **16**(7), 1262–1269 (2005)