УДК 338.242.2

ПОШУК ПОСЛІДОВНИХ ШАБЛОНІВ У ІНДЕКСАХ ЯКОСТІ СТАНУ ТЕХНОГЕННОГО РЕГІОНАЛЬНОГО ВИРОБНИЦТВА

А.В. Сергієнко. Східноукраїнський національний університет імені Володимира Даля, м. Сєвєродонецьк

Стаття отримана редакцією 16.12.2016 р.

Вступ. Більшість з відомих на сьогоднішній день способів інтелектуального аналізу даних полягають у визначенні корисних закономірностей, які часто зустрічаються у великих наборах даних. Одним з найбільш популярних методів інтелектуального аналізу даних є пошук послідовних шаблонів, що являють собою впорядковані послідовності, які найбільш часто зустрічаються, у наборах даних. Наприклад, послідовні шаблони можуть бути використані для аналізу споживчого кошика, у моделях управління запасами, для аналізу шаблонів доступу до веб-сайтів, виявлення певних часових послідовностей у наукових експериментах, стихійних лихах, лікуваннях захворювань, аналізу послідовностей ДНК та ін. У статті розглянуто пошук послідовних шаблонів в індексах якості стану техногенного регіонального виробництва.

Огляд останніх джерел досліджень і публікацій. Задача пошуку послідовних шаблонів може мати деякі варіації, серед яких багатовимірний пошук асоціативних правил [11], пошук закритих шаблонів [5, 15], пошук максимальних шаблонів [2], пошук обмежених шаблонів [4, 6], пошук наближених шаблонів [13].

Проблема пошуку послідовних шаблонів у наборах даних була вперше описана в роботі [1]. Статистична значимість шаблону, яка має назву підтримки, була визначена як відсоток послідовностей даних, що містять необхідний шаблон. У роботі [16] введено такі поняття, як мінімальна підтримка, максимальна підтримка і ковзне вікно.

У роботі [12] виявленим послідовним шаблонам надано назву так званих «епізодів», вони могли мати різний тип упорядкування: повний, частковий або нульовий. «Епізоди» були отримані протягом аналізу однієї послідовності подій, і їх статистична значимість була виміряна як відсоток кількості входжень від загальної кількості «епізодів».

У роботі [9] був запропонований алгоритм для пошуку послідовних шаблонів у розширеній базі даних. Іншим внеском стала пропозиція методу для обробки інтерактивних запитів з метою пошуку послідовних шаблонів. Мета була реалізована шляхом додавання до алгоритму «попереднього» кроку, під час якого визначалися шаблони з низьким рівнем підтримки, котрі надалі зберігались у формі, придатній для ефективного пошуку відповідно до заданих користувачем умов запиту.

Інший підхід до виявлення послідовних шаблонів був представлений у роботі [18]. Алгоритм, наведений у ній, міг застосовуватися не тільки для розширених, але й для зменшених баз даних. Він вимагав деякої додаткової інформації, яка буде зберігатися разом зі знайденими шаблонами.

У роботі [14] розглянуто питання інтерактивного пошуку асоціативних правил, а також запропоновано концепцію «сховища». Воно призначалося для зберігання наборів даних, котрі часто зустрічались і були виявлені під час обробки інших запитів. Важливим внеском був алгоритм, який використовує набори даних, виявлених для високого рівня підтримки у процесі розв'язання задачі, але з більш низьким порогом підтримки. Набори даних, що часто зустрічаються і були виявлені у попередніх завданнях, зберігалися у «сховищі» й використовувалися для визначення рівня підтримки деяких кандидатів без їх додаткової звірки з базою даних. Хоча метод був запропонований для наборів даних, які часто зустрічаються, він також може бути використаний і для пошуку послідовних шаблонів.

Ідея попереднього обчислення наборів даних, що часто зустрічаються, у багатовимірних базах даних і їх використання під час виявлення асоціативних правил у всій базі даних або в її частині обговорювалась у роботі [19]. Запропонований метод використовував таку властивість, що одна послідовність міститься в іншій послідовності тоді й тільки тоді, якщо всі набори даних першої послідовності містяться в наборах даних іншої послідовності.

Постановка завдання. Корисні знання, які були отримані під час пошуку послідовних шаблонів для індексів якості стану техногенного регіонального виробництва, дозволяють прийняти рішення про залежність одних індексів від інших, причому таку залежність можна виявити не тільки в межах одного регіону чи одного індексу, але й визначити міжрегіональні або міжіндексні часові зв’язки.

Основний матеріал і результати. З початку 1980-х рр. у промислово розвинених країнах намітилася тенденція підкреслювати роль регіональних суб'єктів політики в питаннях економічного розвитку. Термін «регіон» широко використовується в багатьох сферах, будь то економіка, політика чи громадська сфера. Територія регіону зазвичай менша, ніж територія країни, до якої він входить.

Регіони в межах країни можуть бути визначені на основі ряду характеристик, що відрізняють їх від інших регіонів, починаючи з адміністративних меж і закінчуючи загальними географічними, культурними й соціально-економічними особливостями, такими як ландшафт, клімат, мови, етнічне походження, спільна історія та ін. Межі регіонів, засновані на цих характеристиках, рідко збігаються з більш точними межами, що визначаються законодавчими актами, тому межі регіонів зазвичай історично були визначені як компроміс між адміністративними та іншими характеристиками. Ідентифікація меж регіонів залежить від обраних критеріїв. Як правило, в основі обстеження в регіональній географії визначаються гомогенні функціональні області.

Техногенне регіональне виробництво – економічна діяльність суб'єктів господарювання всіх видів і форм власності у межах одного регіону з використанням різних технологій виробництва, корисним результатом якої є випуск продукції або надання послуг з урахуванням екологічного та соціального ефекту, тобто позитивним або негативним впливом економічної діяльності на навколишнє середовище та соціум, а кращі результати у забезпеченні сталості зростання техногенного регіонального виробництва досягаються завдяки поєднанню економічної ефективності з екологічною та соціальною результативністю.

Індекс якості стану техногенного регіонального виробництва – агрегований багатоступеневий відносний показник, який розраховується за статистичними даними, змінюється в межах від 0 до 1 та відображає місце досліджуваного об'єкта серед інших подібних йому об'єктів.

Визначення вимірюваних індексів якості стану техногенного регіонального виробництва є важливим пунктом у реалізації концепції сталого розвитку, бо ці індекси можуть пов'язувати всі складові техногенного регіонального виробництва [3, 8, 10, 17, 20 – 22]. Міжнародні організації, такі як ООН, Світовий санк, Організація країн економічного співробітництва та розвитку, Європейська комісія, Науковий комітет з проблем навколишнього середовища тощо активно займаються розробленням вимірюваних критеріїв сталого розвитку. Світовий банк можна назвати світовим лідером за індикаторами сталого розвитку. Розроблення критеріїв сталого розвитку вимагає збору й обробки значної кількості статистичних даних, отримати які буває дуже складно, а часом і неможливо, тому ця процедура є комплексною та дорогою. До того ж обробка великих масивів інформації, агрегування цієї інформації в індикатори й індекси інколи бувають неможливими через відсутність певних вхідних статистичних даних, а одним з основоположних принципів проведення моделювання є відсутність пропусків значень у масивах вхідних даних.

Концептуальну модель формування індексів якості стану техногенного регіонального виробництва можна зобразити у вигляді багаторівневої структури, де на найнижчому рівні знаходяться «сирі» вхідні дані, на найвищому рівні – безпосередньо індекс сталого розвитку, а для агрегування та згортки величин використовуються вагові коефіцієнти відповідного рівня (рис. 1). До індексів якості стану техногенного регіонального виробництва належать економічний, екологічний, соціальний, гуманітарний індекси, а також індекс сталого розвитку та індекс гармонійності, які формуються з трьох попередніх індексів.

Рис. 1. Концептуальна модель формування індексу сталого розвитку

Послідовність – це впорядкований список [7]. Аналіз послідовних шаблонів є одним з найбільш поширених методів інтелектуального аналізу даних, що часто використовується у багатьох

дослідженнях. Найбільш серйозною проблемою в інтелектуальному аналізі даних завжди був довгий час обробки запитів користувачів, і коли йдеться про обробку значних інформаційних масивів, то сучасні системи інтелектуального аналізу даних витрачають хвилини або навіть години, щоб відповісти на простий запит. Задача пошуку послідовних шаблонів аналогічна до задачі виявлення асоціативних правил, але, на відміну від асоціативних правил, у послідовних шаблонах наявні часові залежності. Пошук послідовних шаблонів у часових рядах допомагає збільшити кількість і якість знань, вилучених з даних, підвищити їхню якість, продуктивність та отримати кращий ефект від даних під час дослідження. Метод пошуку послідовних шаблонів у своєму використанні є досить простим, зручним і легким у розумінні та інтерпретації.

Завдання пошуку послідовних шаблонів є логічним продовженням завдання пошуку асоціативних правил, тобто його логічним ускладненням: під час пошуку асоціативних правил залишається невизначеним часовий аспект послідовностей. Якщо в асоціативних правилах розглядаються предметні набори, то в послідовних шаблонах досліджуються вже впорядковані списки предметних наборів. До того ж, завдання пошуку послідовних шаблонів можна розглядати на різних часових рівнях, послідовності не обов'язково мають бути безперервними. Для характеристики послідовних шаблонів зазвичай використовуються такі поняття, як підтримка, достовірність та довжина правила, або кількість елементів, з яких воно складається. Послідовності, що задовольняють мінімальним значенням підтримки, називають частими.

Задача пошуку послідовних шаблонів для визначення зв'язків між індексами якості стану може розглядатися на різних рівнях, коли задаються різні вхідні умови.

Для пошуку послідовних шаблонів в індексах якості стану було сформовано таблиці, які містять значення приросту кожного з восьми індексів: економічного, екологічного, соціального, гуманітарного, індексу сталого розвитку та індексу гармонійності. Нові таблиці охоплюють значення за період з 2001 р. по 2014 р. Усі значення в них було прологарифмовано, помножено на коефіцієнт, що дорівнює 100, та ранжовано у зростаючій послідовності, від меншого до більшого. Крок інтервалу зміни значень прийнято таким, що дорівнює 0,1, й усім інтервалам надано значення, починаючи з 1. Таким чином, кожне логарифмоване значення з таблиці має ціле числове значення інтервалу, до якого воно входить. Вхідні дані, модифіковані за такою схемою, зведено в таблиці: вони придатні для програмного пошуку послідовних шаблонів. Ці таблиці також містять часову змінну, котра приймає значення в інтервалі від 1 до 14, де 1 – значення для 2001 року, 14 – значення для 2014 року відповідно.

У табл. 1 – 2 наведені отримані послідовні шаблони у міжіндексному розрізі, а також у розрізі кожного індексу. Також у таблицях подано значення підтримки до достовірності знайдених послідовних шаблонів. Зазначимо, що довжина всіх правил, зазначених у таблицях, дорівнює одному, тобто вони складаються з одного елемента.

Таблиця 1

Послідовні шаблони для індексів у розрізі кожного індексу

№ з/п	Індекс	Інтервал зміни індексів						Підтримка правила	Достовірність правила
		нижня межа інтервалу			верхня межа інтервалу				
		тип зміни	від	до	тип зміни	від	до		
1	IDX_R^{SOC}	збільш.	0,4%	0,5%	зменш.	0,2%	0,3%	14,8	57,1
2	IDX_R^{SOC}	зменш.	1,1%	1,2%	збільш.	1,8%	1,9%	11,1	100,0
3	IDX_R^{SD3}	збільш.	0,7%	0,8%	зменш.	0,5%	0,6%	11,1	50,0
4	IDX_R^{SD3}	збільш.	0,9%	1,0%	зменш.	0,3%	0,4%	11,1	60,0
5	IDX_R^{SD3}	збільш.	0,1%	0,2%	збільш.	0,1%	0,2%	18,5	71,4
6	IDX_R^{SD4}	зменш.	0,8%	0,9%	зменш.	0,8%	0,9%	11,1	50,0

Таблиця 2

Послідовні шаблони для індексів у міжіндексному розрізі

№ п/п	Інтервал зміни індексів						Підтримка правила	Достовірність правила
	нижня межа інтервалу			верхня межа інтервалу				
	тип зміни	від	до	тип зміни	від	до		
1	зменшення	1,2%	1,3%	зменшення	0,7%	0,8%	100,0	100,0
2	зменшення	1,2%	1,3%	збільшення	1,1%	1,2%	100,0	100,0

Продовження таблиці 2

2	зменшення	1,2%	1,3%	збільшення	1,1%	1,2%	100,0	100,0
3	зменшення	1,1%	1,2%	збільшення	1,1%	1,2%	100,0	100,0
4	зменшення	1,0%	1,1%	зменшення	1,1%	1,2%	100,0	100,0
5	зменшення	1,0%	1,1%	зменшення	0,7%	0,8%	100,0	100,0
6	зменшення	0,1%	0,2%	зменшення	1,1%	1,2%	100,0	100,0
7	зменшення	0,1%	0,2%	зменшення	0,7%	0,8%	100,0	100,0
8	зменшення	0,1%	0,2%	збільшення	0,3%	0,4%	100,0	100,0
9	зменшення	0,1%	0,2%	збільшення	1,1%	1,2%	100,0	100,0
10	збільшення	0,3%	0,4%	зменшення	1,1%	1,2%	100,0	100,0
11	збільшення	0,9%	1,0%	зменшення	1,1%	1,2%	100,0	100,0
12	збільшення	0,9%	1,0%	зменшення	0,9%	1,0%	100,0	100,0
13	збільшення	0,9%	1,0%	зменшення	0,7%	0,8%	100,0	100,0

Висновки. Таким чином, послідовні шаблони дозволяють виявити часові зв'язки між індексами якості стану, що надалі може бути використано при моделюванні. Пошук послідовних шаблонів для індексів якості стану техногенного регіонального виробництва дозволяє виявити корисні закономірності, які як додаткова вхідна інформація можуть бути використані під час прогнозування. Послідовні шаблони дозволяють виявити часові зв'язки між змінними не тільки в межах певного індексу, але за їх допомогою можна також визначити міжіндексні часові зв'язки.

ЛІТЕРАТУРА:

1. Agrawal R. Mining Sequential Patterns / R. Agrawal, R. Srikant // Proc. of the 11th ICDE Conf. / R. Agrawal, R. Srikant. – Washington, DC, USA: IEEE Computer Society, 1995. – P. 3 – 14.

2. Bayardo R. J. Efficiently mining long patterns from databases / R. J. Bayardo // In Proceedings of ACM-SIGMOD international conference on management of data (SIGMOD'98) / R. J. Bayardo. – Seattle, WA, 1998. – P. 85 – 93.

3. Becker G. S. The Quantity And Quality Of Life And The Evolution Of World Inequality / G.S. Becker, T. J. Philipson, R. R. Soares. // American Economic Review. – 2005. – №95. – P. 277 – 291.

4. Bonchi F. On closed constrained frequent pattern mining / F. Bonchi, C. Lucchese // In Proceedings of International conference on data mining (ICDM'04) / F. Bonchi, C. Lucchese. – Brighton, UK, 2004. – P. 35 – 42.

5. Discovering frequent closed itemsets for association rules / N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal // Proceeding of the 7th international conference on database theory (ICDT'99) / N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal. – Jerusalem, Israel, 1999. – P. 398 – 416.

6. Grahne G. Efficient mining of constrained correlated sets / G. Grahne, L. Lakshmanan, X. Wang // Proceeding of the 2000 international conference on data engineering (ICDE'00) / G. Grahne, L. Lakshmanan, X. Wang. – San Diego, CA, 2000. – P. 512 – 521.

7. Han J. Frequent pattern mining: current status and future directions / J. Han, H. Cheng, D. Xin. // Data Mining and Knowledge Discovery. – 2007. – №15. – P. 55 – 86.

8. Human Development Reports [Електронний ресурс]. – Режим доступу: http://www.hdr.undp.org/en.

9. Incremental and Interactive Sequence Mining / S. Parthasarathy, M. J. Zaki, M. Ogihara, S. Dwarkadas // Proceedings of the eighth international conference on Information and knowledge management / S. Parthasarathy, M. J. Zaki, M. Ogihara, S. Dwarkadas. – 1999. – P. 251 – 258.

10. Index of Economic Well-Being. Introduction and methodology [Електронний ресурс]. – Режим доступу: http://www.csls.ca/iwb.asp.

11. Kamber M. Metarule-guided mining of multi-dimensional association rules using data cubes / M. Kamber, J. Han, J. Y. Chiang // Proceeding of the 1997 international conference on knowledge discovery and data mining (KDD'97) / M. Kamber, J. Han, J. Y. Chiang. – Newport Beach, CA, 1997. – P. 207 – 210.

12. Mannila H. Discovering frequent episodes in sequences / H. Mannila, H. Toivonen, A.I. Verkamo // First International Conference on Knowledge Discovery and Data Mining (KDD'95) / H. Mannila, H. Toivonen, A. I. Verkamo. 1995. – P. 210 – 215.

13. Mining approximate frequent itemsets in the presence of noise: algorithm and analysis / [J. Liu, S. Paulsen, X. Sun and etc.] // Proceeding of the 2006 SIAM international conference on data mining (SDM'06) / [J. Liu, S. Paulsen, X. Sun and etc.]. – Bethesda, MD, 2006. – P. 405 – 416.

14. Nag B. Using a Knowledge Cache for Interactive Discovery of Association Rules / B. Nag, P.M. Deshpande, D.J. DeWitt // Proc. of the 5th KDD Conference / B. Nag, P. M. Deshpande, D. J. DeWitt. 1999.

15. Pei J. CLOSET: an efficient algorithm for mining frequent closed itemsets / J. Pei, J. Han, R. Mao // Proceedings of ACM-SI0MOD international workshop data mining and knowledge discovery (DMKD'00) / J. Pei, J. Han, R. Mao. – Dallas, 2000.

16. Srikant R. Mining Sequential Patterns: Generalizations and Performance Improvements / R. Srikant, R. Agrawal // Proc. of the 5th EDBT Conference / R. Srikant, R. Agrawal. – 1996.

17. Talberth J. The Genuine Progress Indicator 2006. A Tool for Sustainable Development [Електронний ресурс] / J. Talberth, C. Cobb, N. Slattery. – 2007. – Режим доступу: http://rprogress.org/publications/2007/GPI%202006.pdf.

18. Wang K. Incremental Discovery of Sequential Patterns / K. Wang, J. Tan // ACM-SIGMOD's 96 Data Mining Workshop: On Research Issues on Data Mining and Knowledge Discovery / K. Wang, J. Tan. – 1996.

19. Wojciechowski M. Itemset Materializing for Fast Mining of Association Rules / M. Wojciechowski, M. Zakrzewicz // Proc. of the 2nd ADBIS Conference / M. Wojciechowski, M. Zakrzewicz. – 1998.

20. Артеменко В. Б. Методи інтегральної оцінки якості життя населення в управлінні регіональним розвитком / В. Б. Артеменко. // Регіональна економіка. – 2002. – №1. – С. 166 – 176.

21. Данилишин Б. М. Екологічна складова політики сталого розвитку: Монографія / Б.М. Данилишин. – Донецьк: ТОВ «Юго-Восток, Лтд», 2008. – 256 с.

22. Порохня В. М. Моделювання людського потенціалу держави: монографія / В. М. Порохня, В.В. Бирський. – Запоріжжя : КПУ, 2008. – 200 с.

УДК 338.242.2

Аліса Валентинівна Сергієнко, здобувач. Східноукраїнський національний університет імені Володимира Даля. **Пошук послідовних шаблонів в індексах якості стану техногенного регіонального виробництва.** Розглянуто проблему використання послідовних шаблонів для отримання нових знань із послідовностей індексів якості стану. Обґрунтовано доцільність використання послідовних шаблонів для визначення часових зв'язків усередині послідовностей індексів або між ними. Визначено перспективи використання послідовних шаблонів при моделюванні управління регіоном.

Ключові слова: послідовні шаблони, індекси якості стану, техногенне регіональне виробництво, управління регіоном.

UDC 338.242.2

Alice Serhiienko, applicant. East Ukrainian Volodymyr Dahl National University. **Sequential Patterns Searching in the Quality Status Indexes of Technogenic Regional Production**. It is considered the problem of use of sequential patterns for obtaining the new knowledge from the sequences of the quality status indexes. The article deals also with the feasibility of using of sequential patterns to determine the time connections between the sequences of indexes or inside them. The author defined the prospects of use of sequential patterns during the modeling the management for the region are.

Keywords: sequential patterns, the quality status index, technogenic regional production, management for the region.

УДК 338.242.2

Алиса Валентиновна Сергиенко, соискатель. Восточноукраинский национальный университет имени Владимира Даля. **Поиск последовательных шаблонов в индексах качества состояния техногенного регионального производства.** Рассмотрена проблема использования последовательных шаблонов для получения новых знаний из последовательностей индексов качества состояния. Определена целесообразность использования последовательных шаблонов для определения временных связей внутри последовательностей индексов или между ними. Доказаны перспективы использования последовательных шаблонов при моделировании управления регионом.

Ключевые слова: последовательные шаблоны, индексы качества состояния, техногенное региональное производство, управление регионом.