EKSPLORACJA WZORCÓW SEKWENCJI I INTER-SEKWENCJI W DUŻYCH BAZACH DANYCH

Anh Nguyen

Katedra Informatyki Stosowanej, Politechnika Wrocławska, Polska

1 Wprowadzenie

Eksploracja wzorców inter-sekwencji to rozszerzenie eksploracji sekwencyjnych wzorców, które polega na odkrywaniu wspólnych wzorców, powiązań i zależności między sekwencjami w bazie danych sekwencyjnych. Identyfikuje wzorce wspólne nie tylko w ramach tej samej transakcji, ale także między transakcjami. Problem eksploracji wzorców inter-sekwencji po raz pierwszy został zaproponowany przez Wang et al. w 2009 roku (Wang & Lee, 2009), a od tego czasu opracowano kilka algorytmów, w tym DBV-ISP (Vo et al., 2012), ISP-IC, iISP-IC i pISP-IC (Le et al., 2018), DBV-ISPMIC oraz pDBV-ISPMIC (Nguyen et al., 2023).

2 Problem badawczy

Problem eksploracji wzorców inter-sekwencji stanowi rozszerzenie eksploracji sekwencyjnego, obejmujące trzy rozszerzenia, w tym zbiorów elementów, sekwencje i inter-sekwencje. Podczas pracy z dużymi bazami danych sekwencyjnych i dużą wartością maxspan określoną przez użytkownika, generowana jest duża liczba wzorców kandydackich. Dlatego problem eksploracji wzorców inter-sekwencji wymaga optymalizacji procesu generowania kandydatów oraz optymalizacji przestrzeni przechowywania, co przedstawiono w Tabeli 1.

Tabela 1. Porównanie Zalet i Wad Proponowanego Eksploracji Wzorców Inter-sekwencji

Referencje	Struktura danych	Zaleta	Wada
Wang & Lee (2009)	PatternList	Problem wzorców inter-sekwencji zostaje przedstawiony po raz pierwszy. EISP-Miner wymaga tylko jednego skanowania bazy danych i może ograniczyć operacje łączenia i liczenia wsparcia do niewielkiej liczby list wzorców, unikając tym samym kosztownego dopasowania inter-sekwencji podzbiorów. Wykazuje także większą efektywność w porównaniu do M-Apriori.	Generowane jest wiele wzorców, co wymaga znacznej ilości przestrzeni magazynowej.

Vo et al. (2012)	DBV-PatternList	Algorytm DBV-ISP wykorzystuje strukturę danych DBV-PatternList do przechowywania informacji o częstych wzorcach inter-sekwencji, co prowadzi do mniejszego zużycia pamięci w porównaniu z algorytmem EISP-Miner. Dodatkowo obliczanie jest przyspieszone, ponieważ opiera się na bitach.	Algorytm wykorzystuje zestaw sąsiednich bitów do przechowywania informacji o pozycji wzorca. Dlatego, gdy wzorzec jest rzadko spotykany w bazie danych, pozycje, które są nieobecne wewnątrz wzorca, muszą być reprezentowane przez wartość bitową 0. Skutkuje to wykorzystaniem większej ilości pamięci niż jest to konieczne.
T. Le et al. (2018)	DBV-PatternList	Algorytm ISP-IC stosuje warunek elementu do problemu ekstrakcji wzorców inter-sekwencji, co pozwala na ograniczenie liczby generowanych wzorców inter-sekwencji.	Algorytm nie rozwiązał jeszcze ograniczeń struktury danych DBV-PatternList.

2.1 Cel Pracy Doktorskiej

Celem pracy doktorskiej jest rozwiązanie problemów eksploracji wzorców inter-sekwencji pod względem czasu przetwarzania i przestrzeni przechowywania. W rozprawie doktorskiej zaproponowano nową strukturę danych do przechowywania wzorców inter-sekwencji, mając na celu zminimalizowanie powielania danych w trakcie procesu eksploracji. Dodatkowo wprowadza model eksploracji wzorców inter-sekwencji z ograniczeniami zbiorów elementów w celu zredukowania liczby generowanych kandydatów, co przyspiesza wyszukiwanie i przetwarzanie istotnych informacji. Ponadto praca doktorska przedstawia dodatkowe propozycje mające na celu zwiększenie efektywności proponowanych metod i algorytmów.

2.2 Cele Pracy Doktorskiej

Cele tej pracy doktorskiej to:

1. Propozycja rozwiązania problemu eksploracji wzorców inter-sekwencji z ograniczeniami zbiorów elementów, przedstawienie algorytmu DBV-ISPMIC.
2. Opracowanie optymalnego podejścia do rozwiązywania problemu eksploracji wzorców inter-sekwencji z ograniczeniami zbiorów elementów, prezentacja algorytmu *p*DBV-ISPMIC.

3. Wprowadzenie metody optymalizacji przestrzeni przechowywania w kontekście problemu eksploracji wzorców inter-sekwencji, wykorzystując algorytm ISP-PI (Eksploracja Wzorców Inter-Sekwencji na podstawie Pseudo-Indeksu).
4. Propozycja techniki przycinania kandydatów w problemie eksploracji wzorców inter-sekwencji, włączając metodę ISP-IC (Eksploracja Wzorców Inter-Sekwencji z Kontrolą Przecięcia Indeksów).
5. Przeprowadzenie oceny zaproponowanych algorytmów i metod poprzez eksperymenty przy użyciu baz danych rzeczywistych pochodzących z magazynu danych społeczności zajmującej się eksploracją danych. Porównanie wyników przeprowadzonych eksperymentów na podstawie czasu pracy algorytmu i wymagań dotyczących zużycia pamięci.

3 Wkład Pracy Doktorskiej

Na podstawie celów pracy doktorskiej główne wkłady są przedstawione w dwóch sekcjach i krótko przedstawione poniżej:

3.1 Pierwszy Wkład Obejmuje Cele 1 i 2 Pracy Doktorskiej

Na podstawie problemów eksploracji wzorców inter-sekwencji, takich jak EISP-Miner (Wang & Lee, 2009), DBV-ISP (Vo et al., 2012) oraz ISP-IC (Le et al., 2018), a także problemów eksploracji sekwencyjnych wzorców z ograniczeniami zbiorów elementów, takich jak MSPIC-DBV (Van et al., 2018), wprowadzony został problem eksploracji wzorców inter-sekwencji z ograniczeniami zbiorów elementów o nazwie DBV-ISPMIC (Nguyen et al., 2023). Algorytm wykorzystuje strukturę danych o nazwie DBV-PatternList do przechowywania kandydatów, a także strukturę drzewa o nazwie ISP-Tree do przechowywania częstych wzorców. Ponadto zaproponowana metoda szybkiej weryfikacji warunków generowanego zbioru elementów kandydujących, stosuje eksploracja równoległe, w celu przyspieszenia algorytmu.

Struktura danych DBV-PatternList optymalizuje przechowywanie informacji o kandydatach. Zamiast używać typu danych liczbowych do reprezentowania informacji o wzorcach, DBV-PatternList wykorzystuje strukturę danych bitową typu bit-wektor. Informacja o wzorcu jest wskazywana poprzez włączanie lub wyłączanie bitów, co pozwala na przechowywanie większej ilości informacji o kandydacie za pomocą typu danych liczbowych, co redukuje przestrzeń potrzebną do przechowywania kandydatów.

Sprawdzanie ograniczeń zbiorów elementów dla wszystkich generowanych próbek jest czasochłonne dla algorytmu. Zaproponowana metoda szybkiej weryfikacji, czy generowany kandydat spełnia ograniczenia zbiorów elementów, wykorzystując informacje o warunkach pochodzące od wzorców nadrzędnych, które go wygenerowały. To zmniejsza czas pracy algorytmu.

Problem eksploracji wzorców inter-sekwencji wykorzystuje strukturę ISP-Tree do przechowywania generowanych częstych wzorców, przetwarzając algorytm zgodnie z metodą przechodzenia w głąb drzewa. Ponieważ obsługa gałęzi drzewa jest oddzielna, przedstawiana jest technika przetwarzania równoległego gałęzi drzewa. Pozwala to na optymalizację czasu pracy algorytmu poprzez przetwarzanie wielu gałęzi jednocześnie.

Ewaluacja wyników eksperymentalnych jest przedstawiona w artykule [R2].

3.2 Drugi Wkład Obejmuje Cele 3 i 4 Pracy Doktorskiej

W oparciu o algorytmy eksploracji wzorców inter-sekwencji, takie jak EISP-Miner i DBV-ISP, proponowane jest wprowadzenie nowego algorytmu o nazwie ISP-PI. Algorytm ma na celu optymalizację modeli eksploracji danych w kontekście eksploracji wzorców inter-sekwencji przy użyciu struktury danych znanej jako Pseudo-IDList. ISP-PI rozwiązuje problemy poprzednich algorytmów związane z powielaniem danych. Zamiast wymagać przechowywania wszystkich informacji o kandydacie, możliwe jest odzyskiwanie jego informacje z oryginalnego wzorca. Ta metoda kompresuje wartości pozycji generowanych kandydatów, umożliwiając odtworzenie wartości z oryginalnych wzorców tworzących kandydatów, co eliminuje konieczność zapisywania wszystkich pozycji.

Ponadto algorytm wykorzystuje metodę przycinania o nazwie ISP-IC, aby skutecznie zmniejszyć liczbę generowanych kandydatów. Ta optymalizacja poprawia czas pracy i przestrzeń przechowywania, co jest istotne ze względu na rosnący wolumen gromadzonych danych. Algorytm ISP-PI efektywnie kompresuje dane w celu zminimalizowania przestrzeni przechowywania i wykorzystuje przycinanie kandydatów do przyspieszenia czasu pracy algorytmu w wydobywaniu wzorców inter-sekwencji.

Ewaluacja wyników badawczych, a także proponowanego algorytmu, jest przedstawiona w artykule [R1].

4 Publikacje

[R1]. **Nguyen, A.**, Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). An Efficient Pruning Method for Mining Inter-sequence Patterns based on Pseudo-IDList. *Expert Systems with Applications* (Accepted, Impact Factor: 8.5, Punktacja MEiN: 200)

[R2]. **Nguyen, A.**, Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). Mining inter-sequence patterns with Itemset constraints. *Applied Intelligence*, *53*(17), 19827–19842. https://doi.org/10.1007/S10489-023-04514-7 (Impact Factor: 5.3, Punktacja MEiN: 70)

[R3]. Nguyen, T. T. D., Nguyen, L. T. T., **Nguyen, A.**, Yun, U., & Vo, B. (2021). A method for efficient clustering of spatial data in network space. *Journal of Intelligent & Fuzzy Systems*, *40*(6), 11653–11670. https://doi.org/10.3233/JIFS-202806 (Impact Factor: 1.737, Punktacja MEiN: 70)

[R4]. Huynh, H. M., Nguyen, L. T. T., Vo, B., **Nguyen, A.**, & Tseng, V. S. (2020). Efficient

methods for mining weighted clickstream patterns. *Expert Systems with Applications*, *142*, 112993. https://doi.org/10.1016/j.eswa.2019.112993 (Impact Factor: 6.954, Punktacja MEiN: 140)

[R5]. Nguyen, L. T. T., Nguyen, T. D. D., **Nguyen, A.**, Tran, P.-N., Trinh, C., Huynh, B., & Vo, B. (2020). Efficient Method for Mining High-Utility Itemsets Using High-Average Utility Measure. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12496 LNAI*, 305–315. https://doi.org/10.1007/978-3-030-63007-2_24 (Core ranking: C, Punktacja referatów z wykazu MEiN: 20)

[R6]. Nguyen, L. T. T., Vo, B., Nguyen, T. N., & **Nguyen, A.** (2019). Mining class association rules on imbalanced class datasets. *Journal of Intelligent and Fuzzy Systems*, *37*(6), 7131–7139. https://doi.org/10.3233/JIFS-179326 (Impact Factor: 1.851, Punktacja MEiN: 70)

[R7]. Le, T., **Nguyen, A.**, Huynh, B., Vo, B., & Pedrycz, W. (2018). Mining constrained inter-sequence patterns: a novel approach to cope with item constraints. *Applied Intelligence*, *48*(5), 1327–1343. https://doi.org/10.1007/s10489-017-1123-9 (Impact Factor: 2.882, Q2)

5 Referencje

Le, T., Nguyen, A., Huynh, B., Vo, B., & Pedrycz, W. (2018). Mining constrained inter-sequence patterns: a novel approach to cope with item constraints. *Applied Intelligence*, *48*(5), 1327–1343. https://doi.org/10.1007/s10489-017-1123-9

Nguyen, A., Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). Mining inter-sequence patterns with Itemset constraints. *Applied Intelligence*, 1–16. https://doi.org/10.1007/S10489-023-04514-7

Van, T., Vo, B., & Le, B. (2018). Mining sequential patterns with itemset constraints. *Knowledge and Information Systems*, *57*(2), 311–330. https://doi.org/10.1007/s10115-018-1161-6

Vo, B., Tran, M. T., Hong, T. P., Nguyen, H., & Le, B. (2012). A dynamic bit-vector approach for efficiently mining inter-sequence patterns. *Proceedings - 3rd International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA 2012*, 51–56. https://doi.org/10.1109/IBICA.2012.31

Wang, C. S., & Lee, A. J. T. (2009). Mining inter-sequence patterns. *Expert Systems with Applications*, *36*(4), 8649–8658. https://doi.org/10.1016/j.eswa.2008.10.008

7/11/2023