***Гаврилюк А.М. магістрант***

***Науковий керівник – к.т.н. доцент Левус Є.В.***

*Національний університет «Львівська політехніка», м. Львів*

*Кафедра програмного забезпечення****.***

**ДОСЛІДЖЕННЯ АЛГОРИТМУ APRIORI ТА ЙОГО МОДИФІКАЦІЙ**

Розглядається задача аналізу великих даних у предметній області – виникнення дорожньо-транспортних пригод, з метою зниження їх появи у майбутньому. Застосування методів пошуку асоціативних правил надає можливість знаходити приховані закономірності виникнення ДТП, базуючись на інформації про транспортні пригоди, які вже відбулися.

Для пошуку асоціативних правил використовується алгоритм, запропонований Agrawal and Srikant (1994), який ідентифікує елементи/одиниці, що часто повторюються у базі, і розширює їх список до все більших множин з дотриманням правила достатньої частотності [1].

Для реалізації алгоритму пошуку асоціативних правил[2] для аналізу ДТП, приймемо такі позначення.

Нехай *I = {i1, i2, …, in }* – множина всіх можливих характеристик, що аналізуються, де *in –* характеристика, що призвела до виникнення ДТП, *n* – потужність множини всіх можливих характеристик транспортних пригод. *D = {d1, d2, …, dm}* – множина транзакцій, яка піддається аналізу, де *di* – транзакція, яка є підмножиною *I(di* ⊆ *I)* та описує окреме ДТП, де *m* – потужність множини транзакцій.

Правило *X => Y* справедливе з вірогідністю *C = conf(X => Y)* відсотку транзакцій з *D*(які містять *X* та *Y*), що можна визначити як

*conf(X => Y) = supp(X ∪ Y)/sup(X)*

Тоді задача пошуку асоціативних правил для аналізу ДТП полягає у знаходженні всіх асоціативних правил *X => Y*, де *X* та *Y* – набори характеристик з множини всіх можливих ДТП *I*, що мають задані користувачем коефіцієнти.

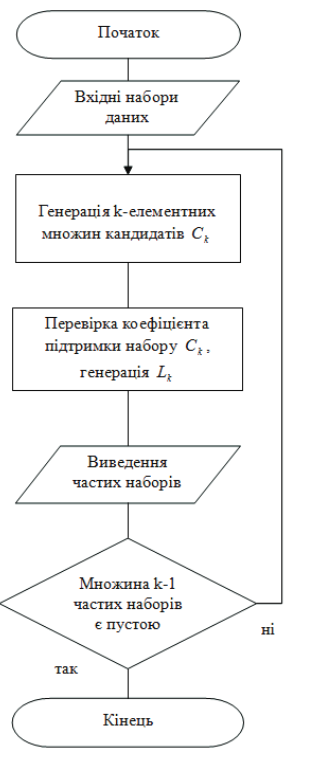


Рис. 1. Схема алгоритму apriori

При застосуванні пошуку асоціативних правил для аналізу ДТП виникають проблеми:

1. Збільшення потужності множини всіх можливих ДТП.
2. Збільшення часу для аналізу даних.
3. Низька інформаційність отриманого результату.

Запропоновано наступні модифікації для розв'язання цих проблем:

1. Розбиття множини всіх можливих транспортних пригод *I = {Is, Id*} на дві підмножини:[3]

* *Is* – множина характеристик, дані якої для нашого випадку є абсурдними або загальновідомими фактами.
* *Id* – множина характеристик, дані якої для нашого випадку будуть цікавими.

Асоціативні правила, що містять лише підмножину із множини Is не містять корисної інформації, тому на кожному кроці після генерації кандидатів можна відсікати кандидатів, для яких *Ck ∩ Id = null.*

У результаті застосування такого підходу збільшується швидкість роботи алгоритму та ступінь інформаційності отриманих асоціативних правил.

1. Додати ще одну характеристику асоціативних правил Ліфт[4] (англ. lift) - це відношення частоти появи умови в транзакціях, які містять й умову, і наслідок, до частоти появи наслідку в цілому. Та обчислювати її на кожному кроці генерації кандидатів. Обраховуючи Ліфт на кожному кроці алгоритму можна здійснити більше відсікання абсурдних або випадкових груп правил, отже швидкість алгоритму повинна зрости адже не потрібно здійснювати наступні кроки по вже відкинутих правилах.

Література

1. Agrawal R., Imielinski T., Swami A. Minning Association Rules Between Sets of Items in Large Databases // SIGMOD Conference 1993: 207-216

2. Опис асоціативних правил studfiles.net – Режим доступу: https://studfiles.net/preview/4494757/page:2/ (дата звернення: 09.09.2018).

3. Т.О.Савчук, К.В.Щепановський Застосування алгоритму apriori для аналізу надзвичайних ситуацій на залізничному транспорті. Пулікація 2010, ст. 1-6.

4. Tong Qiang, Zhou Yuanchun, Wu Kaichao, Yan Baoping. A quantitative association rules mining algorithm[J]. Computer engineering. 2007