ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 004.89+656.2

Т. О. САВЧУК, К. В. ЩЕПАНОВСЬКИЙ

Вінницький національний технічний університет, Вінниця

ЗАСТОСУВАННЯ АЛГОРИТМУ APRIORI ДЛЯ АНАЛІЗУ НАДЗВИЧАЙНИХ СИТУАЦІЙ НА ЗАЛІЗНИЧНОМУ ТРАНСПОРТІ

Анотація: В даній роботі проаналізовано проблеми, що виникають при застосуванні пошуку асоціативних правил для аналізу надзвичайних ситуацій на залізниці. Запропоновано модифікацію алгоритму Apriori, що базується на розділенні множини характеристик надзвичайної ситуації на дві підмножини, яка зменшує час роботи алгоритму.

Ключові слова: асоціативні правила, apriori, надзвичайні ситуації, пошук закономірностей

Вступ

При збільшенні об'ємів вантажоперевезень за допомогою залізничного транспорту виникає проблема ефективного аналізу причин виникнення надзвичайних ситуацій (НС). Тому важливою є розробка засобів, що орієнтовані на виявлення причин виникнення надзвичайних ситуацій на залізничному транспорті для зниження загрози їх появи в майбутньому. Використання новітніх інформаційних технологій надає можливість виконувати пошук закономірностей в інформації про наявні надзвичайні ситуації, що є особливо актуальним і передбачає знаходження інформації про сукупність характеристик, що підвищують ризик виникнення надзвичайних ситуацій.

Застосування методів пошуку асоціативних правил надає можливість знаходити приховані закономірності виникнення надзвичайних ситуацій, базуючись на інформації про надзвичайні ситуації, що вже відбулись. Такий підхід не потребує виконання складних аналітичних розрахунків. Необхідною є лише наявність формалізованої інформації про надзвичайні ситуації, які вже відбулись.

Актуальність

Однією із проблем при застосуванні алгоритмів пошуку асоціативних правил є їх обчислювальна складність. Дана задача є NP-повною [2], що виключає можливість розробки ефективних алгоритмів із поліноміальним часом роботи. Тому для зменшення часу роботи алгоритмів використовуються різного роду евристики. Відповідно, в задачі пошуку асоціативних правил для аналізу надзвичайних ситуації можна виділити характеристики, що дозволяють збільшити ефективність алгоритму.

Мета

Об'єктом дослідження є процес формування логічних залежностей на основі формалізованих даних про надзвичайні ситуації на залізничному транспорті при перевезенні шкідливих та небезпечних вантажів у вигляді асоціативних правил. Предметом дослідження є алгоритми генерації асоціативних при аналізі надзвичайних ситуацій на залізничному транспорті. Метою дослідження є підвищення ефективності пошуку логічних залежностей в базах даних надзвичайних ситуацій на залізничному транспорті на основі розробки алгоритмів пошуку асоціативних правил з урахуванням властивостей існуючих алгоритмів при їх застосуванні для аналізу надзвичайних ситуацій на залізниці.

Постановка задачі

Для розробки алгоритму пошуку асоціативних правил для аналізу надзвичайних ситуацій на залізничному транспорті, приймемо такі позначення.

Нехай $I = \{i_1, i_2, \ldots, i_n\}$ – множина всіх можливих характеристик надзвичайних ситуацій, що аналізуються, де i_j – j-а характеристика, $j = \overline{1,n}$, де n – потужність множини всіх можливих характеристик надзвичайних ситуацій. $D = \{d_1, d_2, \ldots, d_m\}$ – множина транзакцій, яка піддається аналізу, де d_i – транзакція, яка є підмножиною $I (d_i \subseteq I)$ та описує окрему надзвичайну ситуацію на залізничному транспорті, $i = \overline{1,m}$, де m – потужність множини транзакцій [2].

Правило $X \Rightarrow Y$ справедливе з достовірністю $C = conf(X \Rightarrow Y)$ відсотку транзакцій з D (які містять X та Y), що можна визначити як

$$conf(X \Rightarrow Y) = supp(X \cup Y) / supp(X).$$

Тоді, задача пошуку асоціативних правил для аналізу надзвичайних ситуацій на залізничному транспорті полягає у знаходженні всіх асоціативних правил $X \Rightarrow Y$, де X та Y - набори характеристик з множини всіх можливих характеристик I, що мають задані користувачем коефіцієнти

підтримки S_0 та достовірності C_0.

Розв'язання задачі

Розглянемо найбільш поширені алгоритми пошуку асоціативних правил (табл. 1).

Таблиця 1 – Порівняльна характеристика алгоритмів пошуку асоціативних правил

Назва алгоритму	Опис	Переваги	Недоліки
Apriori[2,3,4]	Алгоритм призначений для знаходження всіх частих наборів елементів множини транзакцій, що аналізується.	Проста структура, висока ефективність	Ускладнена реалізація алгоритму для роботи із базами транзакцій, що не поміщаються в оперативну пам'ять
Eclat [5]	Алгоритм побудований на основі пошуку в глибину використовуючи перетин наборів елементів для знаходження частих наборів елементів.	Висока швидкодія на малих об'ємах даних	Значне збільшення часу роботи при збільшенні потужності множини характеристик НС.
FP-growth [6]	Алгоритм використовує розширене префіксне дерево [7] для збереження бази даних у стислому вигляді. Він застосовує метод «розділяй і володарюй» для декомпозиції і видобутку знань із бази даних.	Не використовується процес генерації частих кандидатів, що є основою в алгоритмі Apriori.	Важко застосовувати оптимізації для використання з специфічними наборами даних.
OPUS[6]	Алгоритм, що на відміну від більшості альтернативних алгоритмів не потребує визначення ступіні мінімальної підтримки асоціативних правил.	Працює ефективніше, ніж алгоритм Apriori	Практична реалізація алгоритму для обробки великих об'ємів інформації ускладнена

Отже, враховуючи переваги та недоліки розглянутих алгоритмів пошуку асоціативних правил, як основу для аналізу та удосконалення обрано алгоритм Apriori.

Проте, при застосуванні пошуку асоціативних правил для аналізу надзвичайних ситуацій на залізниці виникають наступні проблеми:

1. Збільшення потужності множини всіх можливих характеристик НС I.
2. Збільшення часу необхідного для аналізу даних.
3. Низька інформативність отриманого результату.

Збільшення часу роботи алгоритму безпосередньо витікає із збільшення потужності множини всіх можливих характеристик I. Низька інформативність отриманих асоціативних правил пояснюється тим, що в результаті роботи алгоритму буде отримано велику кількість асоціативних правил, які не будуть містити причинно-наслідкових зв'язків.

Для вирішення означених проблем запропоновано модифікацію алгоритму Apriori: розбиття множини всіх можливих характеристик НС $I = \{I_S, I_D\}$ на дві підмножини:

1. I_S –характеристики, сукупність яких могла стати причиною виникнення НС.
2. I_D – характеристики, сукупність яких описує НС що виникла.

В результаті застосування такого підходу збільшується швидкість роботи алгоритму та ступінь інформативності отриманих асоціативних правил.

Вхідними даними для алгоритму Apriori при аналізі НС на залізниці є множина транзакцій D, що описує наявні в базі даних надзвичайні ситуації, заданий коефіцієнт підтримки для правил S' та коефіцєнт достовірності C'. Множина всіх можливих характеристик НС I подається разом із множиною транзакцій D, або обчислюється на її основі [3].

Робота алгоритму відбувається ітеративно. Множинами кандидатів вважаються множини характеристик НС, які формуються для подальшої генерації на їх основі асоціативних правил. На кожній ітерації відбувається генерація k-елементних множин кандидатів $C_k \in I$, $k = \overline{1, N}$, де N – потужність I. Для згенерованоих наборів відбувається перевірка коефіцієнта підтримки $S = supp(C_k)$, і відсікаються ті набори, коефіцієнт підтримки для яких менше заданого коефіцієнта S'. Результатом є множина частих наборів L_k, де k – потужність множини L на поточній ітерації алгоритму, яка задовольняє поставленим вимогам. Після цього виконується процедура виведення частих наборів, яка на

основі L_k генерує асоціативне правило, приведене до вигляду $X \Rightarrow Y$. Якщо згенерована множина частих наборів L_k є пустою, алгоритм завершує свою роботу, оскільки при продовженні роботи всі наступні згенеровані множини кандидатів $C_k \in I$, $k = \overline{1, N}$ матимуть потужність k більшу ніж для поточної ітерації, і відповідно жоден набір гарантовано не пройде перевірку $suðð(C_k) \geq S'$. Це ґрунтується на таких властивостях підтримки:

1. Ступінь підтримки будь-якої множини елементів не може перевищувати підтримку будь-якої її підмножини.
2. Для будь-якого набору L_k її підтримка буде менше, ніж підтримка наборів із множини L_{k-1}.

Означені умови підвищують швидкість алгоритмом роботи алгоритму до прийнятної для оперативної обробки рівня, не знижуючи при цьому точність отриманого результату [4] . Схема алгоритму Apriori наведена на рисунку 1.

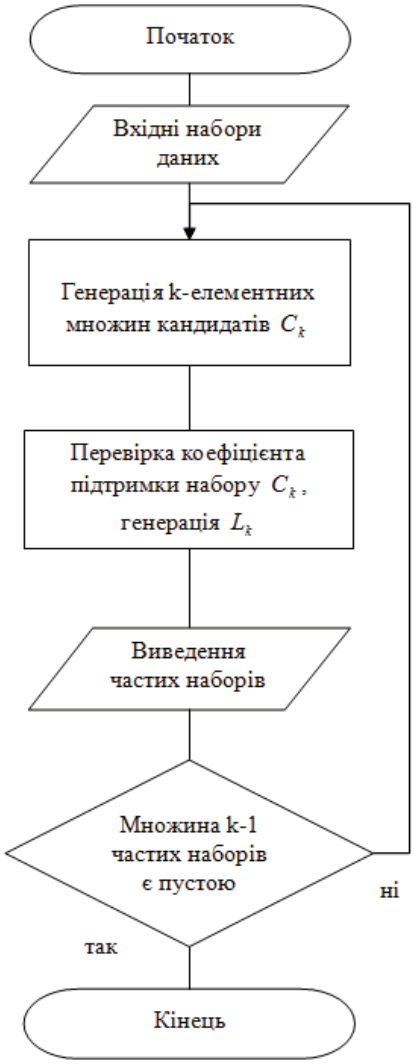

Рисунок 1 – Схема алгоритму Apriori

Актуальним залишається відсікання відомих або неінформативних результатів на етапі пошуку частих наборів кандидатів. Для цього в структуру алгоритму вводяться модифікації, пов'язані із особливостями представлення даних для опису надзвичайних ситуацій на залізничному транспорті. Можна ввести обмеження щодо параметрів та результатів роботи алгоритму, що дозволяє збільшити наглядність отриманого результату та зменшити час обробки даних. Для цього застосовується розбиття множини всіх можливих характеристик НС $I = \{I_S, I_D\}$ на дві підмножини, де I_S – сукупність

характеристик, що що могли стати причинами виникнення НС, I_D – сукупність характеристик, що описують результат надзвичайної ситуації.

Асоціативні правила, що містять лише підмножину характеристик із множини I_S, не містять корисної інформації для виконання аналізу надзвичайних ситуацій на залізничному транспорті, завдяки чому можна прискорити алгоритм. Для цього необхідно змінити процес генерації наборів кандидатів C_k, в якому відсікати набори, що не містять характеристик із множини I_D. В результаті значно зменшиться кількість наборів кандидатів C_k, що генеруватиметься на кожній ітерації алгоритму, а також це впливатиме на генерацію частих наборів L_k. Час обробки множини наборів C_k та генерації L_k лінійно залежить від потужності даних множин і загальної потужності множини I, яка є сталою в процесі роботи алгоритму. Тому завдяки зменшенню кількості елементів множини C_k зменшиться загальний час роботи алгоритму. Схема модифікованого алгоритму Apriori для аналізу надзвичайних ситуацій на залізничному транспорті наведена на рисунку 2.

Для порівняння ефективності алгоритму Apriori та модифікованого нами алгоритму виконаємо оцінку кількості наборів характеристик НС, що генеруються алгоритмами. Час роботи алгоритму Apriori визначається числом наборів елементів (характеристик НС при аналізі НС на залізничному транспорті), що мають заданий рівень підтримки і на їх основі генеруються асоціативні правила (успішні набори: набір зустрічається щонайменше у k транзакціях), та числом наборів елементів що підраховуються але не використовуються (невдалі набори: всі підмножини набору зустрічаються щонайменше у k транзакціях, проте весь набір зустрічається менше ніж у k транзакціях). Кількість успішних наборів залежить лише від вхідних даних. Кількість невдалих наборів визначається як вхідними даними, так і алгоритмом.

Припустимо, що потужність найбільшого набору дорівнює l, а загальна кількість елементів - m, $m = |I|$. При цьому значення l можна вважати номером ітерації, оскільки на кожній ітерації l збільшується на одиницю. Нехай існує така ітерація l, що задовольняє двом вимогам:

1. На даній ітерації l виконано перевірки всіх наборів, і знайшлось вдалі набори
2. На ітерації ($l+1$) всі набори виявились невдалими

Відповідно до умови роботи алгоритму, після завершення даної ітерації відбувається завершення процедури генерації частих наборів. Оцінимо кількість наборів, що буде згенерована на ітерації l:

$$N_l = C_l^m$$

На ітерації $l+1$ буде згенеровано така кількість наборів:

$$N_{l+1} = C_{l+1}^m.$$

Визначимо кількість згенерованих наборів елементів у найгіршому випадку, коли всі набори є успішними:

$$N = \sum_{i=1}^{k} C_i^m. \quad (1)$$

Однак дана оцінка не є точною, оскільки не враховуються евристики, що застосовуються в алгоритмі Apriori, і відповідно отримане значення буде значно перевищувати результат, отриманий за допомогою моделювання на реальних даних про НС на залізничному транспорті. Проте дану оцінку можна застосувати для порівняння базового алгоритму із наступними модифікованими версіями із застосуванням евристик.

Оцінимо кількість наборів елементів, які будуть згенеровані модифікованим алгоритмом Apriori. В загальному випадку зменшується як кількість успішних, так і кількість невдалих наборів – завдяки тому, що відсікаються набори-кандидати що не містять елементів із I_D. Кількість невдалих наборів також зменшується залежно від характеру вхідних даних. Слід зазначити, що зменшення кількості успішних наборів не впливає на кількість наборів, на основі яких генеруються інформативні асоціативні правила, оскільки відсікаються набори, на основі яких будуть згенеровані завідомо неінформативні асоціативні правила.

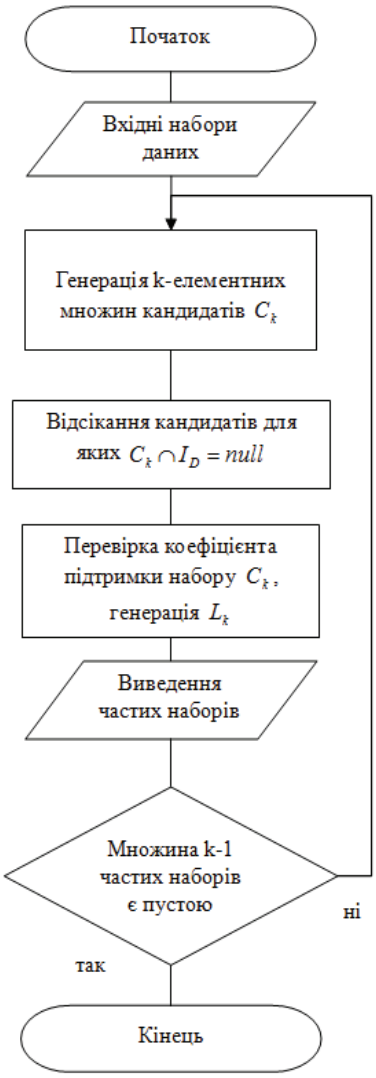

Рисунок 2 – Схема модифікованого алгоритму Apriori для аналізу надзвичайних ситуацій на залізниці

Введемо змінні $m_S = |I_S|$ та $m_D = |I_D|$. Нехай існує така ітерація l, що задовольняє таким вимогам:

1. На даній ітерації l виконано перевірки всіх наборів, і знайшлось вдалі набори C_l
2. На ітерації $l+1$ всі набори виявились невдалими
3. Існує $C_l \cap I_D \neq \varepsilon$
4. На ітерації $l+1$ не існує наборів $C_{l+1} \cap I_D \neq \varepsilon$

Відповідно до умови роботи алгоритму, після завершення даної ітерації відбувається завершення процедури генерації частих наборів. Оцінимо кількість наборів, що буде згенерована на ітерації l:

$$N_l = C_l^{m_S+m_D} - C_l^{m_S}$$

На ітерації $l+1$ буде згенеровано така кількість наборів:

$$N_{l+} = C_l^{m_S+m_D} - C_l^{m_S}$$

Визначимо кількість згенерованих наборів елементів у найгіршому випадку, коли всі набори є успішними:

$$N = \sum_{i=1}^{m_S+m_D} C_i^{m_S+m_D} - \sum_{i=1}^{m_S+m_D} C_l^{m_S} \qquad (2)$$

Використовуючи оцінки (1) та (2), можна коефіцієнт, в скільки разів збільшується швидкість

обробки даних при застосуванні модифікованого алгоритму Apriori в порівнянні із базовим алгоритмом:

$$K = \frac{\sum_{i=1}^{m_S+m_D} C_i^{m_S+m_D}}{\sum_{i=1}^{m_S+m_D} C_i^{m_S+m_D} - \sum_{i=1}^{m_S+m_D} C_l^{m_S}} \quad (3)$$

В результаті, в формулі (3) доведено збільшення швидкості роботи алгоритму пошуку асоціативних правил для аналізу надзвичайних ситуацій на залізниці у найгіршому випадку, оскільки знаменник формули гарантовано не може перевищувати значення чисельника. Відповідно, на будь-яких інших вхідних даних час їх обробки даних зменшиться.

Вивновки

Отже, при виконанні аналізу надзвичайних ситуацій на залізничному транспорті застосування пошуку асоціативних правил дає можливість виявляти сукупності характеристик, які підвищують ризик виникнення надзвичайної ситуації. При цьому виконується пошук частих наборів характеристик за допомогою модифікованого алгоритму Apriori, на основі яких будуються асоціативні правила, з якими в подальшому працюватиме аналітик. Щоб збільшити швидкість роботи алгоритму Apriori для аналізу надзвичайних ситуацій на залізничному транспорті використано особливості подання інформації про надзвичайні ситуації, а саме введено евристику яка дозволяє зменшити кількість обчислень та збільшити інформативність знайдених асоціативних правил. Також показано, що завдяки введеним модифікаціям зростає швидкість роботи базового алгоритму пошуку асоціативних правил.

Список використаної літератури

1. Аветисян В.Г., Сенчихін Ю.М., Кулаков С.В., Куліш Ю.О., Тригуб В.В. Організація аварійно-рятувальних робіт – навч. посіб. для студ. вузів III—IV рівнів акр. за напр. підг. «Пожежна безпека» – Харків, Університет цивільного захисту України, 2006.
2. Purdom P. W., Gucht D. V., Groth D. P. Average case performance of the apriori algorithm – SIAM Journal on Computing, 33(5):1223–1260, 2004.
3. Барсегян А. А., Купріянов М. С., Степаненко В. В., Холод І. І. Методи і моделі аналізу даних: OLAP і DATA MINING – БХВ-Петербург, 2004.-336 с.
4. Agrawal R., Imielinski T., Swami A. Mining Association Rules Between Sets of Items in Large Databases // SIGMOD Conference 1993: 207-216
5. Mohammed J. Zaki. Scalable algorithms for association mining – IEEE Transactions on Knowledge and Data Engineering, 12(3):372-390, May/June 2000.
6. Brin S., Rajeev Motwani, Ullman J., Tsur S.. Dynamic itemset counting and implication rules for market basket data // SIGMOD Conference 1997, Proceedings ACM SIGMOD International Conference on Management of Data, pages 255-264, Tucson, Arizona, USA, May 1997.
7. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы: построение и анализ, 2-е изд. : Пер с англ. – М. : Издательский дом «Вильямс», 2009. – 1296с.

Стаття надійшла до редакції 07.10.2010.

Відомості про авторів

Савчук Тамара Олександрівна – к.т.н., професор, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, e-mail: savchtam@vstu.vinnica.ua

Щепановський Костянтин Валентинович – магістр кафедри комп'ютерних наук, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, email: kostya.vntu@gmail.com.