УДК 004.65:004.85

ВИЯВЛЕННЯ ПРИХОВАНИХ ЗАКОНОМІРНОСТЕЙ У ТРАНЗАКЦІЙНИХ ДАНИХ МЕТОДАМИ АСОЦІАТИВНОГО АНАЛІЗУ

Крісілов В. А., д.т.н., професор
Національний університет «Одеська політехніка»
Україна, Одеса
softwareengineering121.2@gmail.com

Анотація. *Розглянуто виявлення асоціативних закономірностей у транзакційних даних за допомогою алгоритмів Apriori та FP-Growth. Проаналізовано відмінності способів пошуку частих наборів елементів, вплив порогів підтримки та особливості обробки розріджених і щільних транзакцій. Обґрунтовано необхідність оцінювати правила не лише за підтримкою та достовірністю, а й за показником підйому. Запропоновано критерії коректного порівняння алгоритмів.*
Ключові слова: *інтелектуальний аналіз даних; асоціативні правила; транзакційні дані; Apriori; FP-Growth.*

Вступна частина. Транзакційні дані відображають сукупності об’єктів, які з’являються разом у межах окремої події: товарів у кошику покупця, послуг в одному замовленні або операцій у сеансі користувача. Значний обсяг таких записів ускладнює виявлення повторюваних комбінацій звичайними засобами перегляду таблиць. Асоціативний аналіз дає змогу виявляти закономірності спільної появи елементів і формувати правила, придатні для подальшої інтерпретації.

Одним із базових підходів є алгоритм Apriori, який послідовно формує кандидатні набори та відсіює комбінації, що не досягають установленої мінімальної підтримки [1]. Альтернативний алгоритм FP-Growth використовує компактне подання частотних зв’язків і дає змогу уникати явного формування великої кількості кандидатів [2]. Проте практичний вибір між ними залежить від кількості транзакцій, різноманітності елементів, середньої довжини транзакції та порогових параметрів.

Мета роботи – здійснити порівняльний аналіз Apriori та FP-Growth для пошуку закономірностей у транзакційних даних, визначити умови доцільності кожного підходу й обґрунтувати критерії відбору змістовних асоціативних правил. Акцент зроблено на алгоритмічних особливостях і методиці оцінювання; неперевірені числові результати продуктивності не наводяться.

Основна частина. Транзакційну базу можна подати як множину записів, кожен із яких містить набір елементів. Частим називають набір, який зустрічається щонайменше в установленій частці транзакцій. Асоціативне правило описує зв’язок між двома неперетинними наборами: наявність елементів у лівій частині пов’язана з появою елементів у правій. Такий зв’язок є статистичним співіснуванням і сам по собі не доводить причинно-наслідкової залежності.

Алгоритм Apriori спирається на властивість антимонотонності підтримки: якщо набір не є частим, жоден його наднабір також не може бути частим. Завдяки цьому простір пошуку поступово звужується. Водночас для кожного рівня довжини наборів потрібні генерація кандидатів і підрахунок їхньої підтримки. За низького порога підтримки або великої кількості поширених елементів число кандидатів може швидко зростати, що збільшує час обробки та потребу в пам’яті.

FP-Growth застосовує іншу організацію обчислень. Часті елементи впорядковуються за частотою, після чого транзакції стискаються в дерево префіксів. Спільні частини транзакцій зберігаються разом, а пошук частих наборів виконується через умовні структури. Перевагою є відсутність явного перебору кандидатів у стилі Apriori. Однак компактність дерева залежить від подібності транзакцій: за великої різноманітності їхніх комбінацій ефект спільних префіксів може зменшуватися.

Порівняння алгоритмів не слід зводити до твердження про безумовну перевагу одного з них. Для коротких розріджених транзакцій із відносно високим порогом підтримки Apriori може залишатися достатньо простим і практичним рішенням. У наборах із багатьма частими комбінаціями FP-Growth зазвичай має переваги завдяки скороченню витрат на генерацію кандидатів. Водночас конкретний результат залежить від реалізації, способу зберігання даних і доступної оперативної пам'яті. Результати сучасного порівняльного дослідження підтверджують, що неоднорідність транзакційних даних істотно впливає на продуктивність алгоритмів пошуку частих наборів [4].

Важливою характеристикою є поріг мінімальної підтримки. Його підвищення скорочує кількість знайдених наборів і час пошуку, але може приховати рідкісні, проте корисні поєднання. Зниження порога збільшує охоплення закономірностей, водночас ускладнюючи обчислення та подальшу інтерпретацію. Отже, налаштування підтримки має відповідати завданню аналізу: масовим типовим комбінаціям або менш поширеним спеціалізованим зв'язкам.

Після виявлення частих наборів формується множина асоціативних правил. Для їх оцінювання використовують підтримку, достовірність і підйом (lift). Підтримка відображає частку транзакцій, у яких разом присутні всі елементи правила. Достовірність показує частку транзакцій із лівою частиною, що також містять праву. Підйом порівнює цю умовну частоту з поширеністю правої частини в усій базі. Значення підйому, близьке до одиниці, вказує на відсутність вираженої позитивної асоціації, навіть якщо достовірність здається високою [3].

Для пояснення розглянемо умовну сукупність із 100 транзакцій. Елемент A трапляється у 40 записах, B – у 50, а обидва разом – у 30. Підтримка правила $A \rightarrow B$ становить 30 %, достовірність – 75 %, підйом – 1,5. Це означає, що B зустрічається серед транзакцій із A частіше, ніж у вибірці загалом. Приклад ілюстративний і не є результатом експерименту на реальному наборі даних. Крім того, показник підйому не усуває впливу випадкових коливань у малих вибірках.

Правило з високою достовірністю не завжди має аналітичну цінність. Якщо правий елемент зустрічається майже в кожній транзакції, його висока умовна частота може бути тривіальною. Тому рекомендується застосовувати сукупність порогів підтримки, достовірності та підйому, а також оцінювати зрозумілість і практичну значущість правил. Для невеликих підгруп бажано додатково перевіряти кількість спостережень, щоб не надавати надмірної ваги випадковим збігам.

Під час аналізу великих наборів даних виникає проблема надлишковості результатів: кілька правил можуть описувати близькі або вкладені комбінації елементів. Їх механічне включення до звіту ускладнює інтерпретацію. Доцільно групувати правила за спільними елементами, відбирати ті, що мають найбільший підйом за достатньої підтримки, та окремо позначати правила, зміст яких дублює інші. Важливо не ототожнювати статистичну значущість із корисністю для прийняття рішень. Новітні огляди також наголошують на важливості рідкісних і негативних закономірностей, які можуть залишатися поза увагою стандартного пошуку частих комбінацій [5].

Для коректного експериментального порівняння Apriori та FP-Growth необхідно використовувати однакові набори транзакцій, ідентичні пороги підтримки та єдині правила попередньої обробки. Доцільно змінювати кількість транзакцій, число унікальних елементів і середню довжину запису, фіксуючи час пошуку частих наборів, пікове споживання пам'яті та кількість знайдених наборів. Окремо потрібно вимірювати час генерації асоціативних правил, щоб не змішувати витрати на два різні завдання. Сучасні дослідження паралельних алгоритмів підтверджують актуальність окремого оцінювання масштабованості обчислень [6].

Додатковим чинником є розрідженість даних. У розріджених наборах більшість елементів трапляється рідко, тому значна частина комбінацій відсіюється вже на ранніх кроках пошуку. У щільних наборах із повторюваними довгими транзакціями зростає кількість частих підмножин, і переваги стисненого представлення можуть проявлятися сильніше. Саме тому результати тестування на одному наборі даних не варто автоматично переносити на інші предметні області.

Практичні сценарії застосування охоплюють аналіз спільних покупок, комбінацій послуг, наборів подій інформаційної системи та взаємопов'язаних ознак у реєстрах. Однак отримані правила необхідно перевіряти з позиції предметної області: асоціація може відображати сезонність, особливості формування вибірки або вплив третього чинника. Використання правил для рекомендацій потребує додаткової перевірки на даних, які не застосовувалися для їх початкового виявлення. Для аналізу транзакцій, що містять чутливі відомості, додатковим обмеженням є захист приватності під час пошуку правил [7].

Отже, алгоритмічна ефективність і аналітична якість є двома окремими критеріями. Apriori та FP-Growth за однакових умов повинні знаходити ту саму множину частих наборів, якщо реалізують однакове визначення підтримки. Відмінності стосуються переважно організації пошуку та ресурсних витрат. Натомість змістовність правил визначається порогами, статистичними показниками, обсягом даних і можливістю предметної інтерпретації.

Висновки. Проведено порівняльний аналіз алгоритмів Apriori та FP-Growth для пошуку частих наборів у транзакційних даних. Установлено, що Apriori характеризується простотою та прозорим відсіканням кандидатів, тоді як FP-Growth уникає їх явного масового формування завдяки деревоподібному поданню транзакцій. Доцільність застосування алгоритмів залежить від розрідженості даних, довжини транзакцій і мінімальної підтримки.

Обґрунтовано, що відбір корисних асоціативних правил має ґрунтуватися на спільному аналізі підтримки, достовірності й підйому, а не на одному показнику. Визначено умови коректного майбутнього експерименту: однакові дані й пороги, вимірювання часу та пам'яті, окрема оцінка якості правил. Запропоновані критерії можуть використовуватися під час побудови систем інтелектуального аналізу транзакційних даних.

Використання інструментів ШІ: не використовувались.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499.

[2] Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 1–12. https://doi.org/10.1145/342009.335372

[3] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 255–264. https://doi.org/10.1145/253260.253325
[4] Trasierras, A. M., Luna, J. M., Fournier-Viger, P., & Ventura, S. (2024). Data heterogeneity's impact on the performance of frequent itemset mining algorithms. Information Sciences, 678, 120981. https://doi.org/10.1016/j.ins.2024.120981
[5] Gangaramani, D., & Londhe, R. (2025). Survey on association rule analysis: Exploration using mining analysis. Human Systems Management, 21(1). https://doi.org/10.3233/HIS-240015
[6] Mokkadem, A., Pelletier, M., & Raimbault, L. (2024). SufRec, an algorithm for mining association rules: Recursivity and task parallelism. Expert Systems with Applications, 121321. https://doi.org/10.1016/j.eswa.2023.121321
[7] Barua, E., & Dutta, M. (2024). Privacy-preserving association rule mining: A survey of techniques for sensitive rule identification and enhanced data protection. International Journal of Computers and Applications, 46(4), 252–265. https://doi.org/10.1080/1206212X.2024.2307086

Krisilov V. A.

Discovering hidden patterns in transactional data using association analysis methods

Abstract. *The paper compares Apriori and FP-Growth for discovering frequent itemsets in transactional datasets. It discusses candidate generation, prefix-tree compression, data sparsity, and the effect of minimum support. Association rules are assessed using support, confidence, and lift, with attention to misleading high-confidence relationships. Consistent experimental criteria for runtime, memory consumption, and interpretability are proposed without presenting unverified benchmark results.*

Keywords: *data mining; association rules; transactional data; Apriori; FP-Growth.*