УДК 0004.8

ОБГРУНТУВАННЯ ВИБОРУ МЕТОДУ ГЕНЕРАЦІЇ ЧАСТИХ ПРЕДМЕТНИХ НАБОРІВ ДЛЯ ПОШУКУ АСОЦІАТИВНИХ ПРАВИЛ ПРИ РОЗРОБЦІ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Савчук Тамара, Приймак Наталія

Вінницький національний технічний університет

Анотація

У даній роботі здійснено порівняльний аналіз методів генерації частих предметних наборів, що використовуються для пошуку асоціативних правил.. Для кожного методу визначено його переваги та недоліки та обрано метод, який варто застосовувати для пошуку асоціативних правил при розробці програмного забезпечення.

Abstract

In this article a comparative analysis of the methods of generating frequent subject sets used to search for associative rules was done. For each method the advantages and disadvantages were determined and also the method, which should be used to search for associative rules in software development, was chosen.

Використання засобів та методів штучного інтелекту під час розробки програмного забезпечення (ПЗ), а саме методів пошуку асоціативних правил є актуальною задачею, що може бути застосована до основних етапів даного процесу, що допоможе розробити якісне програмне забезпечення у заданий термін та у межах виділеного бюджету, за рахунок використання знайдених залежностей.

Метою даної роботи є обґрунтування вибору методу генерації частих предметних наборів, серед яких здійснюється пошуку асоціативних правил при розробці ПЗ.

Асоціативні правила (АП) описують закономірності виду $X \to Y$ для яких виконується умова $X \cap Y \to \varnothing$. Задачу пошуку асоціативних правил при розробці ПЗ можна розділити на дві підзадачі: підзадачу генерації наборів даних, що часто зустрічаються – так званих частих предметних наборів, та підзадачу генерації правил $X \to Y$, що мають рівень достовірності не нижче заданого експертом порогового значення $minconf\,(X \to Y)$ [1]. Підзадача генерації частих предметних наборів даних вирішується з використанням методів, що можуть бути класифіковані за способом цієї генерації (рисунок 1).

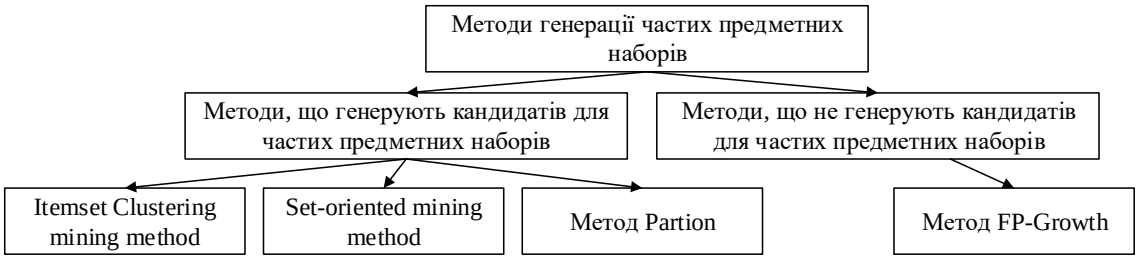

Рисунок 1 – Методи генерації частих предметних наборів

1) Методи, що генерують кандидатів для частих предметних наборів.

Set-oriented mining method – наперший запропонований метод генерації частих переметних наборів, серед яких можна здійснювати пошук АП при розробці ПЗ. Перевагами даного методу є його простота для розуміння та можливість застосування до великих БД. Серед недоліків можна виділити те, що даний метод потребує багато часу,

місця та пам’яті комп’ютера для процесу генерації можливих частих кандидатів, а також необхідне багаторазове сканування БД [2].

Метод Partion дозволяє здійснювати пошук асоціативних правил у великих БД, що містять інформацію про розробку ПЗ. Особливістю даного методу є розподіл БД на декілька частин, кожна з яких обробляється окремо. Серед переваг можна виділити швидкий пошук у великих БД та лише два сканування БД для генерації частих предметних наборів. Недоліки: виконання даного методу потрібно здійснювати на комп’ютері з великою оперативної пам’яттю, недосконалий підхід поділу БД на частини на першому етапі [3].

В методі Itemset Clustering mining method генеруються потенційні максимальні часті предметні набори, для чого використовуються методи кластеризації: на основі класів еквівалентності або maximal Hypergraph Clique. Для генерації кінцевих частих предметних наборів, серед яких можна здійснювати пошук АП при розробці ПЗ, використовуються методи проходження по графу: з низу до верху та гібридний (з низу до верху та зверху до низу). Переваги методу: використовується мало оперативної пам’яті, не потрібно будувати складну хеш-структуру для генерації частих предметних наборів, відбувається лише одне сканування БД. Серед недоліків виділяють необхідність у здійсненні вибору параметрів для алгоритму кластеризації [4].

2) Методи, що не виконують генерація кандидатів перед пошуком частих предметних наборів

Першим етапом реалізації методу FP-Growth є процес перетворення БД в деревовидну структуру, що називається FP-дерево під час чого підраховується значення підтримки для кожного елемента. На другому етапі здійснюється добування частих предметних наборів із FP-дерева, серед яких буде здійснюватися пошук АП при розробці ПЗ. Перевагами даного методу є те, що розмір FP-дерева досить малий, що дозволяє уникнути затратної процедури генерації кандидатів, швидкість пошуку частих предметних наборів вища ніж в попередніх методах. До недоліків відносять не можливість побудувати FP-дерево, що матиме розмір більший ніж основна пам'ять комп’ютера [5].

Отже, в результаті проведеного аналізу методів генерації частих предметних наборів, що можуть бути використані для пошуку асоціативних правил при розробці програмного забезпечення, було з’ясовано, що варто застосовувати метод FP – Growth, оскільки він відповідає таким вимогам: дозволяє обробляти потужні БД; швидкість генерації частих предметних наборів досить висока.

Список використаних джерел:

1.Zayko T.A. Association rules in data mining / T.A. Zayko, A.A. Oliinyk // Herald of the National University "KhPI". Subject issue: Information science and Modeling. – Kharkov: NTU "KhPI". – 2013. – № 39 (1012). – p. 82-96.

2.Swami A. Mining Associations between Sets of Items in Massive Databases / A. Swami, R. Agrawal, T. Imielinski // Proc. of the Intel Conference on Management of Data. – Washington. – 1993. – p. 134-141.

3.Savasere A. An Efficient Algorithm for Mining Association Rules in Large Databases / A. Savasere, E. Omiecinski, Shamkant B. Navathe // VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases – San Francisco. – 1995. – p. 432-444.

4.Zaki M. New Algorithms for Fast Discovery of Association Rules / M. J Zaki // Proceedings of the Third International Conference on Knowledge Discovery and DM. – New York. – 1997. – p. 283-286.

5.Han J. Mining frequent patterns without candidate generation/J. Han, J. Pei // Proceedings of the 2000 ACM SIGMOD international conference on Management of data. – New York. – 2000. – №2. – p. 1-12.