Лабораторна робота №1

МЕТОДИ ВИДОБУВАННЯ АСОЦІАТИВНИХ ПРАВИЛ

З ВЕЛИКИХ МАСИВІВ ДАНИХ

Мета роботи: вивчити основні методи видобування асоціативних правил з великих масивів даних, навчитися використовувати спеціалізовані програмні засоби для видобування знань з масивів даних.

### Завдання до роботи

1. Ознайомитися з конспектом лекцій та рекомендованою літературою за темою роботи, а також з додатком А, що містить опис програмного забезпечення для видобування асоціативних правил з великих масивів даних.
2. Сформувати набір даних для обробки та аналізу.
3. Розробити за допомогою середовища Matlab програмне забезпечення для видобування асоціативних правил з великих масивів даних або вивчити рекомендоване програмне забезпечення (пакет Armada модулю Matlab) та здійснити обробку набору даних з метою виділення асоціативних правил.
4. Оформити звіт з роботи.
5. Відповісти на контрольні питання.

### Зміст звіту

1. Тема та мета роботи.
2. Короткі теоретичні відомості.
3. Набір даних для обробки (якщо він великий – навести фрагмент).
4. Детальний опис процесу використання програмного забезпечення для обробки набору даних, що має бути проілюстрований зображеннями екранних форм з описом кожного їх елементу та коментарями до кожної дії.
5. Текст програми з коментарями.
6. Результати роботи програмного забезпечення (набір отриманих правил, інші характеристики).
7. Висновки, що містять відповіді на контрольні запитання, а також відображують результати виконання роботи та їх критичний аналіз.

### Контрольні питання

1. Що таке асоціативне правило?
2. Для чого призначені асоціативні правила?
3. Дати визначення понять підтримки та достовірності правила.
4. Яке призначення алгоритмів пошуку асоціативних правил?
5. На які підзадачі розбивається задача знаходження асоціативних правил?
6. Які методи використовуються для знаходження асоціативних правил?
7. Яким чином обираються значення параметрів minsupport та minconfidence?
8. Що таке числові асоціативні правила?
9. Поясніть поняття “узагальнене асоціативне правило”.
10. Що називається ієрархією елементів?
11. Які переваги дає введення додаткової інформації про групування елементів?
12. Поясніть, які проблеми можуть виникнути при безпосередньому застосуванні алгоритмів знаходження асоціативних правил.
13. В чому полягає сутність виявлення узагальнених асоціативних правил?
14. Яким чином визначають “цікаві” правила? В чому полягає актуальність такого процесу?
15. Дати визначення понять батьківського правила (пращура) та найближчого батьківського правила.
16. Порівняйте поняття цікавого та частково цікавого правила.
17. Які проблеми усуває алгоритм обчислення узагальнених асоціативних правил?
18. З яких етапів складається процес обчислення узагальнених асоціативних правил?
19. Проаналізуйте базовий алгоритм пошуку множин, що зустрічаються часто.
20. Опишіть алгоритм генерації кандидатів.
21. Яким чином використовується хеш-дерево для підрахунку підтримки кандидатів? Як відбувається процес побудови такого дерева?
22. Виконайте порівняльний аналіз базового та покращеного алгоритмів пошуку множин, що зустрічаються часто.
23. За рахунок яких оптимізацій відбувається покращення базового алгоритму пошуку множин, що зустрічаються часто?
24. В чому полягає сутність маштабованого алгоритму пошуку асоціативних правил Apriori?
25. Яким чином перетворюються дані для можливості використання алгоритму Apriori?
26. Яка властивість використовується в алгоритмі Apriori? Для чого вона використовується?
27. Наведіть послідовність виконання алгоритму Apriori.
28. Опишіть функцію генерації кандидатів в алгоритмі Apriori.
29. Як відбувається підрахунок підтримки для кожного кандидату в алгоритмі Apriori? Для чого в цій процедурі використовують хеш-дерево?
30. Як здійснити добування правил з набору, що часто зустрічається?
31. Проаналізуйте внутрішню структуру модулю Armada пакету Matlab: основні змінні, параметри, методи та функції, їх призначення та використання.

**Додаток A**

**АСОЦІАТИВНІ ПРАВИЛА У ПАКЕТІ MATLAB**

Операції з асоціативними правилами у пакеті MATLAB дозволяє виконувати модуль *ARMADA*. Він дозволяє створювати асоціативні правила по заданим користувачем даним в рамках середовища MATLAB.

Для запуску модулю *ARMADA* необхідно зробити папку, де знаходиться цей модулю, поточною, після чого в командному рядку середовища MATLAB написати наступну команду:

armada

В результаті з’явиться головна інтерфейсна форма модулю, в якій необхідно обрати файл з вхідними даними та параметри для аналізу даних (рис. А.1).

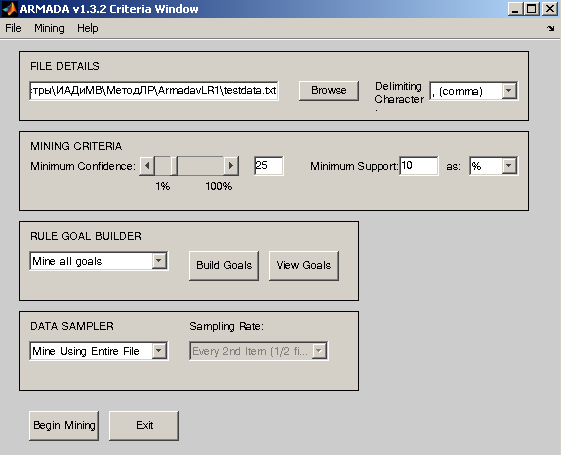


Рисунок А.1 – Головна інтерфейсна форма модулю *ARMADA*

Для інтелектуального аналізу даних за допомогою асоціативних правил в полі FILE DETAILS необхідно ввести шлях до файлу з даними для аналізу. Файл з даними для аналізу можна також обрати у віконному режимі, натиснувши кнопку Browse. В полі зі списком Delimiting Character необхідно вказати символ, що відокремлює дані одне від одного.

Компоненти головної форми Minimum Confidence Minimum Support призначені для введення параметрів minsupport та minconfidence, відповідно.

Для виконання аналізу даних та отримання вихідної інформації у вигляді асоціативних правил необхідно натиснути кнопку Begin Mining, після чого відбудеться процес аналізу даних, та з’явиться форма з результатами роботи програми (рис. А.2). Вихід з програми виконується за допомогою кнопки Exit.

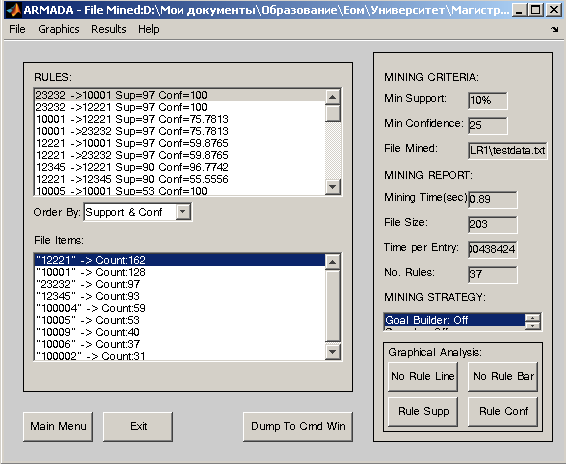


Рисунок А.2 – Результати аналізу даних за допомогою модулю *ARMADA*

В області RULES наведено отримані асоціативні правила, в області File Items – елементи файлу для аналізу та кількість разів, які вони зустрічаються у файлі.

Компоненти в області MINING CRITERIA відображають параметри аналізу: значення параметрів minsupport та minconfidence, а також файл із даними для аналізу.

В області MINING REPORT відображається інформація про хід виконання аналізу:

– час, витрачений на виконання аналізу (Mining Time);

– розмір файлу (File Size);

– час, затрачений на один запис (Time per Entry);

– кількість отриманих асоціативних правил (No. Rules)

За допомогою головного меню отримані дані можна зберегти на диск (File –> Save) та отримати графіки, що відображають процес аналізу даних (Graphics).