УДК: 004.832.32

Кравець І.О., Пархоменко Д.Ю

Дослідження ефективності алгоритмів DATA MINING стосовно пошуку закономірностей поведінки користувачів web-вузла

Проведено дослідження ефективності алгоритмів Data Mining для аналізу соціально-економічних показників. Для аналізу неструктурованої, неоднорідної, розподіленої і значної по обсягу інформації о користувачів Web-вузлів розроблено та реалізовано алгоритми пошуку дерев рішень для задач класифікації, кластерного аналізу. Запропоновано алгоритм кластеризації з визначенням як центрів, так і числа кластерів.

The efficiency of Data Mining algorithms, which are applied for the analysis of the WEB-servers users' information, is researched. The algorithm of searching decision's tree for classification, the cluster analyses algorithm and the algorithm of searching associated groups are developed.

У зв'язку зі стрімким зростанням інформаційних ресурсів, доступних у Світовій мережі Internet, для користувачів стало все більш і більш необхідним використовувати автоматизовані інструменти для пошуку бажаних інформаційних ресурсів, їх відстеження та аналіз зразків використання. Ці чинники підвищують необхідність створення серверних (serverside) та клієнтських (clientside) інтелектуальних систем, здатних ефективно видобувати знання. У широкому розумінні термін Web mining може бути визначений як видобуття та аналіз корисної інформації з World Wide Web.

Технологія Web Mining охоплює методи інтелектуальної обробки даних Data Mining, використаних для аналізу неструктурованої, неоднорідної, розподіленої і значної за обсягом інформації, яка зберігається на Web-вузлах. У Web Mining можливо виділити три основні напрямки: Web Structure Mining, Web Content Mining и Web Usage Mining.

Перший напрямок пов'язаний з автоматичним пошуком якісної інформації з різних джерел Інтернету, перевантажених "інформаційним шумом".

Другий – опрацьовує структуру гіперпосилань у мережі, щоб застосувати алгоритми аналізу мереж. Це робиться з метою побудови ретельної моделі Web-простору. Деякі алгоритми можуть бути застосовані для побудови топології Web-мереж.

Web Usage Mining виконує пошук закономірностей поведінки користувачів Web-вузла. Аналізується інформація:

- які сторінки проглядав користувач;
- яка послідовність перегляду;

- яка географія користувачів;
- як розбити користувачів Web-вузла на групи за даними анкети.

Усі напрямки є дуже актуальними, але особливо актуальний третій напрямок, пов'язаний з електронною комерцією.

Був проведений аналітичний огляд існуючих програмних продуктів, робота яких пов'язана з інтелектуальною обробкою даних (OLAP, ORACLE, статистичні пакети). Ці програмні продукти виконують поставлені задачі і є досить потужним інструментом для їх вирішення. Їх можна використовувати в різних сферах, для різних задач, і вони є комерційним продуктом, впровадження якого вимагає великих коштів та спеціальних навиків у користуванні.

Внаслідок аналітичного огляду існуючих методів статистичного та інтелектуального аналізу даних було виявлено, що аналіз за допомогою методів інтелектуального аналізу даних (Data Mining) є більш ефективним, ніж той, що проводиться з використанням традиційних статистичних методів обробки інформації, особливо при неструктурованій, неоднорідній, розподіленій та значній за обсягом інформації, на Web-вузлах.

Було розроблено та реалізовано такі алгоритми інтелектуального аналізу:

- алгоритм пошуку класифікаційних правил, який дозволяє розв'язати задачу класифікації користувачів Web-вузла при якісному вигляді інформації;
- алгоритм кластерного аналізу, який дозволяє розбити користувачів на групи за рядом показників, причому було досліджено його жорстку та ймовірнісну модифікації;
- алгоритм пошуку асоціативних груп та правил, який дозволяє групувати ресурси Web-вузлів, щодо їх використання користувачами.

Створена інформаційно-аналітична система, яка поєднує базу даних та інформаційний блок, може слугувати демонстраційним прикладом використання методів статистичного та інтелектуального аналізу для задач WEB Usage Mining.

Методами Data Mining розв'язуються три основні задачі: класифікація та регресія, пошук асоціативних правил, кластеризація. [1, 2]. В задачі класифікації та регресії потрібно визначити значення залежної змінної об'єкта на підставі значень інших змінних, що характеризують його. Якщо залежна змінна приймає безперервні чисельні значення, то це задача регресії, якщо залежна змінна приймає фіксований набір якісних або чисельних значень, то це задача класифікації.

Для розв'язання задач регресії використовуються такі статистичні методи, як кореляційний, однофакторний дисперсійний аналіз (перевірка впливу вхідних факторів) та множинний регресійний аналіз [4, 6]. Однак ці методи можуть бути використані при достатньому обсязі однорідних статистичних даних у чисельному вигляді.

Найбільш розповсюджені моделі, що відбувають результати класифікації, – це класифікаційні правила, дерева рішень, розподіляючи математичні (лінійні і нелінійні) функції [1, 2]. Розподіляючи математичні функції, використовуються при чисельному представленні однорідних вхідних даних.

Найпоширенішими алгоритмами побудови класифікаційних правил є алгоритми 1R та Баєсовські алгоритми. Однак вони мають суттєві недоліки: алгоритм 1R будує правило за однієї змінною, Баєсовські алгоритми працюють з незалежними вхідними змінними. Тому для побудови класифікаційних правил було обрано побудову дерев рішень, а з алгоритмів – алгоритм покриття.

Метою алгоритму покриття є розбивка навчальної вибірки таким чином, щоб одержати

підмножини, що відповідають класам. Даний підхід полягає в побудові дерев рішень для кожного класу окремо. На кожному кроці алгоритму обирається значення змінної, яке поділяє всю множину на дві підмножини. Така розбивка робиться доти, доки не буде побудована підмножина, що містить тільки об'єкти одного класу. Блок-схема алгоритму наведена на рис.1.

У алгоритмі використані такі позначення:

$D[N,M+1]$ – вхідна матриця значень змінних; N – число об'єктів; M – число вхідних змінних; M+1 стовпець – значення вихідної змінної;

$P^{k}{}_{i,j} = \dfrac{n^{k}_{i,j}}{n_{i,j}}$ – умовна ймовірність віднесення до *k-го* класу при *i*-му значенні *j*-ої змінної;

$n^{k}_{j,i}$ – число даних з *i*-ми значенням j-ої змінної при наявності класу k, n_{ji} – число даних з *i*-им значенням *j*-ої змінної.

Вхідні дані – результати анкети користувача (освіта, діапазон віку, країна, стать), а також дані журналів Web-сервера (IP-адрес клієнта, ім'я Login, час звернення, шлях та тип запитаного мого ресурсу). Вихідні дані – тип запитаною мого ресурсу (програмні файли, аудіофайли, текстова інформація). Алгоритм дозволяє формувати класифікаційні правила при якісному вигляді неоднорідної інформації, не вимагає незалежності змінних та великого обсягу статистичних даних.

Задача кластеризації полягає в пошуку незалежних груп (кластерів) та їхніх характеристик у всій безлічі аналізованих даних. Рішення цієї задачі допомагає краще зрозуміти дані. Крім того, угруповання однорідних об'єктів дозволяє скоротити їхнє число, полегшити аналіз. Так, можливо розбити користувачів Web-вузла на групи за даними анкети.

Найпоширенішими алгоритмами кластеризації є дентограми та алгоритми пошуку оптимальної розбивки на кластери за мінімізацією функції відстані об'єктів від центрів кластерів [1, 2]. Було проаналізовано декілька алгоритмів ієрархічних та неієрархічних і був обраний неієрархічний алгоритм Fuzzy C-Means [1]. Ідея алгоритму полягає у визначенні центрів кластерів і віднесенні до кожного кластера об'єктів, що найбільш близько підходять до даного кластера. Його відмінність полягає в тому, що кластери тепер є нечіткими множинами, і кожна точка належить різним кластерам з різним ступенем належності. Точка належить до того або іншого кластера за критерієм максимуму належності даному кластеру.

Запропоновано модифікацію алгоритму з визначенням як центрів кластерів, так і числа кластерів. Блок-схему алгоритму наведено на рис. 2.

У алгоритмі використані такі позначення:

- навчальна множина $M = \{m_j\}^T$, T – кількість точок (векторів) даних об'єктів, *j*-координата;
- матриця належності $U = \{u_{ij}\}$ *j*-го об'єкта *i*-му кластеру.
- *j*-координата центра *i*-го кластера c_{ij}
- Евклідова відстань *i*-го об'єкту від центру *j*-го кластера $d(x_j, c_i) = \sqrt{\sum_{t=1}^{m} (x_{it} - c_{jt})^2}$

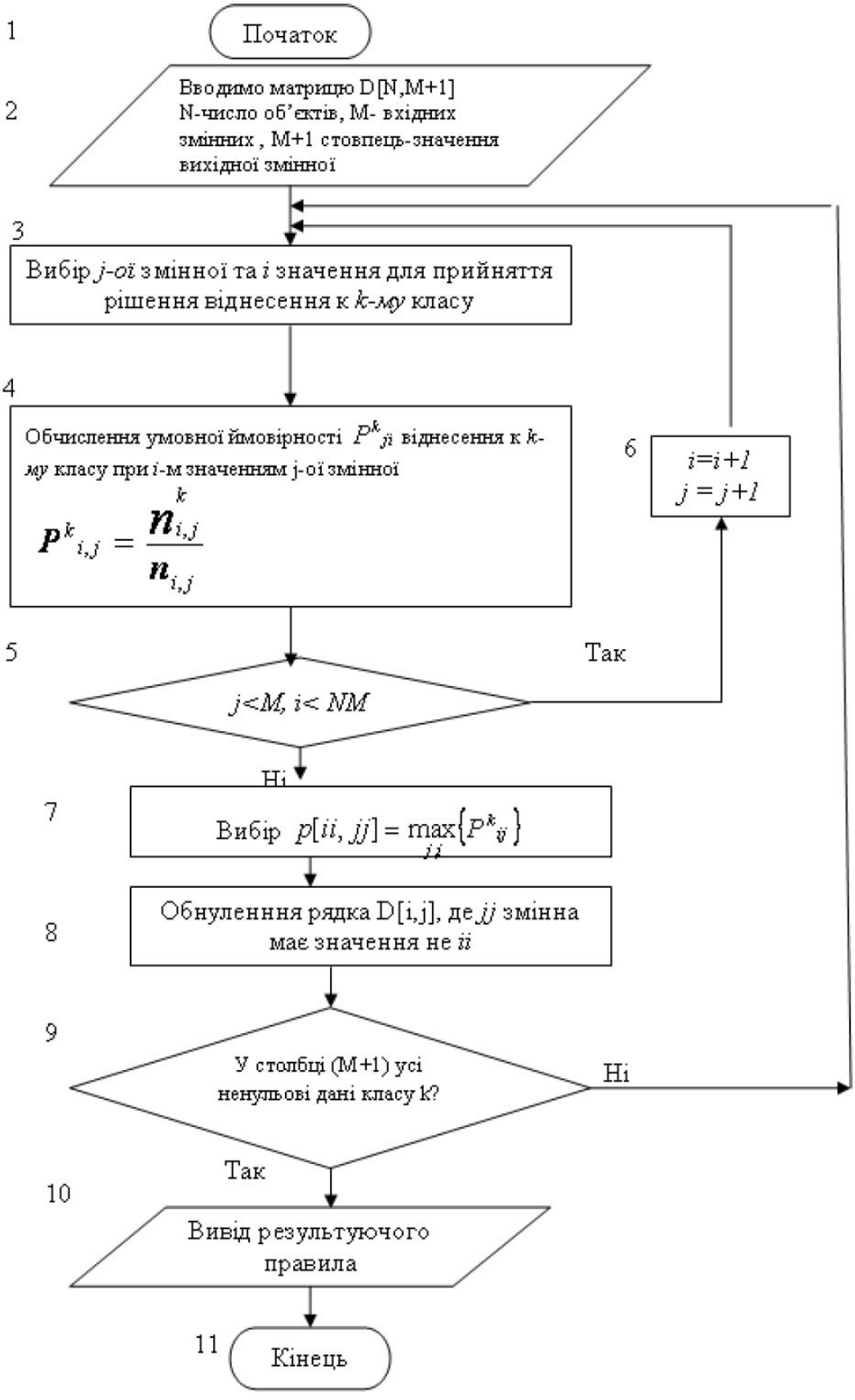

Рис. 1. Блок-схема алгоритму виводу правила

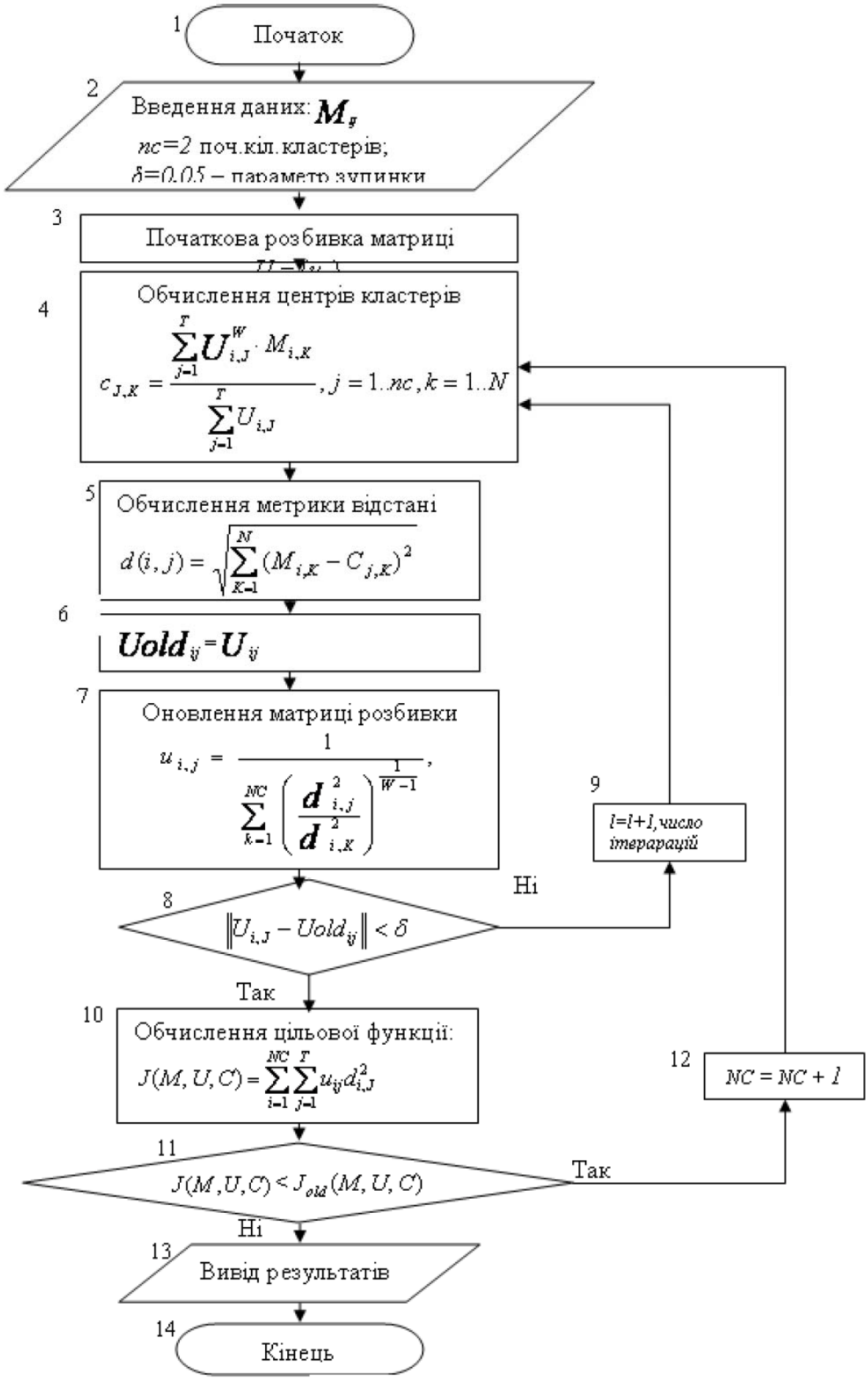

Рис. 2. Блок-схема модифікованого алгоритму

Алгоритм дозволяє розбивати користувачів на групи, не вимагаючи незалежності показників та великого обсягу статистичних даних.

Алгоритми пошуку асоціативних груп визначають набори об'єктів, які часто зустрічаються. У даному випадку об'єкти інформації, яку обрав користувач (програмні файли, аудіофайли, текстова інформація). Алгоритми визначають набори, що часто зустрічаються за кілька етапів. На i-му етапі визначаються всі i-елементні набори, що часто зустрічаються. Кожен етап складається з двох кроків: формування наборів кандидатів і підрахунку підтримки кандидатів. На кроці формування кандидатів i-го етапу алгоритм створює множину кандидатів з i-елементних наборів, чия підтримка поки не обчислюється. На кроці підрахунку кандидатів i-го етапу алгоритм сканує множину трансакцій (наборів, що аналізуються), обчислюючи підтримку (Supp) наборів-кандидатів.

Відношення кількості транзакцій, куди входить набір F, до загальної кількості транзакцій називається підтримкою (support) набору F і позначається Supp (F):

$$Supp(F) = \frac{|D_F|}{|D|}.$$

Після сканування відкидаються кандидати, підтримка яких менша за визначений користувачем мінімум, і зберігаються тільки i-елементні набори, що часто зустрічаються. Під час 1-го етапу обрана множина наборів-кандидатів містить усі 1-елементні частини набору. Алгоритм обчислює їхню підтримку під час кроку підрахунку. Блок-схема алгоритму наведена на рис. 3.

Таким чином, алгоритми Data Mining дуже ефективно працюють при аналізі даних про користувачів Web-вузла. За їх допомогою можливо розв'язувати задачі класифікації та регресії, пошуку асоціативних груп, кластеризації та прогнозування. Для аналізу найбільш ефективними є такі алгоритми інтелектуального аналізу як:

- алгоритм пошуку дерев рішень для задач класифікації (алгоритм покриття);
- неієрархічний алгоритм Fuzzy C-Means для задачі кластерного аналізу;
- алгоритм пошуку асоціативних груп.

Запропоновано алгоритм кластеризації Fuzzy C-Means з визначенням як центрів кластерів, так і числа кластерів. Усі алгоритми показали високу ефективність.

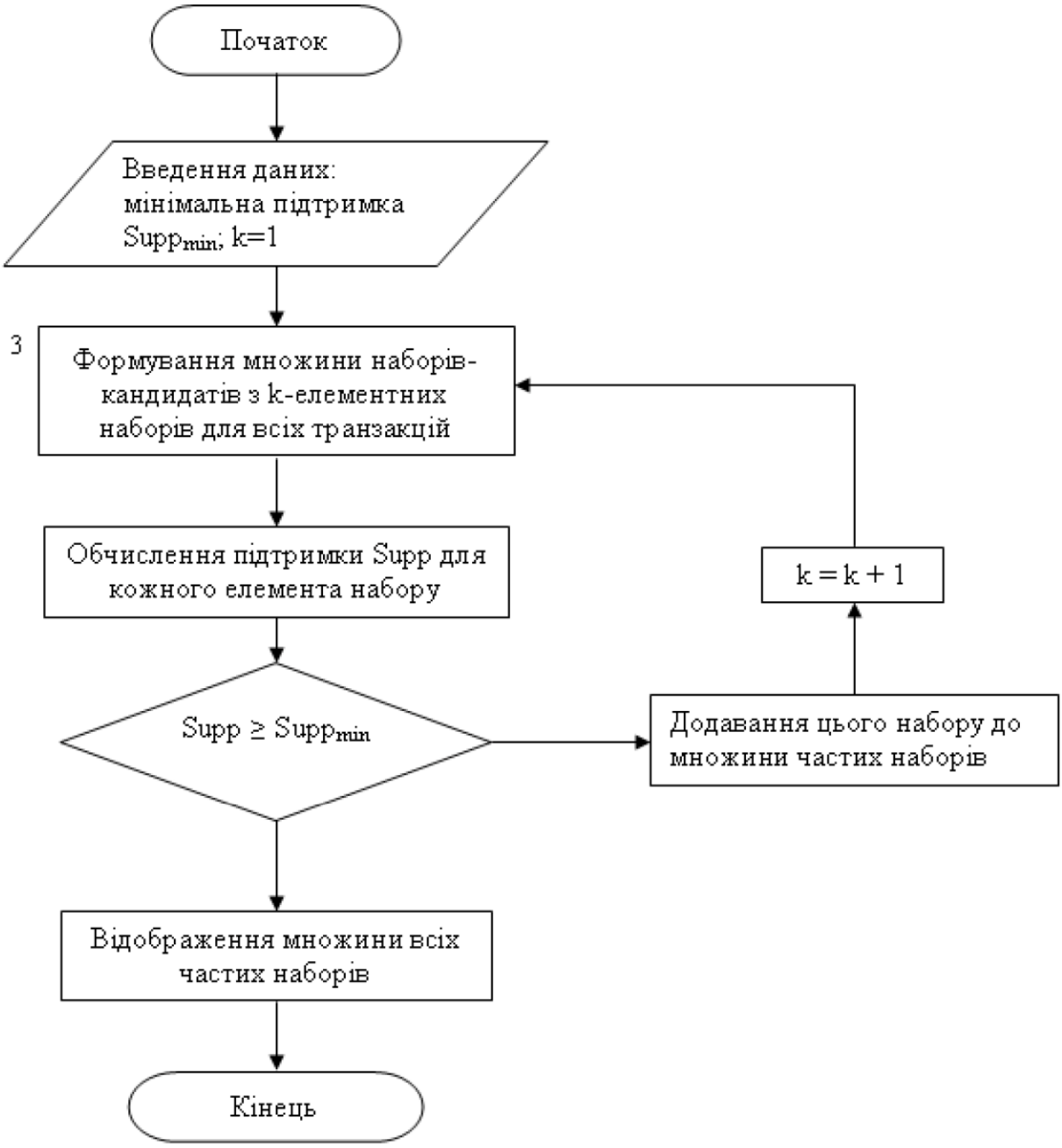

Рис. 3. Блок-схема алгоритму обирання наборів-кандидатів

Література

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining.– СПб.: БХВ-Петербург, 2004. – 336 с.: ил.
2. Жамбю М. Иерархический кластерный анализ и соответствия. – М.: Финансы и статистика, 1988. – 236 с.
3. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.
4. Кравець І.О., Ромакін В.В. Статистичний аналіз даних з використанням статистичних пакетів та MS Excel: Навч. посібник. – Миколаїв: МДГУ, 2002. – 58 с.
5. Лук'яненко І.Г., Краснікова Л.І. Економетрика / Практикум з використанням комп'ютера. – К.: Товариство "Знання", ККО, 1998. – 231 с.
6. Статистические методы анализа информации в социологических исследованиях. – М., 1979, Наука, 194 с.
7. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА-М, 1998. – 528 с.