КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

УДК 681.3.016

О.П. КУРГАЄВ, І.В. САВЧЕНКО

ДОСЛІДЖЕННЯ МЕТОДІВ ПОШУКУ В СЛОВНИКУ ТЕРМІНІВ

1. Актуальність обраної теми та аналіз останніх досліджень і публікацій. На сьогоднішній день системи, які працюють із знаннями (інтелектуальні системи), відіграють значну роль у житті людства, оскільки здатні запропонувати необхідне та кваліфіковане рішення проблеми за короткий час. Такі системи використовуються під час контролю великих промислових процесів, приймають рішення, спираючись на дані сотень периферійних пристроїв, керують великими інформаційними мережами, відіграють роль консультанта в процесі прогнозування процесу, до складу якого входить велика кількість факторів. Тобто розв'язують задачі, які під силу тільки групі експертів.

Розробку експертних систем проводить велика кількість компаній та корпорацій, серед яких: XpertRule® [1] – розробка експертних систем, що використовуються в бізнесі (XpertRule Knowledge Builder, XpertRule Servise, ін.); Production Systems Technologies, Inc.® [2], OPSJ – розробка високоефективних експертних систем, з використанням мови Java; A I Developers, Inc.® [3], EZ-Xpert – розробка систем, що здатні генерувати код процедури на мові C++; Togai InfraLogic, Inc.® [4], FuzzyCLIPS – розробка нечітких експертних систем; Gensym® [5], G2 – платформа використовується для прийняття рішень в системах реального часу (наприклад, моніторинг космічного корабля, процесів автомобілебудування фірми Toyota); Red Hat, Inc.® [6], JBoss Enterprise BRMS – автоматизована система керування підприємством; Jess® [7], Jess – невелика за розмірами та швидка система, що використовує

A.F. Kurgaev, I.V. Savchenko

SEARCH METHODS RESEARCHING IN A DICTIONARY OF TERMS

It is investigated search method researching in a dictionary of terms.

Выполнен анализ эффективности методов поиска в словаре базы знаний.

Проведено дослідження ефективності методів пошуку в словнику бази знань.

мову Java, та інше.

Проблемам створення інтелектуальних систем присвячені праці багатьох вітчизняних та закордонних науковців. Теорії створення та практичному застосуванню експертних систем присвячені роботи [8, 9]. Проблемам отримання, структурування даних, знань та технологічним аспектам розробки систем, основаних на знаннях, присвячені роботи [10–12].

2. Постановка проблеми. У роботі [12] зазначено, що основною проблемою сучасних систем обробки знань (СОЗ) є непродуктивні витрати пам'яті та часу апаратних ресурсів на підтримку програмного забезпечення. Все це зумовлено невідповідністю існуючих СОЗ людським можливостям обробки, накопичення та представлення знань [12]. Тому пропонується створення процесора бази знань, архітектура якого призначена для ефективної роботи з базою знань (БЗ).

Структура БЗ – це гіперграф над системою понять деякої прикладної теорії. Для інтерпретації понять СОЗ використовує словник, терміни якого структурно зв'язані із підграфами гіперграфа. Словник БЗ (СБЗ), зважаючи на його розміри зберігається у вигляді файла.

Оскільки час пошуку в словнику є невід'ємною складовою загального часу обробки знань у процесі вирішення задач, доцільним є проведення дослідження методів пошуку в файлових словниках.

Мета роботи – провести аналіз ефективності існуючих методів пошуку даних у файлових словниках та вибрати такі, що дозволяють підвищити продуктивність СОЗ.

3. Виклад основного матеріалу дослідження. Аналіз робіт [13, 14] свідчить про те, що на сьогоднішній день використовуються такі методи пошуку: послідовний, бінарний, пошук за бінарними та збалансованими деревами, цифровий пошук та хешування.

3.1. Вихідні теоретичні дані пошуку в СБЗ. В табл. 1 наведені результати теоретичного дослідження ефективності основних методів пошуку, які описані в [13]. В табл. 1 прийняті такі позначення:

N – кількість заповнених записів в індексній таблиці чи в файлі;

C, C_1 – середня кількість порівнянь ключів записів;

D – кількість незбалансованих вузлів;

S – дорівнює 1 при вдалому пошуку, та 0 при невдалому пошуку;

S_1 – дорівнює 1, якщо перший пошук вдалий, 0 в іншому випадку;

A – коефіцієнт заповнення індексної таблиці;

M – сумарна кількість заповнених та порожніх записів в індексній таблиці чи в файлі;

u – константа, яка залежить від параметрів системи пошуку.

3.2. Емпіричне дослідження методів пошуку. Для здійснення емпіричного дослідження розроблено програмне забезпечення, функціональна структура якого подана у вигляді наборів модулів (рис. 1). Керування роботою системи відбувається за допомогою інтерфейсу користувача. Проект програми пошуку реалізовано в програмному середовищі Borland C++ Builder 5.0. Усі модулі, в кінце-

вому вигляді, представлені в виконавчому файлі Search.exe даного проекту. В роботі розглянуто СБЗ з кількістю записів 10^5.

ТАБЛИЦЯ 1. Результати теоретичного дослідження ефективності методів пошуку

Метод пошуку	Середній час пошуку даних в СБЗ	Визначення констант
1. Послідовний	$t \cong (5C - 2S + 3) \cdot u$	$C = \frac{N+1}{2}$
2. Бінарний	$t \cong (18C - 10S + 12) \cdot u$	$C = \log_2 N$
3. Пошук по бінарному дереву	$t \cong (7.5C - 2.5S + 4) \cdot u$	$C = 2\ln N$
4. Пошук по збалансованому дереву	$t \cong (10C + C_1 + 2D + 2 - 3S) \cdot u$	$D \approx \frac{1}{3}C$, $C_1 \approx \frac{1}{2}(C+S)$, $C + S \approx 1.01 \cdot \log_2 N + 0.1$
5. Метод роздільного зв'язування	$t \cong (7C + 4A + 17 - 3S + 2S_1) \cdot u$	$C \approx \frac{1}{4}(2.718^2 + 1) = 2.1$, $A = \frac{N}{M}$
6. Метод лінійного зондування	$t \cong (7C + 9E + 21 - 4S) \cdot u$	$C \approx \frac{1}{4}(2.718^2 + 1) = 2.1$, $E = \frac{(C-1)}{M}$

Результати емпіричного дослідження ефективності методів пошуку в СБЗ наведені в табл. 2, в якій прийняті наступні позначення: $t_{\text{розрах.}}$ – розрахунковий час пошуку (див. табл. 1); $t_{\text{експер.}}$ – середній час пошуку, отриманий експериментально; $N_{\text{розрах.}}$ – розрахункове значення кількості зчитаних записів з СБЗ, яке відповідає параметру C (див. табл. 1); $N_{\text{експер.}}$ – кількість зчитаних записів з СБЗ, яка отримана експериментально.

Параметри за якими оцінено ефективність кожного із методів пошуку наведені далі:

- кількість прочитаних записів з файла БЗ;
- кількість прочитаних записів в індексній таблиці;
- розмір файла СБЗ, Mb;
- середній час пошуку, мкс.

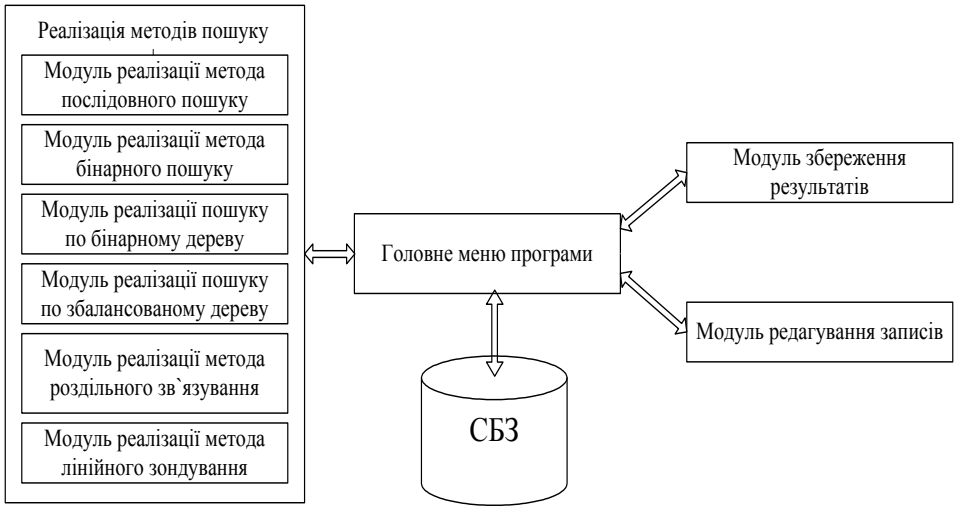

РИС. 1. Функціональна структура програмного забезпечення системи дослідження ефективності алгоритмів пошуку інформації в СБЗ

Результати оцінки вищезазначених параметрів показані на рис. 2 – 5. На рис. 2 – 4 найменування стовпців відповідають найменуванням методів, які зазначені в табл. 2.

ТАБЛИЦЯ 2. Результати емпіричного дослідження ефективності методів пошуку в СБЗ

Метод пошуку	$t_{\text{розрах.}}$	$t_{\text{експер.}}$, мкс	$N_{\text{розрах.}}$	$N_{\text{експер.}}$
1. Послідовний	250003u	1710.21	50000	56122.16
2. Бінарний	301u	691.81	16.61	15.48
3. Пошук по бінарному дереву	34.5u	511.87	23.02	20.65
4. Пошук по збалансованому дереву	82.73u	365.23	15.87	13.45
5. Метод роздільного зв'язування	32.5u	172.26	2.1	2.16
6. Метод лінійного зондування	31.7u	176.00	2.1	2.35

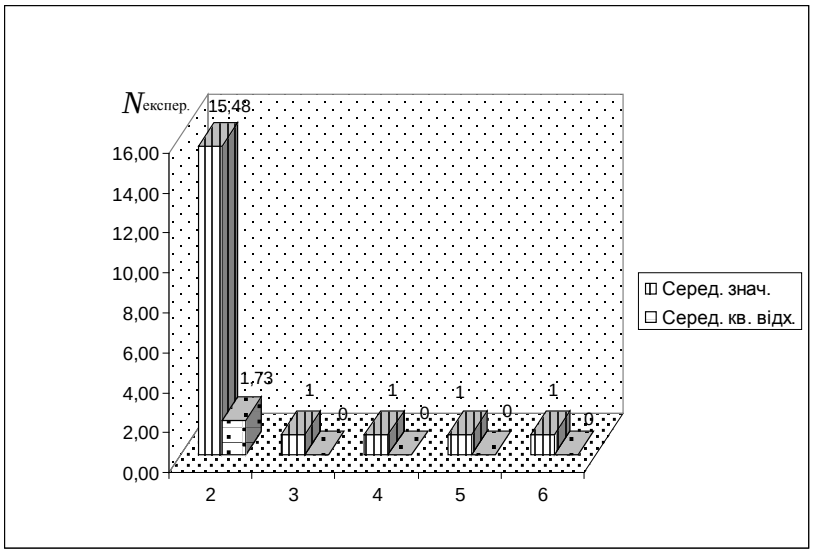

РИС. 2. Кількість прочитаних записів у файлі СБЗ

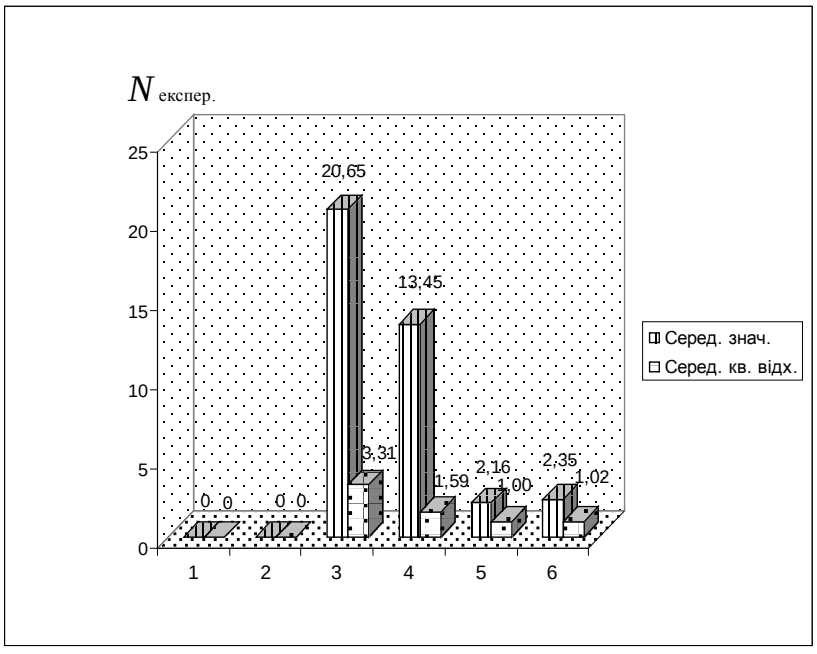

РИС. 3. Кількість прочитаних записів в індексній таблиці СБЗ

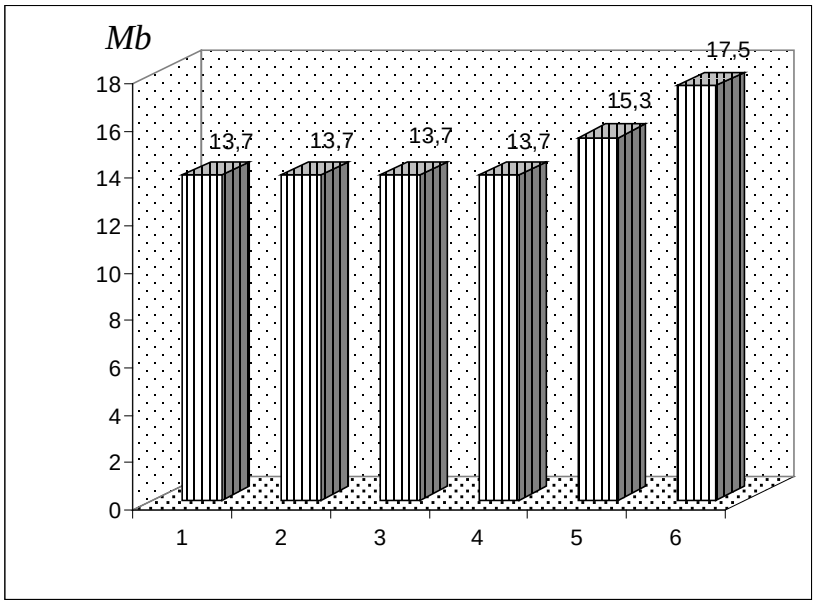

РИС. 4. Розмір файла СБЗ

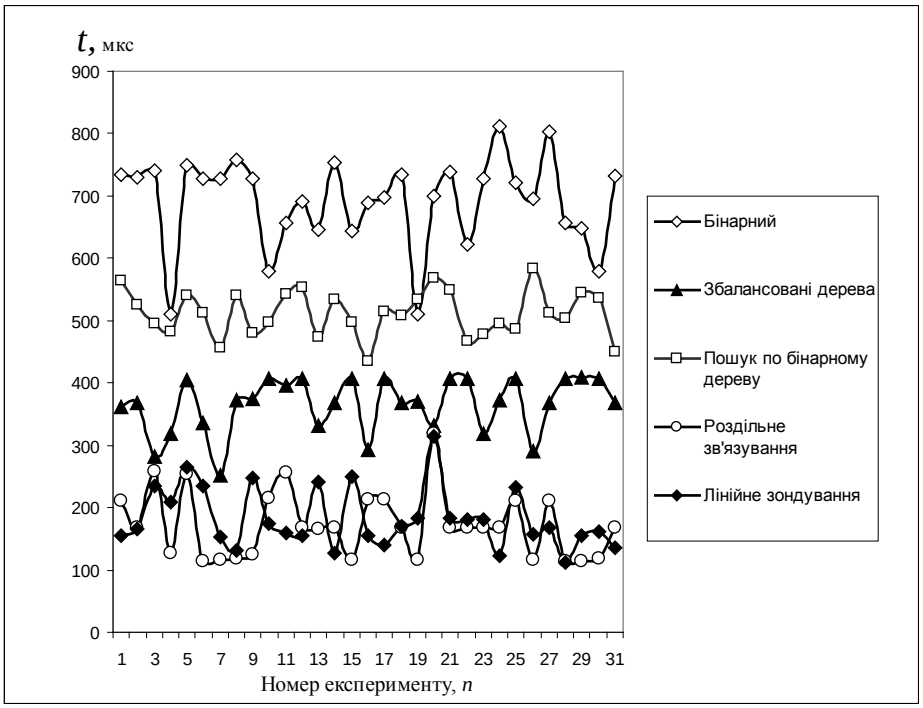

РИС. 5. Результати емпіричного дослідження методів пошуку в СБЗ

Висновки. У роботі проаналізовано наступні методи пошуку в СБЗ: послідовний, бінарний, пошук по деревам та пошук методом хешування.

Розбіжність теоретичних даних та результатів емпіричного дослідження пояснюється тим, що теоретичний метод не враховує практичну реалізацію алгоритмів, в яке входить час на доступ до зовнішніх даних, та особливості архітектури системи.

Аналіз результатів, наведених в табл. 1 та 2 свідчить про те, що методи хешування (роздільне зв'язування, лінійне зондування) є найшвидші. Пошук по деревам (бінарні дерева, збалансовані дерева) дає дещо гірші результати (див. табл. 1 та 2).

Однак при роботі із СБЗ пошук по деревах у порівнянні з методами хешування має переваги, оскільки: 1) дерева не потребують розрахунку хеш-функції, який може бути досить громіздким, якщо ключ запису великий; 2) дерева мають простішу абстрактну структуру представлення даних; 3) дерева можуть забезпечити гарантовану продуктивність пошуку; 4) дерева підтримують більш широкий діапазон операцій над даними СБЗ (сортування, пошук).

На підставі викладеного можна зробити висновок, що для роботи із СБЗ доцільно використовувати методи пошуку по деревах.

1. http://www.xpertrule.com
2. http://www.pst.com
3. http://www.ez-wpert.com
4. http://www.ortech-engr.com
5. http://www.gensym.com
6. http://www.redhat.com
7. http://www.jessrules.com
8. *Джарратано Д.* Экспертные системы: принципы разработки и программирование. 4-е. изд. – М.: Изд. дом Вильямс, 2007. – 1152 с.
9. *Джексон П.* Введение в экспертные системы. 3-е. изд. – М.: Изд. дом Вильямс, 2001. – 624 с.
10. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем: Учебник для вузов. – СПб: Питер, 2000. – 384 с.
11. *Частиков А., Белов Д., Гаврилова Т.* Разработка экспертных систем. Среда CLIPS. – БХВ-Петербург, 2003. – 608 с.
12. *Кургаев А.Ф.* Проблемная ориентация архитектуры компьютерных систем. – Киев: Сталь, 2008. – 540 с.
13. *Кнут Д.* Искусство программирования для ЭВМ. Т. 3. Сортировка и поиск. – М.: Изд. дом Вильямс, 2000. – 832 с.
14. *Седжвик Р.* Фундаментальные алгоритмы на С++. Анализ/Структуры данных/Сортировка/Поиск: Пер. с англ. – К.: Изд-во «ДиаСофт», 2001. – 688 с.

Отримано 25.08.2009