УДК 004.632: 004.75

А.О. ШЕМЕТ
Академія митної служби України, м. Дніпропетровськ

ЗАСТОСУВАННЯ МЕТОДІВ ПРОГРАМНОГО РЕЗЕРВУВАННЯ ІНФОРМАЦІЇ: АСПЕКТ РЕЗЕРВНОГО КОПІЮВАННЯ. МЕТОД ПАРАЛЕЛЬНОГО РЕЗЕРВНОГО КОПІЮВАННЯ

У даній статті зроблено огляд основних методів програмного резервування інформації у аспекті резервного копіювання, окреслено критерії оцінки методів резервного копіювання. Наведено основні недоліки методів інкрементального та диференціального резервного копіювання. Сформульовано метод паралельного блочного резервного копіювання.

This article provides an overview basic techniques of the software reservation: data backup aspect, outlines criteria for evaluating of backup methods. There identified the main disadvantages of the method of differential and incremental backups. Formulated a parallel block backup method.

Ключові слова: методи резервного копіювання, інформаційні системи, алгоритм, паралельна обробка.

Вступ

Надійне збереження інформації є однією з найважливіших задач функціонування будь-якої інформаційної інфраструктури. Обсяг заходів, щодо надійного зберігання залежить від багатьох факторів, серед них: цінність інформації, стійкість інформації до викривлення, обсяг фінансування інфраструктури. Одним з напрямків надійного збереження інформації є резервування. Резервування, як метод підвищення характеристик надійності технічних пристроїв за допомогою введення надмірності, розділяється на декілька видів: апаратне, програмне, часове, функціональне [1]. В рамках цієї статті розглядається програмне резервування інформації як сукупність дій, які направлені на створення надлишковості інформації на серверах та персональних комп'ютерах з метою створення можливості її відновлення в разі пошкодження або втрати.

Основний матеріал

Методи програмного резервування. Перед формулюванням методів програмного резервування, сформулюємо складові резервування. Нехай існує деяка сукупність даних, яку потрібно надійно зберігати – джерело інформації, та сховище даних до якого відбувається резервне копіювання – резервна копія. На даний момент, в галузі систем резервного копіювання, виділяють наступні методи [2]:

- метод повного резервного копіювання – дублювання всіх зазначених даних з джерела інформації повністю до резервної копії, без урахування змін, що відбулися в проміжках часу між копіюванням;
- метод диференціального резервного копіювання – дублювання даних з джерела інформації до резервної копії з урахуванням змін даних від часу останнього дублювання методом повного резервного копіювання;
- метод інкрементального резервного копіювання – дублювання даних з джерела інформації до резервної копії з урахуванням змін даних від часу останнього копіювання. Метод має декілька варіацій реалізації: блочний інкремент, багаторівневий інкремент, зворотній інкремент;

До основних критеріїв оцінки методів програмного резервування необхідно віднести час, затрачений на дублювання інформації (фізичне перенесення до резервної копії); час, затрачений на аналіз файлової структури (порівняння джерела з резервними копіями відповідно до методу), які відбулися у джерелі інформації; фізичний об'єм, який займає резервна копія; кількість резервних копій, потрібних для проведення резервування, відновлення; можливість відновлення станом на проміжний час між резервуваннями; можливість паралельної обробки джерела інформації та розподіленого зберігання резервних копій, складність (трудомісткість алгоритмів методів).

Порівняємо вищенаведені методи більш детально, відповідно до означених критеріїв.

Метод повного резервного копіювання (Рис. 1.). За часовими характеристиками метод є найефективнішим тому, що під час проведення дублювання, не відбувається аналіз файлової структури джерела інформації та резервних копій. Формула часових затрат буде наступною:

$$T_{full} = T_c, \tag{1}$$

де T_{full} – час, затрачений на проведення процедур резервування методом повного резервного копіювання

T_c – час, затрачений на перенесення (копіювання) даних з джерела інформації до резервної копії

За характеристикою об'єму резервної копії метод є найменш ефективним, оскільки кожного наступного резервування проводиться перенесення точної копії джерела інформації, без урахування того, відбувалася модифікація даних в ньому чи ні. Формула об'єму резервної копії буде наступною:

$$S_{full} = \sum_{1}^{x} S_{rez}, \tag{2}$$

де S_{full} – повний об'єм зарезервованих даних методом повного резервного копіювання,

x – кількість резервних копій,

S_{rez} – об'єм резервної копії при x - резервуванні

Використовуючи метод повного резервного копіювання, можливо провести відновлення даних станом на дату резервування. Для проведення резервування, попередні резервні копії не потрібні. Для проведення відновлення необхідна наявність лише однієї резервної копії. Дані, зарезервовані наведеним методом, можливо зберігати як централізовано, так й розподілено. Створення паралельної системи обробки джерела інформації недоцільно. Це пов'язано з відсутністю потреби в аналізі джерела при резервуванні. Складність методу можливо оцінити за складністю його алгоритму [3]. Алгоритм метода є лінійним, тому його складність складає – $O(n)$.

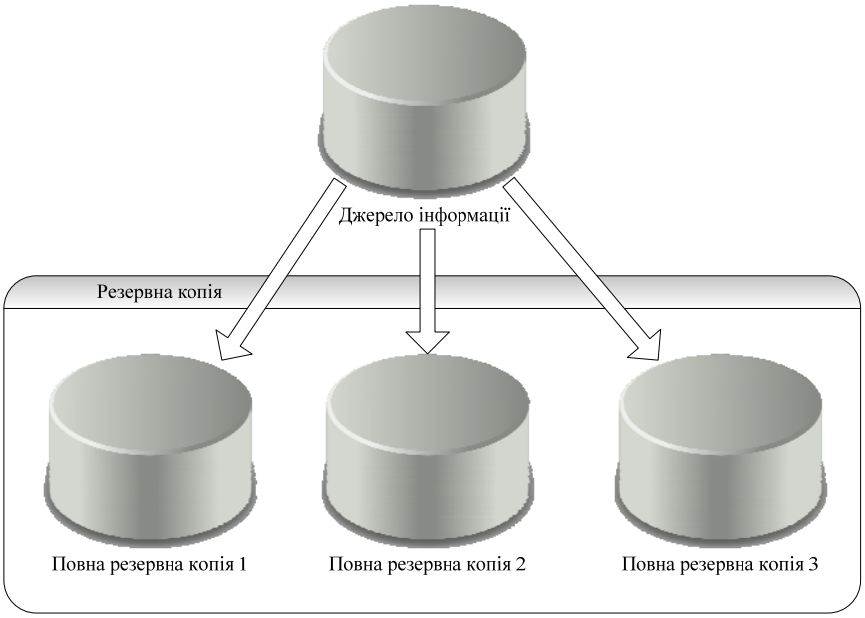

Рис. 1. Візуалізація методу повного резервного копіювання

Метод диференціального резервного копіювання (Рис. 2.). Резервування за цим методом потрібно розділити на два етапи. При першому резервуванні, дублювання відбувається без файлової структури джерела інформації (метод повного резервного копіювання). Кожне наступне резервування супроводжується проведенням аналізу блоків даних джерела інформації та порівнянням його з блоками даних повної резервної копії. Отже, оцінка затрат часу при першому резервуванні за цим методом буде наступною:

$$T_{diff_1} = T_{full}, \tag{3}$$

де T_{diff_1} – час, затрачений на проведення процедур резервування методом диференціального резервного копіювання (перше резервування),

При кожній наступній процедурі резервування затрати часу будуть наступними:

$$T_{diff_x} = T_a + T_c, \tag{4}$$

де T_{diff_x} – час, затрачений на проведення процедур резервування методом диференціального резервного копіювання (x – резервування, $x > 1$),

T_a – час, затрачений на порівняння джерела інформації з першою (повною) резервною копією.

Механізми методу дозволяють резервувати тільки ті ланцюги даних, які були змінені з часу останньої повної копії. Отже при достатній частоті резервувань можна досягти помірного росту резервної копії та забезпечити можливість відновлення до потрібної дати. Формула об'єму резервної копії буде наступною:

$$S_{diff_{(f)}} = S_{full} + \sum_{2}^{x} S_{ad}, \tag{5}$$

де $S_{diff_{(f)}}$ – об'єм резервної копії методом диференціального резервного копіювання, в разі необхідності зберігання всіх станів джерела інформації,

S_{ad} – об'єм додаткової інформації при x – резервуванні

В разі, коли потрібно зберігати лише останній стан системи, формула об'єму резервної копії буде наступною:

$$S_{diff_{(last)}} = S_{full} + S_{ad_x}, \tag{6}$$

де $S_{diff_{(last)}}$ – об'єм резервної копії при резервуванні останнього стану джерела інформації.

Аналізуючи метод, треба зауважити, що можливість відновлення залежить від обраного механізму резервування. Якщо обрано резервування останніх станів джерела інформації, то відновлення можливе лише на час першого та останнього. Для проведення процедури резервування потрібна наявність тільки першої (повної) резервної копії. В разі виконання відновлення до x – стану потрібна повна резервна та x – копія. Складність алгоритму метода у кращому випадку складає – $O(n)$, у гіршому – $O(n^3)$.

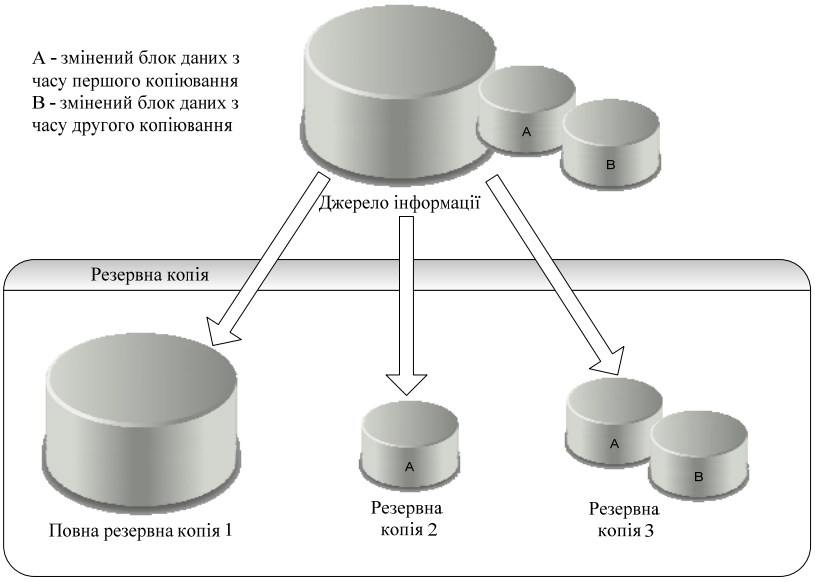

Рис. 2. Візуалізація методу диференціального резервного копіювання

Метод інкрементального резервного копіювання (Рис. 3.). Резервування даних за цим методом розділяється на три етапи, а починаючи з третього етапу – на дві стадії: на першому етапі (при першому резервуванні) відбувається дублювання даних за методом повного резервного копіювання. На другому етапі (при другому резервуванні) – методом диференціального резервування. Починаючи з третього резервування, на першій стадії, відбувається аналіз файлової структури всіх попередніх резервних копій, на другій фазі – безпосереднє перенесення даних. Сформулюємо формулу затрат часу для метода:

$$T_{inc_1} = T_{full}, \tag{7}$$

де T_{inc_1} – час, затрачений на проведення процедур резервування методом інкрементального резервного копіювання (перше резервування).

Якщо спростити формулу затрат часу на друге резервування, то загальна формула затрат часу буде наступною:

$$T_{inc_x} = \sum_{2}^{x-1} T_a + T_c, \tag{8}$$

де T_{inc_x} – час, затрачений на проведення процедур резервування методом інкрементального резервного копіювання (x – резервування, $x > 1$).

Резервування цим методом дозволяє оптимально використовувати простір резервної копії, за рахунок дублювання лише змінених ланцюгів даних (або блоків – у видку використання блочного інкрементального методу). Формула об'єму резервної копії буде наступною:

$$S_{inc} = S_{full} + \sum_{2}^{x} S_{ad}, \tag{9}$$

де S_{inc} – об'єм резервної копії при дублюванні даних методом інкрементального резервного копіювання.

Однак потрібно зазначити, що для проведення резервування чи відновлення потрібні всі попередні резервні копії. В разі втрати або пошкодження однієї з частин, відновлення буде неможливим. Метод дозволяє зробити відновлення станом на дату проведення резервування. Складність алгоритму вищезазначеного метода у кращому випадку складає – $O(n)$, у гіршому – $O(n^4)$.

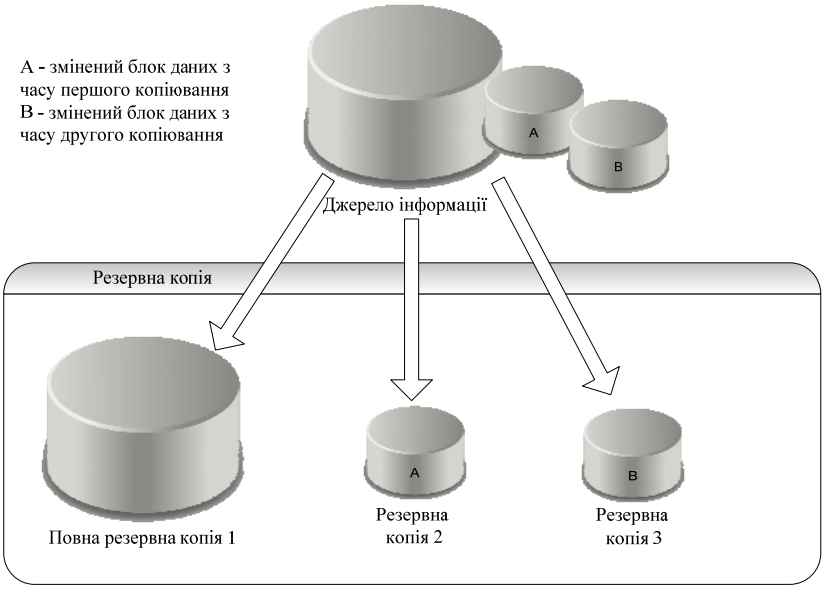

Рис. 3. Візуалізація методу інкрементального резервного копіювання

Вибір методу для програмного резервування інформаційних ресурсів залежить від періодичності змін у файловій системі, вимог до відновлюваності (збереження попередніх станів джерела інформації), вартості резервних копій тощо. Одним, та водночас, важливим недоліком цих методів є їх залежність від платформи на який проводиться резервування. Тобто сервер або ПЕОМ, на якому проводиться резервування, відповідає за обробку резервних копій, проводить обчислення відповідних змін (при резервуванні або при відновленні). Вищезазначений факт накладає обмеження на використання інкрементального та диференціального методів у системах з високим рівнем доступності (за рахунок порівняно великих затрат часу проведення резервування). Одним з рішень проблем значної обчислювальної складності є розпаралелення задачі аналізу файлової структури джерел інформації та резервних копій при проведенні резервного копіювання. Для цього пропонується адаптувати метод інкрементального резервування, до паралельної обробки. Треба зауважити, що перехід до паралельної обробки доцільний лише за умови достатнього рівня розпаралелюваності алгоритмів та порівняно великої кількості процесорів (вузлів обчислення) [4, 5].

За основу візьмемо блочний інкремент. Розпаралелення можливе як на рівні локальної машини (ПЕОМ або сервера), так і на рівні робочої групи або домену. У випадку паралельної обробки, затрати часу будуть обчислюватись за наступною формулою:

$$T_{inc_x}(p) = \sum_{2}^{x-1} \frac{T_a}{p} + \sum_{1}^{z} \frac{T_c}{k \cdot r}, \tag{10}$$

де $T_{inc_x}(p)$ – час, затрачений на проведення процедур резервування методом паралельного інкрементального резервного копіювання (x – резервування, $x > 1$),

p – кількість процесорів (вузлів обробки) в системі резервування,

z – кількість блоків, на яку розбивається джерело інформації (відповідно до методу блочного інкременту),

k – коефіцієнт затримок при передачі даних між вузлами,

r – кількість вузлів зберігання, між якими розподілена резервна копія.

Розпаралелення задачі резервування інкрементальним методом значно прискорює аналіз файлової структури джерела інформації та порівняння його з резервними копіями. Застосовуючи такий підхід можливо зберігати резервні копії як централізовано ($k = 1, r = 1$), так й розподілено. При цьому формула об'єму резервної копії в загальному випадку буде такою ж самою, як й для звичайного методу інкрементального резервного копіювання.

Висновки

Використання методів резервного копіювання для забезпечення надійного збереження даних у інформаційних системах, як альтернатива апаратним засобам резервування, зменшує витрати на утримання інфраструктури. Інтеграція методу паралельного інкрементального резервного копіювання до програмно-інформаційних комплексів (в тому числі й тих, що використовуються в Державній митній службі України), систем керування базами даних, у якості модулю дублювання даних, дозволяє: мінімізувати час простоїв серверного обладнання підчас виконання завдань обслуговування масивів даних, досягти оптимального розміщення даних у резервних копіях.

Література

1. Острейковский В. А. Теория надежности / В.А. Острейковский. – М.: Высшая школа, 2003. – 463 с.
2. K.A. Anderson and B. H. Kirouac. A Simple and Free System for Automated Network Backups // In The Third Annual System Adminstration, Networking and Security Conference (SANS III), 1994, p. 63– 68.
3. Introduction to Algorithms. 2-nd edition / T. Cormen, C. Leiserson, R. Rivest, C. Stein. – London: The MIT Press, 2001. – 1180 p.
4. J. da Silva, O. Gudmundsson, and D. Mosse. Performance of a Parallel Network Backup Manager // In USENIX Conference Proceedings, 1992, p. 17 – 26.
5. Воеводин В.В. Параллельные вычисления / В.В. Воеводин, Вл. В. Воеводин. – СПб.: БХВ-Петербург, 2002. – 608 с.

Надійшла 11.4.2011 р.

УДК 004.912

Р.С. ЯРМОЛЮК
Хмельницький національний університет

ЗАДАЧА АНАЛІЗУ ТЕКСТОВИХ АТРИБУТІВ В ЕЛЕКТРОННОМУ КАТАЛОЗІ

Розглянута одна з актуальних проблем: задача аналізу текстових атрибутів запису в електронних каталогах. Проведено огляд та аналіз основних задач, що виникають при обробці та аналізі текстової інформації. Запропоновано ефективні алгоритми розв'язку кожної із задач.

Considered one of the critical problems: the problem of analysis of text attribute record in electronic catalogs. The review and analysis of the main problems arising in the processing and analysis of text information. The efficient algorithms for solving each problem.

Ключові слова: текстовий рядок, зіставлення рядків, відстань між рядками, нечітке порівняння рядків, текстові алгоритми, наївний алгоритм.

Постановка проблеми

Бібліотека у інформаційному суспільстві є невід'ємною складовою життя. Електронний каталог, як інформаційна система, що забезпечує доступ до бібліографічних баз даних бібліотеки повинен відповідати ряду вимог [1,2]. Одним із основних критеріїв якості електронного каталогу бібліотеки є наявність механізмів верифікації та пошуку помилок у бібліографічних записах. Якість електронного каталогу безпосередньо залежить від його інформаційного наповнення (наявності різного типу помилок у записах). Основні атрибути бібліографічного запису (назва, автор, видання тощо) мають текстовий тип даних. Тому основними задачами, що виникають при верифікації записів електронного каталогу, є задачі аналізу та обробки текстової інформації.

Аналіз останніх досліджень і публікацій

Теоретичні і практичні основи проблеми автоматичного пошуку та корекції помилок в записах електронного каталогу розробляли Вершинин М. И., Белоногов Г. Г., Бабко-Малая О. Б., Крауш А. С., Randall B. N., Ballard T, та інші.

Алгоритми аналізу текстових рядків представлені у роботах: Stephen G.A., Knuth D.E., Morris J.H., Pratt V.R., Karp R.M., Rabin, M.O., Boyer R.S., Moore J.S., Fischer M.J., Hirschberg D.S., Hunt J.W., Szymanski T.G., Landau G.M., Vishkin U. Хемминг Р.В., Левенштейн В.И.

Формулювання цілей статті та актуальність досліджень

На даний час існує багато методів та алгоритмів аналізу текстових рядків. Проблема полягає у практичній реалізації даних алгоритмів у системах пошуку та корекції помилок в текстових даних. Аналіз описових можливостей переважної більшості популярних автоматизованих бібліотечних інформаційних систем (АБІС) показав відсутність ефективних засобів для аналізу та обробки текстових даних в атрибутах бібліографічного запису. Тому проблема аналізу текстових даних в системах електронного каталогу бібліотек є актуальною.

Виклад основних матеріалів дослідження

До основних задач аналізу текстових рядків, що виникають при розробці засобів верифікації інформації в електронних каталогах, відносяться:

- задача зіставлення текстових рядків;
- задача розрахунку відстані між текстовими рядками;
- задача нечіткого порівняння текстових рядків;
- задача пошуку найдовшого повторюваного текстового підрядка.

Дамо означення основним поняттям. Під текстовим рядком будемо розуміти послідовність символів певного скінченного алфавіту. Текстовий рядок x довжини $|x| = m$ записується, як $x_1 x_2 \ldots x_m$, де x_i представляє собою i-й символ текстового рядка x. Текстовий підрядок $x_i x_{i+1} \ldots x_j$ рядка x, де $i \le j \le m$,