Polskie Towarzystwo Badań Operacyjnych i Systemowych
Instytut Badań Systemowych Polskiej Akademii Nauk
Wojskowa Akademia Techniczna

Redaktorzy:
Zbigniew Nahorski
Marian Chudy
Andrzej Straszak

OPTYMALIZACJA

Warszawa 1991

POLSKIE TOWARZYSTWO
BADAŃ OPERACYJNYCH I SYSTEMOWYCH

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

WOJSKOWA AKADEMIA TECHNICZNA

O P T Y M A L I Z A C J A

Z A D A N I A, M E T O D Y, A L G O R Y T M Y

Redaktorzy

Zbigniew Nahorski, Marian Chudy, Andrzej Straszak

WARSZAWA 1991

OPTYMALIZACJA:
Zadania, metody, algorytmy
PTBOiS, Warszawa 1991.

OPTYMALIZACJA STRATEGII KOMPRESJI DANYCH W BAZIE DANYCH

Tadeusz Nowicki
Wojskowa Akademia Techniczna, Warszawa

Streszczenie: Rozważany jest problem wyznaczania optymalnych wartości parametrów strategii sekwencyjnych kompresji zbiorów danych w bazie danych. Zastosowano znaną z literatury metodę wyznaczania sekwencyjnej strategii optymalnej w profilaktycznych obsługach systemów z odnawialnymi elementami dla szczególnych założeń wynikających z uwarunkowań eksploatacji baz danych.

1. Wstęp

Jednym z parametrów baz danych, uważanym za podstawowy z punktu widzenia użytkownika jest nieprzekraczalny czas reakcji systemu z bazą danych na żądanie dowolnej danej. Przekroczenie przyjętego progowego czasu reakcji można uznać w sensie niezawodnościowym jako awarię takiego systemu. Ciągła aktualizacja baz danych powoduje, że dane z poszczególnych zbiorów danych są rozmieszczone w oddzielnych obszarach fizycznych pamięci systemu komputerowego. Ich logiczna spójność zachowana jest poprzez zapis adresów kolejnych w uszeregowaniu danych. Poszukiwanie konkretnej danej powoduje serię fizycznych odwołań do pamięci systemu komputerowego. W ten sposób, wraz z liczbą aktualizacji bazy danych, rośnie czas dostępu do danych w poszczególnych zbiorach. Proces aktualizacji zbiorów danych oraz rozmieszczenia danych w różnych obszrach fizycznych jest w swej naturze losowy. W celu uniknięcia wzrostu czasu reakcji systemu z bazą danych na żądanie użytkownika przeprowadza się co jakiś czas tzw. kompresję zbiorów danych. Kompresja pojedynczego zbioru danych polega na przepisaniu danych ze zbioru, rozrzuconych w różnych miejscach pamięci, w jeden fizycznie spójny obszar. Ponadto dane ułożone są wtedy w kolejności ustalonej kluczem ich wyszukiwania. Zmniejsza się w ten sposób istotnie czas reakcji systemu. W związku z ograniczeniami kosztowymi kompresja bazy danych nie może być przeprowadzana zbyt często. Ponadto, ograniczenia te powodują, że w normalnych warunkach, tzn. przy niezaistnieniu awarii systemu, kompresji podlegać może jedynie część zbiorów danych.

Eksploatację bazy danych podzielimy na sekwencję rozłącznych okresów, dla których na początku każdego z nich wyznacza się jego długość oraz zbiory danych, które podlegać będą kompresji po tym okresie. Założymy, że jeśli w trakcie któregoś z okresów nastąpi awaria systemu, rozumiana jako przekroczenie progowej wartości czasu reakcji systemu, to kompresji podlegać będą wszystkie zbiory danych i chwila zakończenia pełnej kompresji będzie początkiem nowego okresu eksploatacji systemu. Tak opisaną strategię postępowania nazywać będziemy strategią sekwencyjnych kompresji zbiorów danych w bazie danych.

2. Model matematyczny.

Załóżmy, że mamy do czynienia z bazą danych o N zbiorach danych. Jeśli w pewnej chwili czas dostępu do danej z konkretnego zbioru danych przekroczy zadaną, progową wartość , to uznamy, że nastąpiła awaria całego systemu. Zewnętrzne przyczyny aktualizacji zbiorów danych pozwalają przyjąć upraszczające założenie o niezależności awarii związanych z poszczególnymi zbiorami. Mamy zatem N-elementowy system o szeregowej strukturze niezawodnościowej.

Przez kompresję wymuszoną rozumieć będziemy usunięcie przyczyn awarii systemu, zatem zgodnie z założeniem, natychmiast po awarii wykonana będzie kompresja wszystkich zbiorów danych. Kompresją profilaktyczną nazywać będziemy kompresję wyznaczonych uprzednio zbiorów danych po ustalonym wcześniej okresie eksploatacji systemu nie przerwanym awarią tego systemu.

Strategią sekwencyjnych kompresji profilaktycznych zbiorów bazy danych nazywać będziemy poniższy sposób postępowania :

a/ przed każdym kolejnym okresem eksploatacji bazy danych, zwanym krótko krokiem, wyznaczamy czas do kolejnej kompresji profilaktycznej i zbiory danych, które podlegać mają tej kompresji,

b/ jeśli w wyznaczonym okresie czasu system nie ulegnie awarii, to dokonujemy kompresji wybranych zbiorów danych i powracamy do czynności określonych w punkcie a/,

c/ jeśli przed upływem czasu do kompresji profilaktycznej nastąpi awaria systemu, to dokonuje się kompresji wymuszonej i kończy się w ten sposób kolejny okres eksploatacji bazy danych oraz powracamy do czynności określonych w punkcie a/.

Przyjmijmy, że jesteśmy na początku k-tego okresu (kroku) w eksploatacji bazy danych. Jeśli przez E oznaczymy zbiór numerów zbiorów danych, to Ω symbolizować będzie dowolny podzbiór zbioru E . Niech T będzie czasem od ostatniej kompresji, profilaktycznej lub wymuszonej, do kolejnej planowanej kompresji profilaktycznej. Modelem strategii sekwencyjnych kompresji profilaktycznej w danym kroku będzie zatem para (T,Ω) , którą należy wyznaczyć z punktu widzenia przyjętego kryterium jakości.

Załóżmy, że zmienne losowe τ_n oznaczające czasy pomiędzy kompresjami n-tego zbioru danych, a wystąpieniem najwcześniejszych awarii związanych z tym zbiorem są zmiennymi losowymi o identycznych rozkładach prawdopodobieństwa i intensywnościach uszkodzeń [1] $\lambda_n(t)$, $t \geq 0$, $n = \overline{1,N}$. W praktyce założenie to można przyjąć, a funkcje $\lambda_n(\cdot)$, $n = \overline{1,N}$, są monotonicznie rosnące.

Przyjmijmy, że znamy wielkość $a_n^{(k)}$ przedziału czasu, jaki upłynął od ostatniej kompresji n-tego zbioru danych do początku k-tego kroku, $n = \overline{1,N}$. Przez $\tau_n^{(k)}$ oznaczmy zmienną losową będącą czasem do wystąpienia awarii n-tego zbioru danych liczonym od początku k-tego kroku, a przez

$\tau^{(k)}$ oznaczmy analogiczną zmienną losową odnoszącą się do całego systemu. Zachodzi równość :

$$(1) \qquad \tau^{(k)} = \min\left\{ \tau_n^{(k)},\ n = \overline{1,N} \right\} .$$

Związek pomiędzy zmiennymi losowymi τ_n i $\tau_n^{(k)}$ jest następujący :

$$(2) \qquad P\left\{ \tau_n^{(k)} \geq x \right\} = P\left\{ \tau_n \geq a_n^{(k)} + x \ / \ \tau_n \geq a_n^{(k)} \right\} .$$

Przez $F^{(k)}(t)$ oznaczmy dystrybuantę zmiennej losowej $\tau^{(k)}$, a przez $\lambda^{(k)}(t)$ jej funkcję intensywności. Przez $F_n^{(k)}(t)$ i $\lambda_n^{(k)}(t)$ oznaczymy analogiczne funkcje charakteryzujące zmienne losowe $\tau_n^{(k)}$, $n = \overline{1,N}$.

Prawdziwe są następujące zależności:

$$(3) \qquad F^{(k)}(t) = 1 - \prod_{n=1}^{N} (1 - F_n^{(k)}(t)) \ , \quad \lambda^{(k)}(t) = \sum_{n=1}^{N} \lambda_n^{(k)}(t) ,$$

ponieważ system z bazą danych ma szeregową strukturę niezawodnościową zbiorów danych. Wprowadzając dla uproszczenia dalszych zapisów tzw. funkcje niezawodności zdefiniowane następująco:

$$(4) \qquad R^{(k)}(t) = 1 - F^{(k)}(t) \ , \quad R_n^{(k)}(t) = 1 - F_n^{(k)}(t) \ , \quad n = \overline{1,N}$$

uzyskujemy równość

$$(5) \qquad R^{(k)}(t) = \prod_{n=1}^{N} R_n^{(k)}(t)$$

co wynika również z faktu, że rozpatrywany system ma szeregową strukturę niezawodnościową. Z (2) wynika, że :

$$(6) \qquad \lambda_n^{(k)}(x) = \lambda_n(a_n^{(k)} + x)$$

Kryterium jakości wyboru pary (T,Ω) w k-tym kroku są oczekiwane, jednostkowe koszty poniesione w tym kroku z powodu kompresji zbiorów danych i ewentualnej awarii systemu. Wprowadźmy oznaczenia :

d_n – koszt kompresji n-tego zbioru danych (prognozuje się go na podstawie wartości $a_n^{(k)}$ oraz przeprowadzonych testów na zbiorach danych) , $n = \overline{1,N}$,

h- koszt związany z testowaniem zbiorów danych,

c- koszt związany z usunięciem skutków awarii,

$K_1^{(k)}(T,\Omega)$ - oczekiwane koszty eksploatacji systemu w k-tym kroku przy założeniu, że krok ten zakończył się kompresją profilaktyczną,

$K_2^{(k)}(T,\Omega)$ - analogiczne koszty przy założeniu, że krok k-ty zakończył się kompresją wymuszoną,

$K^{(k)}(t,\Omega)$ - oczekiwane koszty eksploatacji systemu z bazą danych.

Ponieważ

$$K_1^{(k)}(T,\Omega) = h + \sum_{n\in\Omega} d_n = K_1^{(k)}(\Omega) \tag{7}$$

oraz

$$K_2^{(k)}(T,\Omega) = c + h + \sum_{n=1}^{N} d_n = K_2^{(k)} , \tag{8}$$

to

$$K^{(k)}(t,\Omega) = R^{(k)}(T)\ K_1^{(k)}(\Omega) + F^{(k)}(T)\ K_2^{(k)} . \tag{9}$$

Oczekiwany czas pomiędzy kolejnymi kompresjami wyznacza się według wzoru

$$E\left\{ \min\ (T,\tau^{(k)}) \right\} = \int_0^T R^{(k)}(t)dt . \tag{10}$$

Funkcja kryterium wyboru pary (T,Ω) w k-tym kroku ma zatem postać:

$$C^{(k)}(T,\Omega) = \frac{K^{(k)}\ (T,\Omega)}{\int_0^T R^{(k)}(t)dt} . \tag{11}$$

Istnieje obawa, że para (T^*,Ω^*) spełniająca minimum funkcji (11) zapewni jedynie minimum lokalne, a chcemy, aby uzyskane w k-tym kroku rozwiązanie miało związek z eksploatacją systemu w dłuższym okresie. Propozycję takiego związku przedstawiono w [1].

Jeśli $Z^{(k)}(T,\Omega)$ jest oczekiwanym czasem poprawnej pracy systemu po zakończeniu k-tego kroku, a τ_0 jest wartością progową , która musi być osiągnięta przez $Z^{(k)}(T,\Omega)$ dla $(T,\Omega) = (T^*,\Omega^*)$, to związek, o którym mowa, ma postać:

4. Metoda rozwiązania zadania optymalizacji.

Efektywna metoda rozwiązania zadania (17)-(18) przy wykluczeniu warunku (16) ze zbioru ograniczeń (18) została zaproponowana w [1]. Jej idea oparta została na następujących własnościach funkcji $C^{(k)}(T,\Omega)$ i $Z^{(k)}(T,\Omega)$:

a/ jeśli $\lambda^{(k)}(t)$ jest monotonicznie rosnąca, spełnienie tego warunku w naszym przypadku łatwo jest pokazać, to funkcja $C^{(k)}(T,\Omega)$ ma co najwyżej jedno minimum ze względu na argument T

b/ dla dowolnego $\Omega \subset E$ funkcja $Z^{(k)}(T,\Omega)$ ma co najwyżej jedno minimum ze względu na argument T,

c/ jeśli $\Omega_1 \subset \Omega_2$, to $C^{(k)}(T_1,\Omega_1) \le C^{(k)}(T_2,\Omega_2)$, gdzie T_1 i T_2 są optymalnymi długościami okresu czasu do kompresji profilaktycznejodpowiednio dla danych Ω_1 i Ω_2.

Z b/ oraz z (12) wynika, że obszar dopuszczalnych wartości zmiennej decyzyjnej T dla danego Ω jest postaci

$$(19) \qquad B_\Omega = (0,T^1_\Omega] \cup [T^2_\Omega,\infty), \qquad \Omega \subset E.$$

Z własności c/ wynika natomiast wniosek, że jeśli dla pewnych Ω_1 i Ω_2 zachodzi

$$(20) \qquad C^{(k)}(T_1,\Omega_1) < C^{(k)}(T_2,\Omega_2),$$

to każdy zbiór zawierający zbiór Ω_2 na pewno nie jest elementem rozwiązania optymalnego zadania (17)-(18).

Reasumując, własności a/ i b/ pozwalają efektywnie wyznaczyć optymalne T^* dla ustalonego Ω. Własność c/ jest postawą szybkiej eliminacji zbiorów Ω w algorytmie wyznaczania pary (T^*,Ω^*). Wprowadźmy pojęcia:

zbiory pozytywne - istnieje przypuszczenie, że mogą się one zawierać w zbiorze optymalnym Ω^*,

zbiory negatywne - jesteśmy pewni, że nie zawierają się one w zbiorze Ω^*.

Zbiorem negatywnym, zgodnie z wcześniej opisanymi własnościami, jest zbiór Ω, dla którego optymalny okres kompresji profilaktycznej leży wewnątrz obszaru dopuszczalności B_Ω i jednocześnie minimalny koszt jednostkowy w danym kroku jest większy niż minimalny koszt

(12) $$Z^{(k)}(T,\Omega) \geq \tau_o \quad .$$

W [1] zaprezentowano postać funkcji $Z^{(k)}(T,\Omega)$ i można pokazać, że dla naszego przypadku opisana jest zależnościami:

(13) $$Z^{(k)}(T,\Omega) = F^{(k)}(T)\ Z_1^{(k)}(T,\Omega) + R^{(k)}(T)\ Z_2^{(k)}(T,\Omega)$$

gdzie

(14) $$Z_1^{(k)}(T,\Omega) = Z_1^{(k)} = \int_0^\infty \exp\left[-\int_0^t \sum_{n=1}^{N} \lambda_n(u)du\right]dt \quad ,$$

(15) $$Z_2^{(k)}(T,\Omega) = \int_0^\infty \exp\left[-\int_0^t \sum_{n\in\Omega} \lambda_n(u) + \sum_{n\in\Omega} \lambda_n(u+a_n^{(k)}+T)du\right]dt$$

gdzie $Z_1^{(k)}$ jest oczekiwanym czasem poprawnej pracy systemu po k-tym kroku, jeśli krok ten zakończył się awarią, a $Z_2^{(k)}(T,\Omega)$ jest analogicznym czasem, przy założeniu, że k-ty krok zakończył się kompresją profilaktyczną.

W [1] pokazano również jak wyznaczyć można praktycznie użyteczną wartość τ_o lecz dla jasności rozważań nie będziemy tego przytaczać.

W praktyce istnieje jeszcze jedno ograniczenie na zmienną decyzyjną. Kompresje profilaktyczne nie powinny zazwyczaj przekraczać pewnego założonego kosztu d, co daje nam ograniczenie:

(16) $$\sum_{n\in\Omega} d_n \leq d \quad .$$

3. Sformułowanie zadania optymalizacji.

Zadanie wyznaczania optymalnej strategii sekwencyjnych kompresji zbiorów w bazie danych w k-tym kroku ma postać:

wyznaczyć parę (T^*,Ω^*) taką, aby

(17) $$C^{(k)}(T^*,\Omega^*) = \min_{(T,\Omega)\in D} C^{(k)}(T,\Omega)$$

gdzie

(18) $$D = \left\{(T,\Omega)\in R_+ x 2^E \ : \ Z^{(k)}(T,\Omega) \geq \tau_o \ , \sum_{n\in\Omega} d_n \leq d \right\} ,$$

a 2^E jest symbolem zbioru wszystkich podzbiorów zbioru E.

jednostkowy odpowiadający dowolnemu innemu zbiorowi. Algorytm wyznaczania pary (T^*,Ω^*) po uwzględnieniu modyfikacji ograniczeniem (16) ma następującą postać:

1°. Rozpoczynamy od zbioru pustego Ω_o; obliczamy dla niego T^*_o oraz $C^{(k)}(T^*_o,\Omega_o)$.

2°. Rozpatrujemy wszystkie jednoelementowe zbiory Ω obliczając dla nich również T^*_Ω oraz $C^{(k)}(T^*_\Omega,\Omega)$.

3°. Spośród zbiorów jednoelementowych wybieramy jedynie zbiory pozytywne. Dodając do uzyskanych zbiorów inne pozytywne zbiory jednoelementowe tworzymy zbiory dwuelementowe.

4°. Eliminujemy spośród nich dwuelementowe zbiory negatywne po uprzednim obliczeniu dla zbiorów optymalnego okresu kompresji i minimalnych kosztów jednostkowych.

5°. Do powstałych zbiorów dodajemy jednoelementowe zbiory pozytywne tworząc zbiory trzyelementowe, eliminując z nich zbiory negatywne itd.

Postępujemy tak aż do wyczerpania wszystkich zbiorów pozytywnych w zbiorze E.

6°. Ze wszystkich uzyskanych zbiorów pozytywnych eliminujemy te, które nie spełniają ograniczenia (16).

7°. Wybieramy z pozostałych zbiorów optymalny zbiór Ω^* i odpowiadający mu optymalny okres T^* kompresji profilaktycznej.

Literatura

[1] Karpiński J., Firkowicz S. - Zasady profilaktyki obiektów technicznych. PWN, Warszawa 1981.

ISBN 83-900412-1-9