INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 20 • 2011

ISSN 1507-3858

Jakub Swacha
Uniwersytet Szczeciński

METODA WYCENY WARTOŚCI ZBIORÓW DANYCH NA POTRZEBY ZARZĄDZANIA CYKLEM ŻYCIA INFORMACJI

Streszczenie: W artykule opisano nową metodę wyceny wartości zbiorów danych, która może być wykorzystana na potrzeby zarządzania cyklem życia informacji. Metoda uwzględnia liczbę i moment przeprowadzenia operacji wykonywanych na zbiorach danych w okresie poprzedzającym wycenę w celu określenia relatywnej wartości zbiorów danych przechowywanych w systemie. W stosunku do istniejących metod tego typu proponowana metoda charakteryzuje się prostotą pod względem zarówno koniecznych do wykonania obliczeń, jak i samej implementacji. W artykule omówiono krótko cel opracowania metody, podstawowy obszar jej stosowania (zarządzanie cyklem życia informacji), przeprowadzono dyskusję literatury, opisano samą metodę i przedstawiono przykład jej zastosowania.

Słowa kluczowe: wartość informacji, wycena wartości zbiorów danych, zarządzanie cyklem życia informacji, zarządzanie przechowywaniem danych.

1. Wstęp

Według opublikowanego przez IDC raportu Gantza wielkość światowych zasobów informacyjnych sięgnie w 2011 r. 1,8 zettabajta ($1{,}8 \cdot 10^{21}$ bajtów) [Gantz i in. 2008]. Z kolei prawdopodobnie największy obecnie na świecie system przechowywania danych Centrum Superkomputerowego San Diego ma pojemność 36 petabajtów [*San Diego Supercomputer*… 2009].

Tak olbrzymia ilość zgromadzonych informacji wymaga nie tylko odpowiednio rozwiniętej infrastruktury technicznej, ale także efektywnych metod zarządzania – zarówno samą przechowywaną informacją, jak i zasobami pamięciowymi służącymi do jej przechowywania [Swacha 2009]. Imponującemu postępowi w technologii przechowania danych [Song, Zhu 2009] towarzyszy mniej głośny, lecz nie mniej ważny postęp w dziedzinie zarządzania przechowywaniem danych. Obecnie wielce obiecującą koncepcją jest zarządzanie cyklem życia informacji (*Information Lifecycle Management*, ILM), które stawia sobie za cel przyporządkowanie informacji, z uwzględnieniem jej użytkowej wartości, najbardziej właściwej i efektywnej ekonomicznie infrastruktury informatycznej, począwszy od chwili pojawienia się informacji w systemie, aż do jej ostatecznego zeń usunięcia [Turczyk i in. 2006].

Kluczowy element zarządzania cyklem życia informacji stanowi wyznaczenie wartości przechowywanej informacji. Jakkolwiek istnieje szereg podejść do rozwiązania tego problemu, które zostaną przedstawione w dalszej części artykułu, trudno wskazać wśród nich jedno, które spełniałoby wszystkie potrzeby omawianego zastosowania. Stąd uzasadnienie do poszukiwania nowych metod wyceny zgromadzonych zasobów informacyjnych. Taką właśnie nową metodę opisano w punkcie 4 niniejszego artykułu. Wcześniej, w punkcie 2, przybliżono istotę zarządzania cyklem życia informacji, a w punkcie 3 omówiono znane metody wyceny wartości informacji. W artykule zawarto także prosty przykład praktyczny zastosowania proponowanej metody i podsumowanie uzyskanych wyników.

2. Zarządzanie cyklem życia informacji

Cykl życia informacji określa sekwencję faz, przez które przechodzi zbiór danych – od jego utworzenia aż po usunięcie [*Storage Management*… 2008]. Poszczególne fazy charakteryzują się różnym rodzajem i stopniem aktywności. W fazie tworzenia dane podlegają częstej modyfikacji; później przechodzą w fazę wysokiej aktywności, kiedy przestają być modyfikowane, a stają się często wykorzystywane. Wraz z upływem czasu aktywność staje się coraz niższa, aż wreszcie zanika niemal zupełnie. Dane przechodzą w fazę archiwalną, a powodem dalszego ich przechowywania mogą być jedynie wymogi prawa lub wewnętrzne regulacje obowiązujące w organizacji. Ostatecznym końcem cyklu życia informacji jest zniszczenie danych.

Różne rodzaje informacji charakteryzują się różnymi cyklami życia. Pewne rodzaje informacji mogą trwać stale w jednej fazie (np. terminarz w fazie tworzenia), inne wykazują nawroty do poprzednich faz (np. informacje dotyczące okresowo powtarzających się procesów). Absolutna długość cyklu życia jest zależna również od rodzaju informacji, może trwać chwilę (w przypadku informacji dotyczącej pomiaru czasu) albo wiele lat (w przypadku danych statystycznych). Na długość i przebieg cyklu życia informacji istotnie wpływają też obowiązujące uregulowania dotyczące wymogów jej przechowywania, szczególnie regulacje prawne [Volonino i in. 2007].

Zarządzanie cyklem życia informacji ma za zadanie zapewnić, że na każdym etapie cyklu życia informacji jest ona przechowywana w sposób spełniający wszystkie warunki istotne dla realizacji procesach biznesowych. Z perspektywy przechowywania danych do najważniejszych procesów zarządzania cyklem życia informacji należy zaliczyć:

- przyporządkowanie docelowych poziomów usług (SLO, *Service Level Objectives*) dla poszczególnych klas zbiorów danych, zależnie od ich wartości i specyficznych wymagań co do warunków przechowywania;
- określenie oferowanego poziomu usług dla poszczególnych urządzeń (ODSL, *Offered Data Service Level*), wyrażającego techniczne możliwości urządzenia przy wykorzystaniu miar użytych do zdefiniowania docelowych poziomów usług;

- wzajemne przyporządkowanie docelowych (SLO) i oferowanych (ODSL) poziomów usług w celu ustalenia najbardziej stosownego miejsca przechowywania danych.

3. Wycena wartości informacji

Określenie wartości informacji zazwyczaj nie jest łatwe. W literaturze przedmiotu dominuje pogląd, że jej wartość daje się w praktyce wyznaczyć tylko w pewnym kontekście, a co za tym idzie, przeznaczenie informacji decyduje o sposobie wyznaczania jej wartości [Repo 1989; Oppenheim i in. 2001; Hubbard 2007]. Na przykład:

- jeżeli traktuje się informację jako przeznaczony na sprzedaż produkt informacyjny (np. w postaci raportu rynkowego), za jej wartość należy przyjąć możliwą do uzyskania cenę sprzedaży, kształtowaną przez istniejące na rynku popyt i podaż,
- jeżeli informacja stanowi parametr wejściowy dla realizowanego wewnątrz organizacji procesu decyzyjnego (np.: ile warte są dla producenta samochodów wyniki badań marketingowych wśród potencjalnych nabywców nowego modelu samochodu, którego rozpoczęcie produkcji jest rozważane?), jej wartość może zostać obliczona opartymi na rachunku prawdopodobieństwa metodami wyceny wpływu wykorzystania informacji na wynik procesu decyzyjnego,
- jeżeli informacja zostanie wykorzystana wewnątrz organizacji do sterowania procesem materialnym (np. produkcyjnym), wartość informacji wyznacza efekt ekonomiczny zmiany przebiegu realizacji procesu w wyniku dostarczenia informacji [Cypryjański 2007].

Wszystkie wymienione wyżej metody niestety nie nadają się do wykorzystania w zarządzaniu cyklem życia informacji. W jego przypadku potrzebne jest bowiem określenie wartości wszystkich przechowywanych zbiorów danych, niezależnie od ich przeznaczenia, a nawet wtedy, gdy jest ono nieznane. Abstrahując nawet od zbyt wąskiego zakresu zastosowań lub zbyt małej szczegółowości wymienionych metod, wszystkie one wymagają istotnego udziału ludzi. Trudno sobie wyobrazić zespół ludzi wykonujący na bieżąco wycenę w systemie, w którym przechowuje się miliony zbiorów danych, a ich zawartość podlega ciągłym zmianom. Oczywistą koniecznością jest użycie metod wyceny, które mogą być zautomatyzowane.

Specyfika zarządzania cyklem życia informacji dopuszcza możliwość znacznego uproszczenia procesu wyceny dzięki temu, że nie wymaga użycia miernika pieniężnego; zadowala się dowolnym miernikiem pozwalającym na uporządkowanie przechowywanych zbiorów danych według ich relatywnej wartości i poprawne ich przyporządkowanie do sprzętowych komponentów systemu najbardziej adekwatnych do ich wartości.

W dotychczasowej ogólnodostępnej literaturze można znaleźć opisy trzech metod wyceny wartości zbiorów danych stworzonych na potrzeby zarządzania cy-

klem życia informacji, autorstwa odpowiednio: Y. Chena [Chen 2005], L. Turczyka i współpracowników [Turczyk i in. 2007] oraz H. Jina i współpracowników [Jin i in. 2008]. Metody te opierają się na dwóch prawach informacji (spośród siedmiu zdefiniowanych przez D. Moody'ego i P. Walsha) [Moody, Walsh 1999] mówiących, że: (II prawo) wartość informacji rośnie z każdym użyciem i (III prawo) wartość informacji maleje z czasem.

Metoda Y. Chena opiera się na zliczaniu liczby wykonanych na zbiorze danych operacji dostępu (zwanych użyciami) w następujących po sobie okresach czasu (zwanych etapami życia). Liczba użyć w poszczególnych etapach jest normalizowana, ważona (tak, by ostatnie etapy życia miały większy wpływ na wycenę informacji niż wcześniejsze), sumowana i skalowana do pewnego założonego przedziału wartości (np. 1,10).

Wartość zbioru danych *d* według Chena można wyznaczyć według poniższej formuły [Chen 2005]:

$$V_t(d) = \left(\sum_{i=1}^{N_t} \left(w(i) \cdot f(U_i(d)) \right) \right) \cdot us + ls, \tag{1}$$

gdzie: $V_t(d)$ – wartość zbioru danych *d* w czasie *t*,
N_t – liczba etapów życia,
d – numer etapu życia (mniejszy numer oznacza późniejszy etap),
$w(i)$ – znormalizowana waga *i*-tego etapu życia:

$$w(i) = \frac{\left(\frac{1}{x}\right)^i}{\sum_{j=1}^{N_t} \left(\frac{1}{x}\right)^j}, \quad \sum_{i=1}^{N_t} w(i) = 1, \tag{2}$$

gdzie: x – współczynnik wpływający na relatywne wagi poszczególnych etapów życia (im wyższy, tym wagi stają się bardziej oddalone od siebie),
$Ui(d)$ – liczba użyć zbioru danych $\underline{d}$ w etapie życia *i*,
$f(U_i(d))$ – funkcja normalizująca liczbę użyć do przedziału <0, 1>, np.:

$$f(U_i(d)) = \begin{cases} 1 \Leftrightarrow U_i(d) > sf \\ \dfrac{U_i(d)}{sf} \Leftrightarrow U_i(d) \le sf \end{cases}, \tag{3}$$

gdzie: sf – współczynnik skalowania liczby użyć, np. 10,
us – zakres skali wartości, np. 9,
ls – minimum skali wartości, np. 1.

Jako główne wady metody Chena należy wskazać, że jest dość skomplikowana obliczeniowo, dla każdego zbioru danych wymaga utrzymywania wielu liczników użyć, a liczba, długość i wagi etapów życia, jak i współczynniki skalujące, są wyznaczane arbitralnie, przyjęcie zaś dla nich nietrafionych wartości może doprowadzić do otrzymania wyników wyceny bez przydatności praktycznej (w skrajnym przypadku – przypisujących tę samą wartość wszystkim przechowywanym zbiorom danych).

Metoda L. Turczyka [Turczyk i in. 2007] zaczyna się od podzielenia przechowywanych zbiorów danych na grupy według wartości jednego z trzech atrybutów:

- przedziału wieku pliku (w dniach),
- przedziału liczby zarejestrowanych operacji dostępu do pliku,
- typu pliku.

Dla każdej grupy uzyskanej w wyniku takiego podziału określa się następnie parametry opartego na rozkładzie prawdopodobieństwa gamma lub Weibulla modelu probabilistycznego dla liczby dni, które upłynęły od ostatniej operacji dostępu do zbioru danych. Model weryfikuje się, wykorzystując test dopasowania χ^2 z poziomem istotności 0,001, co oznacza, że nie dla każdej grupy zbiorów danych uda się znaleźć adekwatny model. Po podstawieniu liczby dni, które upłynęły od ostatniej operacji dostępu do zbioru danych do uzyskanego modelu (właściwego dla grupy, do której ten zbiór danych należy), otrzymuje się wyrażone w procentach prawdopodobieństwo ponownego użycia zbioru danych w przyszłości, które można traktować jako miarę wartości tego zbioru danych.

Jako główne wady metody Turczyka należy wskazać następujące:

- jest jeszcze bardziej skomplikowana obliczeniowo niż metoda Chena,
- jest zawodna: niezależnie od kryterium grupowania, nie dla każdej grupy zbiorów danych udaje się znaleźć właściwy im model, co podaje w wątpliwość praktyczną przydatność tej metody do zarządzania cyklem życia informacji.

H. Jin dąży w swej metodzie [Jin i in. 2008] do uwzględnienia podaży informacji i popytu na nie. Stronę podażową reprezentują w jego modelu takie parametry, jak:

- rozmiar zbioru danych $s(d)$,
- szybkość (B), z jaką mogą być realizowane operacje na nim (wynikająca z miejsca jego przechowywania).

Stronę popytową zaś reprezentują:

- spodziewany okres do kolejnego dostępu do zbioru danych (Δt), który może zostać przybliżony średnim czasem pomiędzy dwiema kolejnymi operacjami dostępu do zbioru danych w rozpatrywanym okresie przeszłości,
- liczba użytkowników (N), którzy używali zbioru danych w rozpatrywanym okresie przeszłości,
- stopień zależności pomiędzy zbiorami danych $M \cdot \cos\theta$, gdzie:

$$\cos\theta = \frac{\sum_{i} \cos\theta_{ij}}{l}, \quad i \neq j \tag{4}$$

$$\cos\theta_{ij} = \frac{\sum_k a_{ik} a_{jk}}{\sqrt{\sum_k a_{ik}^2} \cdot \sqrt{\sum_k a_{jk}^2}}, \quad i \neq j, \tag{5}$$

gdzie: l – liczba przechowywanych w systemie zbiorów danych,

a_{ik} – liczba operacji dostępu do zbioru danych i w k-tym dniu rozpatrywanego okresu,

M wskazuje, dla ilu par i, j wartość cos θ przekroczyła pewien próg th.

Wartość zbioru danych d według Jina można wyznaczyć według poniższej formuły:

$$V_t(d) = f(t) \cdot N^2 \cdot \left(1 + M \cdot \cos\theta\right) \cdot \frac{s(d)}{B} \tag{6}$$

Główne wady metody Jina są następujące:

- jest nieco bardziej skomplikowana obliczeniowo niż metoda Chena,
- dla każdego zbioru danych wymaga utrzymywania wielu liczników użyć (dla poszczególnych dni okresu oceny),
- długość rozpatrywanego okresu i wartość progu th są wyznaczane arbitralnie, a przyjęcie dla nich nietrafionych wartości może negatywnie wpłynąć na praktyczną przydatność wyników wyceny,
- ujęcie w formule parametru zależnego od miejsca przechowywania zbioru danych (B) wydaje się mocno problematyczne, zważywszy na to, że celem wykorzystania wyceny w zarządzaniu cyklem życia informacji jest właśnie ustalenie odpowiedniego miejsca przechowywania zbioru danych.

Nową metodę wyceny wartości zbiorów danych na potrzeby zarządzania cyklem życia informacji, przy której opracowywaniu starano się uniknąć wad metod omówionych wyżej, przedstawiono w kolejnym punkcie artykułu.

4. Propozycja nowej metody wyceny wartości zbiorów danych

Przed projektowaną metodą wyceny wartości zbiorów danych postawiono następujące wymagania:

- wartość zbioru danych zależy wyłącznie i jednocześnie od czasu użycia oraz częstotliwości używania zbioru danych, zgodnie z II i III prawem informacji D. Moody'ego i P. Walsha [Moody, Walsh 1999]:
  - zbiory danych używane często mają wyższą wartość od zbiorów używanych rzadko,
  - zbiory danych używane w niedawnej przeszłości mają wyższą wartość od zbiorów używanych w odległej przeszłości;
- możliwe jest określenie wartości dla każdego zbioru danych przechowywanego w systemie, w tym także zbioru danych, który nie był używany ani razu w okresie uwzględnionym w wycenie,

- jest nieskomplikowana obliczeniowo,
- może być używana w trybie ciągłym (dla dowolnego okresu działania systemu).

Gdyby wartość zbioru danych miała zależeć wyłącznie od czasu ostatniego użycia, można by ją określić jednoparametrową funkcją długości okresu od ostatniego użycia zbioru danych. Z kolei gdyby miała zależeć wyłącznie od częstotliwości używania czasu ostatniego użycia, można by ją określić jednoparametrową funkcją liczby użyć w rozpatrywanym okresie. Postulowane uzależnienie wartości zbioru danych jednocześnie od czasu ostatniego użycia oraz częstotliwości używania do zbioru danych wymaga bardziej złożonego podejścia, zainspirowanego rozwiązaniem stosowanym w wariancie D metody predykcji z częściowym dopasowaniem (PPMD) [Howard 1993].

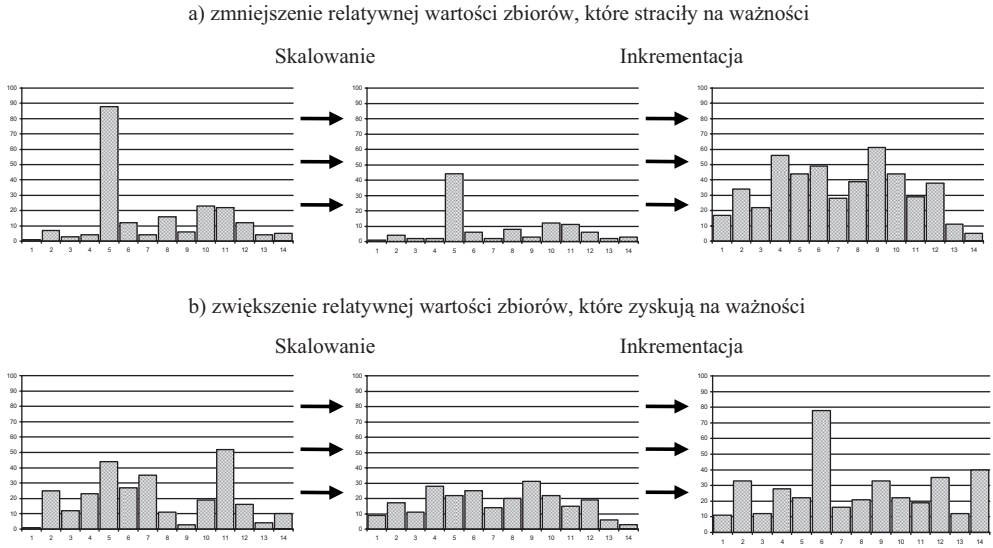
Proponowane rozwiązanie opiera się na wykorzystaniu liczników częstotliwości użycia. Utworzenie zbioru danych uważa się za jego pierwsze użycie, a zatem liczniki częstotliwości użycia inicjowane są wartością 1 (również taką wartością inicjowane są liczniki wszystkich zbiorów danych, które istniały w systemie w momencie rozpoczęcia wyceny).

Użyty zbiór danych	Liczniki częstotliwości użycia zbiorów danych			
	A	B	C	D
A	**2**	1	1	1
A	**3**	1	1	1
A	**4**	1	1	1
B	4	**2**	1	1
A	**5**	2	1	1
Skalowanie, licznik zbioru A osiągnął $c_{max} = 5$				
A	**4**	1	1	1
B	4	**2**	1	1
B	4	**3**	1	1
C	4	3	**2**	1
Skalowanie, suma liczników osiągnęła $c_{summax} = 10$				
B	2	**3**	1	1
B	2	**4**	1	1
C	2	4	**2**	1
C	2	4	**3**	1
Skalowanie, suma liczników osiągnęła $c_{summax} = 10$				
A	**2**	2	2	1

Rys. 1. Przykładowy przebieg procesu inkrementacji liczników częstotliwości użycia dla $c_{max} = 5$ i $c_{summax} = 10$

Źródło: opracowanie własne.

Każdorazowe użycie zbioru danych *d* zwiększa przypisany mu licznik częstotliwości użycia *c*(*d*) o 1. Jeżeli licznik częstotliwości użycia zbioru danych osiągnie wartość maksymalną $c_{\max}$ lub suma liczników częstotliwości użycia wszystkich zbiorów danych osiągnie wartość maksymalną $c_{sum\max}$, następuje procedura skalowania. Skalowanie polega na redukcji o połowę (z zaokrągleniem w górę) wartości wszystkich liczników w modelu, tak by relatywnie zmniejszyć wpływ na wartość licznika użyć wcześniejszych względem użyć późniejszych. Przykład przebiegu procesu inkrementacji liczników częstotliwości użycia pokazano na rys. 1, natomiast korzyści wynikające ze skalowania obrazuje rys. 2.



Rys. 2. Ilustracja korzyści wynikających ze skalowania liczników częstotliwości użycia

Źródło: opracowanie własne.

Należy zwrócić uwagę, że częstotliwość skalowania zmienia się proporcjonalnie do liczby operacji dostępu realizowanych w systemie: w okresach niskiej aktywności skalowanie będzie wykonywane częściej niż w okresach wysokiej aktywności. Oczywiście, w dowolnym momencie można dokonać ponownego rozpoczęcia wyceny poprzez nadanie wartości początkowej (1) wszystkim licznikom częstotliwości użycia.

Wartość maksymalna $c_{\max}$ wyznacza górną granicę zakresu skali wartości, dolną granicą jest 1. Ustalając wartość maksymalną $c_{\max}$, określa się więc, ile razy maksymalnie obliczona wartość najbardziej wartościowego zbioru danych przechowywanego w systemie będzie wyższa od wartości najmniej wartościowego zbioru danych. Wartość maksymalna $c_{\max}$ nie powinna być zatem zbyt mała, w praktyce dowolna wartość wyższa lub równa 10 daje wyniki przydatne z punktu widzenia zarządza-

nia cyklem życia informacji (proponowana przez autora wartość zalecana wynosi $c_{\max} = 100$). Wartość maksymalna dla sumy liczników $c_{sum\,\max}$ nie może być mniejsza od dwukrotności liczby zbiorów danych przechowywanych w systemie (zalecana przez autora wartość minimalna $c_{sum\,\max}$ wynosi czterokrotność tej liczby).

Chwilowa wartość zbioru danych d jest zatem stanem licznika $c_t(d)$ w momencie czasu t. Można ją przybliżyć poniższą formułą:

$$V_t(d) = c_t(d) \approx \sum_{i=1,2,\ldots,j+1} \frac{c_{t(i)}(d) - c'_{t(i-1)}(d)}{2^{i-1}}, \tag{7}$$

gdzie: i – numer porządkowy kolejnej operacji skalowania (liczony od 1),
j – liczba przeprowadzonych do tej pory operacji skalowania,
$c_{t(i)}(d)$ – stan licznika zbioru danych d przed przeprowadzeniem i-tego skalowania,
$c'_{t(i)}(d)$ – stan licznika zbioru danych d po przeprowadzeniu i-tego skalowania,
$c'_{t(0)}(d)$ – stan licznika zbioru danych d sprzed rozpoczęcia wyceny ($\forall d : c'_{t(0)}(d) = 0$).

Faktycznie realizowaną inkrementację liczników można zaś ująć w następujący sposób:

$$c_{t+1}(d) = \begin{cases} c_t(d) + 1 \Leftrightarrow c_t(d) < c_{\max} \wedge \sum_{e \in D} c_t(e) < c_{sum\,\max} \\ \left\lceil \frac{c_t(d)+1}{2} \right\rceil \Leftrightarrow c_t(d) \geq c_{\max} \vee \sum_{e \in D} c_t(e) \geq c_{sum\,\max} \end{cases}, \tag{8}$$

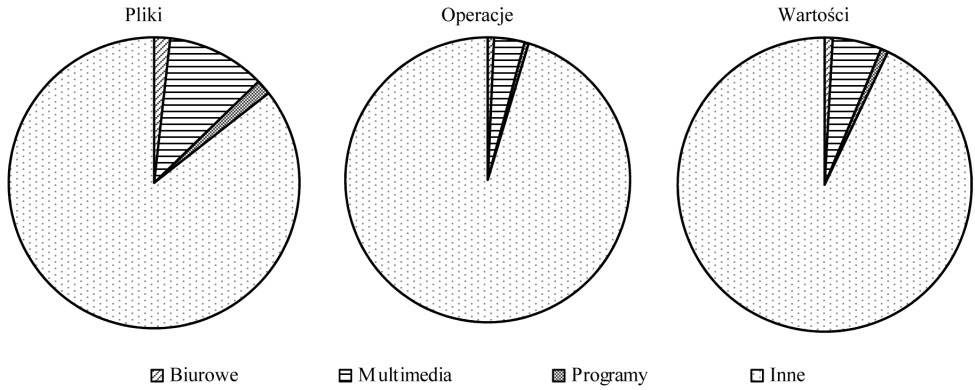
gdzie: D – kolekcja wszystkich zbiorów danych przechowywanych w systemie.

5. Przykład zastosowania nowej metody

Aby zilustrować metodę praktycznym przykładem użycia, dokonano próbnej wyceny wartości zbiorów danych zgromadzonych na pojedynczej stacji roboczej działającej w systemie informatycznym Uniwersytetu Szczecińskiego. Wyceny dokonano w oparciu o plik zdarzeń przechowujący operacje o dostępie do danych w okresie 12 miesięcy poprzedzających moment wyceny; wykorzystano implementujący proponowaną metodę skrypt napisany w języku Python.

Dla lepszej orientacji w rodzaju danych wykorzystanych w eksperymencie pogrupowano je w cztery klasy: dokumenty biurowe (m.in. dokumenty tekstowe i arkusze kalkulacyjne), multimedia (obrazy, dźwięki, filmy), pliki programów (w tym bibliotek) i inne (m.in. archiwa, relacyjne bazy danych, ale także pliki tymczasowe różnego typu). Rysunek 3 pokazuje udział plików poszczególnych klas odpowiednio

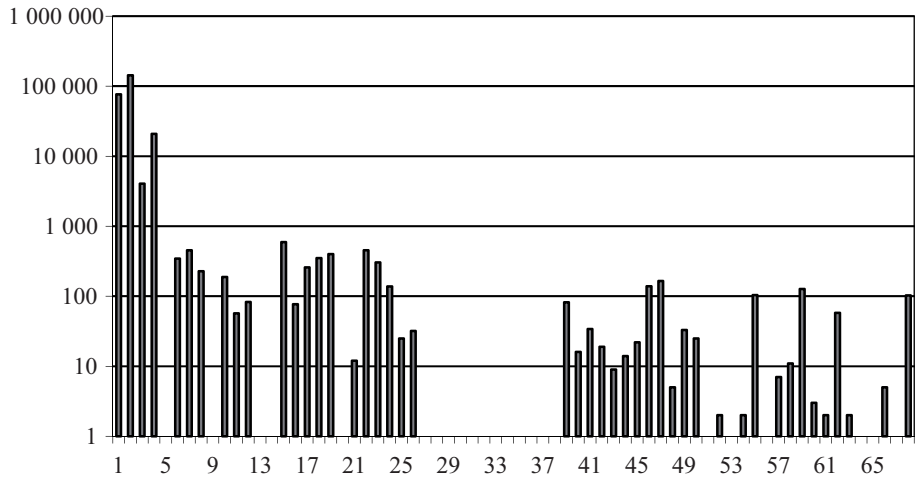
w: całkowitej liczbie zbiorów danych (łącznie ponad 250 tys. o sumarycznej objętości blisko 150 GB), całkowitej liczbie wykonanych operacji (niemal 4 mln) oraz wyliczonej wartości (z parametrami: $c_{\max} = 100$, $c_{sum\ \max} = 1\ 000\ 000$).



Rys. 3. Udziały zbiorów danych poszczególnych klas

Źródło: opracowanie własne.

Jak należało przypuszczać, biorąc pod uwagę specyfikę proponowanej metody, wykres dla wartości najbliższy jest wykresowi liczby wykonywanych operacji, choć dostrzegalny jest wyraźny wpływ procedury skalowania (np. na zbiory multimedialne przypada 3,5% wykonywanych operacji, a 5,4% wyliczonej wartości).



Rys. 4. Liczba zbiorów danych o określonej wartości

Źródło: opracowanie własne.

Histogram na rys. 4 ukazuje liczbę zbiorów danych przypadających na poszczególne poziomy wartości. Ze względu na bardzo duże różnice w liczebności zbiorów o różnych poziomach wartości na wykresie posłużono się skalą logarytmiczną.

Przedział wartości wynosi od 1 do 68, a maksymalna wartość jest konsekwencją odległości czasowej momentu zakończenia wyceny od momentu przeprowadzenia ostatniej procedury skalowania. Jak widać, wartość 98% przechowywanych zbiorów danych nie przekracza 4 (średnia wartość zbioru danych wynosi 2,28). Z perspektywy zarządzania cyklem życia informacji można ten wynik zinterpretować tak, że wystarczyłoby przeniesienie na szybsze pamięci masowe zaledwie 2% przechowywanych zbiorów danych, aby użytkownicy odczuli istotną poprawę wydajności.

6. Podsumowanie

Wyznaczenie wartości przechowywanych informacji stanowi istotny element zarządzania cyklem życia informacji, pozwalający na możliwie jak najbardziej pożądane rozlokowanie zbiorów danych pomiędzy podsystemy pamięciowe o różnych poziomach kosztów i wydajności. Ze względu na różnorodność i wielką liczbę zbiorów danych przechowanych we współczesnych systemach przechowywania danych zadanie to wymaga użycia prostych i dających się w pełni zautomatyzować metod wyceny.

Przedstawiona w niniejszym artykule metoda całkowicie spełnia te wymogi. Opiera się na modelu wartości uwzględniającym jedynie dwa łatwo mierzalne parametry: częstotliwość używania zbioru danych i czas, w którym użycie miało miejsce. W pracy przedstawiono technikę wzorowaną na rozwiązaniu stosowanym w predykcyjnych metodach bezstratnej kompresji danych, pozwalającą na wspólne ujęcie obu tych parametrów w postaci pojedynczego licznika przypisanego każdemu przechowywanemu zbiorowi danych.

W porównaniu z istniejącymi metodami tego typu proponowana metoda jest znacznie prostsza obliczeniowo i łatwiejsza w implementacji, nie nastręcza także trudności w określaniu parametrów stałych modelu.

Literatura

Chen Y., *Information valuation for Information Lifecycle Management*, [w:] *Second International Conference on Autonomic Computing*, IEEE Computer Society, Seattle 2005.

Cypryjański J., *Metodyczne podstawy ekonomicznej oceny inwestycji informatycznych przedsiębiorstw*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2007.

Gantz J.F., Chute C., Manfrediz A., Minton S., Reinsel D., Schlichting W., Toncheva A., *The Diverse and Exploding Digital Universe. An Updated Forecast of Worldwide Information Growth Through 2011*, IDC, Framingham, USA, March 2008, http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.

Howard P.G., *The Design and Analysis of Efficient Lossless Data Compression Systems*, Brown University, Providence, USA 1993.

Hubbard D.W., *How to Measure Anything. Finding the Value of Intangibles in Business*, John Wiley & Sons, Hoboken, USA 2007.

Jin H., Xiong M., Wu S., *Information value evaluation model for ILM*, [w:] *Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, IEEE Computer Society, Phuket, Tajland 2008.

Moody D., Walsh P., *Measuring the value of information: An asset valuation approach*, [w:] *Seventh European Conference on Information Systems*, Copenhagen Business School, Copenhagen 1999.

Oppenheim Ch., Stenson J., Wilson R., *The attributes of information as an asset, its measurement and role in enhancing organizational effectiveness*, "New Library World" 2001, no. 1170/1171.

Repo A.J., *The value of information: Approaches in economics, accounting, and management science*, "Journal of the American Society for Information Science" 1989, no. 40(2).

San Diego Supercomputer Center Completes Major Storage Upgrade, San Diego Supercomputer Center, San Diego, CA, USA 2009, http://www.sdsc.edu/News%20Items/PR022409_storage.html.

Storage Management Technical Specification. Version 1.3.0, Rev 5. Part 7: Information Lifecycle Management, Storage Networking Industry Association, 2008, http://www.snia.org/tech_activities/standards/curr_standards/smi/SMI-S_Technical_Position_v1.3.0r5.zip.

Song Y., Zhu D. (ed.), *High Density Data Storage: Principle, Technology, and Materials*, World Scientific Publishing, Singapur 2009.

Swacha J., *Zarządzanie przechowywaniem danych – Metodyka oceny efektywności*, Placet, Warszawa 2009.

Turczyk L., Gröpl M., Liebau N., Steinmetz R., *A method for file valuation in information lifecycle management*, [w:] *Americas Conference on Information Systems*, Association for Information Systems, Keystone, USA 2007.

Turczyk L., Heckmann O., Berbner R., Steinmetz R., *A formal approach to information lifecycle management*, [w:] *Emerging Trends and Challenges in Information Technology Management*, Information Resources Management Association, Washington, USA 2006.

Volonino L., Sipior J.C., Ward B.T., *Managing the lifecycle of electronically stored information*, "Information Systems Management" 2007, no. 24(3).

METHOD OF DATA SET VALUATION FOR INFORMATION LIFECYCLE MANAGEMENT

Summary: Estimating the value of stored information is an important element of information lifecycle management, allowing to distribute stored data sets to the most appropriate storage subsystems, characterized by different levels of costs and performance. Because of the huge amounts of data kept in contemporary storage systems, this task has to be accomplished using methods which are simple and capable of being automated. The method proposed in this paper meets these requirements. It uses a value model based on only two easily measurable parameters: data set usage frequency, and the moment of time the usage took place. Compared to the existing methods of similar type, the proposed method is much simpler, both in preparation and use.

Key words: information value, data set valuation, information lifecycle management, data storing management.