Kompresja danych i formaty plików graficznych

Tomasz Lewicki

WWSIS, Wrocław

maj 2007

Kompresja i dekompresja

W znaczeniu informatycznym **kompresja** to zmniejszenie objętości danych przy zachowaniu „ładunku informacyjnego", czyli sensu tych danych. Celem kompresji jest zatem możliwie dokładna reprezentacja informacji przy użyciu możliwie małej ilości bitów. Kompresja ma zazwyczaj na celu oszczędność nośnika i/lub łącza sieciowego, którym są przesyłane dane, czyli redukcję kosztów.

Proces odwrotny do kompresji nazywamy **dekompresją**.

Dlaczego kompresja jest możliwa?

- duża część danych cechuje się znaczną redundancją (nadmiarowością), tzn. pewne informacje powtarzają się z różną częstością
- dane są prezentowane w rozmaity sposób (np. grafika może być rastrowa lub wektorowa)
- człowiek ma ograniczone zdolności percepcyjne — mózg można do pewnego stopnia „oszukać”

Dane poddawane kompresji

Kompresji można poddawać dane (zbiory informacji) różnego rodzaju, m.in.:

- tekst
- dźwięk (mowa, muzyka)
- obraz ruchomy i nieruchomy
- pliki wykonywalne

Zalety i wady kompresji

+ przesyłanie większej ilości danej w jednostce czasu
+ przesyłanie tej samej ilości danych w krótszym czasie
+ zmniejszenie rozmiarów danych przechowywanych na nośnikach
– przed użyciem danych należy je rozpakować
– w pewnych sytuacjach dekompresja w czasie rzeczywistym lub *quasi*-rzeczywistym może pochłaniać sporo zasobów systemu komputerowego

Algorytmy kompresji

Algorytmem kompresji nazywamy schemat postępowania przy zmniejszaniu objętości pliku; najczęściej zależy on od charakteru kompresowanych danych.

Algorytmy kompresji można podzielić ze względu na różne kryteria, np. stopień zmiany „ładunku informacyjnego”:

- kompresja bezstratna
- kompresja stratna

Kompresja bezstratna

Algorytmy kompresji bezstratnej umożliwiają takie przechowanie danych, by w procesie dekompresji uzyskać dane w postaci identycznej z tą, jaką miały przed poddaniem ich kompresji. Dzieje się to jednak zazwyczaj kosztem gorszego współczynnika kompresji. Nadają się do danych charakteryzujących się dużą redundancją (nadmiarowością) informacji.

Niektóre obszary zastosowań: tekst, bazy danych, pewne rodzaje obrazów statycznych (np. do zastosowań medycznych).

Przykłady algorytmów bezstratnych: Deflate, Huffman, LZW, RLE, BZIP2.

Kompresja stratna

Przy użyciu algorytmu kompresji stratnej część oryginalnych danych zostaje utracona, chociaż „ładunek informacyjny” zazwyczaj jest zachowany. Algorytmy kompresji stratnej oferują lepsze współczynniki kompresji niż algorytmy kompresji bezstratnej.

Niektóre obszary zastosowań: obraz ruchomy i nieruchomy, muzyka, mowa.

Przykłady algorytmów stratnych: DCT, metoda falkowa, JPEG, MPEG, Vorbis, MP3.

Współczynnik i stopień kompresji

Współczynnik kompresji (ang. *compression ratio*) jest definiowany jako stosunek objętości danych skompresowanych (wyjściowych) do objętości danych oryginalnych (wejściowych), czyli jego wartość zawiera się w przedziale od 0 do 1.

Stopień kompresji (ang. *compression factor*) jest odwrotnością współczynnika kompresji i przyjmuje wartości większe od 1.

Algorytmy kompresji w zastosowaniach

Algorytmy można podzielić również w inny sposób, np. ze względu na obszar zastosowań. Pewne algorytmy są **uniwersalne** i mogą zostać użyte do kompresji danych o różnej zawartości, inne z kolei są **specjalizowane** do konkretnych, czasami bardzo wąskich obszarów.

Przypadek szczególny: grafika

Coraz większa ilość informacji jest przekazywana w postaci obrazów (statycznych i ruchomych). Według magazynu *NetWorld* w 2006 r. na świecie wygenerowano 161 miliardów gigabajtów danych (czyli około 150 eksabajtów; $\approx 1,5 \cdot 10^{20}$ B). Zasadnicza część pochodzi od plików graficznych produkowanych przez rozmaite urządzenia do cyfrowej rejestracji obrazu (cyfrowe aparaty fotograficzne, cyfrowe kamery filmowe, skanery, zdjęcia satelitarne, systemy obrazowania medycznego i inne). Szacuje się, że w 2010 r. zostanie wygenerowanych około 1000 miliardów gigabajtów danych (czyli $\approx 10^{21}$ B; liczbę tę w układzie SI określa się przedrostkiem „zetta").

Taka ilość danych wymaga projektowania i użycia efektywnych metod gromadzenia, indeksowania, przeglądania i wymiany informacji w postaci graficznej.

Rodzaje plików graficznych

- rastrowe (bitmapy) — obraz zapisany jako siatka punktów (pikseli) opisanych przez ich położenie na płaszczyznie oraz przez bity koloru (np. JPEG, GIF, PNG, TIFF)
- wektorowe (obiektowe) — obraz jest tworzony przez obiekty matematyczne: punkty i krzywe (np. SVG, Flash)
- metapliki — swego rodzaju „kontenery”; pliki zawierające w sobie inne pliki i/lub informacje opisujące te pliki (np. WMF, EMF)
- pliki opisu strony — obraz zapisany jest w specyficznym języku programowania (np. PCL, PostScript)

Wybrane formaty graficzne

JPEG — Joint Photographic Experts Group

+ szeroko rozpowszechniony
+ szeroka paleta barw: do $(2^8)^3 = 2^{24} \approx 16,8$ mln (tzw. *True Color*)
+ dobrze sprawdza się w przypadku obrazów o łagodnych przejściach tonalnych, zarówno wielobarwnych, jak i w odcieniach szarości (np. fotografie)
+ obsuguje tryb progresywny
+ istnieją rozszerzenia standardu i modyfikacje algorytmu kompresującego (JPEG-LS, JPEG 2000)
+ istnieją wolne (niekomercyjne i nieopatentowane) implementacje
– kompresja stratna
– w procesie kompresji z obrazu usuwane są drobne szczegóły
– istnieje wiele formatów plików, w których wykorzystuje się algorytm JPEG, który sam w sobie nie został dokładnie sprecyzowany — niektóre są niezgodne ze sobą lub używane przez nieliczne programy

Wybrane formaty graficzne

GIF — Graphics Interchange Format

+ szeroko rozpowszechniony
+ kompresja bezstratna (algorytm LZW)
+ małe rozmiary plików
+ dobrze sprawdza się w przypadku obrazów składających się z dużych obszarów o jednolitej barwie (np. wykresy „tortowe") i/lub szczegółowych oraz z ostrymi krawędziami (np. wykresy, rysunki, siatki, szkice)
+ można zapisać informację o przezroczystości wybranego koloru
+ istnieją rozszerzenia standardu pozwalające na zapis animacji
– niewielka paleta barw: maksymalnie $2^8 = 256$ (tzw. tryb indeksowany)
– obrazy o dużej rozpiętości tonalnej są „redukowane" do 256 kolorów przed wykonaniem kompresji, przez co tracą na jakości
– format był do niedawna objęty patentami

Wybrane formaty graficzne

PNG — Portable Network Graphics

- \+ kompresja bezstratna
- \+ łączy zalety JPEG i GIF: dobrze sprawdza się zarówno w przypadku obrazów o płynnych przejściach tonalnych wielobarwnych i w odcieniach szarości (paleta barw do do $\approx 16,8$ mln kolorów), jak i GIF (pełna 8-bitową przezroczystość [tzw. kanał *alfa*] oraz paleta barw od 2 do 256 kolorów)
- \+ obsługuje korekcję gamma, tryb progresywny i kontrolę poprawności pliku
- \+ obsługuje różne głębokości bitowe (do 48 bitów na piksel)
- \+ algorytm i format są wolne od patentów
- \+ zalecany przez W3C (*World Wide Web Consortium*) jako do reprezentacji grafiki rastrowej w sieci Web
- – nie obsługuje animacji (istnieje osobny format do tego celu [MNG] oparty na algorytmie PNG)
- – nieprawidłowo obsługiwany przez przeglądarkę Internet Explorer < 7.0
- – niektóre programy nie obsługują wszystkich właściwości PNG
- – występują problemy z obsługą korekcji gamma w przeglądarkach

Wybrane formaty graficzne

TIFF — Tag(ged) Image File Format

+ szeroko rozpowszechniony
+ kompresja bezstratna (brak kompresji, algorytm LZW, algorytm CCITT Group 4) i stratna (algorytm JPEG)
+ szeroka paleta barw: od 2 do $\approx 16,8$ mln
+ obsługuje różne głębokości bitowe (do 48 bitów na piksel) oraz przestrzenie barw
+ obsługuje tryb wielostronnicowy (*multipage*) — zostało to wykorzystane np. w faksach
+ obsługuje przezroczystość i profile barw
+ specyfikacja formatu umożliwia jego rozszerzanie o dodatkowe znaczniki (za zgodą właściciela praw autorskich do formatu TIFF)
– mogą występować problemy z odczytem plików TIFF w niektórych programach
– format jest objęty patentami

Wybrane formaty graficzne

SVG — Scalable Vector Graphics

+ brak kompresji ze względu na sposób zapisu obiektów w postaci wektorów
+ oparty na języku XML — grafikę można tworzyć nawet w edytorze tekstowym
+ skalowanie bez utraty jakości (właściwość grafiki wektorowej)
+ obsługuje przezroczystość, gradienty i filtry
+ możliwość zagnieżdżania obiektów (np. odnośników, grafiki rastrowej) w pliku oraz umożliwienie wyszukiwarkom indeksowania tekstu zawartego w grafice (na razie za pomocą wtyczek do przeglądarek)
+ stanowi doskonałe uzupełnienie nowoczesnych stron WWW zgodnych ze standardami i zaleceniami W3C
+ wolny od patentów
+ zalecany przez W3C (*World Wide Web Consortium*) do reprezentacji grafiki wektorowej w sieci Web
– nie wspiera treści multimedialnych (dźwięku i obrazu ruchomego) — potrafi to konkurencyjny format Flash
– niektóre przeglądarki (m.in. Internet Explorer) wymagają wtyczek do obsługi formatu SVG

Wybrane formaty graficzne

SWF — Shockwave Flash (obecnie Adobe Flash)

+ szeroko rozpowszechniony
+ zapis obiektów w postaci wektorów
+ skalowanie obiektów bez utraty jakości
+ możliwość włączania do pliku obiektów rastrowych
+ umożliwia stworzenie atrakcyjnie wyglądających prezentacji przy niewielkiej objętości pliku
+ obsługuje transmisję strumieniową
+ nieformalny standard prezentacji multimedialnych w sieci Web
– przeglądarki nie obsługują natywnie formatu Flash, wymagana jest instalacja wtyczki
– format jest objęty patentami

Wybrane formaty graficzne

DjVu

- \+ zaprojektowany specjalnie do kompresji skanowanych dokumentów
- \+ w założeniu ma służyć do tworzenia bibliotek cyfrowych, tzn. przechowywania zeskanowanych książek itp. materiałów na nośnikach elektronicznych
- \+ pliki wynikowe mają niewielką objętość przy zachowaniu jakości wiernej oryginałowi
- \+ pliki wynikowe są znacznie mniejsze niż w przypadku grafik w formacie JPEG czy TIFF przy tej samej jakości
- \+ obsługuje tryb progresywny
- \+ obraz jest zapisywany w warstwach, np. tło, kolor tła i tekst poddany OCR
- \+ oddzielenie warstwy tekstowej od graficznej umożliwia indeksowanie i przeszukiwanie tekstu
- – przeglądarki nie obsługują natywnie formatu DjVu, wymagana jest instalacja wtyczki
- – format jest objęty patentami

Inne ciekawe formaty

PS — Post Script

+ nie jest *stricte* formatem graficznym, ale językiem programowania i językiem opisu strony
+ czcionki opisane są krzywymi Béziera (krzywe trzeciego stopnia), czyli są obiektami wektorowymi (skalowalnymi)
+ język opisu strony jest niezależny od urządzenia drukującego (np. drukarki, plotera, naświetlarki) — urządzenie otrzymuje program w języku Post Script, który jest wykonywany przez interpreter języka wbudowany w urządzenie
– jest objęty patentami
– wbudowany interpreter języka Post Script zawierają jedynie urządzenia „z wyższej półki”

Inne ciekawe formaty

PDF — Portable Document Format

+ podobnie jak Post Script nie jest *stricte* formatem graficznym
+ jest oparty na okrojonej wersji języka Post Script i uzupełniony o elementy hipertekstowe
+ dokument zapisany w formacie PDF będzie wyglądał identycznie na każdym komputerze
+ istnieją bezpłatne czytniki plików PDF; wiele programów ma możliwość eksportu plików do formatu PDF
+ format obsługuje szyfrowanie i podpisywanie plików oraz umożliwia ograniczanie dostępu do pewnych właściwości dokumentu
+ specyfikacja formatu jest otwarta
+ format aspiruje do miana oficjalnego standardu
– można napotkać problemy przy obsłudze pliku zapisanego w formacie PDF, np. kopiowanie tekstu do innego dokumentu
– niektóre starsze narzędzia mają kłopoty z obsługą zaawansowanych właściwości plików PDF i wyszukiwaniem słów