Streszczenie

Entropia jest jednym z głównych paradygmatów kompresji danych, a kompresory entropijne są jedynymi z najczęściej używanych. Wykorzystują one obserwację, że większość danych wejściowych ma niejednostajny rozkład prawdopodobieństwa wystąpienia poszczególnych symboli, np. w języku polskim litera „a" występuję znacznie częściej niż „ź". Innym podejściem do kompresji jest kompresja słownikowa, która opiera się na spostrzeżeniu, że pewne bloki danych często się powtarzają.

Powyższe podejścia wydają się na pierwszy rzut oka niezwiązane ze sobą, jednak wiadomo, że można je połączyć: dla niektórych kompresorów słownikowych (np. LZ78, LZ77) pokazano ograniczenia względem entropii [Manzini 2001; Kosaraju, Manzini 1999; Szpankowski 1993; Plotnik et al. 1992]. Zatem rozmiar wyjścia tych metod kompresji słownikowej może być ograniczony zarówno za pomocą entropii jak i za pomocą miar powtarzalności danych. Konsekwencją tego jest ich uniwersalność — działają dobrze dla różnych typów danych.

Kompresory gramatykowe dla wejściowego ciągu S konstruują gramatykę bezkontekstową, która generuje tylko jedno słowo: S. Są szczególnym przypadkiem kompresji słownikowej, eksperymenty pokazują, że w praktyce osiągają dobre współczynniki kompresji, szczególnie dla danych z dużą liczbą powtarzających się bloków.

Pierwsza część prezentowanej pracy skupia się na ograniczeniach entropijnych dla kompresorów gramatykowych. Pokazujemy, że kompresory te osiągają $\alpha|S|H_k(S)$ bitów, plus asymptotycznie małe składniki, dla pewnej stałej α zależnej od konkretnego kompresora, przez $H_k(S)$ oznaczamy entropię k-tego rzędu ciągu wejściowego S. Nasze wyniki pokazują, że kompresory gramatykowe, podobnie jak niektóre inne metody słownikowe, także mają szeroki zastosowanie. Porównując z poprzednimi wynikami dla kompresorów gramatykowych, nasze wyniki pokazują ograniczenia dla znacznie obszerniejszej grupy kompresorów, ponadto dotyczą praktycznie używanych wariantów.

W drugiej części pokazujemy, że otrzymane ograniczenia są optymalne. Ten wynik dotyczy nie tylko kompresji gramatykowej, ale także innych metod kompresji słownikowej, takich jak LZ78 czy LZ77.

Trzecia część pokazuje jak narzędzia stworzone w celu poprzedniej analizy mogą być wykorzystane do zbudowania praktycznej i małej reprezentacji danych, która obsługuje szybki dostęp do dowolnego elementu. Wierzymy, że przedstawione wyniki mają zastosowanie praktyczne.

W ostatniej części rozszerzamy prezentowane metody tak, by zastosować je dla etykietowanych drzew. W efekcie uzyskujemy bardzo zwięzłą i relatywnie prostą strukturę danych dla etykietowanych drzew, wierzymy, że także ona może być zaadaptowana do użycia w praktyce.