Kierownik projektu: **dr hab. Jakub Radoszewski**
Tytuł projektu: „Poszukiwanie optymalności w wariantach problemu indeksowania tekstu i zagadnieniach pokrewnych”

Popularnonaukowe streszczenie projektu

Projekt badawczy jest poświęcony poszukiwaniu efektywnych rozwiązań problemów indeksowania tekstów. Podstawowy scenariusz w indeksowaniu tekstu jest następujący. Dysponujemy dużych rozmiarów kolekcją danych przechowywanych cyfrowo w formacie tekstowym. Tekst traktujemy tu raczej jako sekwencję znaków lub liczb niż jako tekst w języku naturalnym, co pozwala stosować nasze metody do dowolnych danych występujących w formacie tekstowym, jak choćby do ciągów danych dostarczonych przez określone urządzenie pomiarowe czy sekwencji DNA. Naszym celem jest stworzenie indeksu, czyli struktury danych pozwalającej – w najprostszym przypadku – odpowiadać na pytania o to, czy dany wzorzec ma wystąpienie będące fragmentem jakiegoś napisu z tej kolekcji. Kluczowym dla nas jest, aby indeks zużywał możliwie mało pamięci ponad to, co jest potrzebne do przechowywania kolekcji danych, aby można go było zbudować w rozsądnym czasie oraz aby czas odpowiadania na pytania o wystąpienia wzorca był minimalny.

Powyższy klasyczny problem indeksowania doczekał się wielu efektywnych rozwiązań. W szczególności dostępne są gotowe biblioteki programistyczne implementujące bardzo szybkie i zarazem małych rozmiarów indeksy, a podstawowe struktury danych do indeksowania tekstu zazwyczaj są zawarte w materiale studiów z informatyki. W praktyce jednak problem indeksowania w wersji klasycznej jest często niewystarczający, gdyż dane mogą zawierać błędy czy przekłamania; czasem poszukiwane są tylko niektóre wystąpienia, spełniające pewne dodatkowe warunki; może wreszcie być tak, że poszukujemy niekoniecznie fragmentów dokładnie pasujących do określonego wzorca, lecz raczej fragmentów, które możliwie go przypominają. Istniejące indeksy potrafią sobie radzić w takich sytuacjach lepiej lub gorzej, jednak w większości przypadków indeksy te nie są optymalne lub też do tej pory nie wiadomo, czy da się je ulepszyć. Celem tego projektu jest udoskonalenie tych indeksów. Będziemy również rozważali problemy tekstowe powiązane z problemem indeksowania, które pozwalają spojrzeć na ten problem w szerszej perspektywie.

Przykładowy problem, jaki będziemy rozważać, to konstrukcja tzw. przybliżonego indeksu, który pozwala sprawdzać w sposób efektywny, czy kolekcja tekstów zawiera fragment, który różni się od wzorca na co najwyżej k pozycjach, dla ustalonego parametru k. Przykładowo, wzorzec „efektowny” występuje w poprzednim zdaniu z $k = 1$ niedopasowaniem (jako słowo „efektywny”). Rozmiar i czas działania istniejących przybliżonych indeksów rośnie wykładniczo względem parametru k. W 2019 roku udowodniono, że dla dostatecznie dużego parametru k, przy pewnych dodatkowych założeniach takiej zależności nie da się uniknąć. Jednak dla małych k problem ten jest na tyle istotny z praktycznego punktu widzenia, że taka odpowiedź nie jest wystarczająca. W szczególności istnieją efektywniejsze przybliżone indeksy w prostym przypadku, gdy $k = 1$. W 2021 roku zespół, którego członkiem był kierownik proponowanego projektu, rozpatrywał problem wyznaczania najdłuższego wspólnego podsłowa dwóch słów z co najwyżej k błędami, dla małej wartości parametru k. W wariancie klasycznym ($k = 0$) problem ten ma to samo rozwiązanie co problem budowy indeksu. Z kolei w przypadku przybliżonym udało nam się uzyskać rozwiązanie problemu najdłuższego wspólnego podsłowa, które działa (trochę) szybciej niż wynikałoby to z zastosowania przybliżonego indeksu. Ten wynik nasuwa pytania o to, czy dla małego k można skonstruować bardziej efektywny przybliżony indeks niż dotychczas znane oraz jak duże usprawnienie w obu problemach można uzyskać. Jednym z celów projektu jest próba uzyskania odpowiedzi na to pytanie.

Wynikiem prac nad projektem będą publikacje, w których opisane zostaną najlepsze z zaproponowanych przez nas indeksów, a także być może uda się w nich udowodnić, że lepsze indeksy (najprawdopodobniej) nie mogą zostać skonstruowane. W ocenie indeksów bazujemy na tzw. przypadku pesymistycznym, czyli chcemy, aby indeks miał określone parametry niezależnie od tego, dla jakiego tekstu będzie miał zostać skonstruowany. W niektórych przypadkach mogą powstać gotowe programy implementujące zaproponowane indeksy, jednak dla nas ważniejsze będzie określenie, czy daną teoretyczną barierę da się w ogóle przełamać.