Nr rejestracyjny: 2023/51/B/ST6/01505; Kierownik projektu: dr Paweł Maciej Gawrychowski

Przetwarzanie olbrzymich danych tekstowych i ich uogólnień: algorytmy i warunkowe ograniczenia dolne

Paweł Gawrychowski

Proponowany projekt dotyczy algorytmów tekstowych, gdzie tekst to dowolny ciąg znaków z pewnego ustalonego alfabetu. Podstawowe zagadnienia rozważane w tej dziedzinie to wyszukiwanie wzorca, indeksowanie i porównywanie tekstów. Wyszukiwanie wzorca to wykrywanie wystąpień jednego tekstu (wzorca) w innym. Indeksowanie oznacza zbudowanie dla zadanego tekstu struktury danych, która następnie umożliwia szybkie znajdowanie w tym tekście wystąpień dowolnego wzorca. Natomiast w problemie porównywania tekstów celem jest sprawdzenie, jak podobne do siebie są dwa zadane ciągi. Wszystkie te zagadnienia można rozważać w wielu wersjach, na przykład zakładając, że dane wejściowe potencjalnie zawierają błędy, a także wybierając odpowiednie definicje 'występowania' i 'podobieństwa'.

Praktyczną motywację dla tego typu zagadnień zapewnia między innymi bioinformatyka, w której tekst to sekwencja nukleotydów składająca się ze znaków A, C, G, T lub chromosom składający się z genów oznaczanych liczbami (potencjalnie mających znak oznaczający kierunek). Zastosowania algorytmów tekstowych nie ograniczają się do bioinformatyki, jest to dojrzała dziedzina z tematycznymi konferencjami naukowymi CPM (Combinatorial Pattern Matching) oraz SPIRE (String Processing and Information Retrieval), które odbywają się corocznie od wczesnych lat 90. Każdego roku publikacje poświęcone tematyce tekstowej pojawiają się także na wiodących konferencjach algorytmicznych, takich jak SODA (Symposium on Discrete Algorithms), ICALP (International Colloquium on Automata, Languages and Programming), ESA (European Symposium on Algorithms), a także na najważniejszych konferencjach z zakresu informatyki teoretycznej: STOC (Symposium on Theory of Computing) oraz FOCS (Foundations of Computer Science).

Ten projekt dotyczy teoretycznych aspektów algorytmów tekstowych i ma na celu pogłębienie wiedzy na temat kilku głównych problemów otwartych, motywowanych między innymi potencjalnymi zastosowaniami w bioinformatyce. Zamierzam uzyskać nowe wyniki, stosując narzędzia związane z kombinatoryką na słowach, algorytmami i strukturami danych w modelu Word RAM, a także złożonością drobnoziarnistą. Model Word RAM zapewnia elegancki formalizm dla projektowania algorytmów i (w niektórych przypadkach) granic dolnych, jednocześnie dobrze uwzględniając możliwości współczesnych procesorów. Z kolei stosunkowo młoda dziedzina złożoności drobnoziarnistej zajmuje się klasyfikacją problemów rozwiązywalnych w czasie wielomianowym według ich złożoności, znajdując redukcje pomiędzy poszczególnymi problemami i wyjaśniając, dlaczego w niektórych przypadkach znalezienie efektywnego algorytmu może być niezwykle trudne.

Chciałbym skoncentrować się na pięciu kierunkach, z których pierwsze trzy są motywowane przez ważne nierozstrzygnięte dotąd problemy, a pozostałe dwa kierunki mają charakter bardziej otwarty. Pierwszym tematem jest wyszukiwanie wzorca w skompresowanych tekstach, którym zajmowałem się już w swojej rozprawie doktorskiej. Drugim kierunkiem są zwięzłe i tak zwane kodujące struktury danych, stosowane do indeksowania tekstów, z którymi jestem dość dobrze zaznajomiony ze względu na kilka moich wcześniejszych publikacji dotyczących tej tematyki. Trzecim kierunkiem jest złożoność wyszukiwania wzorca w danych strumieniowych, na temat której opublikowałem kilka prac, a także pracowałem nad najbardziej fundamentalnym pytaniem w tej dziedzinie. Czwartym tematem są różne pojęcia podobieństwa ciągów. Kilka lat temu opublikowałem pracę na temat odległości edycyjnej z przesunięciami, a niedawno pracowałem nad sortowaniem skierowanych permutacji za pomocą odwróceń. Ostatnim kierunkiem jest wyszukiwanie wzorca w uogólnieniach tekstów. Jestem współautorem kilku prac na temat problemów równoważnych zliczaniu cykli w grafach rzadkich, a także pracy na temat wyszukiwania wzorca w pewnej formie uogólnionych tekstów.

W przypadku wszystkich wymienionych obszarów badawczych, moim celem jest zaprojektowanie nowych efektywnych algorytmów, a także (zwłaszcza w przypadku ostatniego kierunku) znalezienie wyjaśnienia, dlaczego takie algorytmy są mało prawdopodobne dla niektórych problemów. Otrzymane wyniki mogą być publikowane na wiodących konferencjach w dziedzinie algorytmiki.