

Algorytmiczne aspekty przetwarzania i analizy dużych zbiorów danych w systemach rozproszonych

Wielkoskalowe rozproszone systemy informacyjne odgrywają obecnie bardzo istotną rolę, znajdując coraz liczniejsze zastosowania w wielu dziedzinach, w których zachodzi potrzeba szybkiego przetwarzania ogromnych ilości informacji. Źródła tych danych stanowią zarówno złożone, zdecentralizowane systemy pomiarowe, gromadzące i przetwarzające wyniki stale przeprowadzanych obserwacji wartości interesujących nas parametrów monitorowanego środowiska, jak i dynamiczne sieci składające się z komunikujących się ze sobą urządzeń mobilnych takich jak smartfony, elektroniczne zegarki czy opaski śledzące aktywność fizyczną. Złożone systemy rozproszone, obejmujące m.in. sieci bezprzewodowe oraz sieci sensorowe składające się z niewielkich urządzeń o ograniczonych zasobach, znajdują także zastosowania w zakresie wykrywania różnego rodzaju zagrożeń, detekcji ruchu, monitorowania infrastruktury przemysłowej czy w systemach implementujących ideę internetu rzeczy, takich jak inteligentne systemy transportowe, inteligentne domy („smart home”) czy systemy zdalnego monitorowania stanu zdrowia.

U podstaw każdego z wymienionych zastosowań leży gromadzenie, przesyłanie między urządzeniami oraz przetwarzanie ogromnych ilości danych, obejmujące wykonywanie różnorodnych obliczeń o zróżnicowanym stopniu złożoności. Wobec coraz wyższych oczekiwań względem prędkości działania, skuteczności i niezawodności tego typu systemów, konieczne jest projektowanie efektywnych metod umożliwiających dostatecznie szybkie i precyzyjne wykonywanie pewnych elementarnych obliczeń (stanowiących zwykle kluczowy element bardziej złożonych procedur) na wielkich ilościach rozproszonych danych, w tym danych strumieniowych, generowanych dynamicznie z dużą intensywnością.

W projekcie badawczym planujemy rozważanie tego typu zagadnień, związanych z wybranymi problemami zdecentralizowanego przetwarzaniem dużych ilości danych, obejmujących zagadnienia agregacji, estymacji ich rozkładu, detekcji anomalii oraz ogłaszania alarmu w sieciach bezprzewodowych. Rozpatrywane problemy stanowią jedne z najistotniejszych w dziedzinie teorii obliczeń rozproszonych. W przypadku ogromnych ilości rozproszonych danych, niewielkie urządzenia tworzące sieć nie są w stanie zgromadzić ich w ograniczonej pamięci. Wobec powyższego konieczna staje się ich agregacja polegająca na wyznaczeniu pewnych charakterystyk istotnych z perspektywy danych zastosowań, takich jak wartość średnia, mediana czy liczba wartości spełniających określone warunki. Obliczanie dokładnych rezultatów w zdecentralizowanych systemach rozproszonych zwykle stanowi bardzo złożony problem. Na szczęście w wielu przypadkach wystarczająca jest znajomość pewnych przybliżonych wartości (dla przykładu, korzystając z systemu gromadzącego i przetwarzającego dane o natężeniu ruchu, interesuje nas jedynie pewna ogólna informacja, a nie dokładna wartość liczbowa), niemniej precyzja uzyskiwanych oszacowań jest istotną kwestią. Przykładowo, system monitorujący temperaturę czy stężenie niebezpiecznych substancji (jak np. czujnik czadu), po wykryciu już niewielkiego przekroczenia dopuszczalnych wartości w jakimkolwiek punkcie obserwowanego otoczenia powinien natychmiast wszczać alarm, w jak najkrótszym czasie informując inne urządzenia o wykrytej anomalii (jednocześnie zależy nam na tym, aby zminimalizować prawdopodobieństwo wystąpienia fałszywego alarmu).

W planowanym projekcie zamierzamy skupić się na algorytmicznych aspektach omawianych problemów, poszukując dla nich nowych, szybkich i energetycznie efektywnych procedur, dostosowanych do wciąż rosnącej ilości przetwarzanych danych, w tym danych strumieniowych. Skupimy się również na aspektach odporności projektowanych algorytmów na różnego rodzaju zakłócenia, będące nieodłącznym elementem komunikacji bezprzewodowej. Zaproponowane metody poddane zostaną szczegółowym badaniom eksperymentalnym obejmującym przeprowadzenie szeregu symulacji komputerowych. Celem pracy badawczej będzie także formalne udowodnienie poprawności analizowanych algorytmów oraz określenie precyzji zwracanych przez nie rezultatów. W ten sposób będziemy mogli dostarczyć przekonujących argumentów odnośnie ich efektywności i niezawodności. W badaniach wykorzystywać będziemy różne zaawansowane techniki algorytmiczne i matematyczne, obejmujące metody z zakresu teorii prawdopodobieństwa, teorii grafów, analizy kombinatorycznej i statystyki.