Rozpoznawanie języka XPath w czasie liniowym

Paweł Parys (parys@mimuw.edu.pl)

20 marca 2008

XPath jest notacją pozwalającą opisywać własności wierzchołków w dokumentach XML. W XPath można odnosić się zarówno do etykiet wierzchołków (które pochodzą ze skończonego zbioru), jak i do danych (które mogą być dowolne). Z jednej strony w XPath nie wyrażają się nawet pewne języki regularne. Z drugiej strony możemy porównywać dane zawarte w różnych miejscach drzewa dokumentu XML, czego żaden automat skończony nie potrafi.

Przykładowo, rozważmy dokument, którego jedna część zawiera dane pracowników firmy, a druga część listę płac mówiącą: taki pracownik w takim miesiącu zarobił tyle. Wówczas w XPath możemy opisać własność mówiącą: każdy pracownik pojawiający się w drugiej części, znajduje się także w pierwszej części.

Praktycznym problemem jest sprawdzenie, czy dokument XML spełnia własność opisaną daną formułą XPath (lub też stwierdzenie, które jego wierzchołki spełniają taką własność). Obecnie w tym zastosowaniu stosuje się algorytmy kwadratowe ze względu na rozmiar dokumentu i potencjalnie wykładnicze względem rozmiaru formuły. Istnieją także algorytmy wielomianowe względem obu tych czynników. Ja przedstawię algorytm rozwiązujący to zagadnienie działający w czasie liniowym względem rozmiaru dokumentu (również potencjalnie wykładniczy względem rozmiaru formuły).