Wyszukiwanie z wildcardami w tekście i wzorcu

Kamil Rajtar

Czerwiec 2020

Opis problemu

Na wejściu do algorytmu dostajemy tekst i wzorzec w których oprócz liter moga wystepować wildcardy (?) pasujace do dowolnego znaku. Na wyjściu ma sie znaleźć lista pozycji w których wzorzec pasuje do tekstu.

Wersja bez wildcardów

W rozwiazaniu problemu użyty jest splot obliczany za pomoca szybkiej transformaty Furiera (FFT) zdefiniowany nastepujaco:

$$p \otimes t \stackrel{\text{def}}{=} (\sum_{j=0}^{m-1} p_j t_{i+j}, 0 \leq i \leq n-m)$$

Normalne (bez wildcardów) dopasowanie za pomoca splotu tekstu (t) i odwróconego wzorca (p) możemy obliczyć za pomoca wzoru:

$$\sum_{j-0}^{m-1} (p_j - t_{i+j})^2 = \sum_{j-0}^{m-1} (p_j^2 - 2p_j t_{i+j} + t_{i+j}^2)$$

Zobaczmy że lewa cześć równania równa 0 to znaleźliśmy dopasowanie. (Tekst nie różni sie od wzorca na żadnej z m kolejnych pozycji). Prawa strone potrafimy szybko obliczyć ponieważ podniesienie każdej pozycji do kwadratu wykonywane jest w czasie stałym a środkowy składnik liczmy za pomoca FFT w O(n log m).

Wersja z wildcardami

Zdefiniujmy ciagi:

$$p'_j = \begin{cases} 0 & \text{p}_j = '?' \\ 1 & \text{wpp} \end{cases}$$

$$t'_j = \begin{cases} 0 & \text{t}_j = '?' \\ 1 & \text{wpp} \end{cases}$$

Wtedy łatwo jest widać że nastepujace równanie jest naturalnym rozszerzeniem rozwiazania poprzedniego problemu.

$$\sum_{j-0}^{m-1} p'_j t'_{i+j} (p_j - t_{i+j})^2 = 0$$

Po rozwinieciu dostajemy forme:

$$\sum_{j-0}^{m-1} (p'_j p_j^2 t'_{i+j} - 2p'_j p_j t_{i+j} t'_{i+j} + p'_j t_{i+j}^2 t'_{i+j})$$

Takie rozwiniecie łatwo jest obliczyć wykorzystujac 3xFFT oraz podstawowe operacje na ciagach.

Złożoność

Analiza złożoności jest trywialna ponieważ algorytm działa w czasie FFT.