Modelowanie Internetu

Projekt 2 Robot internetowy i graf WWW

Jakub Mamelski 137339

# 1.Koncepcja

Moim zadaniem było zaimplementowanie robota internetowego, który przejrzy wybraną domenę, pobierze jej zawartość na dysk, dokona analizy powstałego grafu oraz obliczy szereg statystyk. Dodatkowo uruchomiona zostanie własna implementacja algorytmu PageRank. Wybrana przeze mnie strona to:

<http://www.sportowapolitechnika.pl>

Dodatkowo przeprowadzałem testy na kilku mniejszych stronach ze względu na szybkość działania programu.

Program zaimplementowałem w języku C# ze względu na dobrą znajomość tego języka oraz łatwość pisania. Jako, że C# jest językiem zarządzany prawdopodobnie straciłem trochę na wydajności działania programu.

# 2. Analiza grafu

Podczas analizowania grafu połączeń domeny <http://www.sportowapolitechnika.pl> mój program uzyskał następujące statystyki grafu połączeń dokumentów w domenie:

a) Liczba wierzchołków

237

b) Liczba krawędzi

12193

c)Rozkład stopni wejściowych:

Zapisany w pliku z raportem

d)Rozkład stopni wyjściowych:

Zapisany w pliku z raportem

e) Najkrótsze ścieżki :

Zapisane w pliku z raportem.

Do wyznaczenia najkrótszych ścieżek zaimplementowałem algorytm Floyda-Warshalla. Posiłkowałem się pseudokodem ze strony: <https://en.wikipedia.org/wiki/Floyd–Warshall_algorithm>

f) Średnia odległość między wierzchołkami w grafie :

5,4256489705817

g) Średnica grafu :

16

h) Promień grafu :

6

i) Średni stopień wejściowy :

51,4472573839662

j) Średni stopień wyjściowy:

51,4472573839662

k) Średnia wartość PageRank :

0,003055240317245

j) Czas analizy dokumentów:

00:00:06.4465581

k) Czas działania algorytmu Floyda-Warshalla :

00:00:06.6717503

# 3. Analiza wielowątkowego parsowania dokumentów

Program jest napisany w technologii .NET i korzysta z TPL(Task Parallel Library). Zrównoleglone zostały 2 rzeczy:

* Liczba dopuszczalnych wątków w programie:

Jak widać różnice nie są duże. Optymalną liczbą wątków w programie to powyżej 100.

Ciekawy jest wysoki wynik (długi czas) dla 50 wątków, ale może to być spowodowane losowymi zaburzeniami.

* Stopień zrównoleglenia pętli obliczeniowej

Widać, że zrównoleglenie, czyli przeskok z jednego wątku do dwóch dało widoczne przyśpieszenie. Przy tak małej ilości obliczeń zwiększanie ilości wątków nie ma większego znaczenia a wraz ze wzrostem ich liczby rośnie czas poświęcane na ich obsługę.

# 4. Ataki i awarie

Podczas implementacji programu musiałem uodpornić się na potencjalne awarie. Aby to osiągnąć:

* Każdy link, który potencjalnie program mógłby odwiedzić jest analizowany pod kątem poprawności składni.
* Uwzględniam reguły z pliku robots.txt ( przykładowa domena <https://pomoc.pg.gda.pl> )
* Analizowane są tylko pliki html.
* Nie analizowane są pliki pdf, zdjęcia itp.
* Wszystkie wyjątki są łapane i odpowiednio obsługiwane
* Linki prowadzone poza domenę są ignorowane (nie dodawane do grafu)