STUDIA INFORMATICA 2011
Volume 32 Number 3A (98)

Ziemowit NOWAK, Dawid SMORAWSKI
Politechnika Wrocławska, Instytut Informatyki

POMIARY WYDAJNOŚCI INTERNETU – STUDIUM PRZYPADKU

Streszczenie. Analizie poddano dane pomiarowe, zarejestrowane w trakcie eksperymentu prowadzonego przez system MWING. Cztery agenty systemu rejestrowały czasy etapów transakcji WWW z 52 serwerami na całym świecie. Wyniki obrazują charakter ruchu sieciowego WWW oraz prezentują występujące zależności i zjawiska, mające wpływ na wydajność Internetu.

Słowa kluczowe: badanie wydajności Internetu, analiza danych pomiarowych, eksperyment aktywny, MWING

INTERNET PERFORMANCE MEASUREMENTS – A CASE STUDY

Summary. We analyzed the measurement data recorded during an experiment conducted by the MWING system. Four agents accounts the transaction times of 52 Web servers around the world. The results illustrate the nature of Web traffic and present relationships and phenomena that affect the Internet performance.

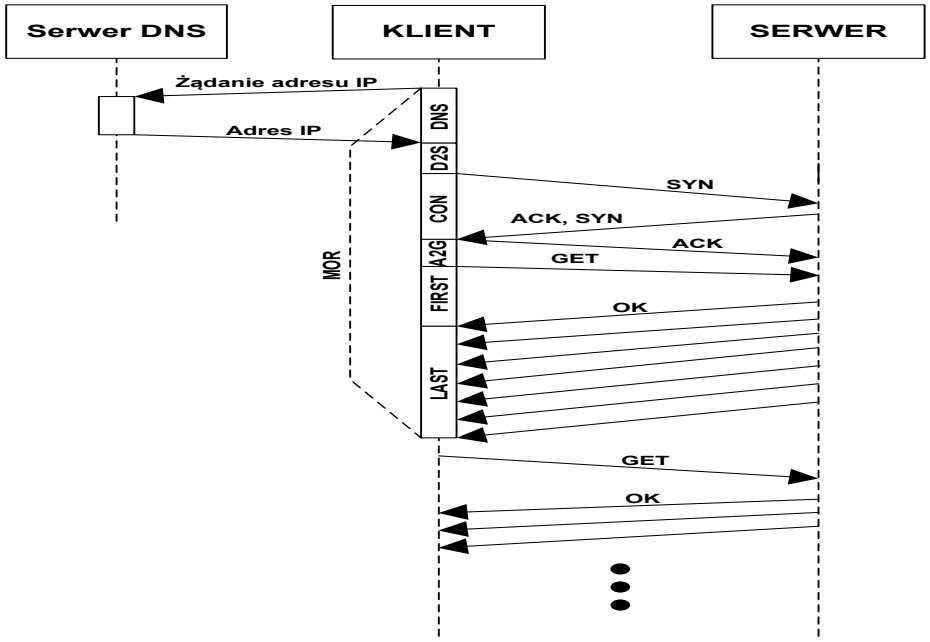
Keywords: Internet performance test, data analysis, active experiment, MWING

1. Wprowadzenie

Artykuł dotyczy analizy danych, będących wynikami pomiarów przeprowadzonych w celu sprawdzenia wydajności działania usługi WWW. Przez pojęcie wydajności Internetu rozumiemy efektywność funkcjonowania usługi WWW, czyli czasu oczekiwania użytkownika na zasób, którego zażądał. Czas oczekiwania na żądane zasoby jest jednym z podstawowych kryteriów oceny jakości usług WWW przez użytkowników. Liczba użytkowników sieci Internet stale rośnie, zatem jej wydajność ulega zmianom [1]. Pomiary wydajności Internetu i badania dostępności usługi WWW prowadzone są już od kilkunastu lat i mają na celu zapobieganie spadkom wydajności poprzez tworzenie nowych rozwiązań polepszających funkcjo-

nowanie usługi WWW. Zakłócenia i przerwy występujące w sieci mogą prowadzić do nieprawidłowego działania usługi WWW lub do znacznego zwiększenia czasu oczekiwania na pobranie zasobów.

Przeanalizowano dane pomiarowe zebrane przez system MWING (ang. *Multiagent Web pING)*. Agenty pomiarowe systemu były zlokalizowane w czterech ośrodkach akademickich, we Wrocławiu, Gdańsku, Gliwicach oraz Las Vegas. Zastosowano eksperyment aktywny poprzez generowanie i wysyłanie żądań do próbki rozproszonych po całym świecie serwerów WWW. Celem pomiarów było zbadanie czasów poszczególnych etapów transakcji WWW [2]. Agenty na przestrzeni piętnastu miesięcy, co pół godziny wysyłały żądania do 52 serwerów w celu pozyskania zasobu, którym był plik rfc1945.txt. Jednym z głównych założeń prowadzonego badania była obserwacja wydajności Internetu podczas pobierania zasobu tej samej wielkości. Na rys. 1 zostały zobrazowane mierzone interwały czasowe, wchodzące w skład transakcji WWW.



Rys. 1. Diagram prostej transakcji WWW
Fig. 1. Diagram of a simple Web transactions

W trakcie wykonywania pomiarów zmierzone zostały następujące parametry:

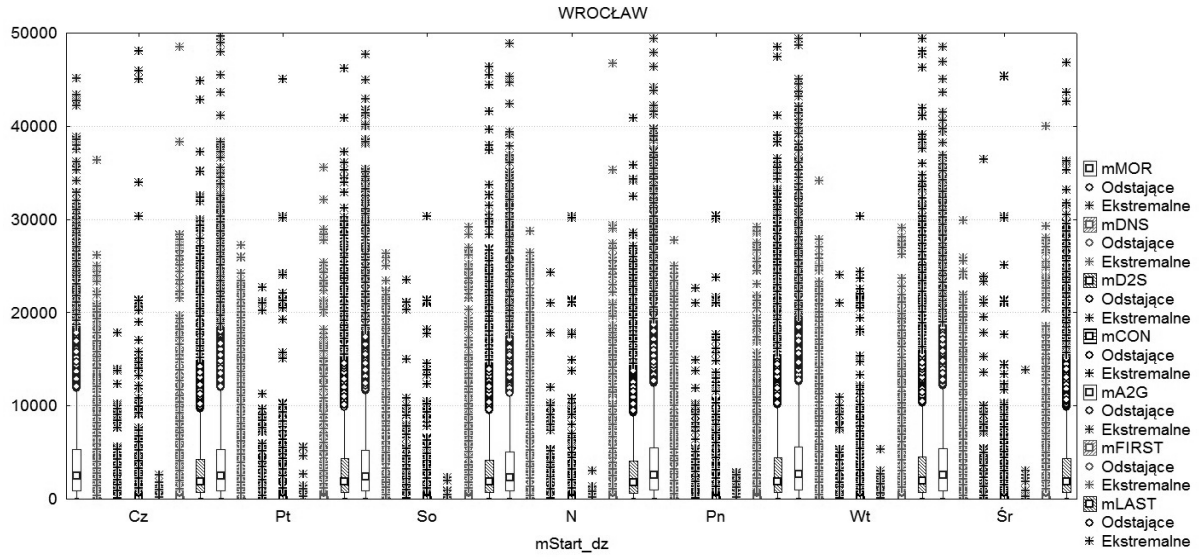
- DNS – czas potrzebny do uzyskania adresu IP serwera WWW,
- D2S – czas pomiędzy otrzymaniem adresu IP serwera WWW a wysłaniem żądania o ustanowienie z nim połączenia,
- CON – czas trwania zestawiania połączenia klienta z serwerem, nazywany również parametrem RTT,

- A2G – czas pomiędzy nawiązaniem połączenia z serwerem WWW a wysłaniem żądania GET do serwera WWW,
- FIRST – czas pomiędzy wysłaniem żądania GET a otrzymaniem pierwszego pakietu z odpowiedzią,
- LAST – czas pomiędzy odebraniem pierwszego pakietu z odpowiedzią a resztą pobieranych danych (właściwy czas transmisji danych),
- MOR – całkowity czas pobierania zasobu, stanowiący sumę wyżej wymienionych parametrów, czyli (DNS+D2S+CON+A2G+FIRST+LAST).

System MWING rejestrował czasy trwania etapów transakcji w nanosekundach. Na potrzeby analizy, czasy zostały zamienione na milisekundy i zaokrąglone do liczb całkowitych. W datach rozpoczęcia pojedynczych pomiarów i serii pomiarów wyodrębniono też dni tygodnia. W nazwach przekształconych parametrów dodany został przedrostek „m", np. mMOR, mCON.

2. Przygotowanie danych do analizy

Zanim przystąpiono do szczegółowej analizy danych pomiarowych, utworzone zostały wykresy typu Box-plot oraz histogramy rozkładów empirycznych wszystkich mierzonych parametrów z czterech miast: Wrocławia, Gdańska, Gliwic i Las Vegas.



Rys. 2. Box-plot przedstawiający rozkład wartości odstających i ekstremalnych
Fig. 2. Box-plot showing the distribution of outliers and extreme

Analizując rys. 2 można zauważyć, że poszczególne parametry mają bardzo zróżnicowane wartości. Łatwo można również dostrzec dużą liczbę wartości odstających i ekstremalnych. Zauważyć można także, że rozkład wartości parametrów dla dni tygodnia jest równomierny.

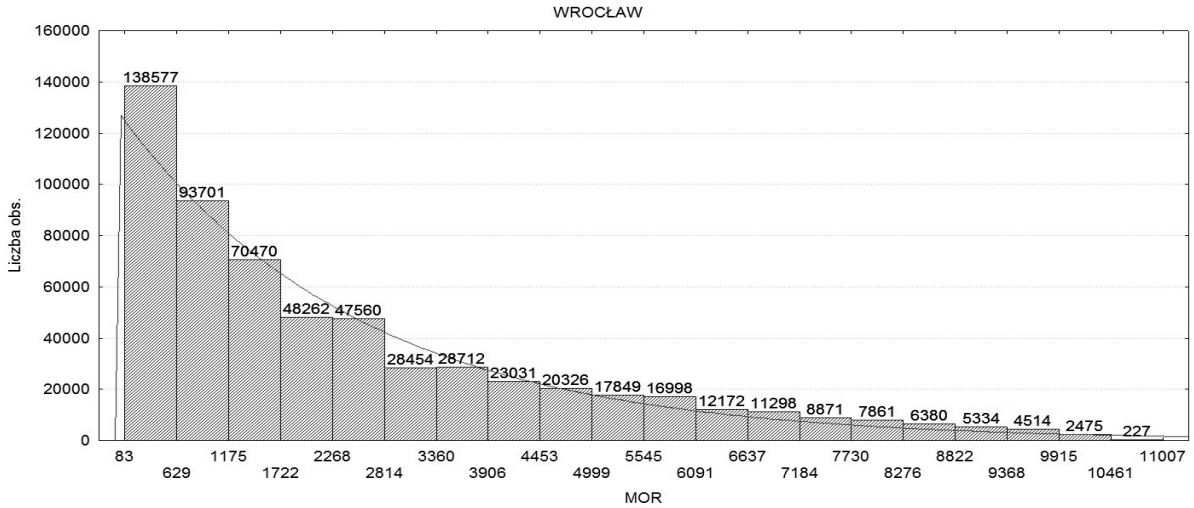
W kolejnym etapie, za pomocą programu Statistica [3], odfiltrowane zostały wartości odstające i ekstremalne. Tabela opisuje zakres wartości nieodstających, na podstawie których zostały przygotowane dane do analizy.

Tabela 1

Przedziały wartości nieodstających dla badanych parametrów

Miasto	Przedziały wartości nieodstających [ms]						
	MOR	DNS	D2S	CON	A2G	FIRST	LAST
Wrocław	83 – 12006	0 – 222	0 – 2	2 – 632	0 – 0	0 – 666	57 – 9801
Gdańsk	90 – 6296	0 – 255	0 – 2	2 – 649	0 – 0	0 – 669	59 – 4267
Gliwice	60 – 6294	0 – 197	0 – 2	0 – 557	0 – 0	1 – 657	40 – 4298
Las Vegas	58 – 3759	0 – 428	0 – 2	1 – 395	0 – 0	8 – 373	40 – 2420

Ograniczenie zbioru do wartości nieodstających umożliwiło uzyskanie lepszego rozkładu empirycznego na histogramie oraz zagwarantowało uzyskiwanie wiarygodnych wyników podczas dalszych analiz.



Rys. 3. Rozkład empiryczny parametru MOR po oczyszczeniu z wartości odstających
Fig. 3. Empirical distribution of parameter MOR after treatment with outliers

Na rys. 3 widać wyraźnie, że histogram dla parametru MOR przyjmuje rozkład podobny do rozkładu wykładniczego. Z tak przygotowanych danych można uzyskać bardziej klarowne wyniki, dlatego też opisane dalej analizy i obliczenia statystyczne zostały przeprowadzone na danych oczyszczonych z wartości odstających. Dla pozostałych parametrów, mierzonych z poziomu innych miast, sytuacja wygląda podobnie. Po oczyszczeniu, zbiór danych był znacznie zróżnicowany, ale nie zawsze przyjmował rozkład wykładniczy. W przypadku parametrów D2S oraz A2G zróżnicowanie wartości było tak znikome, że pominięto je w dalszych analizach.

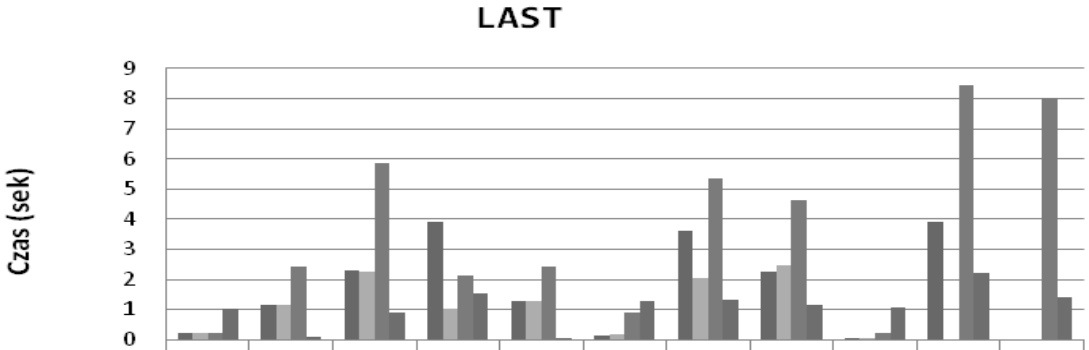
3. Analiza czasów transmisji

Niniejsza analiza ma na celu pokazanie podobieństw i różnic pomiędzy parametrami zarejestrowanymi dla badanych serwerów. Na potrzeby analizy przeprowadzono selekcję serwerów według następujących kryteriów:

- dwa serwery, z których pobieranie zasobu trwało najdłużej,
- dwa serwery, z których pobieranie zasobu trwało najkrócej.

Innymi słowy, dla każdego agenta wybrane zostały te serwery, dla których na przestrzeni całego okresu trwania eksperymentu mediany parametru LAST przyjmowały najmniejsze lub największe wartości. Przykładowo, jeżeli dla agenta wrocławskiego wybrano serwery A, B, C, D, a dla agenta gliwickiego D, E, F, G, to zbiór serwerów obydwu agentów uzupełniono o wybrane serwery innego agenta, czyli w tym przypadku agent wrocławski uzupełnia swoją listę serwerów o serwery E, F, G agenta gliwickiego, a agent gliwicki uzupełnia listę serwerów serwerami A, B, C agenta wrocławskiego. Mechanizm ten został zastosowany wobec wszystkich agentów. W efekcie, z puli 52 serwerów wybrano 11 wspólnych serwerów, z których każdy agent miał wybrane dwa, z których pobieranie zasobów było najszybsze oraz dwa, z których pobieranie zasobów było najwolniejsze.

Zestawienie wyników przedstawia rys. 4. Analizowanymi wartościami były mediany parametru LAST, zmierzone podczas całego okresu badania.

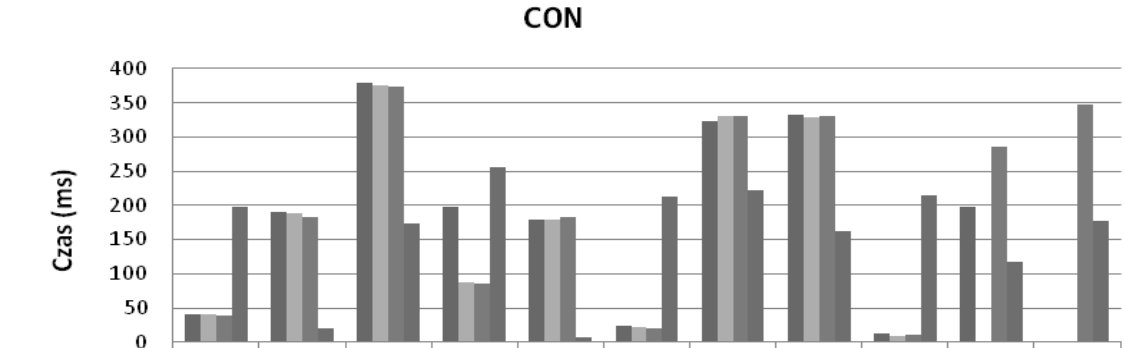


	1 BE	2 US	3 AU	4 IL	5 US	6 AT	7 TW	8 KR	9 PL	10 JP	11 CN
LAST GDAŃSK	238	1143	2303	3930	1269	149	3616	2263	65	3930	
LAST GLIWICE	240	1137	2261	1013	1267	158	2037	2477	52		
LAST WROCŁAW	235	2428	5872	2147	2409	915	5351	4637	225	8458	8041
LAST VEGAS	1023	112	882	1554	58	1273	1343	1150	1075	2196	1396

Rys. 4. Mediany parametru LAST dla wybranych serwerów
Fig. 4. Medians of LAST parameter for the selected servers

Zgodnie z tabelą zestawieniową można zauważyć, że dla wszystkich polskich miast najwydajniejsze było połączenie do polskiego serwera nr 9, drugim najszybszym serwerem dla Gdańska i Gliwic był austriacki serwer nr 6, natomiast dla Wrocławia belgijski serwer nr 1. Przyglądając się czasom połączeń z tymi serwerami można zauważyć różnice dla każdego

z agentów. Największe zróżnicowanie mają czasy zarejestrowane przez agenta wrocławskiego. Czasy odnotowane przez agenty z Gdańska i Gliwic są bardzo podobne. W przypadku agenta z Las Vegas czas połączenia z polskim serwerem jest jednym z najdłuższych zarejestrowanych połączeń.



	1 BE	2 US	3 AU	4 IL	5 US	6 AT	7 TW	8 KR	9 PL	10 JP	11 CN
CON GDAŃSK	41	190	380	198	180	23	324	333	12	198	
CON GLIWICE	40	188	375	87	179	22	330	328	8		
CON WROCŁAW	38	182	373	85	183	20	330	330	10	287	348
CON VEGAS	197	20	173	256	7	212	222	163	214	118	177

Rys. 5. Mediany parametru CON dla wybranych serwerów
Fig. 5. Medians of CON parameter for the selected servers

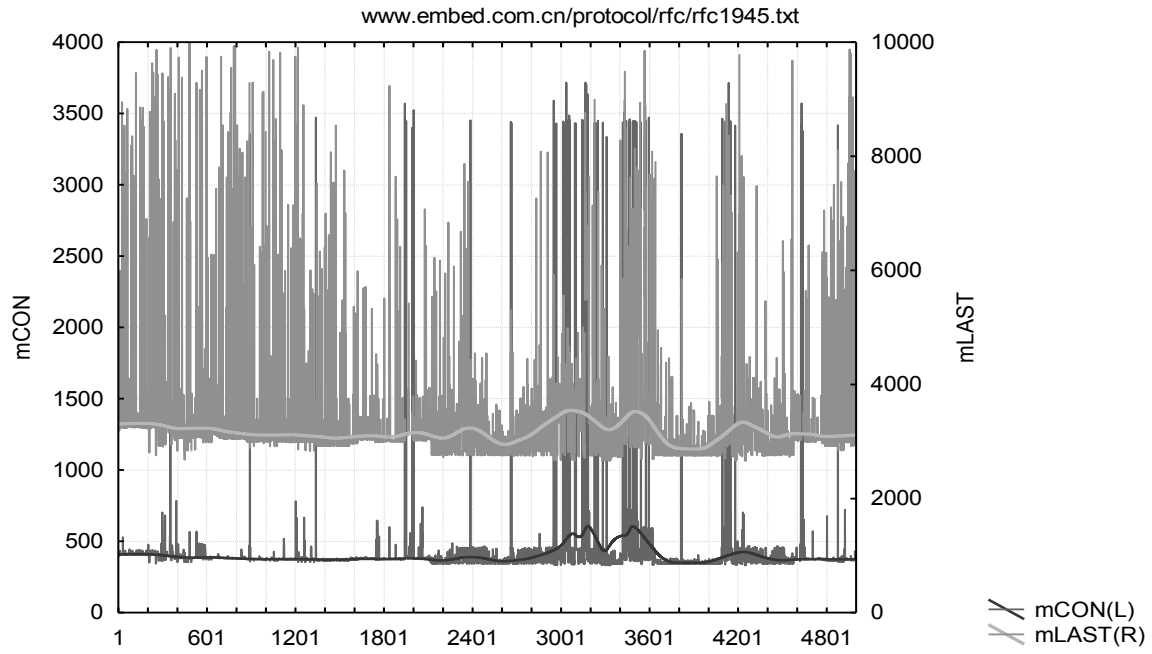
Przyglądając się największym wartościom parametru LAST można zauważyć, że dla Wrocławia, Gdańska i Las Vegas najniższą wydajnością charakteryzowało się połączenie do serwera japońskiego numer 10. Wyjątkiem był agent gliwicki, który najwolniejsze połączenie miał z serwerem nr 8. Na wykresach widać, że połączenia, które są wydajne z punktu widzenia jednych agentów nie są postrzegane, jako wydajne przez inne agenty. Zależność tę widać, np. dla serwerów nr 6 i 10. W przypadku serwera nr 6 różnice te są nawet czterokrotne.

4. Przewidywanie wartości THROUGHPUT

Przepustowość stanowi uniwersalną miarę oceny wydajności Internetu, która mówi o ilości przesłanych informacji w jednostce czasu z jednego węzła do drugiego. Nazywana jest potocznie szybkością transmisji danych. Znając rozmiar pobieranego zasobu oraz wartość parametru LAST, który dotyczy właściwego czasu pobierania zasobu, można w prosty sposób obliczyć daną przepustowość. Przepustowość oznaczona została przez THROUGHPUT i wyrażona w kb/s. Obliczona została na podstawie następującej zależności:

$$THROUGHPUT = \frac{137686 \cdot 8 \cdot 1000}{LAST \cdot 1024} \tag{1}$$

Następnie próbowano odpowiedzieć na pytanie, czy jest możliwe oszacowanie przepustowości na podstawie parametru CON. W tym celu analizie poddany został parametr CON. Jak można zaobserwować na rys. 5, o ile wartości parametru CON są równomierne i adekwatne do wartości LAST w przypadku małych wartości, to sprawa już nie jest tak oczywista w przypadku dużych wartości tego parametru. Widać to wyraźnie na przykładzie serwera nr 10, dla którego wartości parametru LAST przyjmowały najwyższe wartości, zaobserwowane przez każdego agenta.

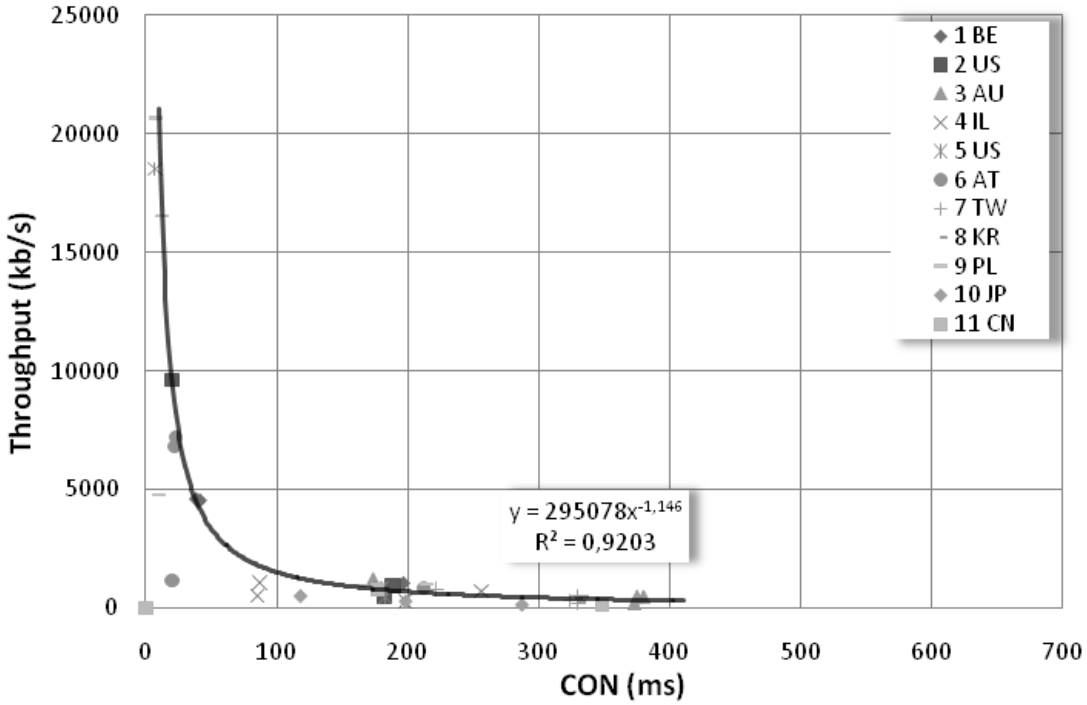


Rys. 6. Zarejestrowane dla jednego z serwerów wartości LAST na tle CON
Fig. 6. Registered for one of the servers values LAST on the background CON

Kolejnym krokiem, w celu sprawdzenia, czy parametr CON może predykować LAST, a tym samym THROUGHPUT, było przeprowadzenie analizy trendu (rys. 6). Żeby znaleźć linie trendu kolejnych pomiarów parametrów CON i LAST, konieczne było zmniejszenie zagęszczenia próbek, przez zmniejszenie skali odciętych do wartości 5000. Bez tej czynności, przy normalnej skali wykresu nie byłoby możliwości dojrzenia wyznaczonej linii trendu. Na lewej osi rzędnych odłożono wartości parametru CON, ona też została ograniczona przedziałem wartości od 0 do 4000. Prawa oś rzędnych dotyczy parametru LAST, z ograniczoną skalą od 0 do 10000. Wyniki pomiarów obydwu parametrów zestawione zostały na jednym wykresie, aby ułatwić znalezienie podobieństwa trendu. Linie trendu uzyskano przez dopasowanie danych metodą regresji Lowess (ang. *locally weighted scatterplot smoothing*) [4].

Dolna linia trendu dotyczy parametru CON, natomiast górna – parametru LAST. Przyglądając się liniom trendu badanych wielkości można zauważyć ich podobieństwo. Wskazuje to na istnienie pewnych zależności pomiędzy tymi parametrami i sugeruje możliwość wykorzystania CON, jako predyktora LAST.

Następnym krokiem było utworzenie wykresu punktowego, przedstawiającego wzajemną korelację wartości parametru CON z wartościami THROUGHPUT (rys. 7). Na osi odciętych odłożono zmierzone dla wybranych serwerów wartości parametru CON, które należy traktować, jako przyczynę (zmienna niezależna, objaśniająca). Na osi rzędnych odłożono obliczone wartości THROUGHPUT, które należy traktować, jako skutek (zmienna zależna, objaśniana). Punkty wyznaczone na wykresie reprezentują wybrane serwery. Na położenie każdego punktu wpływa wartość parametrów CON oraz THROUGHPUT.



Rys. 7. Wykres wzajemnej korelacji wartości CON i THROUGHPUT
Fig. 7. Figure mutual correlation values CON and THROUGHPUT

Przyglądając się wykresowi można stwierdzić, że rozkład wartości CON i THROUGHPUT przedstawia zależność potęgową obydwu parametrów. Model funkcyjny, wyznaczony metodą najmniejszych kwadratów oraz jego współczynnik determinacji R^2=0,9203 wskazują na dobre dopasowanie modelu do wartości pomiarowych. Tylko nieliczne próbki odstają od linii dopasowania. Potwierdza to wcześniejszy wniosek, że parametr CON nie może być w pełni wiarygodnym predyktorem przepustowości. Jednak w większości przypadków znając wartości CON i model zależności z wielkością THROUGHPUT można oszacować przeciętną przepustowość dla wybranych serwerów.

Podsumowując, rozkład punktów na wykresie wskazuje silną zależność pomiędzy parametrami CON i THROUGPUT, polegającą na tym, że im większa jest wartość CON, tym parametr THROUGHPUT będzie mniejszy.

5. Podsumowanie

Badanie wydajności Internetu przynosi wiele korzyści. Poznanie zjawisk i zależności występujących podczas pobierania zasobów umożliwia tworzenie nowych metod i algorytmów, usprawniających ich pobieranie. Na potrzeby dalszych analiz dobrym rozwiązaniem byłoby także badanie wydajności Internetu w trakcie pozyskiwania zasobów o dużych rozmiarach. Pożądane byłyby także badania dla większej liczby serwerów oraz przez większą liczbę agentów, rozmieszonych na innych kontynentach.

Dane pomiarowe przeanalizowano prostymi metodami statystycznymi. Jednak występowanie dużej liczby wartości odstających sugeruje, że badany rozkład może być superpozycją kilku rozkładów. Dlatego w dalszych badaniach wskazane byłoby wykorzystanie do analizy tzw. modeli mieszanin (ang. *mixture models*).

W celu przeprowadzenia szczegółowych analiz, związanych z metodami przewidywania wydajności Internetu, dobrym rozwiązaniem byłoby wykorzystanie metod eksploracji danych (ang. *data mining*) [5]. Zautomatyzowanie procesu analizy danych pozwoliłoby na szybsze otrzymywanie wyników. Można byłoby również operować na znacznie większych zbiorach danych pomiarowych.

BIBLIOGRAFIA

1. Borzemski L., Nowak Z., Starczewski G.: Internet: zmiany po pięciu latach. Techniczne i teoretyczne aspekty współczesnych sieci komputerowych. WKŁ, Warszawa 2009.
2. Zhi J.: Web Page Design and Download Time. CMG Journal of Computer Resource Management, Issue 102, Spring 2001.
3. Luszniewicz A., Słaby T.: Statystyka z pakietem komputerowym Statistica: teoria i zastosowania. Wydawnictwo C.H. Beck, Warszawa 2008.
4. Zey C. i inni: e-Handbook of Statistical Methods. National Institute of Standards and Technology. http://www.itl.nist.gov/div898/handbook/, 2010.
5. Borzemski L., Kliber M., Nowak Z.: Using data mining algorithms in Web performance prediction. Cybernetics and Systems, 2009, Vol. 40, No. 2.

Recenzenci: Prof. dr hab. inż. Bolesław Pochopień
Prof. dr hab. inż. Tadeusz Wieczorek

Wpłynęło do Redakcji 14 kwietnia 2011 r.

Abstract

This article concerns the analysis of data are the results of the measurements made to verify the performance of Web services. Concept of Internet performance should be understood as effectiveness of web services, i.e., user waiting time for a resource.

The authors analyzed survey data collected by the system MWING (Multiagent Web pING). Measurement agents were located at four academic centers in Wroclaw, Gdansk, Gliwice, and Las Vegas. It was active experiment. MWING had to generate and send requests to spread around the world web servers. The purpose of the measurements was to measure the time the various stages of Web transactions (Figure 1). Agents for 15 months every half-hour sent the requests to 52 servers in order to gain a resource that was the rfc1945.txt file.

On the basis of measurement data, Box-plot graphs were created, and histograms of the empirical distributions of all measured by the four agents (Figure 2 and 3). Figure 3 clear that the histogram for the distribution of MOR parameter is similar to the exponential distribution. From such prepared data can be more clear results, therefore, described in the article analysis and statistical calculations were performed on the data cleaned of outliers.

Then, we analyzed transmission times for selected servers. Summary of results are shown in Figure 4. For the analysis we took the medians LAST parameter. Medians were calculated from all measurements.

The next step was to establish whether it is possible to predict the value of THROUGHPUT (1). To this end, we conducted an analysis of the trend (Figure 6 and 7). The distribution of points in Figure 7 indicates a strong correlation between the parameter of CON and THROUGPUT. The higher the value of CON, the value of the THROUGHPUT will be less. This is the so-called power law.

Performance testing the Internet brings many benefits. Thanks to the understanding of phenomena and relationships that occur when downloading resources, you can create new methods and algorithms to improve their collection.

Adresy

Ziemowit NOWAK: Politechnika Wrocławska, Instytut Informatyki,
ul. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Polska, ziemowit.nowak@pwr.wroc.pl
Dawid SMORAWSKI: Politechnika Wrocławska, Instytut Informatyki,
ul. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Polska.