Atrybuty

Dla każdych dwóch urządzeń porównujemy ich atrybuty i otrzymujemy ciąg podobieństwa. Oznaczmy przez D_1, D_2 porównywane urządzenia.

1 Atrybuty podstawowe

Atrybuty podstawowe porównujemy zero-jedynkowo.

2 Czas korzystania

Reprezentujemy przez ciąg 24 - elementowy. Jeżeli w całym badanym okresie, w danej godzinie pojawiła się aktywność, to stawiamy 1, w p.p 0. Stosujemy miarę $\frac{\text{liczba pokrywających się jedynek}}{max\{c_1,c_2\}}$, gdzie c_1, c_2 to liczba jedynek w odpowiadających ciągach. Ciągi są znormalizowane względem stref czasowych.

3 Dominujący kraj

Dla każdego urządzenia tworzymy ciąg (k_1, k_2, c), gdzie $k_1,\ k_2-$ to pierwszy i drugi najczęściej występujący kraj (przy czym kraj drugi wybieramy tylko, gdy występuje w co najmniej 30% zapytań, w przeciwnym przypadku $k_2 = 0$), c - liczba wszystkich krajów. Za tę samą wartość na miejscach 1 i 2 przyznajemy odpowiednio wartości 0.5, 0.2. Liczbę krajów porównujemy następująco: $0.3 \cdot \frac{min\{c_1,c_2\}}{max\{c_1,c_2\}}$, gdzie c_1 liczba krajów dla D_1, c_2 - dla D_2. Wszystkie przyznane wartości sumujemy.

4 Dominujący region

Jak dla kraju.

5 Dominujące IP

Dla każdej godziny porównujemy dominujące IP. Liczbę takich samych, niezerowych wyników dzielimy przez liczbę godzin, w których oba urządzenia były aktywne. Inna wersja: zliczamy wspólne IP, podwajamy i dzielimy przez ilość wszystkich IP dla obu urządzeń.

6 Dominujący ISP

Jak przy IP.

7 Typ użytkownika określony względem URL

Z pliku requests usuwamy wszystkie informacje poza dev i url; grupujemy je po takim samym dev i klastrujemy devices za pomocą programu R, na podstawie podobieństwa zbiorów url przypisanych do jednego device (typ użytkownika). Informacje o przyporządkowaniu do danego klastra porównujemy dla dwóch urządzeń zero-jedynkowo.

8 Liczba odwiedzonych stron w ciągu godziny

Dla urządzenia tworzymy parę (Max, med), gdzie Max - to maksimum, a med - mediana liczby stron otwartych w ciągu godziny. Dla D_1 i D_2 mamy odpowiednio pary (Max_1, med_1), (Max_2, med_2). Niech $M = max\{Max_1, Max_2\}$, $m = min\{Max_1, Max_2\}$. Miara:

$$\frac{1}{(1+M-m)^2}\ .$$

Podobnie porównujemy medianę.

9 Liczba stron startowych

Atrybut określony na podstawie pierwszego połączenia w ciągu dnia. Porównujemy podobnie jak w punkcie 8, z tym że ustalamy dodatkowy parametr *days* - liczba dni aktywności urządzenia. Jeżeli $days < 10$, to pomijamy porównywanie.

10 Typ połączenia (connection type)

Mamy 5 typów połączeń:

1. The database identifies dial-up (modem telefoniczny)
2. cellular (komórkowy)
3. cable
4. DSL
5. corporate connection speeds (wszyscy mają jedno IP)

Miara jak dla kraju.

11 Anonymous

Zliczmy liczbę takich samych wartości w kolumnach anonymous.