WYKŁAD X

O ROZKŁADZIE ODWOŁAŃ DO STRON WWW

1. ZNAJOMOŚĆ STRUKTURY ODWOŁAŃ W SIECI WWW POZWALA ZOPTYMALIZOWAĆ JEJ BUDOWĘ. BADANIA EMPIRYCZNE POKAZAŁY, ŻE ROZKŁAD ODWOŁAŃ DO STRON WWW ODBYWA SIĘ ZGODNIE Z PRAWEM ZIPFA:

NIECH Ω BĘDZIE ZBIOREM ZASOBÓW WWW. $|\Omega| = N$

WPROWADZIMY PORZĄDEK NA TYCH ZASOBACH: $\Omega = \{z_1, z_2, \dots, z_N\}$

PRZY CZYM INDEKS i PRZY z_i OZNACZA MIEJSCE NA LIŚCIE RANKINGOWEJ NAJPOPULARNIEJSZYCH ZASOBÓW (z_1 – NAJBARDZIEJ POPULARNY; z_N – NAJMNIEJ POPULARNY). OZNACZMY PRZEZ P_i PRAWDOPODOBIEŃSTWO, ŻE DANE ODWOŁANIE JEST ODWOŁANIEM DO ZASOBU z_i. ZACHODZI ZALEŻNOŚĆ:

$$P_i = \frac{c}{i^\alpha},$$

GDZIE c JEST STAŁĄ NORMUJĄCĄ I WYNOSI $\frac{1}{H_N^\alpha}$, GDZIE $H_N^\alpha = \sum_{i=1}^{N} \frac{1}{i^\alpha}$.

(DLA UPROSZCZENIA $H_N^1 = H_N$). BADANIA WYKAZAŁY, ŻE $0 < \alpha \leq 1$.

WARTOŚĆ α ZALEŻY OD POPULACJI LUDZI, KTÓRZY ODWOŁUJĄ SIĘ DO STRON.

2. JAKIE JEST ZNACZENIE WSPÓŁCZYNNIKA α?

ZASTANÓWMY SIĘ NAD $\phi(k)$ – PRAWDOPODOBIEŃSTWEM, ŻE NASTĄPI ODWOŁANIE DO JEDNEJ Z k NAJPOPULARNIEJSZYCH STRON

$$\phi(k) = \sum_{i=1}^{k} \frac{c}{i^\alpha} = c \cdot H_k^\alpha = \frac{H_k^\alpha}{H_N^\alpha}$$

ŻEBY MIEĆ LEPSZE WYOBRAŻENIE O TEJ WARTOŚCI, WYKONAMY SZACOWANIE

$$\int_1^N \frac{1}{x^\alpha}\,dx \leq H_n^\alpha \leq 1 + \int_1^N \frac{1}{x^\alpha}\,dx$$

$$\int_1^m \frac{1}{x^\alpha}\,dx = \begin{cases} \ln x \Big|_1^m = \ln m \\ \dfrac{x^{1-\alpha}}{1-\alpha}\Big|_1^m = \dfrac{m^{1-\alpha}}{1-\alpha} - \dfrac{1}{1-\alpha} = \dfrac{m^{1-\alpha}-1}{1-\alpha} \end{cases}$$

CZYLI

$$\frac{k^{1-\alpha}-1}{N^{1-\alpha}-\alpha} = \frac{\frac{k^{1-\alpha}-1}{1-\alpha}}{1+\frac{N^{1-\alpha}-1}{1-\alpha}} \leq \frac{H_k^\alpha}{H_N^\alpha} \leq \frac{1+\frac{k^{1-\alpha}-1}{1-\alpha}}{\frac{N^{1-\alpha}-1}{1-\alpha}} = \frac{k^{1-\alpha}-\alpha}{N^{1-\alpha}-1} \leq (1+\varepsilon)\left(\frac{k}{N}\right)^{1-\alpha}+\delta$$

$$\Lambda$$

$$(1+\varepsilon')\left(\frac{k}{N}\right)^{1-\alpha}-\delta'$$

CZYLI $$\frac{H_k^\alpha}{H_N^\alpha} = \Omega\left(\left(\frac{k}{N}\right)^{1-\alpha}\right).$$

OBSERWACJA:

DLA α BLISKIEGO ZERU PRAWDOPODOBIEŃSTWO, ŻE ODWOŁAMY SIĘ DO KTÓREJŚ Z k PIERWSZYCH STRON JEST MNIEJSZE NIŻ, GDY α JEST BLISKIE 1.

3. JAKI JEST WSPÓŁCZYNNIK TRAFIEŃ?

A GDY CACHE JEST NA TYLE DUŻE, ŻE NIE TRZEBA BYŁO NIC USUWAĆ.

Załóżmy, że wykonaliśmy już R odwołań i zastanawiamy się, jaka jest szansa, że R+1. odwołanie jest trafieniem.

$$H(R) = Pr\left(\bigcup_{i=1}^{N} \text{„wylosowano } i\text{-ty zasób za } R+1 \text{ razem i wcześniej też go wylosowano"}\right)$$

$$= \sum_{i=1}^{N} Pr(\text{wylosowano } i\text{-ty zasób za } R+1 \text{ razem i wcześniej też go wylosowano"}) =$$

$$= \sum_{i=1}^{N} P_i \cdot Pr(\text{„wcześniej wylosowano } i\text{-ty zasób"}) =$$

$$= \sum_{i=1}^{N} P_i \cdot (1 - Pr(\text{„wcześniej nie wylosowano } i\text{-tego zasobu"})) =$$

$$= \sum_{i=1}^{N} P_i \cdot \left(1 - \prod_{k=1}^{R} Pr(\text{„w } k\text{-tej rundzie nie wylosowano } i\text{-tego zasobu"})\right) =$$

$$= \sum_{i=1}^{N} P_i \left(1 - \prod_{k=1}^{R} (1-P_i)\right) = \sum_{i=1}^{N} P_i \left(1 - (1-P_i)^R\right)$$

PODSTAWIAMY TERAZ TO, CO WYNIKA Z PRAWA ZIPFA

$$S = \sum_{i=1}^{N} \frac{1}{i \ln N}\left(1 - \left(1 - \frac{1}{i \ln N}\right)^R\right)$$

OSZACUJEMY TERAZ:

$$1-\left(1-\frac{1}{i\ln N}\right)^R$$

PRZEZ 1 DLA $i \le \frac{R}{\ln N}$ ORAZ PRZEZ $1-\left(1-\frac{1}{R}\right)^R$ DLA $i > \frac{R}{\ln N}$

OTRZYMUJEMY

$$S \le \sum_{i=1}^{\frac{R}{\ln N}} \frac{1}{i\ln N} + \underbrace{\sum_{i=1}^{N} \frac{\left(1-\left(1-\frac{1}{R}\right)^R\right)}{i\ln N} - \sum_{i=1}^{\frac{R}{\ln N}} \frac{\left(1-\left(1-\frac{1}{R}\right)^R\right)}{i\ln N}}_{\frac{1}{\ln N}\left(H_N - H_{\frac{R}{\ln N}}\right)}$$

WARTOŚĆ $\left(1-\frac{1}{R}\right)^R$ DLA ODPOWIEDNIO DUŻYCH R BLISKA JEST $\frac{1}{R}$

MOŻEMY ZATEM PRZYJĄĆ, ŻE $1-\left(1-\frac{1}{R}\right)^R \approx d$. MAMY ZATEM

$$S \approx \frac{\ln \frac{R}{\ln N}}{\ln N} + d - d\left(\frac{\ln \frac{R}{\ln N}}{\ln N}\right) \approx D \ln R + C$$

CZYLI OBSERWUJĄC ZMIENNOŚĆ R POWINNIŚMY DOSTAĆ WYKRES LOGARYTMICZNY

b GDY ROZMIAR CACHE JEST OGRANICZONY.

ZAKŁADAMY, ŻE W CACHE MIEŚCI SIĘ S STRON NIEZALEŻNIE OD ICH ROZMIARU. ZAKŁADAMY TEŻ, ŻE W CACHE ZNAJDUJE SIĘ S NAJPOPULARNIEJSZYCH STRON. WSPÓŁCZYNNIK TRAFIEŃ

$$H(S) = Pr(\text{„WYLOSOWANO ZASÓB SPOŚRÓD } S \text{ NAJBARDZIEJ POPULARNYCH"}) = \sum_{i=1}^{S} P_i$$

MAMY

$$\frac{S^{1-\alpha}-1}{1-\alpha} \leq H_S^{\alpha} \leq 1 + \frac{S^{1-\alpha}-1}{1-\alpha}$$

LUB

$$\ln S \leq H_S^{\alpha} \leq 1 + \ln S$$

CZYLI

$$\frac{\frac{S^{1-\alpha}-1}{1-\alpha}}{\frac{N^{1-\alpha}-\alpha}{1-\alpha}} \leq H(\$) \leq \frac{\frac{S^{1-\alpha}-\alpha}{1-\alpha}}{\frac{N^{1-\alpha}-1}{1-\alpha}}$$

$$\frac{S^{1-\alpha}-1}{N^{1-\alpha}-\alpha} \leq H(\$) \leq \frac{S^{1-\alpha}-\alpha}{N^{1-\alpha}-1}$$

LUB

$$\frac{\ln S}{1+\ln N} \leq H(\$) \leq \frac{1+\ln S}{\ln N}$$

CZYLI $H(\$) = \Omega(c \ln S)$, gdzie $c = H_N$

4. ILE CZASU UPŁYWA MIĘDZY KOLEJNYMI ODWOŁANIAMI DO TEJ SAMEJ STRONY?

OBLICZYMY PRAWDOPODOBIEŃSTWO ŻE DOPIERO ZA k-TYM ODWOŁANIEM DOSTANIEMY ZASÓB

$$d(k) = \bigcup_{i=1}^{N} Pr(\text{UZYSKALIŚMY } i\text{-TY ZASÓB I DOSTANIEMY GO DOPIERO ZA } k\text{-TYM ODWOŁANIEM})$$

$$= \sum_{i=1}^{N} P_i \, Pr(\text{Przez } k-1 \text{ nie uz. } z_i \text{ oraz ze } k\text{-tym razem jest } z_i \,/\, \text{teraz było } z_i) =$$

$$= \sum_{i=1}^{N} P_i (1-P_i)^{k-1} \cdot P_i = \sum_{i=1}^{N} P_i^2 (1-P_i)^{k-1}$$

SPRÓBUJMY DO UZYSKANEGO WZORU WSTAWIĆ PRAWO ZIPFA I OSZACOWAĆ ZA POMOCĄ CAŁKI:

$$\int_1^N \left(\frac{1}{x \ln N}\right)^2 \left(1-\frac{1}{\ln N}\right)^{k-1} dx = (-1)^k \sum_{j=0}^{k-1} \frac{(-1)^j \binom{k-1}{j}}{(j+1)\, x^{j+1} \ln^{j+2} N} \Bigg|_{x=1}^{x=N} = \quad (1)$$

$$\sum_{j=0}^{k-1} \frac{(-1)^j \binom{k-1}{j}}{(j+1)\, x^{j+1} \ln^{j+1} N} z^{j+1} \Bigg|_{z=1} = \int_0^z \left(1-\frac{z}{x \ln N}\right)^{k-1} dz \Bigg|_{z=1} =$$

$$= \left(\frac{(z - x\ln(N)) \left(1-\frac{z}{x\ln N}\right)^k}{k+1} + \frac{x\ln(N)}{k+1} \right) \Bigg|_{z=1} =$$

$$= \frac{(1 - x\ln N)\left(1-\frac{1}{x\ln N}\right)^k + x\ln N}{k+1}$$

PO ZASTOSOWANIU DO (1) MAMY:

$$(-1)^k \frac{1}{(k+1)\ln N} \left((1-N\ln N)\left(1-\frac{1}{N\ln N}\right)^k + N\ln N - (1-\ln N)\left(1-\frac{1}{\ln N}\right)^k - \ln N \right)$$

WNIOSKI:

MOŻEMY ZAŁOŻYĆ, ŻE N DĄŻY STALE WTEDY CAŁOŚĆ ZACHOWUJE SIĘ $\approx \frac{1}{k}$.

5. NA ILE MOŻEMY ZBLIŻYĆ SIĘ DO ROZKŁADU ZIPFA?

NIECH P BĘDZIE CIĄGIEM KOLEJNYCH ODWOŁAŃ DO ZASOBÓW WWW.

NIECH $n_p(i)$ BĘDZIE LICZBĄ WYSTĄPIEŃ ODWOŁANIA DO z_i W P.

NIECH E(i) BĘDZIE WARTOŚCIĄ OCZEKIWANĄ LICZBY WYSTĄPIEŃ ODWOŁANIA DO z_i.

$$E(i) = \sum_{k=0}^{|P|} k \binom{|P|}{k} p_i^k (1-p_i)^{|P|-k} = \frac{|P|}{i \ln N}$$

UŻYJEMY TERAZ NIERÓWNOŚCI CHERNOFFA:

NIECH $X_1, \dots, X_n$ BĘDĄ NIEZALEŻNYMI ZMIENNYMI LOSOWYMI TAKIE

$$Pr(X_i = 1) = p$$
$$Pr(X_i = 0) = 1 - p$$

NIECH $X = \sum_{i=1}^{n} X_i$. NIECH $EX = pn$. WTEDY

$$Pr(|X - pn| > \alpha) \leq 2e^{-\frac{2\alpha^2}{n}}$$

TO PASUJE DO NASZEJ SYTUACJI, BO $X_j^i = \begin{cases} 1 & \text{gdy } j\text{-te odwołanie jest do } i \\ 0 & \text{gdy } j\text{-te odwołanie nie jest do } i \end{cases}$

$$Pr\left(|n_p(i) - \frac{|P|}{i \ln N}| > \alpha\right) \leq 2e^{-\frac{2\alpha^2}{|P|}}$$

MOŻEMY CHCIEĆ SIĘ DOWIEDZIEĆ, JAK DŁUGIEGO P TRZEBA, ABY UZYSKAĆ ODPOWIEDNIO MAŁE PRAWDOPODOBIEŃSTWO, ŻE RÓŻNIMY SIĘ OD WARTOŚCI OCZEKIWANEJ.

$$2e^{-\frac{2\alpha^2}{|P|}} \leq \varepsilon$$

$$-\frac{2\alpha^2}{|P|} \leq \ln \frac{\varepsilon}{2}$$

$$|P| \geq \frac{2\alpha^2}{\ln 2 - \ln \varepsilon}$$

GDYBYŚMY CHCIELI DUŻE PRAWDOPODOB., ŻE $n_p(i)$ JEST BLIŻEJ $E(i)$ NIŻ $E(i+1)$, TO PRZYJĘLIBYŚMY

$$\alpha < \frac{E(i) - E(i+1)}{2} = \frac{|P|}{2 \ln N}\left(\frac{1}{i} - \frac{1}{i+1}\right) = \frac{|P|}{2 \ln |N|} \cdot \frac{1}{i(i+1)}$$

DLA DOWOLNEGO i, CZYLI

$$\alpha < \frac{|P|}{2 \ln |N|} \cdot \frac{1}{N(N+1)}$$

WTEDY

$$|P| \geq \frac{2 \frac{|P|^2}{4 m^2 N \cdot N^2 (N+1)^2}}{\ln \frac{2}{\varepsilon}}$$

$$\ln \frac{2}{\varepsilon} \geq \frac{2|P|}{4 m^2 N \cdot N^2 \cdot (N+1)^2}$$

$$|P| \leq 2 m^2 N \cdot N^2 (N+1)^2 \cdot \ln \frac{2}{\varepsilon}$$

WNIOSEK: PRZY BARDZO DŁUGICH PRÓBKACH MOŻEMY SIĘ POGUBIĆ.