Fundusze Europejskie Wiedza Edukacja Rozwój
Politechnika Wrocławska
Unia Europejska Europejski Fundusz Społeczny
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Materiały do wykładów z analizy algorytmów

Problem przybliżonego zliczania

Jakub Lemiesz

1 Problem zliczania

Multizbiór $\mathfrak{M}$ definiujmy jako parę (S, m), gdzie S jest pewnym zbiorem, często nazywanym zbiorem podstawowym, natomiast m jest funkcją $m : S \to \mathbb{N}_{\geq 1}$. Wartość funkcji $m(s)$ nazywamy licznością elementu $s \in S$. Problem zliczania możemy sformułować następująco: jak wyznaczyć moc n zbioru podstawowego S dla multizbioru $\mathfrak{M}$ dysponując jedynie ograniczoną pamięcią.

W ogólnym przypadku, bez dodatkowych założeń dotyczących multizbioru $\mathfrak{M}$, dokładne wyznaczenie wartości n wymaga przechowania wszystkich elementów zbioru S, a zatem pamięci rzędu $O(n)$. W wielu praktycznych zastosowaniach, gdy zbiór S jest bardzo duży, wyznaczanie dokładnej wartości n może być zatem kosztowne, a często nie jest konieczne. Dopuszczając, że uzyskamy jedynie pewne oszacowanie prawdziwej wartości n możemy jednak w istotny sposób ograniczyć wykorzystanie pamięci.

2 Algorytmy przybliżonego zliczania

Rozważane dalej algorytmy umożliwiające oszacowanie mocy zbioru podstawowego są algorytmami probabilistycznymi. Umożliwiają one ustalenie kompromisu pomiędzy wymaganą pamięcią a dokładnością oszacowania. Algorytmy te opierają się na sprytnym wykorzystaniu tzw. szkiców danych, czyli zwięzłego podsumowania danych np. w postaci tablicy haszy wybranych elementów. Dzięki temu mają bardzo ograniczone wymagania pamięciowe.

Algorytmy przybliżonego zliczania mają zastosowanie w wielu różnego rodzaju scenariuszach. Najczęściej pojawiają się w kontekście analizy strumieni danych, kiedy dokładne zliczanie często nie ma uzasadnienia, a rozmiary przetwarzanych danych są olbrzymie. Algorytmy przybliżonego zliczania stosuje się też m.in. w systemach bazodanowych (np. operacja *distinct count*), w analizie genomu czy też w analizie ruchu sieciowego. Możemy np. myśleć o szacowaniu liczby par komunikujących się urządzeń na podstawie nagłówków pakietów przechodzących przez router.

Jednym z pierwowzorów tego typu algorytmów był licznik Morrisa [7]. Jego idea została rozwinięta w algorytmach *Probabilistic Counting* [4] i *LogLog* [2], a ostatecznie znalazła zastosowanie w popularnym algorytmie *HyperLogLog* [3]. Algorytm *HyperLogLog* jest obecnie wykorzystywany i rozwijany m.in. przez firmy takie jak Google [5] czy Oracle [1].

Inna, popularna rodzina algorytmów przybliżonego zliczania opiera się na statystykach pozycyjnych traktując przechowywane w szkicu danych hasze jak losowe liczby z przedziału $[0, 1]$. Do tej rodziny należy m.in. analizowany dalej algorytm *MinCount*.

3 Algorytm MinCount

Pseudokod algorytmu *MinCount* prezentujemy poniżej. Algorytm otrzymuje na wejściu trzy argumenty - referencję do analizowanego multizbioru $\mathfrak{M}$, funkcję haszującą h i parametr $k \geq 3$ określający rozmiar tablicy przechowującej hasze wybranych elementów. Wykorzystanie funkcji haszującej zapewnia substytut losowości umożliwiający przypisanie do identycznych elementów tej samej pseudo-losowej wartości. Innymi słowy, aplikując funkcję $h : D \to \{0, 1\}^B$ do elementów multizbioru $\mathfrak{M} = (S, m)$ dla $S \subseteq D, |S| = n$ otrzymujemy n unikalnych wartości, które mogą

być traktowane jako realizacje n niezależnych zmiennych losowych o rozkładzie jednostajnym na przedziale $[0, 1)$. Wartość dla parametru B określającego długość hasza powinna być tak dobrana by prawdopodobieństwo, że dla różnych elementów zostanie wygenerowany ten sam hasz było małe (patrz: paradoks urodzinowy).

Algorytm `MinCount` wylicza hasze dla kolejnych elementów multizbioru $\mathfrak{M}$ i w każdej chwili przechowuje k najmniejszych haszy. W dowolnym momencie można na podstawie tych haszy oszacować liczbę zaobserwowanych dotychczas unikalnych elementów. Jeśli liczba unikalnych elementów n jest mniejsza niż k to algorytm zwróci dokładną wartość n (pomijając mało prawdopodobną sytuację kolizji haszy dla różnych elementów). Jeśli $n \geq k$ algorytm zwróci oszacowanie

$$\hat{n} = \frac{k-1}{M[k]} ,$$

gdzie $M[k]$ oznacza największy z k przechowywanych haszy. Wyprowadzenie powyższego estymatora opiera się na statystykach pozycyjnych.

MinCount($\mathfrak{M}$, h , k)

Inicjalizacja: ustaw k pozycji $M[1], \ldots, M[k]$ tablicy M na 1

for all $x \in \mathfrak{M}$ **do**
  if $h(x) < M[k] \wedge h(x) \notin M$ **then** ▷ *hasz* $h(x) \in [0, 1)$
    $M[k] \leftarrow h(x)$
    sort(M) ▷ *posortuj* M *rosnąco*
  end if
end for
if $M[k] == 1$ **then**
  Return: $\hat{n} \leftarrow |\{i : M[i] \neq 1\}|$
else
  Return: $\hat{n} \leftarrow (k-1)/M[k]$
end if

Statystyki pozycyjne

Niech $X_1, X_2, \ldots, X_n$ oznaczają dowolne zmienne losowe. Posortujmy realizację tych zmiennych w porządku rosnącym $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. Wówczas zmienną $X_{i:n}$ nazywamy i-tą statyką pozycyjną. Zauważmy, że w szczególności

$$X_{1:n} = \min\{X_1, X_2, \ldots, X_n\}$$

oraz

$$X_{n:n} = \max\{X_1, X_2, \ldots, X_n\} .$$

Lemat 1 *Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o tym samym rozkładzie opisanym funkcją gęstości $f(x)$ oraz dystrybuantą $F(x)$. Wówczas k-ta statystyka pozycyjna $X_{k:n}$ ma rozkład zadany funkcją gęstości*

$$f_k(x) = nf(x)\binom{n-1}{k-1}F(x)^{k-1}\left(1 - F(x)\right)^{n-k} .$$

Dowód.

Wykazanie prawdziwości lematu pozostawiamy jako ćwiczenie (zobacz: Sekcja 5). Wskazówka: zauważ, że jeśli pewna zmienna X_i jest k-tą statystyką pozycyjną to dokładnie $k-1$ innych zmiennych przyjęło wartości mniejsze od niej i że zmienne te można wybrać na $\binom{n-1}{k-1}$ sposobów. ■

Lemat 2 *Niech $U_1, U_2, \ldots, U_n$ będą niezależnymi zmiennymi losowymi o rozkładzie jednostajnym na odcinku $[0, 1)$. Wówczas zmienna losowa $U_{k:n}$ ma rozkład $Beta(k, n-k+1)$ oraz wartość oczekiwaną*

$$\mathbb{E}\left(U_{k:n}\right) = \frac{k}{n+1} .$$

Dowód.

Szczegółowy dowód lematu pozostawiamy jako ćwiczenie (zobacz: Sekcja 5). Istotne jest, aby skorzystać z Lematu 1 oraz z faktu, że rozkład $Beta(\alpha, \beta)$ jest opisany funkcją gęstości

$$b(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

gdzie $B(\alpha, \beta)$ jest funkcją beta i dla $Re(\alpha) > 0$, $Re(\beta) > 0$ mamy

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt\ .$$

Funkcję beta można również przedstawić jako

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\ ,$$

gdzie $\Gamma(z)$ oznacza *gammę Eulera* i zachodzi zależność $\Gamma(n+1) = n!$ dla $n \in \mathbb{N}$.

Następnie należy pokazać, że wartość oczekiwana zmiennej losowej X pochodzącej z rozkładu $Beta(\alpha, \beta)$ możemy wyrazić jako:

$$\mathbb{E}\,(X) = \frac{\alpha}{\alpha+\beta}.$$

■

Estymator liczności

Zauważmy, że przechowywane w tablicy M algorytmu hasze możemy interpretować jako statystki pozycyjne $U_{1:n}, U_{2:n}, \ldots, U_{k:n}$ dla n niezależnych zmiennych losowych $U_1, U_2, \ldots, U_n$ o rozkładzie jednostajnym $U_i \sim U(0,1)$. W szczególności wartość na pozycji $M[k]$ może być interpretowana jako k-ta statystyka pozycyjna $U_{k:n}$. W oparciu o Lemat 2 możemy dla $k \geq 2$ pokazać następującą równość (ćwiczenie):

$$\mathbb{E}\left(\frac{1}{U_{k:n}}\right) = \int_0^1 \frac{1}{x}\, b(x; k, n+1-k)\, dx = \frac{n}{k-1}\ ,$$

gdzie $b(x; \alpha, \beta)$ jest zdefiniowaną wyżej funkcją gęstości rozkładu beta. Z tej równości możemy wyprowadzić następujący estymator liczby elementów n zależny od ustalonej wartości k (metoda momentów):

$$\hat{n}_k = \frac{k-1}{U_{k:n}}\ .$$

Estymator $\hat{n}_k$ jest nieobciążonym estymatorem parametru n dla $k \geq 2$:

$$\mathbb{E}\,(\hat{n}_k) = (k-1)\frac{n}{k-1} = n.$$

Podobnie, gdy $k \geq 3$, możemy policzyć wariancję i błąd standardowy estymatora (ćwiczenie):

$$\mathbb{SE}\,[\hat{n}_k] = \sqrt{\mathbb{V}\mathrm{ar}\,(\hat{n}_k/n)}\ .$$

Podsumowanie wyników znajduję się w Tabeli 1.

k	$\mathbb{E}\left(\hat{n}_k\right)$	$\mathbb{V}\mathrm{ar}\left(\hat{n}_k\right)$	$\mathbb{SE}\left[\hat{n}_k\right]$
1	∞	∞	–
2	n	∞	∞
3	n	$n(n-2)$	$\sqrt{\frac{n-2}{n}}$
...	...	...	...
k	n	$\frac{n(n-k+1)}{k-2}$	$\sqrt{\frac{1}{k-2}} + \mathcal{O}\left(\frac{1}{n}\right)$

Tabela 1: Średnia, wariancja i błąd standardowy estymatora $\hat{n}_k$ dla różnych wartości k.

4 Koncentracja estymatora

Wariancja oraz błąd standardowy dają pewną informację o koncentracji rozkładu estymatora wokół średniej. Zazwyczaj jesteśmy jednak zainteresowani mocniejszymi gwarancjami, na przykład by z dużym prawdopodobieństwem błąd estymatora był ograniczony:

$$\mathbb{P}\left(\delta_1 n < \hat{n}_k < \delta_2 n\right) \geq 1 - f(\delta_1, \delta_2)\ ,$$

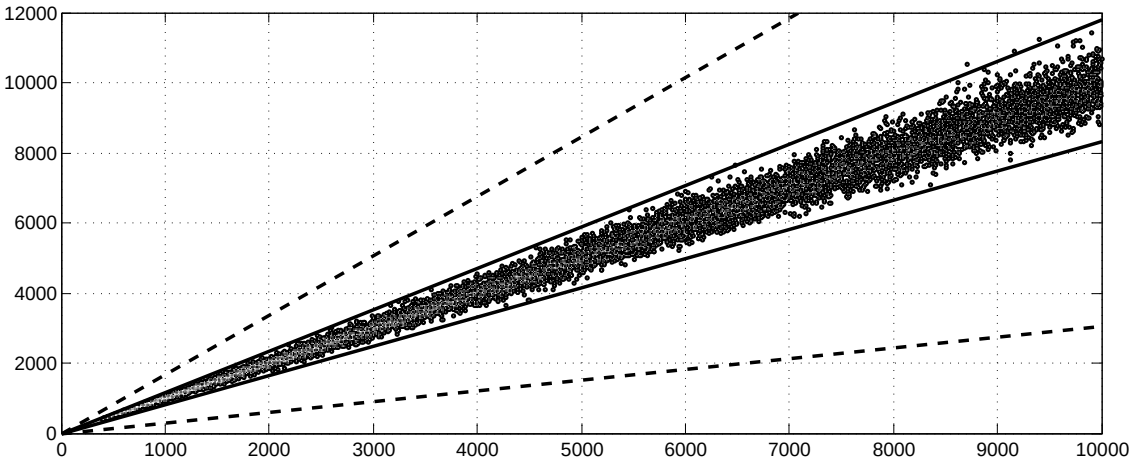
gdzie $f(\delta_1, \delta_2)$ jest pewną ustaloną funkcją.

Mając formułę opisującą wariancję estymatora możemy celu uzyskanie tego rodzaju gwarancji wykorzystać nierówność Czebyszewa. Mianowicie dla dowolnej liczby rzeczywistej $a > 0$ oraz dla zmiennej losowej X o skończonej wartości oczekiwanej i skończonej, niezerowej wariancji mamy (ćwiczenie):

$$\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| < a\right) > 1 - \frac{\mathbb{V}\mathrm{ar}\left(X\right)}{a^2}\ .$$

Ustalając wartość parametru a i podstawiając do powyższej formuły wartość oczekiwaną oraz standardowe odchylenie estymatora $\hat{n}_k$ możemy uzyskać pewne informacje o jego koncentracji (ćwiczenie).

Na Rysunku 1 znajduje się eksperymentalne porównanie wyników zwracanych przez estymator z ograniczeniami uzyskanymi przez zastosowanie nierówności Czebyszewa oraz omawianej dalej nierówności Chernoffa. Ograniczenia wynikające z nierówności Czebyszewa są w tym przypadku niezbyt precyzyjne.



Rysunek 1: Porównanie wyników symulacji algorytmu *MinCount* (kropki) dla $k = 400$ z ograniczeniami uzyskanymi przez zastosowanie nierówności Czebyszewa (linie przerywane) oraz nierówności Chernoffa (linie ciągłe) na poziomie ufności $1 - \alpha$ dla $\alpha = 0.005$. Na osi poziomej znajdują się rzeczywiste wartości $n = 1, \ldots, 10^4$.

Pokażemy w jaki sposób można otrzymać dokładniejsze ograniczenia redukując własności statystyk pozycyjnych do rozkładu dwumianowego oraz wykorzystując nierówność Chernoffa. Niech $B_{p,n}$ oznacza zmienną losową o rozkładzie dwumianowym z parametrami p oraz n :

$$\mathbb{P}\left(B_{p,n}=k\right)=\binom{n}{k}p^k(1-p)^{n-k}\ .$$

W dalszej części wykorzystamy następujący lemat.

Lemat 3 *Niech $k\leq n$ oraz $p\in(0,1)$.*

1. *Jeśli $pn<k$ to $\mathbb{P}\left(U_{k:n}\leq p\right)\leq e^{-pn}\left(\frac{pne}{k}\right)^k$.*
2. *Jeśli $pn>k$ to $\mathbb{P}\left(U_{k:n}\geq p\right)\leq e^{-pn}\left(\frac{pne}{k}\right)^k$.*

Dowód.

Wykorzystamy znane nierówności Chernoffa dla rozkładu dwumianowego. Dla $\delta>0$ oraz $0<\rho\leq 1$ mamy (ćwiczenie):

$$\mathbb{P}\left(B_{p,n}\geq(1+\delta)np\right)\leq\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{np}\ , \tag{1}$$

$$\mathbb{P}\left(B_{p,n}\leq(1-\rho)np\right)\leq\left(\frac{e^{-\rho}}{(1-\rho)^{(1-\rho)}}\right)^{np}\ . \tag{2}$$

Dla niezależnych zmiennych losowych $U_1,\ldots,U_n$ o rozkładzie jednostajnym na odcinku $[0,1)$ zdefiniujmy zmienne $I_1,\ldots,I_n$ takie, że $I_i=1$ jeśli $U_i\leq p$ oraz $I_i=0$ w przeciwnym przypadku. Wówczas zmienna losowa

$$B_{p,n}=I_1+\ldots+I_n$$

ma rozkład dwumianowy z parametrami n i p.

Warunek $U_{k:n}\leq p$ oznacza, że $|\{i:U_i\leq p\}|\geq k$, co oznacza, że $B_{p,n}\geq k$. Zauważmy, że

$$k=pn\left(1+\frac{k-pn}{pn}\right)\ .$$

Stad, jeśli $pn<k$ to z nierówności (1) mamy

$$\mathbb{P}\left(U_{k:n}\leq p\right)=\mathbb{P}\left(B_{p,n}\geq k\right)\leq e^{-pn}\left(\frac{pne}{k}\right)^k\ .$$

Załóżmy teraz, że $pn>k$. Warunek $U_{k:n}\geq p$ jest równoważny $B_{p,n}\leq k$. Możemy zatem użyć nierówności (2) i otrzymać wynik analogicznie jak w poprzednim przypadku. ■

Twierdzenie 1 *Niech $3\leq k\leq n$ oraz $\eta>0$, $0<\varepsilon<1$. Oznaczmy*

$$f_k(x)=e^{xk}(1-x)^k\ .$$

Wówczas dla estymatora $\hat{n}_k=\frac{k-1}{U_{k:n}}$ mamy

$$\mathbb{P}\left(\frac{n}{1+\eta}\frac{k-1}{k}<\hat{n}_k<\frac{n}{1-\varepsilon}\frac{k-1}{k}\right)>1-f_k(\varepsilon)-f_k(-\eta)\ .$$

Dowód.

Niech $0<\varepsilon<1$. Z pierwszej części Lematu 3 mamy

$$\mathbb{P}\left(U_{k:n}\leq(1-\varepsilon)\frac{k}{n}\right)\leq e^{\varepsilon k}\left(1-\varepsilon\right)^k\ .$$

Zauważmy, że $U_{k:n}\leq(1-\varepsilon)\frac{k}{n}$ wtedy i tylko wtedy, gdy $\frac{n}{1-\varepsilon}\frac{k-1}{k}\leq\frac{k-1}{U_{k:n}}$. Stąd

$$\mathbb{P}\left(\frac{n}{1-\varepsilon}\frac{k-1}{k}\leq\frac{k-1}{U_{k:n}}\right)\leq e^{\varepsilon k}\left(1-\varepsilon\right)^k\ .$$

Podobnie pokazujemy, że

$$\mathbb{P}\left(\frac{n}{1+\eta}\frac{k-1}{k} \geq \frac{k-1}{U_{k:n}}\right) \leq e^{-\eta k}\left(1+\eta\right)^{k} \ .$$

■

Możemy zapisać powyższe twierdzenie w bardziej zwięzłej postaci.

Twierdzenie 2 *Niech* $3 \leq k \leq n$ *oraz* $0 < \delta_1 < \frac{k-1}{k} < \delta_2$ *. Oznaczmy* $f_k(x) = xe^{-x+1}$ *oraz*

$$F_k(\delta_1, \delta_2) = f_k\left(1 - \frac{k-1}{k\delta_1}\right) + f_k\left(1 - \frac{k-1}{k\delta_2}\right) \ .$$

Wówczas zachodzi nierówność:

$$\mathbb{P}\left(\delta_1 n < \hat{n} < \delta_2 n\right) > 1 - F_k(\delta_1, \delta_2) \ .$$

Na Rysunku 1 widać, że ograniczenie wynikające z nierówności Chernoffa jest w miarę ścisłe. W ogólności nierówność Chernoffa jest często silniejsza od nierówności Czebyszewa, ale opiera się na silniejszych założeniach, np. o niezależności zmiennych losowych (zobacz ćwiczenie 8).

5 Lista ćwiczeń

1 — Dla ciągłych i niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ o tym samym rozkładzie zadanym funkcją gęstości $f(x)$ i dystrybuantą $F(x)$ pokaż, że k-ta statystyka pozycyjna $X_{k:n}$ ma rozkład opisany funkcją gęstości

$$f_k(x) = \frac{F^{k-1}(x)\left[1-F(x)\right]^{n-k} f(x)}{B(k, n-k+1)} \ ,$$

gdzie $B(\alpha, \beta)$ oznacza funkcję beta. Wskazówka: zobacz dowód Lematu 1.

2 — Dla n niezależnych zmiennych losowych $U_1, \ldots, U_n$ o rozkładzie jednostajnym: $U_i \sim \mathcal{U}(0,1)$, pokaż, że k-ta statystyka pozycyjna ma rozkład $Beta(k, n-k+1)$ i wartość oczekiwaną $k/(n+1)$. Wskazówka: zobacz dowód Lematu 2.

3 — Niech $U_{k:n}$ oznacza k-tą statystykę pozycyjną dla n niezależnych zmiennych losowych o rozkładzie jednostajnym $\mathcal{U}(0,1)$. Pokaż, że dla $\hat{n}_k = \frac{k-1}{U_{k:n}}$ oraz $k \geq 2$ mamy $\mathbb{E}\left(\hat{n}_k\right) = n$ oraz, że dla dla $k \geq 3$ mamy

$$\mathbb{V}\mathrm{ar}\left(\hat{n}_k\right) = \frac{n(n-k+1)}{k-2} \ .$$

Wskazówka: wykorzystaj różne reprezentacje funkcji beta.

4 — (Nierówność Markowa) Niech X oznacza zmienną losową, która przyjmuje tylko nieujemne wartości. Wtedy dla wszystkich $a > 0$

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}\left(X\right)}{a} \ .$$

5 — (Nierówność Czebyszewa) Niech X oznacza zmienną losową o skończonej wartości oczekiwanej i skończonej, niezerowej wariancji. Pokaż, że dla dowolnego $a > 0$ zachodzi nierówność:

$$\mathbb{P}\left(\left|X - \mathbb{E}\left(X\right)\right| < a\right) > 1 - \frac{\mathbb{V}\mathrm{ar}\left(X\right)}{a^2} \ .$$

Wskazówka: wykorzystaj nierówność Markowa.

6 — (Nierównosc Chernoffa dla sumy prób Bernoulliego) Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi próbami Bernoulliego takimi, że $\mathbb{P}(X_i = 1) = p_i$. Niech $X = \sum_{i=1}^{n} X_i$ oraz $\mu = \mathbb{E}(X)$. Pokaż, że

a) dla dowolnego $\delta > 0$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu},$$

b) dla dowolnego $0 < \rho \leq 1$

$$\mathbb{P}(X \leq (1-\rho)\mu) \leq \left(\frac{e^{-\rho}}{(1-\rho)^{(1-\rho)}}\right)^{\mu}.$$

Wskazówka: możesz zajrzeć do książki [6].

7 — Wykorzystując oznaczenia i uzyskane w poprzednim zadaniu nierówności, pokaż, że dla dowolnego $0 < \delta < 1$

$$\mathbb{P}(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

8 — Niech S_n będzie liczbą orłów uzyskanych w n rzutach symetryczną monetą. Pokaż, że

a) stosując nierówność Czebyszewa mamy

$$\mathbb{P}\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{n}{4},$$

b) stosując nierówność Chernoffa z poprzedniego zadania mamy

$$\mathbb{P}\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq 2e^{-n/24}.$$

Literatura

[1] Analytical sql in oracle database 12c. http://www.oracle.com/technetwork/database/bi-datawarehousing/wp-in-database-analytics-12c-2132656.pdf, 2015.

[2] M. Durand and P. Flajolet. Loglog counting of large cardinalities. In *ESA*, 2003.

[3] P. Flajolet, E. Fusy, O. Gandouet, and F. Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *DMTCS Proceedings*, 2007.

[4] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31(2):182–209, 1985.

[5] S. Heule, M. Nunkesser, and A. Hall. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *EDBT'13*, 2013.

[6] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

[7] R. Morris. Counting large numbers of events in small registers. *Commun. ACM*, 21(10):840–842, Oct. 1978.