The space complexity of approximating the frequency moments

- Κωστόπουλος Δημήτριος
- Μπλα – Advanced Data Structures
- June 2007

Εισαγωγή – Ορισμός Frequency moments

- Έστω ακολουθία A = $\{a_1,a_2,...,a_m\}$ με κάθε a_i να ανήκει στο N = {1,2,...,n}.
- Έστω m_l το πλήθος των εμφανίσεων του αριθμού l στην ακολουθία A,
  $m_l = |\{ j : a_j = l \}|$
- Ορίζουμε $F_k = \Sigma_{i=1}^{n} m_i^k$. - Frequency moments
- $F_0 = \Sigma_{i=1}^{n} m_i^0$. Το πλήθος των διαφορετικών στοιχείων της ακολουθίας.
- $F_1 = \Sigma_{i=1}^{n} m_i$. Το μήκος της ακολουθίας. Ορίζουμε $m = F_1$.
- $F_2 = \Sigma_{i=1}^{n} m_i^2$.

Εισαγωγή – Τα πιο σημαντικά Frequency moments

- $F_k = \Sigma_{i=1}^{n} m_i^k$.
- $F_0 = \Sigma_{i=1}^{n} m_i^0$. Το πλήθος των διαφορετικών στοιχείων της ακολουθίας.
- $F_1 = \Sigma_{i=1}^{n} m_i$. Το μήκος της ακολουθίας. Ορίζουμε $m = F_1$.
- $F_2 = \Sigma_{i=1}^{n} m_i^2$.
- $F_{OO}^{*} = \max \{m_i, 1 <= i <= n\}$
- Τα frequency moments βοηθούν στη στατιστική ανάλυση των ακολουθιών.

Εισαγωγή

- Μπορούμε να υπολογίσουμε τα frequency moments εύκολα και γρήγορα απλά κρατώντας ένα μετρητή m_l για κάθε αριθμό l.
- Όμως αυτό απαιτεί χώρο Ω(nlogm), όπου n θυμίζουμε ο μεγαλύτερος αριθμός που μπορεί να εμφανιστεί στην ακολουθία.
- Θέλουμε πολύ λιγότερο χώρο.

Εισαγωγή – Πάνω φράγμα

- Τα F_k προσεγγίζονται με τυχαίο τρόπο χρησιμοποιώντας το πολύ $O(n^{1-1/k} \log n)$ bits , $k > 0$.
- Για $k = 1$: $O(\log n)$ bits
- Για $k = 2$: $O(n^{1/2} \log n)$ bits

Εισαγωγή – Κάτω φράγμα

- Τα F_k προσεγγίζονται με τυχαίο τρόπο χρησιμοποιώντας τουλάχιστον $\Omega(n^{1-5/k})$ bits , $k > 0$.
- Για $k < 5$: $\Omega(n^{-something})$ bits bits
- Για $k = 5$: $\Omega(1)$ bits
- Για $k > 6$: $\Omega(n^{1 - something})$ bits. Άρα για $k > 6$ τα “πράγματα” δεν είναι και τόσο “καλά”!
- Για $k = OO$: $\Omega(n)$ bits!, για το F_{OO}^{*}.

Κάτω φράγμα - Θεώρημα

- Θεώρημα
- Έστω ακολουθία $A = (a_1,...,a_m)$ με πεδίο τιμών το $\{1,...,n\}$.
  Για κάθε $k >= 1$ και κάθε $\lambda > 0$ και κάθε $\varepsilon > 0$ υπάρχει ένα πιθανοτικός αλγόριθμος που με ένα πέρασμα υπολογίζει το F_k.
  Αυτός ο αλγόριθμος αποκλίνει από το F_k πιο πολύ από λF_k με πιθανότητα το πολύ ε.
  Ο χώρος που χρησιμοποιεί ο αλγόριθμος είναι :
  $O(k\log(1/\varepsilon)\, n^{1-1/k}\, (\log n+\log m) / \lambda^2)$

Κάτω φράγμα - Απόδειξη

- Απόδειξη (σχεδόν ίδια με του paper αλλά κάπως πιο ποιοτικά...)
- Θέλουμε να προσεγγίσουμε όσο πιο πολύ γίνεται το F_k και με όσο μεγαλύτερη πιθανότητα μπορούμε.
- Γενικά, σε τέτοια προβλήματα μία τακτική είναι να μαντέψουμε t τυχαίες μεταβλητές τέτοια που $E(X_j)$ να είναι η τιμή που θέλουμε να προσεγγίσουμε, $j = 1,...,t$.
- Μετά ορίζουμε s Y_i τυχαίες μεταβλητές σαν τον μέσο όρο των X_j, $i = 1,...,s$
- Ο μεσαίος των Y_i είναι η προσέγγιση της τιμής μας.

Κάτω φράγμα - Απόδειξη

- Ξέρουμε, $F_k = \Sigma_{i=1}^{n} m_i^k$.
  Επίσης από πιθανότητες, για μια τυχαία διακριτή μεταβλητή : $E(X) = \Sigma_x x \Pr\{X=x\}$.
- Οπότε μπορούμε να πάρουμε για $X = n\, m_i^k$ και να πάρουμε το επιθυμητό αποτέλεσμα.
  $E(X) = \Sigma_x x \Pr\{X=x\} = \Sigma_{i=1}^{n} n\, m_i^k\, (1/n)$.
- Αυτό όμως δε δουλεύει γιατί έχουμε το i να εξαρτάται από το i !!!

Κάτω φράγμα - Απόδειξη

- Για να διώξουμε το n, πρέπει να πάρουμε κάτι παρόμοιο αλλά με διαφορετικό πλήθος μεταβλητών.
- Αναλύοντας το F_k έχουμε :
  $F_k = \Sigma_{i=1}^{n} m_i^k = m_1^k + m_2^k + ... + m_n^k$, n προσθετέοι. Ας δοκιμάσουμε να τους κάνουμε m (υπενθύμιση m = το μήκος της ακολουθίας).

Κάτω φράγμα - Απόδειξη

- Αλλά, $m_1 + m_2 + ... + m_n = m$. Το m είναι και αυτό μεταβλητό, :-(. Αλλά μπορούμε να το θεωρήσουμε σταθερό, ότι και καλά το ξέρουμε από πριν. Θα δείξουμε παρακάτω ότι και σταθερό να μην είναι κλάιν μάιν.
- $F_k = \Sigma_{i=1}^{n} m_i^k = m_1^k +$
  $+ m_2^k +$
  $+ ... +$
  $+ m_n^k$

Κάτω φράγμα - Απόδειξη

- $F_k = \Sigma_{i=1}^{n} m_i^k = m_1^k +$
  $+ m_2^k +$
  $+ \ldots +$
  $+ m_n^k$
- $F_k =$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_1 - 1)^k - (m_1 - 1)^k + m_1^k +$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_2 - 1)^k - (m_2 - 1)^k + m_2^k +$
  $+ \ldots +$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_n - 1)^k - (m_n - 1)^k + m_n^k +$

Κάτω φράγμα - Απόδειξη

- $F_k =$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_1 - 1)^k - (m_1 - 1)^k + m_1^k +$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_2 - 1)^k - (m_2 - 1)^k + m_2^k +$
  $+ \ldots +$
  $1^k - 1^k + 2^k - 2^k + \ldots + (m_n - 1)^k - (m_n - 1)^k + m_n^k$
- $F_k =$
  $(1^k - 0^k) + (2^k - 1^k) + (3^k - 2^k) \ldots + (m_1^k - (m_1 - 1)^k) +$
  $(1^k - 0^k) + (2^k - 1^k) + (3^k - 2^k) \ldots + (m_2^k - (m_2 - 1)^k) +$
  $+ \ldots +$
  $(1^k - 0^k) + (2^k - 1^k) + (3^k - 2^k) \ldots + (m_n^k - (m_n - 1)^k)$

Κάτω φράγμα - Απόδειξη

- Τώρα, αρκεί να ορίσουμε σαν $X = m\,(r^k - (r-1)^k)$.
- Τι μπορεί όμως να συμβολίζει αυτή η νέα μεταβλητή r;
- Θέλουμε το r να έχει m διαφορετικές τιμές. Τι έχει m διαφορετικές τιμές; Το μήκος της ακολουθίας καθώς την εξετάζουμε. Εάν ορίσουμε σαν r να είναι το r-οστό στοιχείο της ακολουθίας τότε το r θα παίρνει τιμές από το 1 έως το m και $E(X) = \Sigma_x x\ Pr\{X=x\} = m^k$. ΛΑΘΟΣ
- Θέλουμε να μπορούμε να διαχωρίσουμε ποιοτικά το κάθε στοιχείο της ακολουθίας, καθώς δεν ενδιαφερόμαστε πραγματικά για τη θέση του στην ακολουθία αλλά για τον αριθμό επανεμφάνισής του.

Κάτω φράγμα - Απόδειξη

- Οπότε θέλουμε το r να συμβολίζει ακριβώς το εάν συναντήσουμε έναν αριθμό l, να ξέρουμε ποια r-οστή φορά μπορεί να είναι αυτή που τον συναντάμε.
- Ορίζουμε, $r = | \{ q : q >= p, a_q = l = a_p \} |$
- Το r συμβολίζει τον αριθμό των εμφανίσεων του $a_p = l$ μετά τις πρώτες p πρώτες θέσεις της ακολουθίας. (Αυτό το “μετά τις πρώτες p πρώτες θέσεις της ακολουθίας” θα χρησιμοποιήσουμε αργότερα για να δείξουμε ότι δε μας πειράζει εάν δε γνωρίζουμε από πριν το μήκος της ακολουθίας).

Κάτω φράγμα - Απόδειξη

- Τι έχουμε καταφέρει μέχρι στιγμής;
- Τίποτα!
- Το καλό είναι ότι έως τώρα χρησιμοποιούμε χώρο το πολύ logn + logm.
- Logn για να κρατάμε το a_p=l και logm για να κρατάμε τον αριθμό των εμφανίσεων του l.

Κάτω φράγμα - Απόδειξη

- Ανισότητα Chebyshev
- Δοθείσης μίας τυχαίας μεταβλητής X τότε για a > 0,
  Pr { |X-E(X)| >= a } <= σ^2/α^2
  ,$\sigma = E(|X-E(X)|^2)$
  ,όπου σ^2 είναι η διασπορά της τυχαίας μεταβλητής X
  (Var(X) – variance of X).

Κάτω φράγμα - Απόδειξη

- Έστω Y_i ο μέσος όρος των μεταβλητών X, με i = 1 ... s.
- $\Pr \{ |Y_i - F_k| > \lambda F_k \}$
  $= \Pr \{ |Y_i - E(X)| > \lambda F_k \}$, $F_k = E(X)$
  $= \Pr \{ |Y_i - E(Y_i)| > \lambda F_k \}$, $E(X) = E(Y_i)$
  $\leq \mathrm{Var}(Y_i) / (\lambda^2 F_k^2)$, ανισότητα του Chebyshev
  $= \mathrm{Var}(X) / (t \lambda^2 F_k^2)$, $\mathrm{Var}(Y_i) = \mathrm{Var}(X) / t$
  $\leq E(X^2) / (t \lambda^2 F_k^2)$, $\mathrm{Var}(X) = E(X^2) - E(X)^2$
  $\leq k F_1 F_{2k-1} / (t \lambda^2 F_k^2)$, θέλει απόδειξη στον πίνακα.
  $\leq k n^{1-1/k} F_k^2 / (t \lambda^2 F_k^2)$, θέλει απόδειξη στον πίνακα.
  $\leq k n^{1-1/k} / (t \lambda^2)$

Κάτω φράγμα - Απόδειξη

- Οπότε, $\Pr\{ |Y_i - F_k| > \lambda F_k \} <= k\, n^{1-1/k} / (t\, \lambda^2)$
- Chernoff Bound :
  'Εστω, $X_1,...,X_m$ ανεξάρτητες δοκιμές Bernoulli έτσι ώστε :
  $\Pr[X_i=1] = p$, $\Pr[X_i=0] = 1 - p$.
  Έστω $X = \Sigma i Xi$ και $E(X) = mp$. Τότε για κάθε $\varepsilon > 0$:
  $\Pr(|X-E(X)|>=E(X)\varepsilon) <= 2\, e^{-E(X)\varepsilon\, \varepsilon / 2}$
- Αποδεικνύεται, ότι εάν πάρουμε,
  για $s = 2 \log(1/\varepsilon)$ και $t = 8k\, n^{1-1/k} / \lambda^2$, τότε από το Chernoff η πιθανότητα περισσότερες από s/2 μεταβλητές Y_i να αποκλίνουν περισσότερο από λF_k από το F_k είναι το πολύ ε.

Κάτω φράγμα - Απόδειξη

- Τι κάνουμε εάν δεν ξέρουμε το μήκος της ακολουθίας m από πριν; (σχεδόν πάντα συμβαίνει αυτό ...)
- Τότε ξεκινάμε με m = 1 καθώς διαβάζουμε την ακολουθία και διαλέγουμε σαν $l = a_p$ το a_1. Κάνουμε το r = 1, καθώς το a_1 είναι το μοναδικό στοιχείο στην ακολουθία.
- Μετά με m = 2 βάζουμε για a_p το a_2 με πιθανότητα ½ και κάνουμε τους υπολογισμούς.
- Γενικότερα σε κάθε βήμα m αντικαθιστάμε το a_p με το a_m με πιθανότητα 1/m. Αν γίνει η αντικατάσταση κάνουμε το r = 1. Εάν δε γίνει, κοιτάμε εάν το $a_m = a_p$ και αυξάνουμε το r κατά 1, αλλιώς αφήνουμε το r έως έχει.

Κάτω φράγμα - Απόδειξη

- Στην εισαγωγή όμως είπαμε ότι τα F_k προσεγγίζονται με τυχαίο τρόπο χρησιμοποιώντας το πολύ $O(n^{1-1/k} \log n)$ bits , $k > 0$.
- Εάν $m >= O(n)$, τότε έχει αποδειχθεί ότι ο χώρος που χρειάζεται είναι :
  $O(k\log(1/\varepsilon) n^{1-1/k} (\log n+\log\log m+\log(1/\lambda)) / \lambda^2)$

Κάτω φράγμα - F_2

$O(\log(1/\varepsilon) (\log n+\log m) / \lambda^2)$

- Έστω ακολουθία $A = (a_1,...,a_m)$ με πεδίο τιμών το $\{1,...,n\}$.
  Για κάθε $\lambda > 0$ και κάθε $\varepsilon > 0$ υπάρχει ένα πιθανοτικός αλγόριθμος που με ένα πέρασμα υπολογίζει το F_2.
  Αυτός ο αλγόριθμος αποκλίνει από το F_2 το πολύ από λF_2 με πιθανότητα το πολύ ε.
  Ο χώρος που χρησιμοποιεί ο αλγόριθμος είναι :
  $O(\log(1/\varepsilon) (\log n+\log m) / \lambda^2)$

Κάτω φράγμα - F_0

O(logn)

- Έστω ακολουθία $A = (a_1,...,a_m)$ με πεδίο τιμών το $\{1,...,n\}$.
  Για κάθε c>2 υπάρχει ένα πιθανοτικός αλγόριθμος που με ένα πέρασμα υπολογίζει το F_0.
  Η πιθανότητα ο λόγος μεταξύ Y (Y=ο υπολογισμός του αλγορίθμου) και F να μην είναι μεταξύ 1/c και c είναι το πολύ 2/c.
  Ο χώρος που χρησιμοποιεί ο αλγόριθμος είναι :
  O(logn)

Άνω φράγμα - F_{OO}^{*}

Ω(n)

- Έστω ακολουθία $A = (a_1,...,a_m)$ με πεδίο τιμών το $\{1,...,n\}$ και το μήκος m να είναι το πολύ 2n.
  Για κάθε $\varepsilon < ½$ και κάθε πιθανοτικό αλγόριθμο που με ένα πέρασμα υπολογίζει το F_{OO}^{*} και αποκλίνει από το F_{OO}^{*} πιο πολύ από $F_{OO}^{*}/3$ με πιθανότητα το πολύ ε :
  ο χώρος που χρησιμοποιεί ο αλγόριθμος είναι :
  Ω(n).

Άνω φράγμα - F_k

$\Omega(n^{1-5/k})$

- Έστω ακολουθία $A = (a_1,...,a_m)$ με πεδίο τιμών το $\{1,...,n\}$.
  Για κάθε $k > 5$ και $\gamma < ½$ και κάθε πιθανοτικό αλγόριθμο με $Pr\{ | Z_k - F_k | > 0.1 F_k\} < \gamma$ (Z_k η έξοδος του αλγορίθμου):
  ο χώρος που χρησιμοποιεί ο αλγόριθμος είναι :
  $\Omega(n^{1-5/k})$.