מבני נתונים ואלגוריתמים

ממ״ן 14

07/01/2021

עמית רוט ועידו לובן

שאלה א

נניח שאיבר a שייך למנה. אז מהגדרת הטבלה T לכל j מתקיים $T(h_j(a)) = 1$.

כאשר נגיע לבדיקת השייכות. נקבל שלכל j $T(h_j(a)) = 1$ כלומר התנאי:

$T(h_1(a)) == 1 \,\&\&\, \ldots T(h_k(a)) == 1$ יתפרש כאמת ונקבל ש a שייך למבנה.

לכן הסיכוי שהוא לא יוכרז כשייך הוא 0.

שאלה ב

נבדוק מה ההסתברות שיתקיים $T[h_i(a)] = 0$ עבור a.

ההסתברות שפונקצית גיבוב h_i תגבב איבר a לתא מסויים היא $\frac{1}{m}$ לכן ההסתברות

שהיא לא תגבב את a לתא המסויים היא $\frac{m-1}{m} = 1 - \frac{1}{m}$. לכל $1 \geq i \geq K$

פונקציה h_i יכולה לגבב את a לאותו תא בהסתברות זהה, לכן ההסתברות שלכל

$K \leq i \leq 1$ h_i לא תגבב את a לתא מסויים היא $(\frac{m-1}{m})^K$. לכל $0 \leq j < N$ לכן

נקבל שההסתברות שאף פונקציה h_i לכל $K \geq i \geq 1$ לא תגבב אף איבר a_j לכל

$N > j \geq 0$ לתא מסויים היא $(\frac{m-1}{m})^{KN}$.

דהיינו ההסתברות שיתקיים $T[h_i(a)] = 0$ היא $(\frac{m-1}{m})^{KN}$.

נמצא את ההתסברות שאיבר שלא שייך למבנה יוכרז כשייך.

נשלול את הפסוק " ההסתברות שיתקיים $T[h_i(a)] = 0$ היא $(\frac{m-1}{m})^{KN}$"

ההסתברות שיתקיים $T[h_i(a)] = 1$ היא $1 - (\frac{m-1}{m})^{KN}$.

עבור איבר b לא שייך למבנה נקבל שההתסברות שנקבל $T[h_i(b)] = 1$ היא

$1 - (\frac{m-1}{m})^{KN}$. כדי שיוכרז שייך למבנה, נרצה שלכל $K \geq i \geq 1$ יתקיים

$T[h_i(b)] = 1$, ואז על פי התנאי $T(h_1(a)) == 1 \; \& \& \; \ldots T(h_k(a)) == 1$ נקבל

ש b איבר במנה. נקבל שההתסברות שזה יקרה היא $(1 - (\frac{m-1}{m})^{KN})^K$.

דהיינו ההסתברות שאיבר b שלא שייך למבנה יוכרז כשייך היא $(1 - (\frac{m-1}{m})^{KN})^K$.

שאלה ג

נחשב את ההסתברות שאיבר שלא שייך לבמנה יוכרז כשייך עבור

$k = 13, \quad m = 32 \cdot 10^6, \quad N = 10^6$.

$$(1 - (\frac{m-1}{m})^{KN})^K = (1 - (\frac{32 \cdot 10^6 - 1}{32 \cdot 10^6})^{13 \cdot 10^6})^{13} = 6.4 \cdot 10^{-7}$$

מסמך מלווה עבור מבנה הנתונים BelongTable ושגרותיו

נתאר את אופן פעולת וזמני ריצת האלגוריתמים המופיעים בקבצים main, BelongTable. האלגוריתם שבקובץ HashFunc לקוח מגיטהאב (קישור בתיעוד) ולא ננתח את יעילותו.

האלגוריתם שבקובץ Statistics משמש לחישובי בסטטיסטיקה הדרושים בממן.

__הקובץ BelongTable:__

Constructor: מאתחל את המאפיינים של הטבלה למספרים שנקבעו. רץ בזמן $\Theta\left(k\right)$ כאשר k הוא מספר פונקציות הגיבוב.

add_element: האלגוריתם מוסיף את המחרוזת שהתקבלה לתוך הטבלה בהתאם להוראות, כלומר, מדליק כל בית שנמצא באינדקס שהוא פלט של פונקציית גיבוב מתוך k הפונקציות. רץ בזמן $\Theta\left(k\right)$.

is_belong: האלגוריתם בודק אם המחרוזת שהתקבלה נמצאת בטבלה בהתאם להוראות, כלומר, מפעיל עליה את כל k פונקציות הגיבוב ובודק האם הבתים באינדקס של הפלטים כבר דלוקים. רץ בזמן $\Theta\left(k\right)$.

__הקובץ main:__

מקבל את הפרמטרים הבאים מתוך שורת הפקודה: m,k, input_to_insert, input_to_check.

הנחות על הקלט: m,k מספרים שלמים, שתי הכתובות של הקבצים תקינות והקבצים הם קבצי טקסט, השמורים במחשב, של איברים המופרדים בפסיק אחד בדיוק.

האלגוריתם בונה את הטבלה, מכניס אליה את האיברים הרצויים ובודק אם האיברים שהתקבלו שייכים אליה בעזרת השיטות הנ"ל. רץ בזמן לא ידוע משום שמשתמש באלגוריתם לקריאת המידע המופרד בפסיקים, שאותו לא ננתח.

ניתוח סיבוכיות

כדי לנתח את אחוזי השגיאה של קבלת "חיובי כוזב" (False Positive) כתבנו שגרה בשם false_positive שיוצרת BelongTable ומכניסה אליה N(1-test) איברים. לאחר מכן false_positive בודקת שייכות עבור N*test איברים שבודאות לא שייכים למבנה, וסופרת כמה פעמים is_belong מחזירה true. לאחר מכן false_positive מחזירה את היחס בין מספר התשובות החביוביות למספר האיברים שנבדקו. כלומר את הערך הנסיויי עבור ההסתברות שאיבר b שלא שייך למבנה יוכרז כשייך, כפי שחשבנו בסעיף ב'.

ניתן להשתמש ב false_positive עם מערכי סטרינגים שמיוצרים באופן עצמאי, או לחילופין להשתמש בשגרה false_positive_send_arrays שמייצרת באופן אוטומטי מערכים כך שכל איברי מערך הבחינה לא שייכים למערך האיברים שהוכנס. כדי לפשט את השגרה ולאפשר שימוש יעיל עבור מספרים גדולים, בחרנו להשתמש רק במחרוזות שמורכבות מספרות, לדוגמת "1232546" ו "990007". מפני שפונקציית הגיבוב מגבבת באופן אחיד וכל הרצה ה seed מוגרל באופן רנדומלי, ניתן להתייחס לצורת המחרוזות הללו כמייצגות עבור כל מקרה של מחרוזות.

כדי לאשש את הערך הנסיוי - נכתוב שגרה שמחשבת באופן יבש את הערך של הביטוי שמצאנו בסעיף ג' $(1-(\frac{m-1}{m})^{KN})^K$ וכך נוכל להשוות ולראות את קירוב ביצועי הטבלה לערך התיאורטי.

ניתן לשים לב שהביטוי התיארותי קרוב מאוד לתוצאה בפועל עבור ערכים גדולים.

מסקנות מהתוצאות:

מפני שפונקציית ההסתברות $P(m,N,K)=(1-(\frac{m-1}{m})^{KN})^K$, הינה פונקציה ב 3 משתנים וניתוח ההתנהגות שלה - או חקירתה אינו במסגרת הקורס. נשתמש בכלים יותר נאיבים ופחות מבוססי הוכחות ובסיס מוצק. נסתכל על התוצאות ונסיק מסקנות.

לפי מדידות 1-5, נשים לב שככל ש K גדל (מספר פוקנציות הגיבוב) כך גדלה ההסתברות לקבלת False Positive. אך לעומת זאת - ממדידות 6-10 נראה שככל ש K גדל כך **קטנה** ההסתברות לקבלת False Positive.

כלומר, נשים לב שעבור ערכי m שונים (גודל הטבלה) התלות ב K תשתנה בהתאם,

ניתן להתסכל על זאת כך - עבור ערכי N ו m קיים K אופטימלי עבורו נקבל הסתברות מינימלית של False Positive.

על סמך מדידות 11-13 נראה שככל ש m גדל כך ההסתברות לקבל False Positive גם קטנה, ממצא זה תואם את ההגיון - ככל שיהיו פחות תאים כך פונקציות הגיבוב יגבבו באופן צפוף יותר וסביר להניח שנקבל יותר שגיאות (False Positive).