**מבני נתונים – פרויקט מעשי 2 - קובץ תיעוד**

**פרטים אודות מגישות-**

הגר פייתן

* שם משתמש - hagarleap
* תעודת זהות - 206825176

גל קריאל

* שם משתמש - galkariel
* תעודת זהות - 318459666

**חלק ניסויי/תיאורטי**

**שאלה 3 – השוואה בין quadratic probing ל-alternating quadratic probing**

**שאלה 3 - סעיף 1**

גדלי הקבוצות:

( נבחין כי הגודל של זה בדיוק וכי הגודל של זה בדיוק q )

**שאלה 3 - סעיף 2**

מדדנו עבור quadratic probing את כמות השגיאות הנובעות משגיאה כי הטבלה מלאה ומשגיאה כי מפתח כבר קיים במערך.

ראינו כי אין שגיאות שנובעות ממפתח שקיים כבר במערך, אבל קיבלנו בין 0 ל-3 שגיאות בכל ניסוי שנובעות מכך שהטבלה מלאה.

לעומת זאת עבור alternating quadratic probing לא נזרקו שגיאות כלל וכל האיברים הוכנסו לטבלה.

הסבר לשוני בין התוצאות-

ראינו מהסעיף הקודם כי כאשר אנחנו משתמשים במודל של quadratic probingאז כתוצאה סדרת הבדיקות שלנו עוברת רק על קבוצת מספרים בגודל חצי מהמספר הראשוני q () . כלומר פונקציית ה-hash במודל של quadratic probing לא נותנת לנו את כל האינדקסים בטבלה ולכן ייתכנו סדרות של בדיקות שלא עוברות על כל האינדקסים בטבלה כלומר נקבל שגיאות כי הטבלה מלאה למרות שבפועל יש בה עוד מוקמות להכניס איברים.

לעומת זאת כאשר אנחנו משתמשים במודל של alternating quadratic probing ראינו בסעיף הקודם שסדרת הבדיקות שלנו עוברת על קבוצת מספרים בגודל q () כלומר אין חזרות של מספקים ואנחנו עוברים **על כל** האינדקסים בטבלה– כלומר בהכרח נמצא מקום פנוי אם קיים לטובת הכנסת מפתח חדש לכן לא נקבל שגיאות ונצליח להכניס את כל המפתחות.

**שאלה 3 - סעיף 3**



* מספר a נקרא שארית ריבועית אם כש-x מספר שלם.
* תוכנות חשובות:
  + אם p מספר ראשוני, יש (למעט השארית 0( בדיוק שאריות ריבועיות, ו- שאריות לא ריבועיות
  + אם אז אם a הינו שארית ריבועית, אז -a הינו שארית לא ריבועית!
* בכל שיטה של סדרת הבדיקות להכנסה\חיפוש איבר, אנחנו בסופו של דבר מחשבות אינדקס בעזרת mod m. בניסוי הזה m=6571 והוא מקיים . לכן זה מסביר את התופעה שראינו עם כך שאנחנו מצליחות להגיע רק לחצי מהאינדקסים בQ1 ולכל האינדקסים עם Q2.
  + ב-Q1 אנחנו מחשבים רק מספרים חיוביים לכן ניתן רק להחזיר מספרים שונים
  + ב-Q2 אנחנו מחשבים גם מספרים חיוביים וגם מספרים שליליים, לכן נחזיר גם שאריות ריבועיות וגם שאריות לא ריבועיות: ולכן יש לנו את הפוטנציאל להגיע לכל האינדקסים בטבלה.
* לפי המשפט, תופעה זו קוראת רק לראשוניים שבסדר גודל 3mod4, אז זה תנאי לקיום התופעה.

**שאלה 4 – נשווה בין המימושים השונים ל-open addressing**

**שאלה 4 - סעיף 1**

|  |  |
| --- | --- |
| Class | Running Time |
| LPHashTable | 0.510 |
| QPHashTable | 0.498 |
| AQPHashTable | 0.514 |
| DoubleHashTable | 0.704 |

הסבר מילולי להבדלים בזמני הריצה בין סוגי הטבלאות-

נבחין כי זמני הריצה של DoubleHashTable יותר ארוכים באופן יחסי לעומת שאר המימושים. ושאר המימושים רצים בזמנים מאוד דומים.

ראינו בהרצאה כי הסיכוי להתנגשויות כאשר משתמשים בפונקציות hash מהמשפחות האוניברסליות הנו נמוך ( כאשר m זה גודל טבלת ה-hash). בנוסף אנחנו ממלאים רק את חצי הטבלה, ולכן נעריך כי אפילו אם יש התנגשויות לא יהיו "רצפים" גדולים מידי של איברים בטבלה - שעלולים היו לפגוע בזמני הריצה עבור LPHashTable למשל). כלומר בעיית הרצפים (הצטברות ראשונית/משנית) עבור סעיף זה לא תשפיע הרבה על זמני הריצה.

בנוסף לכך, נרצה לציין כי ציפינו ש- LPHashTable ייקח פחות זמן מ- QPHashTableu ומ-AQPHashTable כיוון שראינו במבוא מורחב כי חיבור מספר בודד זה פעולה זולה יותר מהעלאה בחזקה. אבל נראה כי אין משמעות גדולה בזמני הריצה לפעולות חישוב אלה (חיבור והעלאה בחזקה).

עבור LPHashTable, QPHashTable ו- AQPHashTable מתבצעת אותה כמות של פעולות מודולו וראינו כי זמני הריצה שלהם דומים.

לעומת זאת ב- DoubleHashTable יש יותר פעולות מודולו (אנחנו מפעילים שתי פונקציות hash עבור כל בדיקה להכנסת איבר) ולכן אנו מעריכות כי השוני בזמני הריצה נובע מפעולת המודולו.

**שאלה 4 - סעיף 2**

|  |  |
| --- | --- |
| Class | Running Time |
| LPHashTable | 5.551 |
| AQPHashTable | 2.776 |
| DoubleHashTable | 3.836 |

ההסבר לכך שאין לבצע סעיף זה עבור הנה - כפי שראינו בשאלה הקודמת, פונקציית ה-hash במודל של quadratic probing לא נותנת לנו את כל האינדקסים בטבלה, ובפרט נותנת לנו רק חצי מהם. לכן הכנסה של כמות איברים שגדולה מחצי מהטבלה עלולה לגרום לשגיאות "table is full" כיוון שסדרת הבדיקות לטובת הכנסה תהיה רק על חצי מהטבלה.

בניסוי זה אנו מכניסים יותר מחצי איברים של הטבלה ולכן ייתכן ויהיו שגיאות (בשאלה זו לא אמורים להיזרק חריגים) ולכן לא נבצע בסעיף את המימוש עבור .

הסבר להבדל בביצועים (זמני הריצה) לעומת הסעיף הקודם:

באופן כללי, נשים לב כי כמות ההתנגשויות שנצפה בניסוי זה משתנה באופן הבא:

תוחלת של ניסוי 1:

תוחלת של ניסוי 2:

כלומר אם לא הייתה בעיית הצטברות וזמן הריצה היה רק תלוי בחישוב יחיד של אינדקס חדש בכל מקרה של התנגשות, אז היינו מצפות לזמן ריצה לגדול פי 3.6 ().

LPHashTable- זמני הריצה של מודל זה הנם הארוכים ביותר.

זמני הריצה גדלו בערך פי 10 לעומת הניסוי הקודם וזו הגדילה הכי גדולה לעומת AQP ו- DoubleHash.

אנו מעריכות כי הדבר נובע מבעיית ההצטברות הראשונית – נוצרים רצפים ארוכים של תאים תפוסים. סדרות הבדיקה להכניס איבר שיש רצפים ארוכים יהיו יקרות (הרבה פעולות מודלו).

AQPHashTable-

זמני הריצה גדלו בערך פי 5 לעומת הניסוי הקודם.

אנו מעריכות כי הדבר נובע מבעיית ההצטברות המשנית – זוהי צורה מתונה יותר של הצטברות.

יפה לראות כי בעיית ההצטברות המשנית, באמת מתונה יותר מבעיית ההצטברות הראשונית (זמני ריצה שגדלו פי 5 לעומת זמני ריצה שגדלו פי 10 ב-LP).

DoubleHashTable-

זמני הריצה גדלו בערך פי 5 לעומת הניסוי הקודם.

בשיטה זו תיאורטית לא אמורים להיות רצפים ארוכים של תאים תפוסים שכן סדרת הבדיקות עבור הכנסה מחושבת על ידי שתי פונקציות hash שונות.

תוצאות אלה מפתיעות אותנו לאור האמור לעיל, שכן היינו מצפות שזמני הריצה בניסוי זה יגדלו פחות מפי 5.

הסבר אפשרי לכך שזמני הריצה גדלו פי 5 בכל זאת יהיה, שכאשר הטבלה לקראת התמלאות יש יותר התנגשויות מכיוון שהתנגשויות ראשוניות יובילו להתנגשויות נוספות. הדבר יגרור יותר חישובים יקרים.

אפשר לראות אימפירית ששיטת הdoubleHash מאכזבת. הרעיות מאחוריה היא להימנע מרצפים גדולים, אבל בסופו של דבר היא איטית יותר מ-AQP בשני הניסויים.

**שאלה 5 – השפעת מחיקת איברים ב-open addressing**

|  |  |
| --- | --- |
| Iterations | Running Time |
| First 3 iterations | 4.513 |
| Last 3 iterations | 19.45 |

נשווה את זמן ביצוע 3 האיטרציות הראשונות לזמן ביצוע 3 האיטרציות האחרונות.

נבחין כי קיים הבדל משמעותי בין זמני הריצה, ב-3 האיטרציות הראשונות זמני הריצה קצרים משמעותית מ-3 האיטרציות האחרונות.

אנו מעריכות כי הסיבה להבדל הנה שאנו לא מאפסות בכל איטרציה מחדש את המבנה. לכן כשעושים יותר איטרציות יש יותר תאים בטבלה שמסומנים כ-deleted (כיוון שאנחנו מוחקים איברים את כל האיברים שהכנסנו למבנה בכל איטרציה) ולא כ-null . בפונקציית ה-Insert שלנו אנחנו מכניסים איבר בתא שמסומן deleted רק אחרי שעברנו על כל הטבלה לוודא שאין מפתח זהה לזה שאנחנו רוצים להכניס (כיוון שיתכן שהמפתח נמצא בהמשך הטבלה), לעומת זאת עבור תא שמכיל- null אנחנו מכניסים את האיבר החדש ישירות לטבלה שלנו ומסיימים. לכן ככל שיהיו יותר תאים שמסומנים deleted נקבל זמני ריצה ארוכים יותר עבור ההכנסות ולכן זמני ריצה ארוכים יותר עבור הניסוי.