בס"ד

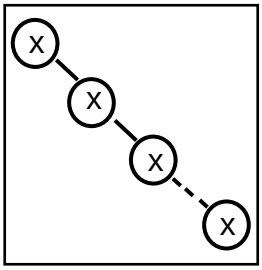
מפתחות זהים בעצי חיפוש בינאריים // יעל לרפלד ונחמה כהן

מחקר זה נעשה במטרה למצוא את הצורה הנכונה לטיפול בבעיית המפתחות הזהים בעץ חיפוש בינארי. זהו עץ שבו בכל צומת, בעת הכנסת איבר, האיבר הקטן נכנס משמאל לצומת ואיבר הגדול נכנס מימין לצומת. וכך נוצר מצב שבהכנסת איבר בעל מפתח זהה לצומת יש אפשרות להכניס את האיבר משמאל לצומת או מימין. בכדי להגיע לסיבוכיות הזמן הטובה ביותר אנו רוצים שהאיבר יכנס לתת העץ הקטן ביותר (על מנת לחסוך השוואות). בכדי לפתור בעיה זו מימשנו מספר גרסאות שונות של הכנסה לעץ חיפוש בינארי עם הוספות שונות וחקרנו כדי לראות האם ההשערות שהעלנו תואמות לתוצאות, ולהגיע לבחירה של הפתרון הטוב ביותר לבעיה - הגרסה עם מספר ההשוואות הקטן ביותר.

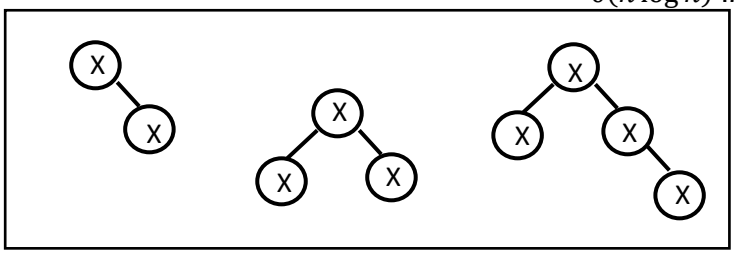
לאחר בדיקת הגרסאות השונות בטווחי מספרים ובגודלי קלט שונים, ראינו כי ככל שטווח המספרים קטן ומספר הקלטים גדול, ז"א שיש מקרים רבים של מפתחות זהים, השוני במספר ההשוואות בין הגרסאות היה משמעותי. כמו כן ראינו כי כאשר טווח המספרים היה גדול ומספר הקלטים קטן ז"א שלא היו הרבה מקרים של מפתחות זהים לא ניכר השוני בין הגרסאות מהסיבה שההבדל במספר ההשוואות בין הגרסאות הוא רק במקרים בהם יש מפתחות זהים.

כעת נסביר את הגרסאות השונות בהינתן n מפתחות זהים, וננתח את סיבוכיות זמן הריצה שלהם:

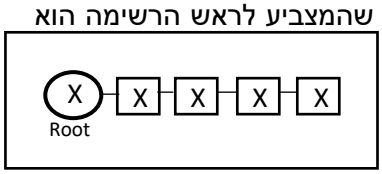
1. **עץ מגרסה 0**: כאשר נכניס n איברים בעלי מפתחות זהים לעץ חיפוש בינארי, המפתחות יפנו כבררת מחדל לצד ימין. כל איבר שנכנס יעבור על כל גובה העץ ויכנס לסוף העץ.
סיבוכיות זמן פעולת ההכנסה היא: $\theta(h)$ (h גובה העץ)
במקרה זה בכל איטרציה גובה העץ הוא מספר האיברים שנכנסו עד כה.
סה"כ סיבוכיות זמן הכנסה בודדת במקרה הגרוע:$\theta(n)$
(n מספר האיברים בעץ)
מכניסים n איברים, ז"א n איברים * $\theta(n)$ = n^2, ולכן זמן הריצה הוא: $\theta(n^2)$



2. **עץ מגרסה a**: במקרה זה קיים דגל בוליאני dir[x] המציין את כיון ההכנסה (direction) במקרה והמפתחות שווים, ז"א המפתחות הזהים נכנסים באופן שווה לשמאל ולימין כך שנוצר עץ (כמעט) מלא, וכך העץ נשאר מאוזן. במקרה זה בכל איטרציה גובה העץ הוא:$\theta(log\ \mathrm{h})$ (h גובה העץ)
ולכן סיבוכיות זמן הכנסה בודדת במקרה הגרוע היא :$\theta(log\ \mathrm{n})$ (n מספר האיברים בעץ)
כאשר יש n איברים, כל הכנסה עוברת על כל הרמות בעץ, ולכן מספר הפעולות המתבצעות הוא $\mathrm{n} \cdot log\mathrm{n} = nlogn$
סה"כ זמן הריצה הוא: $\theta(n \log n)$



עץ מגרסה b: האיבר הראשון שנכנס לעץ יהיה ה- root. מהאיבר השני והלאה, כל איבר זהה שיכנס לעץ יתווסף לראש הרשימה המקושרת שנמצא בצומת בעלת המפתח השווה לו (במקרה זה ה-root) . במקרה זה תיוצר רשימה שהמצביע לראש הרשימה הוא ה root.ז"א שבכל הכנסה של איבר זהה תתבצע פעולה קבועה. ולכן סיבוכיות זמן ההכנסה זהה לסיבוכיות ההכנסה לראש רשימה שהיא פעולה



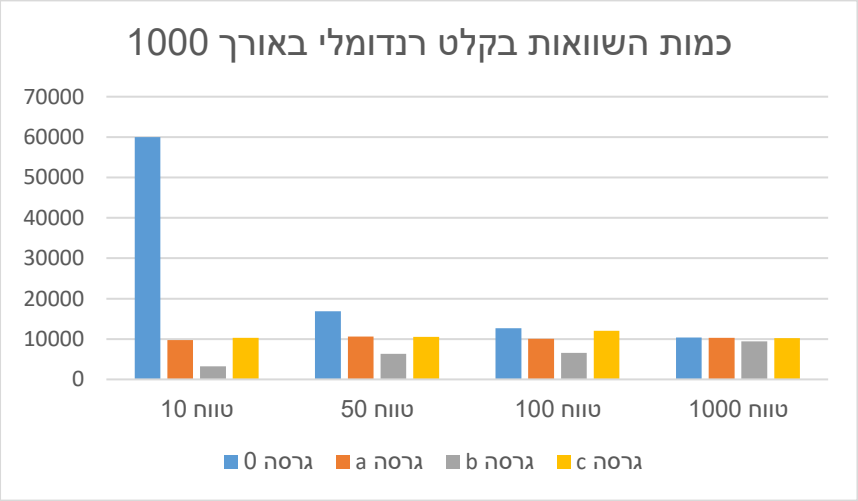
קבועה שלוקחת : $\theta(1)$
סה"כ מתבצעות n הכנסות לרשימה ולכן הסיבוכיות היא: $\theta(n)$

עץ מגרסה c: בגרסה זו בחירת הכיוון היא רנדומלית ולכן נחלק למקרים:

המקרה הממוצע : בכל איטרציה יש 2 אפשרויות להגריל, ימין או שמאל, ההסתברות של כל מקרה היא 0.5. ז"א שמחצית המקרים ימין והשאר שמאל ולכן מקרה זה דומה למקרה a, ובו נבחר פעם ימין ופעם שמאל, וגם במקרה זה העץ יהיה יחסית מאוזן וגובהו יהיה log n. ולכן הסיבוכיות במקרה הממוצע היא: $\theta(n \log n)$

המקרה הגרוע: בכל ההכנסות הוגרל אותו כיוון (הכול לימין / שמאל). מקרה זה דומה למקרה 0 והסיבוכיות היא: $\theta(n^2)$

תוצאות הרצת התוכנית*:



כעת לאחר שיש בידינו את תוצאות ההרצה בטווחים שונים, ננסה להבין מהיכן נובעים ההפרשים בין הגרסאות השונות ובכל גרסה בטווח שונה.

בין הגרסאות, הגרסה היעילה ביותר במקרה של איברים רבים היא גרסה b לאחריה a, c וגרסה 0 היא הגרועה ביותר. ככל שהטווח גדל, ההפרשים קטנים ומסתמנת מגמה של שוויון בין כמות ההשוואות בגרסאות השונות, ובחלק מהמקרים גרסאות a או c יעילות יותר מגרסה b, דבר זה נובע מכיוון שהשוני בין הגרסאות הוא בצורת הכנסת המפתחות הזהים וכשהטווח גדל מספר המפתחות השווים קטן. הוכחה לכן ניתן לראות בהכנסת איברים עם מפתחות שונים, כמות ההשוואות זהה בכל הגרסאות. עם זאת ברוב מוחלט של ההכנסות האפשריות גרסה 0 מבצעת את המספר הגבוה ביותר של ההשוואות.

כפי שניתן לראות מהתרשים, כשהטווח קטן גרסה 0 מבצעת כמות גדולה מאד של השוואות, אך ככל שהטווח גדל מסתמנת ירידה בכמות ההשוואות. וזאת מכיוון שהטווח קטן ישנם איברים רבים שווים וקצת איברים שונים, ז"א שלכל node יש הרבה צאצאים בתת העץ הימני (נוצרת 'שרשרת'), ולכן עומק העץ גדול וממילא בכדי להגיע לעלה במקרה של שוויון יש לעבור השוואות רבות, שהטווח גדל יש יותר איברים שונים וישנם איברם שפונים לשמאל, וההסתברות שהעץ יהיה מאוזן יותר גדלה, ולכן עומק העץ קטן וכתוצאה מכך כמות ההשוואות קטנה. תוצאה זו נצפתה בחלק א' שהרי כפי הידוע סיבוכיות הכנסת n איברים לעץ חיפוש בינארי היא $\theta(n * h)$, כש-h מבטא את גובה העץ שהוא בין n ל log n. וסיבוכיות הזמן שראינו בסעיף א'- שכל האיברים זהים שהיא $\theta(n^2)$ היא הסיבוכיות המקסימלית, וככל שהעץ נהיה מאוזן יותר זמן הריצה מתקרב למקרה הטוב.

בגרסה b אנו רואים מגמה הפוכה, ככל שהטווח גדל כמות ההשוואות גדלה, מכיוון שהטווח קטן כמות האיברים הזהים גדולה ולכן יש פחות node-ים, זה גורם לרשימות ארוכות אך מבחינת ההכנסה זה לא משנה כי מכניסים לראש הרשימה (זוהי פעולה קבועה) ולכן כמות ההשוואות קטנה מאד, ככל שהטווח גדל יש פחות איברים זהים ולכן יש יותר node-ים. ובהכנסת איבר, בכדי למצוא את מיקומו נדרשות השוואות נוספות. גם בגרסה זו התוצאה מתאימה לסעיף א' שבו ראינו שבמקרה של איברים זהים סיבוכיות הכנסת n איברים היא $\theta(n)$ וסיבוכיות זו קטנה מסיבוכיות המקרה הטוב ביותר של הכנסה בגרסה הבסיסית, ולכן ברור שככל שהטווח גדל ומספר ההשוואות נהיה שווה לגרסה הבסיסית, שבה הסיבוכיות במקרה הטוב ביותר היא, כפי הנכתב לעיל, $\theta(n\log n)$.

גרסאות a ו-c הן דומות, השוני ביניהם הוא בצורת בחירת הכיוון, בגרסה a הבחירה היא פעם לימין ופעם לשמאל, ובגרסה c הבחירה אקראית. ואכן אנו רואים תוצאה כמעט זהה, ההפרש ביניהם נובע מהשוני ביניהם- בהגרלת הכיוונים בגרסה c. כך שאפילו באותו טווח פעם a יעילה יותר ופעם c. בגרסאות אלו אין השפעה לגודל הטווח, כמות ההשוואות בטווחים השונים כמעט זהה ואין כיוון ברור של עליה או ירידה בכמות ההשוואות. כמו בסעיפים הקודמים, גם בסעיף זה אנו רואים שהתוצאה מסתדרת עם סעיף א' שהרי ראינו שם שסיבוכיות הכנסת n איברים זהים היא $\theta(n\log n)$ שזו סיבוכיות הזמן במקרה הטוב בהכנסת n איברים לעץ בגרסה הבסיסית, ז"א שבין אם הטווח גדול ובין אם הטווח קטן העץ יהיה יחסית מאוזן ואין משמעות מבחינת זמן הריצה אם קיימים קלטים זהים.

❗ הערה חשובה

כל הנכתב במאמר זה אינו מתייחס לפעולות הנוספות המתבצעות על עץ חיפוש בינארי אלא ליעילות ההכנסה במקרה של איברים זהים. יתכן שגרסה שיעילה להכנסה תצרוך זמן ריצה ארוך לפעולה אחרת.

---

*בכדי לבדוק את נכונות ההשערות נבדקו מספר ההשוואות במקרים הבאים:

- קלט רנדומלי באורך 1000 בטווחים 10, 50, 100, 1000 (מס' הרצות לכל טווח).
- קלט ידני בעל ערכים זהים באורך 10, 50
- קלט ידני בעל ערכים שונים באורך 10,100