הרצאה 8

איתיווייסמן

20 במאי 2019

1 תזכורת - שיבוץ אינטרוולים

טענה 1. בסיס הצעד ה-k של האלגוריתם, קיים פתרון אופטימלי X^*כך ש:

$$\forall 1 \leq j \leq k : I_j \in X^* \iff I_j \in X$$

הוכחה. באינדוקציה על k.
בסיס: עבור $k = 1$. עשינו בשיעור הקודם - על ידי החלפה.
צעד: הנחת האינדוקציה אומרת שיש פתרון אופטימלי X^* כך ש-$\forall 1 \leq j \leq k : I_j \in X^* \iff I_j \in X$.
נסתכל על האינטרוול ה-$k+1$.
נחלק לשני מקרים לפי האם האלגוריתם בחר את I_{k+1}!:

1. האלגוריתם בחר את I_{k+1} כלומר $I_{k+1} \in X$. אם $I_{k+1} \in X^*$ אז סיימנו. אחרת, $I_{k+1} \notin X^*$ ולכן בהכרח קיים אינטרוול ב-X^*שנחתך עם I_{k+1}.
נבחר את האינטרוול I_r בעל האינדקס המינימלי מ-X^*שנחתך עם I_{k+1}.
ניזכר שאנו ממיינים לפי זמן הסיום ולכן בהכרח מתקיים $r \geq k+2$ כי אם $r \leq k$ אז בגלל שהאינטרוול נמצא ב-X^*נקבל ש-$I_r \in X$ אך I_r ו-I_{k+1}נחתכים בסתירה לכך שהאלגוריתם בחר את I_{k+1}.
נשים לב ש-I_r הוא היחיד ב-X^*שנחתך עם I_{k+1}, אחרת X^*לא פיזיבלי.
נסתכל על: $X^* \setminus \{I_r\} \cup \{I_{k+1}\}$. נשים לב שזהו פתרון אופטימלי מבחינת הגודל (כי לא שינינו את הגודל). נרצה להראות שהוא פיזיבלי, כלומר זהו פתרון חוקי.
נראה כי I_{k+1} לא נחתך עם אף אינטרוול ב-$X^* \setminus \{I_r\}$.
   - האם I_{k+1} נחתך עם משימות עם אינדקס $r+1, r+2, \ldots, n$? לא, משום שהם לא יכולים להיחתך עם I_r וI_{k+1} מסתיים לפניו.
   - האם I_{k+1} נחתך עם משימות עם אינדקס $k+2, k+3, \ldots, r-1$? לא. משום שדאגנו לבחור את r בתור המינימלי.
   - האם I_{k+1} נחתך עם משימות עם אינדקסים $1, 2, \ldots, k$? לא. משום שבהנחת האינדוקציה הנחנו שX^*מסכים עם X והוא אופטימלי.
2. האלגוריתם לא בחר את I_{k+1} כלומר $I_{k+1} \notin X$. לפי הנחת האינדוקציה, X^* אינו יכול להכיל את האינטרוול ה-$k+1$ כי אז ב-X^*היו 2 אינטרוולים שנחתכים.

■

1.1 מיקסום רווח

כעת בנוסף לנתוני הבעיה מקודם, נתון כי לכל אינטרוול I_j קיים רווח $p_j \geq 0$.
אנו רוצים למצוא פתרון חוקי שגם ממקסם את סך הרווח מכל האינטרוולים?
מעניין לגלות כי כנראה לא קיים אלגוריתם חמדן שפותר בעיה זו.

2 קידוד Huffman

בגדול, בהינתן קובץ המורכב מאוסף תווים רוצים למצוא איך "לקודד" אותו כך שאורך הקובץ המקודד יהיה כמה שיותר קצר. בקידוד, כל תו בקובץ יהפוך למחרוזת בינארית (מאורך כלשהו).

הגדרה 2. מילת קוד - מחרוזת $w = w_1 w_2 \ldots w_n$ כאשר $w_i \in \{0,1\}$. האורך של מילת קוד יסומן $l(w)$.

הגדרה 3. קוד הוא אוסף של מילות קוד

דוגמה 4. $C = \{c1, c2, c3\}$ וכן $c_1 = 01, c_2 = 0, c_3 = 0$.

הגדרה 5. פעולת הקידוד נֶעְשת ע"י החלפת כל תו במילת הקוד שמתאימה לו (פונקציה).

איך מפענחים קידוד? זה אפשרי רק אם הקוד שלנו מאופיין כך שיש רק דרך אחת לפענחו.
נתמקד בקודים **חסרי רישאות** בהם אין מילת קוד שהיא רישא של מילה אחרת. עבור קודים חסרי רישאות הפענוח פשוט ונעשה על ידי קריאה של הקובץ המקודד.

נתון n תווים, ולתו ה-i נתונה תדירות f_i.

מטרה למצוא קוד שבו התו ה-i מותאם למילת הקוד c_i, שיש דרך אחת לפענח אותו, שמביא למינימום: $\Sigma_{i=1}^{n} l(c) \cdot f_i$

טענה 6. ללא הוכחה. מבין הפתרונות האופטימליים, קיים לפחות אחד שהוא קוד חסר רישאות.

שאלה 7. האם יש דרך נוחה לייצג קוד חסר רישאות?

ניתן לייצר קוד חסר רישאות על ידי עץ בינארי (למשל שמאלה זה 1 וימינה זה 0) ועלי העץ יהיו מילות הקוד, כלומר מבנה של trie.

טענה 8. קוד אופטימלי מיוצג ע"י עץ מלא.

2.1 איך נמצא עץ אופטימלי?

רעיון - אם שני התווים עם התדירות הנמוכה ביותר יהיו עלים אחים עמוקים ביותר בעץ. נאחד אותם לתו יחיד ממשקל $f_i + f_j$ ונפתור אותה בעיה אך כעת על קלט של $n-1$ תווים.

אלגוריתם 9. האלגוריתם של Huffman (1952)

1. נמיין את התווים הנתונים לפי התדירות: $f_1 \geq f_2 \geq \ldots \geq f_{n-1} \geq f_n$.
2. נוציא את התווים ה-n וה-$n-1$ ובמקום נכניס תו "מלאכותי" בעל תדירות $f_{n-1} + f_n$.
3. נפתור (רקורסיבית) את הבעיה עבור הקלט המצומצם. וקיבלנו עץ T'.

4. ב-T' נוסיף לעלה שמייצג את התו המלאכותי שיצרנו, שני ילדים ישירים, אחד עבור התו ה-n והשני עבור התו ה-$n-1$.

5. החזר את T.

6. תנאי העצירה של האלגוריתם הרקורסיבי: אם $n=2$ מחזירים עץ טריוואלי מתאים להם: $f_1 \leftarrow r \rightarrow f_2$.

טענה 10. יהיו x ו-y 2 מילות הקוד בעלות התדירות הנמוכה ביותר. אז קיים פתרון אופטימלי ש-x ו-y עלים אחים נמוכים ביותר.

הוכחה. יהי T עץ אופטימלי. בלי הגבלת הכלליות: $f_x \leq f_y$ וגם $f_a \leq f_b$. ניצור עץ חדש בו נחליף בין x ו-a ובין y ו-b ובכך נקבע כי הם אחים.

מה השינוי בערך של T? $l(x) f_x + l(y) f_y + l(a) f_a + l(b) f_b$.

מתקיים $(l(x) \leq l(a))$ וגם $f_y \leq f_b$. באופן דומה $l(y) \leq l(b)$ וגם $f_y \leq f_b$.

ולכן אם נבצע את ההחלפה הערך של T לא יכול לגדול ולא נפגעת האופטימליות. ■

מסקנה 11. נסמן ב-x וב-y את שני התווים בעלי התדירות הנמוכה ביותר וב-g את התו המלאכותי שהוספנו במקום שניהם.

נסמן ב-$cost(T)$ את ערך העץ T.

נסמן ב-T' את העץ המתקבל מהקריאה הרקורסיבית. מתקיים:

$$\begin{aligned} cost(T) &= cost(T') - l_T(g) f_g + (l(g)+1)(f_x+f_y) \\ &= cost(T') + f_x + f_y \end{aligned}$$

בעזרת טענה 10 והמסקנה אפשר להוכיח נכונות האלגוריתם: ניתן להניח בשלילה.