מבני נתונים ואלגוריתמים - הרצאה 12

6 בדצמבר 2011

דחיסה

תחילה נעסוק בדחיסה שהיא Lossless - לא מאבדים מידע בדחיסה.

טענה

כל דחיסה מגדילה חלק מהקבצים.

הוכחה

נניח שקיימת דחיסה שמטקינה או משאירה שווה כל קובץ:

$$f:\ L\rightarrow K$$

ותמיד $k\leq L$.
ניקח את m להיות גודל הקובץ המינימלי שעובר דחיסה ממש.
קיים קובץ בגודל m שעובר לגודל m_1 כאשר $m_1<m$.
יש 2^{m_1} קבצים אפשריים בגודל m_1. לאחר הדחסיה יש לפחות $2^{m_1}+1$ קבצים אפשריים בגודל m_1.
לפי שובך היונים, לפחות קובץ אחד מופיע פעמיים בקבצים בגודל m_1. לכן קיימים שני קבצים שונים שנדחסים לאותו קובץ, בניגוד לעובדה שהדחיסה ללא הפסד.

אי שוויון גיבס

מכאן והלאה - $p_i,\,q_i$ הן התפלגויות:

$$\begin{aligned} q_i,\,p_i &> 0\\ \sum_i p_i &= 1\\ \sum_i q_i &= 1 \end{aligned}$$

אי שוויון גיבס אומר: לכל $q,\,p$:

$$-\sum_i p_i\log\left(p_i\right)\leq-\sum_i p_i\log\left(q_i\right)$$

נשים לב שצד שמאל הוא אנטרופיה:

$$H=-\sum_i p_i\log\left(p_i\right)$$

אפשר לכתוב אחרת:

$$\begin{aligned} -\sum_i p_i\left[\log\left(p_i\right)-\log\left(q_i\right)\right] &\leq 0\\ -\sum_i p_i\log\left(\frac{p_i}{q_i}\right) &\leq 0 \end{aligned}$$

הוכחה

$$\begin{aligned}-\sum_i p_i\log\left(\frac{p_i}{q_i}\right) &\leq \sum_i p_i\log\left(\frac{q_i}{p_i}\right)\\ &\leq \sum_i p_i\left(\frac{q_i}{p_i}-1\right)\\ &= \sum_i q_i-\sum_i p_i=0\end{aligned}$$

השתמשנו בכך ש$\ln x\leq x-1$:

$$\begin{aligned}f(x) &= x-1-\ln x\\ f^{'}(x) &= 1-\frac{1}{x}=0\\ x &= 1\\ f(1) &= 0\\ f(x) &\geq 0\\ \ln x &\leq x-1\end{aligned}$$

לכן אי השוויון נכון.
אם נגדיר

$$D(p,\,q)=\sum_i p_i\log\left(\frac{p_i}{q_i}\right)\geq 0$$

ומתקיים:

$$D(p,\,q)=0\iff p=q$$

לכן זו סמי-מטריקה (זו לא מטריקה כי אין סימטריות).
$D(p,\,q)$ נקרא מרחק קולבק-לייבלר.
יש גם מקבילה רציפה למרחק זה:

$$D(p,\,q)=\int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

אי שוויון Kraft - קוד רישא

קוד רישא הוא קוד שבו אף מילה אינה רישא של מילה אחרת, שבאלפבית שלו יש D אותיות.
בהינתן קוד רישא יש לכל מילה אורך ℓ_i.
אי שוויון Kraft הוא:

$$\sum_i D^{-\ell_i}\leq 1$$

הוכחה

נכתוב את קוד הרישא בעץ - מילים הן עלים.
נסמן בk את עומק העץ.
נמלא את העץ עד עומק k, כלומר נוסיף לעלים בנים עד עומק k, ו"נסמן" את הבנים שהוספנו ושלא באמת שייכים לעץ המקורי.
מתחת לכל מילה אורך ℓ_i יש $D^{k-\ell_i}$ עלים.
יש בסה"כ D^k עלים בעץ (כי זה עץ מלא בעומק k ולכל צומת יש D בנים).
נעבור על כל המילים וניקח את כל העלים שיש מתחת אליהן בעץ, וזה בהכרח לכל היותר מס' העלים בעץ:

$$\sum_i D^{k-\ell_i}\leq D^k$$

נחלק ב-D^k ונקבל:

$$\sum_i D^{-\ell_i} \le 1$$

מ.ש.ל.

טענה

יהי ℓ_i אורך המילים בקוד רישא, אז:

$$\sum_i p_i \ell_i \ge H$$

כאשר H היא האנטרופיה.

הוכחה

נגדיר

$$\begin{aligned} z &= \sum_i 2^{-\ell_i} \le 1 \\ q_i &= \frac{2^{-\ell_i}}{z} \end{aligned}$$

נשים לב ש-q_i היא התפלגות:

$$\begin{aligned} q_i &\ge 0 \\ \sum_i q_i &= \frac{\sum_i 2^{-\ell_i}}{z} = \frac{z}{z} = 1 \end{aligned}$$

נעביר אגפים ונקבל:

$$\begin{aligned} 2^{-\ell_i} &= z \cdot q_i \\ \ell_i &= -\log_2 z - \log_2 q_i \end{aligned}$$

אזי

$$\begin{aligned} \sum_i p_i \ell_i &= -\sum_i p_i \log_2 z - \sum_i p_i \log_2 q_i \\ &= -\log_2 z - \sum_i p_i \log_2 q_i \\ &\ge 0 - \sum_i p_i \log_2 p_i = H \end{aligned}$$

(המעבר לשורה האחרונה הוא לפי Kraft ו-Gibbs.)
לכן בעצם אלגוריתם הדחיסה הכי טוב בקוד רישא יידחס עד לגבול H.
נשים לב שאם אנו רוצים שהאלגוריתם יפעל כך, אנו צריכים שיתקיים:

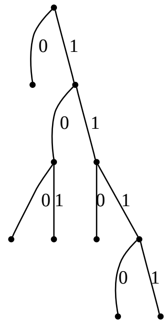
$$\begin{aligned} \log_2(z) &= 0 \\ p_i &\sim q_i \end{aligned}$$

כלומר התנאים שנרצה שיתקיימו הם:

$$\begin{aligned} \log_2(z) &= 0 \\ p_i &\sim \frac{2^{-\ell_i}}{z} \end{aligned}$$

קוד הופמן

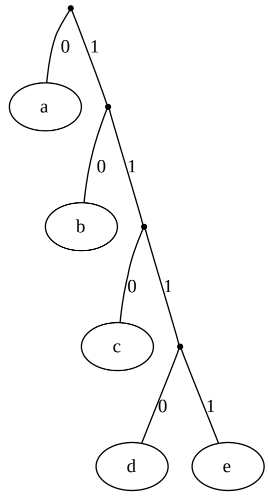
נשים לב שאם יש לנו עץ שלם (כלומר לא קיים צומת עם בן יחיד), יתקיים בהכרח $\log_2 z = 0$, כי $\sum 2^{-\ell_i} = 1$:



קוד הופמן תמיד יתן לנו עץ שלם, ולכן התנאי הראשון תמיד יתקיים בו.
ניקח לדוגמה שאלה המילים והשכיחויות:

מילה	שכיחות
a	$\frac{1}{2}$
b	$\frac{1}{4}$
c	$\frac{1}{8}$
d	$\frac{1}{16}$
e	$\frac{1}{16}$

העץ שנקבל מקוד הופמן הוא:



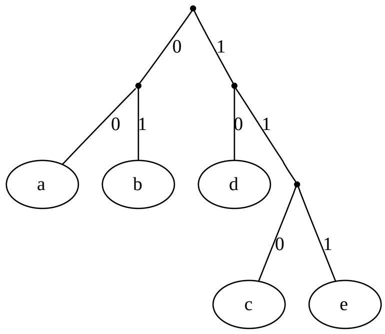
אז נקבל:

מילה	שכיחות	המילה בקוד הופמן	ℓ_i
a	$\frac{1}{2}$	0	1
b	$\frac{1}{4}$	10	2
c	$\frac{1}{8}$	110	3
d	$\frac{1}{16}$	1110	4
e	$\frac{1}{16}$	1111	4

אזי במקרה הזה $z = 2^{-1} + 2^{-2} + 2^{-3} + 2 \cdot 2^{-4} = 1$ וזה יקתיים בכל מקרה בקוד הופמן.
אך התנאי השני לא בהכרח יתקיים, למרות שפה הוא כן מתקיים.
אם השכיחויות לא יהיו חזקות של 2 או משהו קרוב אליהן, יכול להתקבל משהו שונה:

מילה	שכיחות
a	$\frac{1}{3}$
b	$\frac{1}{3}$
c	$\frac{1}{9}$
d	$\frac{2}{9}$
e	0

עץ הופמן שנקבל:



ואז:

מילה	שכיחות	קוד הופמן	אורך
a	$\frac{1}{3}$	00	2
b	$\frac{1}{3}$	01	2
c	$\frac{1}{9}$	110	3
d	$\frac{2}{9}$	10	2
e	0	111	3

ובמקרה זה התנאי השני לא מתקיים.