Haladó informatikai algoritmusok

Mintaillesztés

Rabin-Karp algoritmus

Fogalmak és jelölések

Az egyszerűség kedvéért tegyük fel, hogy a $\Sigma = \{0, 1, 2, \ldots, 9\}$, azaz a jelek a decimális számjegyek.
Ekkor egy k hosszúságú jelsorozatot tekinthetünk egy k jegyű decimális számnak.

Egy adott $P[1..m]$ minta decimális értékét jelöljük p-vel.
Hasonlóan, egy $T[1..n]$ szöveg $T[s+1..s+m]$ m hosszúságú részsorozatának decimális értékét jelöljük t_s-sel, bármely $s = 0, 1, \ldots, n-m$ esetén.
Ekkor $t_s = p$ akkor és csak akkor, ha $T[s+1..s+m] = P[1..m]$, azaz s érvényes eltolási érték akkor és csak akkor, ha $t_s = p$.

Számolás

Ha p értékét $\Theta(m)$ és az összes t_s értékét együttesen $\Theta(n-m+1)$ idő alatt ki tudjuk számolni, akkor az összes s érvényes eltolás meghatározható

$$\Theta(m) + \Theta(n-m+1) = \Theta(n)$$

idő alatt, hiszen p értékét kell összehasonlítani minden t_s értékkel.

p értéke kiszámítható $\Theta(m)$ időben:

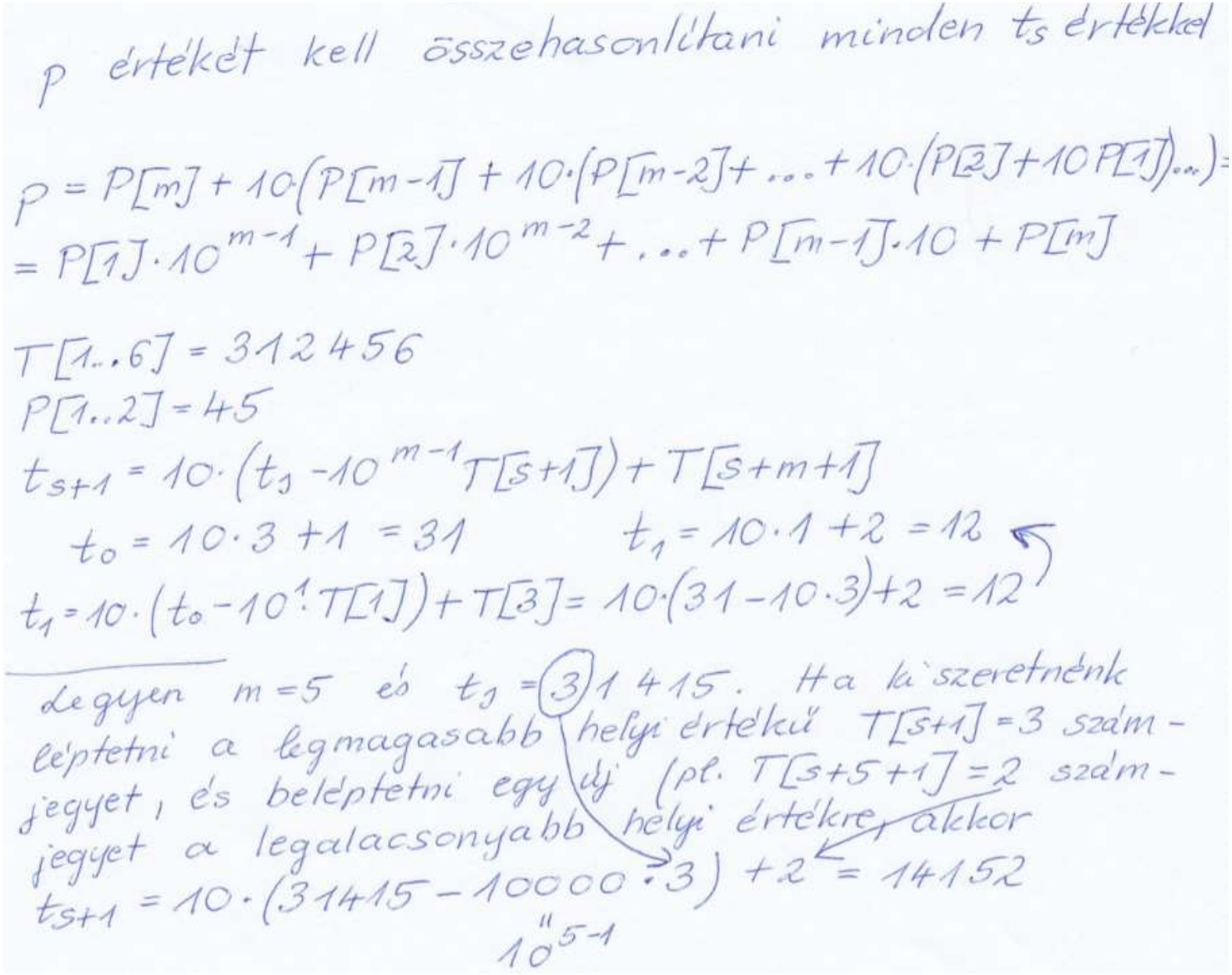
$$p = P[m] + 10(P[m-1] + 10(P[m-2] + \ldots + 10(P[2] + 10P[1])\ldots))$$

t_0 éréke hasonlóan meghatározható $T[1..m]$ segítségével $\Theta(m)$ idő alatt.

Könnyen látható, hogy a további $t_1, t_2, \ldots, t_{n-m}$ értékek $\Theta(n-m)$ idő alatt kiszámíthatóak, ha észrevesszük, hogy t_{s+1} konstans időben megkapható t_s-ből, mivel

$$t_{s+1} = 10(t_s - 10^{m-1}T[s+1]) + T[s+m+1]$$

Számoláshoz



Probléma

$10^{m-1}T[s+1]$ kivonása eltávolítja a legnagyobb helyi értékű számjegyet; az eredményt tízzel szorozva a számot eggyel balra léptetjük; és ehhez hozzáadva $T[s+m+1]$ értékét, a legalacsonyabb helyi értékre a megfelelő számjegy kerül.

Probléma: p és t_s értékek olyan nagyok lehetnek, amelyeket már nem lehet számítógépen a szokásos módon ábrázolni.

Ebben az esetben azonban nem tehetjük fel, hogy az aritmetikai műveletek konstans időben végrehajthatók.

Megoldás

Számítsuk ki p és t_s értékét $modulo\ q$, ahol q egy alkalmas modulus.
$p\ modulo\ q$ meghatározható $\Theta(m)$ idő alatt és az összes t_s érték q-val képzett osztási maradéka kiszámítható $\Theta(n-m+1)$ idő alatt.

A q értékét rendszerint olyan **prímszámnak választják**, hogy $10q$ még éppen ábrázolható legyen előjel nélküli egész számként.

Általában, amikor az ábécé jeleinek száma d, és ezeket a $\{0, 1, \ldots, d-1\}$ értékekkel azonosítjuk, akkor q értékét úgy választjuk meg, hogy $d * q$ ábrázolható legyen a számítógépen előjel nélküli egész számként.

Módosítás

Módosítjuk a korábbi egyenletünket:

$$t_{s+1} = 10(t_s - 10^{m-1}T[s+1]) + T[s+m+1]$$

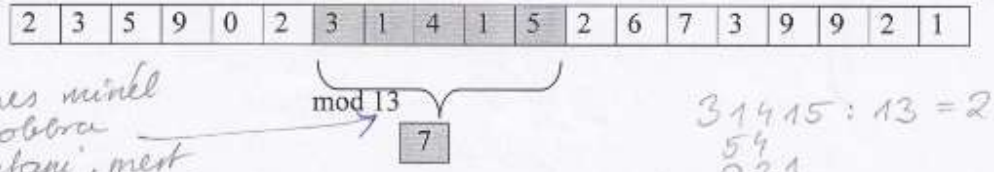
$$\Downarrow$$

$$t_{s+1} = (d * (t_s - T[s+1] * h) + T[s+m+1]) mod\ q$$

ahol $h \equiv d^{m-1}(mod\ q)$ egy m szélességű ablak legmagasabb helyi értékén szereplő érték.

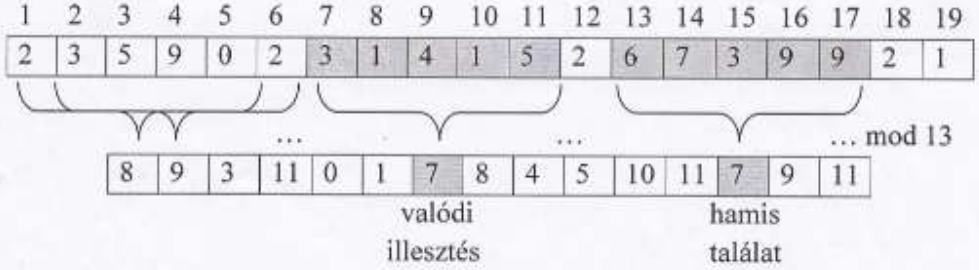
Példa:

Minden jel egy decimális számjegy, és az értékeket modulo 13 számoljuk. A szövegben egy 5 hosszúságú ablak tartalmát szürkével kiemeltük. A kiemelt rész osztási maradéka 7 modulo 13.

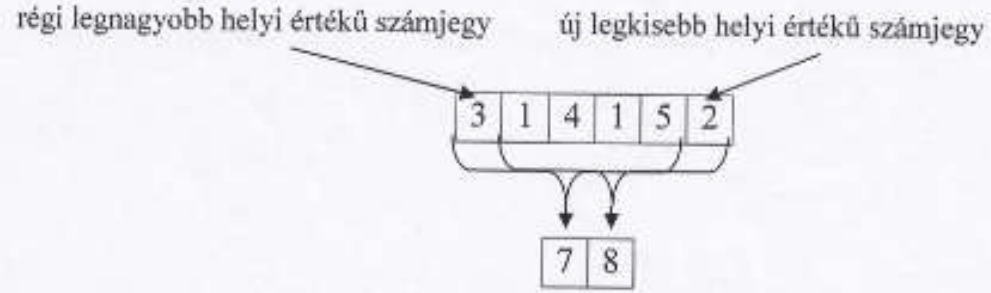


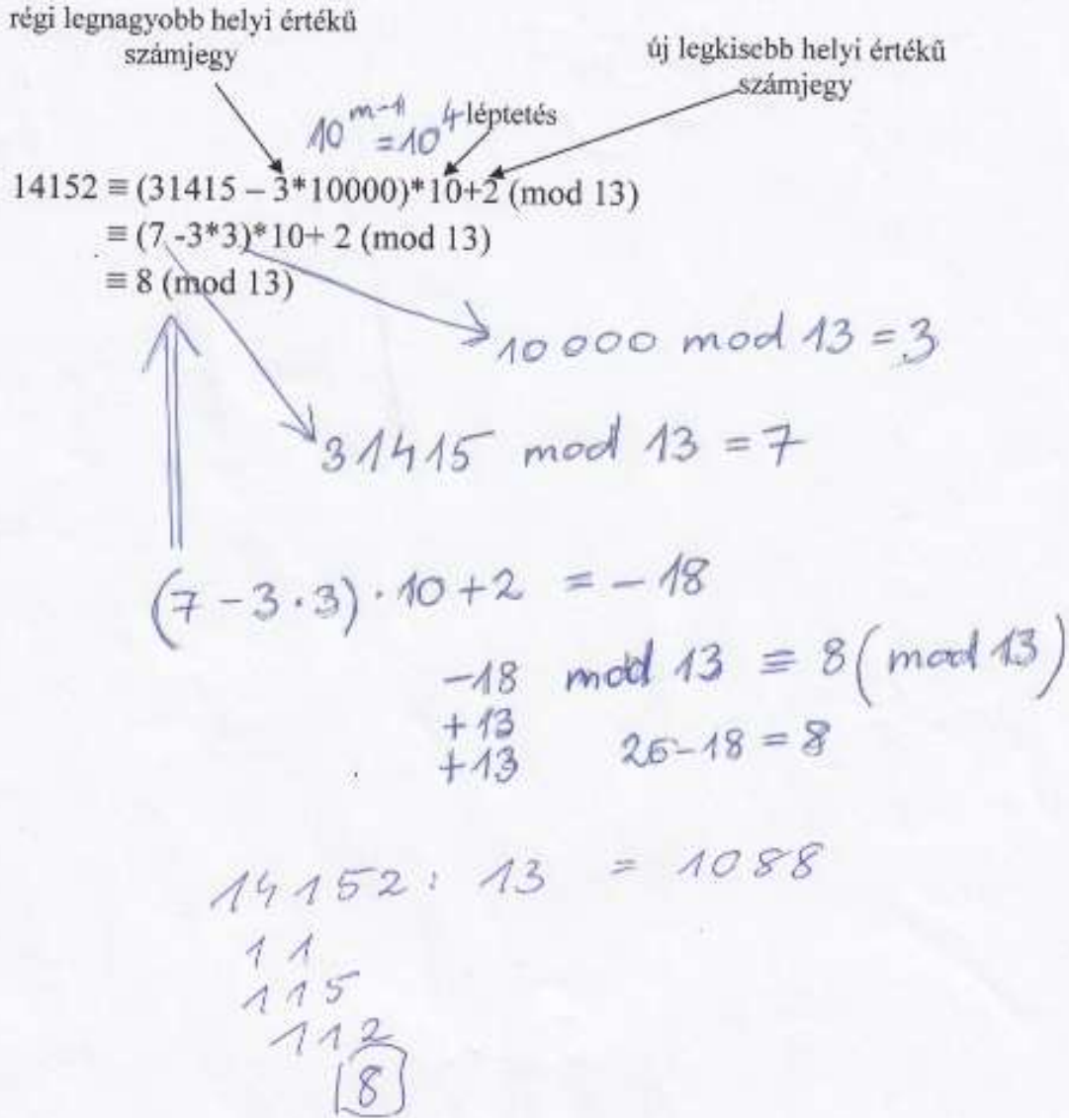
Minden lehetséges 5 hosszúságú ablakra kiszámoltuk az ablakban szereplő szám értékét modulo 13.

Ha a minta P=31415, akkor olyan részeket keresünk, amelyeknek az értéke 7 modulo 13, hiszen $31415 \equiv 7 \pmod{13}$. Két ilyen ablak létezik, ezeket kiemeltük. Az első, amelyik a 7. pozíción kezdődik, a minta egy tényleges előfordulása, ezzel szemben a 13. pozíción kezdődő második ablak egy hamis találat.



Egy ablakhoz tartozó érték konstans idejű kiszámítása a megelőző ablak értékének felhasználásával történhet. Az első ablakban szereplő érték 31415. Elhagyva a rész legmagasabb helyi értékű számjegyét (3), balra léptetve (tízzel szorozva), és végül hozzáadva a következő rész legalacsonyabb helyi értékű jegyét (2), megkapjuk az új rész értékét, ami 14152. Minden számítást modulo 13 végzünk, ezért az első ablaknak megfelelő érték 7, a másodikhoz rendelt pedig 8.


régi legnagyobb helyi értékű számjegy
új legkisebb helyi értékű számjegy
3 1 4 1 5 2
7 8


régi legnagyobb helyi értékű számjegy
új legkisebb helyi értékű számjegy
$10^{m-1} = 10^4$ léptetés
$14152 \equiv (31415 - 3*10000)*10+2 \pmod{13}$
$\equiv (7 - 3*3)*10+2 \pmod{13}$
$\equiv 8 \pmod{13}$
$10000 \bmod 13 = 3$
$31415 \bmod 13 = 7$
$(7 - 3 \cdot 3) \cdot 10 + 2 = -18$
$-18 \bmod 13 \equiv 8 \pmod{13}$
$+13$
$+13$
$26 - 18 = 8$
$14152 : 13 = 1088$
11
115
112
8

Magyarázat

Az ábrázolási korlátokat így sikerült feloldani azzal, hogy $modulo\ q$ számolunk, de ennek következményeként $t_s \equiv p(mod\ q)$ fennállásából nem következik $t_s = p$ teljesülése. Ugyanakkor, ha $t_s \neq p(mod\ q)$, akkor $t_s \neq p$, azaz s érvénytelen eltolás.

Így a $t_s \equiv p(mod\ q)$ feltételt gyors heurisztikaként alkalmazhatjuk érvénytelen eltolási értékek kiszűrésére.
Minden olyan s eltolási érték, amelyre $t_s \equiv p(mod\ q)$ fennáll, további ellenőrzésre szorul, hogy eldönthessük, s valóban érvényes vagy csak egy hamis találatot határoz meg.

Ez az ellenőrzés a $P[1..m] = T[s+1..s+m]$ feltétel explicit vizsgálatát jelenti.
Ha q kellően nagy szám, akkor várhatóan ritkán lépnek fel hamis találatok, így ezek ellenőrzésének költsége alacsony.

Eljárás

Rabin-Karp-Illeszto(T,P,d,q)

1. $n := hossz[T]$
2. $m := hossz[P]$
3. $h := d^{m-1} mod\ q$
4. $p := 0$
5. $t_0 := 0$
6. **for** $i := 1$ **to** m
7. **do** $p := (d * p + P[i]) mod\ q$
8. $t_0 := (d * t_0 + T[i]) mod\ q$
9. **for** $s := 0$ **to** $n - m$
10. **do if** $p = t_s$
11. **then if** $P[1..m] = T[s + 1..s + m]$
12. **then print** "A minta illeszkedik az ("$s + 1$")-edik pozícióra"
13. **if** $s < n - m$
14. **then**
15. $t_{s+1} := (d * (t_s - T[s + 1] * h) + T[s + m + 1])\ \ mod\ q$

1. 2. 3. 4. 5. 6. 7.
2 3 1 4 1 5 2

$n=7$
$m=5$

$q=13$ prímszám $d=10$ számrendszer alapja

$h = 10^{5-1} \bmod 13 = 3$

$p=0 \quad t_0=0$

$p := (d \cdot p + P[i]) \bmod q$

$t_0 := (d \cdot t_0 + T[i]) \bmod q$

$\boxed{p} := (10 \cdot 0 + 3) \bmod 13 = 3 \quad (i=1)$ $\quad \boxed{t_0} := (10 \cdot 0 + 2) \bmod 13 = 2$

$(10 \cdot 3 + 1) \bmod 13 = 5 \quad (i=2)$ $\quad (10 \cdot 2 + 3) \bmod 13 = 10$

$(10 \cdot 5 + 4) \bmod 13 = 2 \quad (i=3)$ $\quad (10 \cdot 10 + 1) \bmod 13 = 10$

$(10 \cdot 2 + 1) \bmod 13 = 8 \quad (i=4)$ $\quad (10 \cdot 10 + 4) \bmod 13 = 0$

$(10 \cdot 8 + 5) \bmod 13 = \boxed{7} \quad (i=5)$ $\quad (10 \cdot 0 + 1) \bmod 13 = \boxed{1}$

23141 minta eleje

$\boxed{s=0}$ $\quad (s := n-m = 7-5 = 2)$

$p \overset{?}{=} t_0 \quad 7 \overset{?}{=} 1$ ✗

$s \overset{?}{<} n-m \quad 0 < 2$ ✓

$t_1 := (10(1 - 2 \cdot 3) + 5) \bmod 13 =$

t_0 $T[s+1]$ $T[s+m+1]$

$= -45 \bmod 3 = \boxed{7}$

$\boxed{s=1}$ $\quad p \overset{?}{=} t_1 \quad 7 \overset{?}{=} 7$ ✓

$P[1..5] \overset{?}{=} T[2..6]$

$31415 = 31415$ ✓

„A minta illeszkedik a 2. pozícióra."

$\boxed{s=2}$ ...