

Union-Find fák felismerése

Gelle Kitti Erzsébet

II. évf. programtervező informatikus

Témavezető: Dr. Iván Szabolcs

SZTE TTIK Számítástudomány Alapjai Tanszék

A Disjoint-set forest adatszerkezetet széles körben alkalmazzák, ha halmazok partícióinak nyilvántartására, illetve azok egyesítésére van szükség. Ilyen például egy minimális feszítőfát építő Kruskal algoritmus, alkalmazzák automaták minimalizálására, a hálózati összekapcsolhatóság ellenőrzésére, a képfeldolgozásban, vagy akár a HEX játékban felismerni, ha egy játékos nyert.

Egy ilyen adatszerkezet egy osztályt egy gyökér felé irányított faként tárol. Ekkor két csúcs egy osztályban van, ha ugyanaz a fájuk gyökere. Ebből adódóan egy ilyen fában való keresés akkor lesz gyors, ha a fák alacsonyak. Erre többféle stratégiát szokás használni: például a *union-by-rank* a fák magasságának, míg a *union-by-size* a fák méretének figyelembevételével építi fel a fákat. Utóbbi esetben egy egyesítés művelet közben a kisebb fát kötjük be a nagyobb méretű fa gyökeréhez, így a keresés egy ilyen fában legrosszabb esetben $\Theta(\log n)$ időigényű lesz. Az így kapható fákat Union fáknak nevezzük. Az adatszerkezet amortizált időkomplexitása $\log n$ -ről $\alpha(n)$ -re javítható az ún. path compression technika alkalmazásával, mely keresés közben a fákat optimalizálja. Az így kapható fákat Union-Find fáknak nevezzük. Itt $\alpha(n)$ az Ackermann függvény rendkívül lassan növekvő inverze, mely még $n = 2^{65536}$ esetén is kisebb, mint 5.

Az adatszerkezetet implementáló szoftvercsomagok tesztelésekor hasznos lehet, ha ellenőrizni tudjuk a futás egy-egy pillanatában, hogy az aktuális adatszerkezetünk fáin előállhatnak-e egyáltalán a vonatkozó algoritmus helyes futása során. Leizhen Cai 1993-as munkájában leír egy Union fák felismerésére szolgáló polinomidejű algoritmust, s megjegyzi, hogy a Union-Find fák felismerése már sokkal nehezebb feladatnak tűnik. Mindazonáltal a fentiek alapján is, a path compression alkalmazása sokat javít az adatszerkezet időkomplexitásán, ezért az úgy előálló fák karakterizálása egy fontos részprobléma ebben a feladatban. Azonban Cai munkája óta nem történt előrelépés ezen a területen.

Jelen dolgozatban megadjuk a Union-Find fák egy leírását a *union-by-size* stratégia alkalmazása mellett, és megmutatjuk, hogy a vonatkozó eldöntési probléma NP-teljes. Ennek egy gyakorlati következménye, hogy az ilyen csomagok debugolása továbbra is csak intenzív loggolással valósítható meg large-scale alkalmazásokban, mint például DNS bankok klaszterezésében.