Lida Ghadamiyan

SkipList, un model de structură de date probabilist

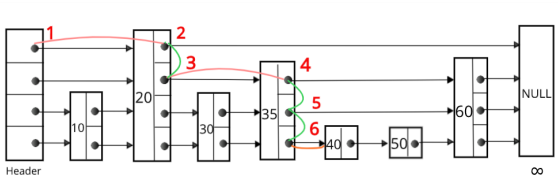
July 19, 2020

Aceasta lucrare prezintă importanța și aplicabilitatea matematicii, mai exact a teoriei probabilităților și a statisticii în informatică, având ca temă analiza probabilistică a algoritmilor unui SkipList.

Lucrarea debutează prin definirea termenilor, noțiunilor și fundamentelor matematice și informatice folosite ulterior. Printre acestea se numără și descrierea succintă a listelor și arborilor, în scopul comparației cu SkipList.

Componenta principală a lucrării se constituie în prezentarea structurii SkipList, și a eficienței acesteia. SkipList reprezintă o structură de date ce are la bază conformația unei liste simplu înlănțuite a cărei noduri au, în plus, pointeri către mai multe noduri decât cel vecin. Numărul pointerilor suplimentari reprezintă variabile aleatoare distribuite Geometric(p), fapt pentru care, spre deosebire de arbori, SkipList nu necesită reechilibrări.

Operațiile ce se pot efectua asupra elementelor acestei structuri sunt căutarea, inserarea și ștergerea. Algoritmul de căutare constă în examinarea structurii începând de la cel mai înalt pointer din header și înaintarea spre dreapta dacă valoarea găsită este mai mică decât cea căutată, sau coborârea unui nivel în caz contrar. În urma căutării unui element, se poate inspecta sau șterge nodul găsit, iar în cazul în care se dorește inserarea unui nou nod, se efectuează căutarea nodului respectiv și inserarea acestuia în locul celui returnat.



Costul unei operații va fi numărul nodurilor examinate până la găsirea nodului dorit. Abordarea lui Pugh pentru a găsi un majorant al costului constă în considerarea drumului de căutare în sens invers, împărțindu-l în mutări în sus și către stânga. În fiecare moment al căutării se va putea merge în sus cu probabilitatea p sau la stânga cu probabilitatea 1-p. Având în vedere probabilitățile fiecărei mutări, costul unei căutări ce conține k nivele este $\frac{k}{p}$. Considerând o structură infinită, acest mod de calcul ne permite să aflăm costul până la un nivel ce conține un număr cunoscut de noduri. Alegem acest nivel, L(n) unde n este numărul elementelor din structură, să conțină $\frac{1}{p}$ pointeri. Cum numărul nodurilor aflate pe nivelele orizontale ale structurii reprezintă variabile aleatoare Binomiale(n,$p^{(l-1)}$) unde l este nivelul, numărul mediu de pointeri aflați pe un nivel este media acestei variabile, anume $np^{(l-1)}$. Astfel, L(n) = $log_{\frac{1}{p}}(n)$. Vom aplica această strategie pentru aflarea costului doar până la nivelul L(n). În continuare vom considera că toate nodurile pe nivele superioare lui L(n) vor fi examinate. Știind că fiecare nod avansează un nivel cu probabilitatea p, numărul de noduri rămase la un nivel k + L(n) este $\frac{1}{p}p^{(k-1)}$. Astfel, se obține numărul total de noduri peste L(n) sumând nodurile de pe fiecare nivel mai mare decât L(n) până la infinit $\sum_{k=1}^{\infty} p^{(l-1)} = \frac{1}{1-p}$. Se obține astfel un majorant al costului $Cost \leq \frac{log_{\frac{1}{p}}(n)}{p} + \frac{1}{1-p} \approx O(log(n))$.

Abordarea lui Papadakis pentru a găsi costul exact constă în împărțirea drumului în mutări orizontale și verticale, obținând astfel costul căutării cheii de pe poziția m într-o structură

cu n elemente ca fiind suma dintre înlățimea structurii și numărul mutărilor orizontale. Din proprietatea de liniaritate a mediei ajungem la $E[C_n^{(m)}] = E[T_n] + E[L_{m-1}]$ unde $C_n^{(m)}$ este costul, T_n înălțimea și L_{m-1} numărul de mutări orizontale. Prin calculul efectiv al acestor medii se ajunge în final la valoarea costul mediu pentru o singură căutare, anume $E[C_n^{(m)}] = V_p(n+1) + \frac{1-p}{p}V_p(m) + 1$ unde $V_p(l) = \frac{1}{l}\sum_{k=2}^{l}\binom{l}{k}(-1)^k\frac{kp^{k-1}}{1-p^{k-1}}$. în această manieră s-a găsit costul mediu al unei singure căutări. Alt rezultat de interes reprezintă costul mediu al căutărilor tuturor nodurilor sau costul căutării "valorii de mijloc", iar în vederea găsirii acestei valori considerăm media empirică a costurilor, S_n, cu $E[S_n] = V_p(n+1) + \frac{1-p}{p}\frac{W_p(n)}{n} + 1$, unde $W_p(l) = \sum_{k=2}^{l}\binom{l}{k}(-1)^k\frac{p^{k-1}}{1-p^{k-1}}$. Prin analiza asimptotică a expresiilor costului mediu al unei căutări și al costului mediu al căutărilor se ajunge la O(log(n)).

În ceea ce privește complexitatea spațiu, pentru aflarea numărului de pointeri din structură, folosim media variabilei binomiale amintite anterior. Sumând numărul de noduri aferente fiecărui nivel orizontal al unei structuri cu n elemente obținem $\sum_{m\geq 1} np^{(m-1)} = \frac{n}{1-p}$ pointeri. Așadar, complexitatea spațiu este O(n).

Există varietăți ale acestei structuri, ele fiind modificate în funcție de scopul pentru care urmează a fi folosite. O atenție deosebită trebuie acordată către SkipList pentru date multidimensionale, deoarece consider că acestă variantă a structurii se va integra cel mai bine în automatizarea sarcinilor de tip computațional, fiind mai practică în circumstanțele realității. Aceste structuri se pretează situațiilor în care datele trebuie sortate în funcție de mai multe criterii, spre exemplu, locațiile în funcție de ambele coordonate, avioanele în funcție de ora sosirii și a decolării sau angajații în funcție de salariu și anul angajării. În întâmpinarea unei situații de această natură, cu un simplu SkipList, ar trebui să rearanjăm structura de fiecare dată, deși este cunoscut faptul că se va ajunge în acest punct. Pentru astfel de situații, se poate folosi un SkipList pentru date multidimensionale. Din punct de vedere al complexității în timp și spațiu, acestea sunt direct proporționale cu valorile găsite pentru SkipList, constanta de proporționalitate fiind numărul de dimensiuni al structurii. Pentru o structură d-dimensională, complexitatea timp va fi O(d log(n)), iar cea spațiu O(d n), unde n este numărul de noduri.

În societatea actuală se dorește automatizarea cât mai multor tipuri de sarcini, în special a celor repetitive, care necesită memorarea datelor. În raport cu alte structuri de date, SkipList este atât mai eficient, cât și mai simplu, la baza acestui rezultat stând conceptele matematice folosite și aplicate. Timpul de execuție este foarte important deoarece, o tendință generala a oamenilor este cuantificarea timpului în bani, drept urmare aceștia fiind foarte preocupați de diminuarea timpului alocat diverselor activități, profitând de progresul tehnologic pentru a atribui aceste sarcini unui program care poate face respectiva activitate în timp scurt și costuri minime. Din aceste motive, este ideală utilizarea unor structuri rapide și ușor de manipulat. Așadar, SkipList, prin eficiență și simplitate se califică utilizării într-o gamă foarte mare de programe. Fiind un subiect recent introdus, sunt șanse mari să se dovedească a fi utile în mai multe domenii, în diverse forme. Datorită rezultatelor de acest tip, matematica devine din ce în ce mai folosită în variate domenii, stând la baza progreselor, economisind timp, resurse și energie.

Lida Ghadamiyan