

Manual del Programador Competitiu

Antti Laaksonen

Draft 18 d'octubre de 2022

Índex

Prefaci	ix
I Tècniques bàsiques	1
1 Introducció	3
1.1 Llenguatges de programació	3
1.2 Entrada i sortida	4
1.3 Treballar amb nombres	6
1.4 Escurçar el codi	8
1.5 Matemàtiques	10
1.6 Concursos i recursos	15
2 Complexitat temporal	19
2.1 Regles de càlcul	19
2.2 Clases de complexitat	22
2.3 Estimació de l'eficiència	23
2.4 Suma màxima d'un subvector	24
3 Ordenació	27
3.1 Teoria de l'ordenació	27
3.2 Ordenació en C++	31
3.3 Cerca binària	33
4 Estructures de dades	37
4.1 Vectors dinàmics	37
4.2 Estructures conjunt	39
4.3 Estructures mapa	40
4.4 Iteradors i intervals	41
4.5 Altres estructures	43
4.6 Comparació amb l'ordenació	47
5 Cerca completa	49
5.1 Generar subconjunts	49
5.2 Generar permutacions	51
5.3 Backtracking	52
5.4 Podar la cerca	54
5.5 Trobar-se al mig	57

6	Algorismes greedy	59
6.1	Problema de les monedes	59
6.2	Scheduling	60
6.3	Tasques i terminis	62
6.4	Minimitzar sumes	63
6.5	Compressió de dades	64
7	Programació dinàmica	67
7.1	Problema de les monedes	67
7.2	Subseqüència creixent més llarga	72
7.3	Camins en una quadrícula	73
7.4	Problemes de motxilla	75
7.5	Distància d'edició	76
7.6	Comptar rajoles	78
8	Anàlisi amortitzada	81
8.1	Mètode dels dos punters	81
8.2	Element menor més propers	83
8.3	Mínim de finestra lliscant	85
9	Consultes d'interval	87
9.1	Consultes de vector estàtiques	88
9.2	Arbre binari indexat	90
9.3	Arbre de segments	93
9.4	Tècniques addicionals	97
10	Manipulació de bits	99
10.1	Representació binària	99
10.2	Operacions de bits	100
10.3	Representació de conjunts	102
10.4	Optimitzacions de bits	104
10.5	Programació dinàmica	106
II	Algorismes de grafs	111
11	Introducció als grafs	113
11.1	Vocabulari de grafs	113
11.2	Representació de grafs	116
12	Recorreguts en grafs	121
12.1	Cerca en profunditat	121
12.2	Cerca en amplada	123
12.3	Aplicacions	125

13 Camins més curts	129
13.1 Algorisme de Bellman–Ford	129
13.2 Algorisme de Dijkstra	132
13.3 Algorisme de Floyd-Warshall	135
14 Algorismes d'arbres	139
14.1 Recorregut d'arbres	140
14.2 Diàmetre	141
14.3 Tots els camins més llargs	143
14.4 Arbres binaris	145
15 Arbres d'expansió	147
15.1 Algorisme de Kruskal	148
15.2 Estructura <i>union-find</i>	151
15.3 Algorisme de Prim	153
16 Grafs dirigits	157
16.1 Ordenació topològica	157
16.2 Programació dinàmica	159
16.3 Camins de successió	162
16.4 Detecció de cicles	163
17 Grafs fortament connexos	167
17.1 Algorisme de Kosaraju	168
17.2 Problema 2SAT	170
18 Consultes d'arbres	173
18.1 Trobar avantpassats	173
18.2 Subarbres i camins	174
18.3 Avantpassat comú més baix	177
18.4 Algorismes <i>offline</i>	180
19 Camins i circuits	185
19.1 Camins eulerians	185
19.2 Camins de Hamilton	189
19.3 Seqüències de De Bruijn	190
19.4 Ruta del cavall	191
20 Fluxos i talls	193
20.1 Algorisme de Ford–Fulkerson	194
20.2 Camins discontinus	198
20.3 Emparellaments màxims	199
20.4 Cobertura de camins	202

III Temes avançats	207
21 Teoria de nombres	209
21.1 Nombres primers i factors	209
21.2 Aritmètica modular	213
21.3 Resolució d'equacions	216
21.4 Altres resultats	217
22 Combinatòria	219
22.1 Coeficients binomials	220
22.2 Nombres de Catalan	223
22.3 Inclusió-exclusió	225
22.4 Lema de Burnside	226
22.5 Fórmula de Cayley	227
23 Matrius	231
23.1 Operacions	231
23.2 Recurrències lineals	234
23.3 Grafs i matrius	236
24 Probabilitat	239
24.1 Càlcul	239
24.2 Esdeveniments	240
24.3 Variables aleatòries	242
24.4 Cadenes de Markov	244
24.5 Algorismes aleatoris	245
25 Teoria de jocs	249
25.1 Estats del joc	249
25.2 Joc de nim	251
25.3 Teorema de Sprague–Grundy	252
26 Algorismes de cadenes	257
26.1 Terminologia de cadenes	257
26.2 Tipus Trie	258
26.3 Hashing de cadenes	259
26.4 Algorisme Z	262
27 Algorismes d'arrel quadrada	265
27.1 Combinació d'algorismes	266
27.2 Particions senceres	268
27.3 Algorisme de Mo	269
28 Arbres de segments revisats	273
28.1 Propagació mandrosa	274
28.2 Arbres dinàmics	277
28.3 Estructures de dades	279
28.4 Bidimensionalitat	280

29 Geometria	283
29.1 Nombres complexos	284
29.2 Punts i línies	286
29.3 Àrea del polígon	289
29.4 Funcions distància	291
30 Algorismes d'escombrat de línies	293
30.1 Punts d'intersecció	294
30.2 Problema de la parella més propera	295
30.3 Problema de l'envolupant convexa	296
Bibliografia	299

Prefaci

L'objectiu d'aquest llibre és donar-vos una introducció completa a la programació competitiva. Se suposa que ja coneixeu els fonaments bàsics de la programació, però no cal cap formació prèvia en programació competitiva.

El llibre està especialment pensat per a estudiants que vulguin aprendre algorismes i possiblement participar en l'Olimpíada Internacional d'Informàtica (IOI) o en el Concurs Internacional de Programació Col·legiata (ICPC). Per descomptat, el llibre també és adequat per a qualsevol altra persona interessada en la programació competitiva.

Es necessita molt de temps per convertir-se en un bon programador competitiu, però també és una oportunitat per aprendre molt. Podeu estar segur que obtindreu una bona comprensió general dels algorismes si passeu temps llegint el llibre, resolent problemes i participant en concursos.

El llibre està en desenvolupament continu. Sempre podeu enviar comentaris sobre el llibre a `ahslaaks@cs.helsinki.fi`.

Helsinki, August 2019
Antti Laaksonen

(Nota del Traductor) L'Antti Laaksonen és l'autor de "Guide to Competitive Programming" (Springer, 2017). Aquest PDF és la traducció al català del seu manual "Competitive Programmer's Handbook" (<https://github.com/p11k/cphb>). Aquesta traducció no hagués estat possible si l'Antti no hagués distribuït les fonts \LaTeX del seu manual sota llicència Creative Common.

Amb aquesta traducció vull permetre que els participants de l'Olimpíada Informàtica de Catalunya *que no saben prou d'anglès* puguin millorar el seu nivell de programació competitiva. Si ets una d'aquestes persones, espero que aquesta traducció et sigui útil! Però recorda que no poder llegir en anglès és literalment una *minusvalia* que t'obliga a dependre de la voluntat i criteri de traductors com jo. Si vols continuar aprenent coses, millora el teu anglès.

Eugene, Desembre 2021
Omer Giménez Llach

Part I

Tècniques bàsiques

Capítol 1

Introducció

La programació competitiva combina dos temes: (1) el disseny d'algorismes i (2) la implementació d'algorismes.

El **disseny d'algorismes** consisteix en la resolució de problemes i el pensament matemàtic. Fan falta habilitats per analitzar problemes i resoldre'ls creativament. Un algorisme per resoldre un problema ha de ser correcte i eficient, i la dificultat consisteix sovint sobre inventar un algorisme eficient.

Els coneixements teòrics d'algorismes són important per als programadors competitius. Normalment, una solució a un problema és una combinació de tècniques conegudes i noves idees. Les tècniques que apareixen en la programació competitiva també constitueixen la base per a la investigació científica d'algorismes.

La **implementació d'algorismes** requereix bones habilitats de programació. En la programació competitiva, les solucions es classifiquen provant un algorisme implementat fent servir un conjunt de casos de prova. Per tant, no n'hi ha prou que la idea del l'algorisme sigui correcte, la implementació també ha de ser correcte.

L'estil correcte de programació en els concursos és ser directe i concís. Els programes s'han d'escriure ràpidament, perquè no hi ha gaire temps disponible. A diferència de l'enginyeria de software tradicional, els programes són curts (normalment com a molt uns pocs centenars de línies de codi) i no cal mantenir el codi un cop acabat el concurs.

1.1 Llenguatges de programació

En l'actualitat, els llenguatges de programació que es fan servir més als concursos són el C++, el Python i el Java. Per exemple, a Google Code Jam 2017, entre els 3.000 millors participants, el 79 % va fer servir C++, el 16 % Python i el 8 % Java [29]. Alguns participants també utilitzaven varis llenguatges.

Molta gent pensa que el C++ és la millor opció per a un programador competitiu, i el C++ gairebé sempre està disponible en els sistemes de concurs. Els avantatges d'utilitzar C++ són que és un llenguatge de programació molt eficient i que la seva biblioteca estàndard conté un gran col·lecció d'estructures de dades i algorismes.

D'altra banda, és bo dominar diversos llenguatges de programació i comprendre els seus punts forts. Per exemple, si es necessiten nombres enters grans en el problema, Python pot ser una bona opció, perquè aquest llenguatge conté operacions integrades per càlculs amb nombres enters grans. Tot i així, la majoria dels problemes en els concursos de programació s'intenten preparar per a que fer servir un llenguatge de programació específic no sigui un avantatge injust.

Tots els programes d'exemple d'aquest llibre estan escrits en C++, i sovint es fa servir les estructures de dades i algorismes de la llibreria estàndard. Els programes segueixen l'estàndard C++11, que es pot fer servir en la majoria dels concursos actuals. Si encara no podeu programar en C++, ara és un bon moment per començar a aprendre.

plantilla de codi C++

Aquesta és una plantilla de codi C++ típica per a la programació competitiva:

```
#include <bits/stdc++.h>

using namespace std;

int main() {
    // Aquí va el codi.
}
```

La línia `#include` al principi del codi és una característica del compilador `g++` que ens permet incloure tota la biblioteca estàndard. Per tant, no cal incloure per separat biblioteques com ara `iostream`, `vector` i `algorithm`, sinó que estan disponibles automàticament.

La línia `using` declara que les classes i funcions de la biblioteca estàndard es poden fer servir directament al codi. Sense la línia `using` tindríem que escriure, per exemple, `std::cout`, però ara n'hi ha prou amb escriure `cout`.

El codi es pot compilar mitjançant l'ordre següent:

```
g++ -std=c++11 -O2 -Wall test.cpp -o test
```

Aquesta ordre produeix un fitxer binari `test` a partir del codi font `test.cpp`. El compilador segueix l'estàndard C++11 (`-std=c++11`), optimitza el codi (`-O2`) i mostra avisos sobre possibles errors (`-Wall`).

1.2 Entrada i sortida

A la majoria de concursos es fan servir els canals estàndard per a la lectura d'entrada i l'escriptura de sortida. En C++, els canals estàndard són cin per a l'entrada i `cout` per a la sortida. També es poden fer servir les funcions de C `scanf` i `printf`.

L'entrada per al programa normalment consisteix en nombres i cadenes de text que es separen amb espais i salts de línia. Es poden llegir des del canal cin

com segueix:

```
int a, b;  
string x;  
cin >> a >> b >> x;
```

Aquest tipus de codi sempre funciona, suposant que hi ha almenys un espai o salt de línia entre cada element de l'entrada. Per exemple, el codi anterior pot llegir les dues entrades següents:

```
123 456 monkey
```

```
123 456  
monkey
```

El canal cout es fa servir per a la sortida:

```
int a = 123, b = 456;  
string x = "monkey";  
cout << a << " " << b << " " << x << "\n";
```

L'entrada i la sortida són de vegades un coll d'ampolla en el programa. Les línies següents al principi del codi fan que l'entrada i la sortida siguin més eficients:

```
ios::sync_with_stdio(0);  
cin.tie(0);
```

Tingueu en compte que el salt de línia "\n" funciona més ràpid que endl, perquè endl sempre provoca una operació de buidat del buffer.

Les funcions C scanf i printf són una alternativa als canals estàndard de C++. Normalment són una mica més ràpides, però també són més difícils d'utilitzar. El codi següent llegeix dos nombres enters de l'entrada:

```
int a, b;  
scanf("%d %d", &a, &b);
```

El codi següent imprimeix dos nombres enters:

```
int a = 123, b = 456;  
printf("%d %d\n", a, b);
```

De vegades, el programa hauria de llegir una línia sencera de l'entrada, possiblement amb espais. Això es pot aconseguir mitjançant l'ús de la funció getline:

```
string s;  
getline(cin, s);
```

Si es desconeix la quantitat de dades, podem fer servir el següent bucle:

```
while (cin >> x) {  
    // codi  
}
```

Aquest bucle llegeix elements de l'entrada un rere l'altre, fins que no n'hi ha hagut més dades disponibles.

En alguns sistemes de concurs es fa servir fitxers per a l'entrada i la sortida. Una solució fàcil per a això és escriure el codi com de costum utilitzant canals estàndard, però afegint les línies següents al començament del codi:

```
freopen("entrada.txt", "r", stdin);  
freopen("output.txt", "w", stdout);
```

Després d'això, el programa llegeix l'entrada del fitxer "input.txt" i escriu la sortida al fitxer "output.txt".

1.3 Treballar amb nombres

Nombres enters

El tipus enter més utilitzat en la programació competitiva és el tipus `int`, que és un tipus de 32 bits amb un rang de valors de $-2^{31} \dots 2^{31} - 1$ o aproximadament $-2 \cdot 10^9 \dots 2 \cdot 10^9$. Si el tipus `int` no és suficient, es pot fer servir el tipus de 64 bits `long long`. Té un rang de valors de $-2^{63} \dots 2^{63} - 1$ o aproximadament $-9 \cdot 10^{18} \dots 9 \cdot 10^{18}$.

El codi següent defineix una variable de tipus `long long`:

```
long long x = 123456789123456789LL;
```

El sufix `LL` significa que el tipus del nombre és `long long`.

Un error comú quan es fa servir el tipus `long long` és fer servir algun tipus `int` en algun lloc al codi. Per exemple, el codi següent conté un error subtil:

```
int a = 123456789;  
long long b = a*a;  
cout << b << "\n"; // -1757895751
```

Tot i que la variable `b` és del tipus `long long`, els dos nombres de l'expressió `a*a` són del tipus `int` i el resultat és també del tipus `int`. Per això, la variable `b` tindrà un resultat incorrecte. El problema es pot resoldre fent que el tipus de `a` sigui `long long` o canviant l'expressió a `(long long)a*a`.

Normalment els problemes de concurs es plantegen de manera que amb el tipus `long long` n'hi ha prou. Tot i així, és bo saber que el compilador `g++` també proporciona un tipus de 128 bits `__int128_t` amb un rang de valors de $-2^{127} \dots 2^{127} - 1$ o aproximadament $-10^{38} \dots 10^{38}$. Tanmateix, aquest tipus no està disponible en tots els sistemes de concurs.

Aritmètica modular

Denotem per $x \bmod m$ el residu de dividir x per m . Per exemple, $17 \bmod 5 = 2$, perquè $17 = 3 \cdot 5 + 2$.

De vegades, la resposta a un problema és a nombre molt gran però es demana que s'escrigui la solució “mòdul m ”, és a dir, el residu quan es divideix la resposta per m (per exemple, “mòdul $10^9 + 7$ ”). D'aquesta manera, encara que la resposta real sigui molt gran, n'hi ha prou amb utilitzar els tipus `int` i `long long`.

Una propietat important del residu és que en la suma, la resta i la multiplicació, el residu es pot prendre abans de l'operació:

$$\begin{aligned} rcr(a + b) \bmod m &= (a \bmod m + b \bmod m) \bmod m \\ (a - b) \bmod m &= (a \bmod m - b \bmod m) \bmod m \\ (a \cdot b) \bmod m &= (a \bmod m \cdot b \bmod m) \bmod m \end{aligned}$$

Així, podem agafar el residu després de cada operació i les xifres mai seran massa grans.

Per exemple, el codi següent calcula $n!$, el factorial de n , mòdul m :

```
long long x = 1;
per (int i = 2; i <= n; i++) {
    x = (x*i)%m;
}
cout << x%m << "\n";
```

Normalment volem que el residu estigui sempre en $0 \dots m - 1$. Tanmateix, en el C++ i altres llenguatges, el residu d'un nombre negatiu és zero o negatiu. Una manera fàcil d'assegurar-s'hi que no tenim residus negatius es cal calcular primer el residu com de costum i després afegeix m si el resultat és negatiu:

```
x = x%m;
if (x < 0) x += m;
```

Això només és necessari quan hi ha restes en el codi i el residu pot arribar a ser negatiu.

Nombres de coma flotant

Els tipus de coma flotant en la programació competitiva són el `double` de 64 bits i, com a extensió al compilador `g++`, el `long double` de 80 bits. En la majoria dels casos, `double` és suficient, però `long double` és més precís.

La precisió requerida de la resposta normalment es dona a l'enunciat del problema. Una manera fàcil d'emetre la resposta és fer servir la funció `printf` i donar el nombre de decimals a la cadena de text que especifica el format. Per exemple, el codi següent escriu el valor de x amb 9 decimals:

```
printf("%.9f\n", x);
```

Una dificultat quan s'utilitzen nombres de coma flotant és que alguns nombres no es poden representar amb precisió com a nombres de coma flotant, i hi haurà errors d'arrodoniment. Per exemple, el resultat del codi següent és sorprenent:

```
double x = 0.3*3+0.1;
printf("%.20f\n", x); // 0.99999999999999988898
```

El valor de x és una mica més petit que 1 degut als errors d'arrodoniment, quan el valor correcte seria 1.

És arriscat comparar nombres de coma flotant amb l'operador `==`, perquè és possible que, encara els valors haurien de ser iguals, en realitat no ho són degut a errors de precisió. Una millor manera de comparar nombres de coma flotant és suposar que dos nombres són iguals si la diferència entre ells és menor que ε , on ε és un nombre petit.

A la pràctica, els nombres es poden comparar de la següent manera ($\varepsilon = 10^{-9}$):

```
if (abs(a-b) < 1e-9) {
    // a i b son iguals
}
```

Tingueu en compte que, tot i que els nombres de coma flotant són inexactes, els enters fins a un cert límit encara es poden representar amb precisió. Per exemple, fent servir `double`, és possible representar amb precisió nombres enters el valor absolut dels quals és com a màxim 2^{53} .

1.4 Esgurçar el codi

El codi curt és ideal en programació competitiva, perquè els programes s'han d'escriure el més ràpid possible. Per això, els programadors competitius sovint defineixen noms més curts per a tipus de dades i altres parts del codi.

Noms dels tipus

Fent servir l'ordre `typedef` és possible donar un nom més curt a un tipus de dades. Per exemple, com que el nom `long long` és llarg, podem definir un nom més curt `ll` d'aquesta manera:

```
typedef long long ll;
```

Després d'això, el codi

```
long long a = 123456789;
long long b = 987654321;
cout << a*b << "\n";
```

es pot escurçar de la següent manera:

```
ll a = 123456789;
```

```
ll b = 987654321;
cout << a*b << "\n";
```

L'ordre typedef també es pot fer servir amb tipus més complexos. Per exemple, el codi següent dona el nom vi per a un vector de nombres enters i el nom pi per a una parella que conté dos nombres enters.

```
typedef vector<int> vi;
typedef pair<int,int> pi;
```

Macros

Una altra manera d'escurçar el codi és definir **macros**. Una macro significa que hi ha determinades cadenes de text que el codi es canviarà abans de la compilació. En C++, les macros es defineixen mitjançant la paraula clau #define.

Per exemple, podem definir les macros següents:

```
#define F primer
#define S segon
#define PB push_back
#define MP make_pair
```

Després d'això, el codi

```
v.push_back(make_pair(y1,x1));
v.push_back(make_pair(y2,x2));
int d = v[i].primer+v[i].segon;
```

esdevé:

```
v.PB(MP(y1,x1));
v.PB(MP(y2,x2));
int d = v[i].F+v[i].S;
```

Una macro també pot tenir paràmetres, la qual cosa fa possible escurçar bucles i altres estructures. Per exemple, podem definir la macro següent:

```
#definir REP(i,a,b) for (int i = a; i <= b; i++)
```

Després d'això, el codi

```
for (int i = 1; i <= n; i++) {
    cerca(i);
}
```

esdevé:

```
REP(i,1,n) {
    cerca(i);
}
```

```
}
```

De vegades, les macros provoquen errors que poden ser difícils per detectar. Per exemple, considereu la macro següent que calcula el quadrat d'un nombre:

```
#define SQ(a) a*a
```

Aquesta macro *no* sempre funciona com s'esperava. Per exemple, el codi

```
cout << SQ(3+3) << "\n";
```

correspon al codi

```
cout << 3+3*3+3 << "\n"; // 15
```

Una versió millor de la macro és la següent:

```
#definir SQ(a) (a)*(a)
```

Ara el codi

```
cout << SQ(3+3) << "\n";
```

correspon al codi

```
cout << (3+3)*(3+3) << "\n"; // 36
```

1.5 Matemàtiques

Les matemàtiques tenen un paper important en la competició programació, i no és possible arribar a ser un programador competitiu exitós sense tenir bones habilitats matemàtiques. En aquesta secció es discuteixen alguns conceptes i fórmules matemàtiques que són necessàries més endavant al llibre.

Fórmules de suma

Cada suma de la forma

$$\sum_{x=1}^n x^k = 1^k + 2^k + 3^k + \dots + n^k,$$

on k és un nombre enter positiu, té una fórmula tancada que és un polinomi de grau $k + 1$. Per exemple¹,

$$\sum_{x=1}^n x = 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$$

¹ Fins i tot hi ha una fórmula general per a aquestes sumes, anomenada **fórmula de Faulhaber**, però és massa complexa per a presentar-la aquí.

i

$$\sum_{x=1}^n x^2 = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Una **progressió aritmètica** és una seqüència de nombres on la diferència entre dos nombres consecutius és una constant. Per exemple,

$$3, 7, 11, 15$$

és una progressió aritmètica amb constant 4. Es pot calcular la suma d'una progressió aritmètica fent servir la fórmula

$$\underbrace{a + \dots + b}_{n \text{ nombres}} = \frac{n(a+b)}{2}$$

on a és el primer nombre, b és l'últim nombre i n és la quantitat de nombres. Per exemple,

$$3 + 7 + 11 + 15 = \frac{4 \cdot (3 + 15)}{2} = 36.$$

La fórmula es basa en el fet que la suma consta de n nombres i el valor de cada nombre és $(a+b)/2$ de mitjana.

Una **progressió geomètrica** és una seqüència de nombres on el rati entre dos consecutius qualsevol nombres és una constant. Per exemple,

$$3, 6, 12, 24$$

és una progressió geomètrica amb constant 2. Es pot calcular la suma d'una progressió geomètrica fent servir la fórmula

$$a + ak + ak^2 + \dots + b = \frac{bk - a}{k - 1}$$

on a és el primer nombre, b és l'últim nombre i el la rati entre nombres consecutius és k . Per exemple,

$$3 + 6 + 12 + 24 = \frac{24 \cdot 2 - 3}{2 - 1} = 45.$$

Aquesta fórmula es pot derivar de la següent manera. Sigui

$$S = a + ak + ak^2 + \dots + b.$$

Quan multipliquem els dos costats per k , obtenim

$$kS = ak + ak^2 + ak^3 + \dots + bk,$$

i resolent l'equació

$$kS - S = bk - a$$

ens dona la fórmula.

Un cas especial de la suma d'una progressió geomètrica és la fórmula

$$1 + 2 + 4 + 8 + \dots + 2^{n-1} = 2^n - 1.$$

Una **suma harmònica** és una suma de la forma

$$\sum_{x=1}^n \frac{1}{x} = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}.$$

Un límit superior per a una suma harmònica és $\log_2(n) + 1$. En concret, podem modificar cada terme $1/k$ de manera que es converteixi en la potència més propera de dos que no superi k . Per exemple, quan $n = 6$, podem estimar la suma de la següent manera:

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} \leq 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}.$$

Aquest límit superior consta de $\log_2(n) + 1$ parts ($1, 2 \cdot 1/2, 4 \cdot 1/4$, etc.), i el valor de cada part és com a màxim 1.

Teoria de conjunts

Un **conjunt** és una col·lecció d'elements. Per exemple, el conjunt

$$X = \{2, 4, 7\}$$

conté els elements 2, 4 i 7. El símbol \emptyset denota un conjunt buit, i $|S|$ denota la mida d'un conjunt S , és a dir, el nombre d'elements del conjunt. Per exemple, en el conjunt anterior, $|X| = 3$.

Si un conjunt S conté un element x , escrivim $x \in S$, i en cas contrari escrivim $x \notin S$. Per exemple, en el conjunt anterior

$$4 \in X \quad \text{i} \quad 5 \notin X.$$

Es poden construir nous conjunts mitjançant operacions de conjunts:

- La **intersecció** $A \cap B$ és el conjunt dels elements que estan tant en A i en B . Per exemple, si $A = \{1, 2, 5\}$ i $B = \{2, 4\}$, llavors $A \cap B = \{2\}$.
- La **unión** $A \cup B$ és el conjunt dels elements que estan en A o en B o en tots dos. Per exemple, si $A = \{3, 7\}$ i $B = \{2, 3, 8\}$, llavors $A \cup B = \{2, 3, 7, 8\}$.
- El **complement** \bar{A} és el conjunt dels elements que no estan en A . La interpretació d'un complement depèn de el **conjunt universal**, que és el conjunt que conté tots els elements possibles. Per exemple, si $A = \{1, 2, 5, 7\}$ i el conjunt universal és $\{1, 2, \dots, 10\}$, aleshores $\bar{A} = \{3, 4, 6, 8, 9, 10\}$.
- La **diferència** $A \setminus B = A \cap \bar{B}$ és el conjunt dels elements que estan en A però no en B . Tingueu en compte que B pot contenir elements que no es troben a A . Per exemple, si $A = \{2, 3, 7, 8\}$ i $B = \{3, 5, 8\}$, aleshores $A \setminus B = \{2, 7\}$.

Si cada element de A també pertany a S , diem que A és un **subconjunt** de S , i ho denotem $A \subset S$. Un conjunt S sempre té $2^{|S|}$ subconjunts, inclòs el conjunt buit. Per exemple, els subconjunts del conjunt $\{2, 4, 7\}$ són

$\emptyset, \{2\}, \{4\}, \{7\}, \{2, 4\}, \{2, 7\}, \{4, 7\}$ i $\{2, 4, 7\}$.

Alguns conjunts fets servir sovint són \mathbb{N} (nombres naturals), \mathbb{Z} (nombres enters), \mathbb{Q} (nombres racionals) i \mathbb{R} (nombres reals). El conjunt \mathbb{N} es pot definir de dues maneres, segons la situació: o $\mathbb{N} = \{0, 1, 2, \dots\}$ o $\mathbb{N} = \{1, 2, 3, \dots\}$.

També podem construir un conjunt utilitzant una expressió de la forma

$$\{f(n) : n \in S\},$$

on $f(n)$ és una funció. Aquest conjunt conté tots els elements de la forma $f(n)$, on n és un element de S . Per exemple, el conjunt

$$X = \{2n : n \in \mathbb{Z}\}$$

conté tots els nombres enters parells.

Lògica

El valor d'una expressió lògica és **cert** (1) o **fals** (0). Les operacions lògiques més importants són \neg (**negació**), \wedge (**conjunció**), \vee (**disjunció**), \Rightarrow (**implicació**) i \Leftrightarrow (**equivalència**). La taula següent mostra el significat d'aquestes operacions:

A	B	$\neg A$	$\neg B$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftrightarrow B$
0	0	1	1	0	0	1	1
0	1	1	0	0	1	1	0
1	0	0	1	0	1	0	0
1	1	0	0	1	1	1	1

L'expressió $\neg A$ té el valor oposat de A . L'expressió $A \wedge B$ és certa si A i B són certes, i l'expressió $A \vee B$ és certa si A o B o tots dos són certes. L'expressió $A \Rightarrow B$ és certa si sempre que A és cert, B també ho és. L'expressió $A \Leftrightarrow B$ és certa quan A i B són les dues certes o les dues falses.

Un **predicat** és una expressió que és certa o falsa en funció dels seus paràmetres. Els predicats solen indicar-se amb majúscules. Per exemple, considerem el predicat $P(x)$ que és cert exactament quan x és un nombre primer. Amb aquesta definició, $P(7)$ és cert però $P(8)$ és fals.

Un **quantificador** connecta una expressió lògica als elements d'un conjunt. Els quantificadors més importants són \forall (**per a tots**) i \exists (**existeix un**). Per exemple,

$$\forall x(\exists y(y < x))$$

significa que per a cada element x del conjunt, existeix un element y al conjunt que compleix que y és més petit que x . Això és cert en el conjunt de nombres enters, però fals en el conjunt dels nombres naturals.

Utilitzant la notació descrita anteriorment, podem expressar molts tipus de proposicions lògiques. Per exemple,

$$\forall x((x > 1 \wedge \neg P(x)) \Rightarrow (\exists a(\exists b(a > 1 \wedge b > 1 \wedge x = ab))))$$

significa que si un nombre x és més gran que 1 i no és un nombre primer, aleshores hi ha els nombres a i b que són més grans que 1 i el producte dels quals és x . Aquesta proposició és certa en el conjunt dels nombres enters.

Funcions

La funció $\lfloor x \rfloor$ arrodoneix (cap avall) el nombre x fins a un nombre enter, i la funció $\lceil x \rceil$ arrodoneix (cap amunt) el nombre x fins a un nombre enter. Per exemple,

$$\lfloor 3/2 \rfloor = 1 \quad \text{i} \quad \lceil 3/2 \rceil = 2.$$

Les funcions $\min(x_1, x_2, \dots, x_n)$ i $\max(x_1, x_2, \dots, x_n)$ donen el valor més petit i més gran dels valors x_1, x_2, \dots, x_n . Per exemple,

$$\min(1, 2, 3) = 1 \quad \text{i} \quad \max(1, 2, 3) = 3.$$

El **factorial** $n!$ es pot definir

$$\prod_{x=1}^n x = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$$

o, recursivament,

$$\begin{aligned} 1! &= 1 \\ n! &= n \cdot (n-1)! \end{aligned}$$

Els **nombres de Fibonacci** sorgeixen en moltes situacions. Es poden definir recursivament de la següent manera:

$$\begin{aligned} f(0) &= 0 \\ f(1) &= 1 \\ f(n) &= f(n-1) + f(n-2) \end{aligned}$$

Els primers nombres de Fibonacci són

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \dots$$

També hi ha una fórmula tancada per calcular els nombres de Fibonacci, que de vegades s'anomena **fórmula de Binet**:

$$f(n) = \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n \sqrt{5}}.$$

Logaritmes

El **logaritme** d'un nombre x es denota per $\log_k(x)$, on k és la base del logaritme. Segons la definició, $\log_k(x) = a$ exactament quan $k^a = x$.

Una propietat útil dels logaritmes és que $\log_k(x)$ és igual al nombre de vegades hem de dividir x per k abans d'arribar el nombre 1. Per exemple, $\log_2(32) = 5$ perquè calen 5 divisions per 2:

$$32 \rightarrow 16 \rightarrow 8 \rightarrow 4 \rightarrow 2 \rightarrow 1$$

Els logaritmes s'utilitzen sovint en l'anàlisi de algorismes, perquè molts algorismes eficients redueixen alguna cosa a la meitat a cada pas. Per tant, podem estimar l'eficiència d'aquests algorismes fent servir logaritmes.

El logaritme d'un producte és

$$\log_k(ab) = \log_k(a) + \log_k(b),$$

i en conseqüència,

$$\log_k(x^n) = n \cdot \log_k(x).$$

A més, el logaritme d'un quocient és

$$\log_k\left(\frac{a}{b}\right) = \log_k(a) - \log_k(b).$$

Una altra fórmula útil és

$$\log_u(x) = \frac{\log_k(x)}{\log_k(u)},$$

i amb això, és possible calcular logaritmes a qualsevol base si hi ha una manera de calcular logaritmes en una base fixa.

El **logaritme natural** $\ln(x)$ d'un nombre x és el logaritme la base del qual és $e \approx 2.71828$. Una altra propietat dels logaritmes és que el nombre de dígitos d'un enter x en base b és $\lfloor \log_b(x) + 1 \rfloor$. Per exemple, la representació de 123 en base 2 és 1111011 i $\lfloor \log_2(123) + 1 \rfloor = 7$.

1.6 Concursos i recursos

IOI

L'Olimpíada Internacional d'Informàtica (IOI, International Olympiad in Informatics) és un concurs de programació anual per a alumnes de secundària. Cada país pot enviar un equip de quatre alumnes al concurs. Normalment hi ha uns 300 participants de 80 països.

L'IOI consta de dos concursos de cinc hores de durada. En ambdós concursos, es demana als participants resoldre tres tasques algorísmiques de diferent dificultat. Les tasques es divideixen en subtasques, cadascuna dels quals té una puntuació assignada. Encara que els concursants estiguin dividits en equips, competeixen com a individus.

El pla d'estudis de l'IOI [41] regula els temes que poden aparèixer a les tasques IOI. Gairebé tots els temes del temari de l'IOI són coberts per aquest llibre.

Els participants de l'IOI es seleccionen mitjançant concursos nacionals. Abans de l'IOI, s'organitzen molts concursos regionals, com l'Olimpíada Bàltica d'Informàtica (BOI), Olimpíada Centro-Europea d'Informàtica (CEOI) i l'Olimpíada d'Informàtica Àsia-Pacífic (APIO).

Alguns països organitzen concursos de pràctiques on-line per als futurs participants de l'IOI, com el Concurs Obert d'Informàtica de Croàcia [11] i l'Olimpíada d'Informàtica dels EUA [68]. A més, hi ha una gran col·lecció de problemes dels concursos polonesos disponibles on-line línia [60].

ICPC

El Concurs Internacional de Programació Col·legial (ICPC, International Collegiate Programming Contest) és un concurs de programació anual per a estudiants universitaris. Cada equip del concurs està format per tres estudiants, i a diferència de l'IOI, els alumnes treballen conjuntament; només hi ha un ordinador disponible per a cada equip.

L'ICPC consta de diverses fases, i només els millors equips estan convidats a la fase final (World Finals). Tot i que hi ha desenes de milers de participants al concurs, només hi ha un petit nombre² d'espais per equips, de manera que fins i tot avançar a la final ja és un gran èxit en algunes regions.

En cada concurs de l'ICPC, els equips disposen de cinc hores de temps per a resoldre uns deu problemes algorísmics. La solució d'un problema només s'accepta si es resolten tots els casos de prova de manera eficient. Durant el concurs, els competidors poden veure els resultats d'altres equips, però durant l'última hora el marcador està congelat i no és possible veure els resultats dels últims enviaments.

Els temes que poden aparèixer a l'ICPC no estan tan bé especificats com els de l'IOI. En tot cas, és evident que calen més coneixements a l'ICPC, sobretot més habilitats matemàtiques.

Concursos on-line

També hi ha molts concursos on-line oberts a tothom. En l'actualitat, la pàgina web amb més concursos actius és Codeforces, que organitza concursos de forma setmanal. A Codeforces, els participants es divideixen en dues divisions: els principiants competeixen a la Div2 i els programadors més experimentats a la Div1. Altres llocs de concurs inclouen AtCoder, CS Academy, HackerRank i Topcoder.

Algunes empreses organitzen concursos on-line amb finals presencials. Exemples d'aquests concursos són Facebook Hacker Cup, Google Code Jam i Yandex.Algorithm. Per descomptat, les empreses també utilitzen aquests concursos per a reclutar programadors: tenir un bon resultat en un concurs és una bona manera de demostrar les teves habilitats.

Llibres

Existeixen alguns llibres (a més d'aquest llibre) que es centren en la programació competitiva i la resolució de problemes algorísmics:

- S. S. Skiena i M. A. Revilla: *Programming Challenges: The Programming Contest Training Manual* [59]
- S. Halim i F. Halim: *Competitive Programming 3: The New Lower Bound of Programming Contests* [33]

²El nombre exacte d'equips a la fase final varia d'un any a un altre; el 2017, hi havia 133 equips.

- K. Diks et al.: *Looking for a Challenge? The Ultimate Problem Set from the University of Warsaw Programming Competitions* [15]

Els dos primers llibres estan pensats per a principiants, mentre que l'últim llibre conté material avançat.

Per descomptat, els llibres generals d'algorismia també són adequats per als programadors competitius. Alguns llibres populars són:

- T. H. Cormen, C. E. Leiserson, R. L. Rivest i C. Stein: *Introduction to Algorithms* [13]
- J. Kleinberg i É. Tardes: *Algorithm Design* [45]
- S. S. Skiena: *The Algorithm Design Manual* [58]

Capítol 2

Complexitat temporal

L'eficiència dels algorismes és important en la programació competitiva. Normalment, és fàcil dissenyar un algorisme que resol el problema lentament, però el veritable repte és inventar un algorisme que sigui ràpid. Si l'algorisme és massa lent només obtindrem punts parcials o cap punt.

La **complexitat temporal** d'un algorisme és el temps, o nombre d'operacions, que l'algorisme necessita per a resoldre entrades. La idea és representar l'eficiència com a una funció de la mida de l'entrada. Quan calculem el cost d'un algorisme podem esbrinar si serà prou ràpid sense implementar-lo.

2.1 Regles de càlcul

La complexitat temporal d'un algorisme es denota per $O(\dots)$ on els punts suspensius representen alguna funció.¹ Normalment, la variable n denota la mida de l'entrada. Per exemple, si l'entrada és una matriu de nombres, és usual prendre n com la mida de la matriu, i si l'entrada és una cadena de text, n serà la mida de la cadena.

Bucles

Sovint els algorismes són lents perquè contenen molts bucles que treballen amb l'entrada. Com més bucles niats (bucles dintre d'altres bucles) contingui l'algorisme, més lent és. Si hi ha k bucles niats, i cadascun d'ells fa n iteracions, la complexitat temporal és $O(n^k)$.

Per exemple, la complexitat temporal del codi següent és $O(n)$:

```
for (int i = 1; i <= n; i++) {  
    // codi  $O(1)$   
}
```

¹Afegit: la complexitat temporal d'un algorisme mesura com de ràpid creix el nombre d'operacions que l'algorisme executa com més fem créixer la mida de l'entrada. Formalment, la notació $O(f(n))$ és el conjunt de funcions que creixen com a molt tan ràpid com $f(n)$, llevat de factors constants: $O(f(n)) = \{g(n) | \exists n_0, c \forall n \geq n_0 \ g(n) \leq c \cdot f(n)\}$.

I la complexitat temporal del codi següent és $O(n^2)$:

```
for (int i = 1; i <= n; i++) {  
    for (int j = 1; j <= n; j++) {  
        // codi  $O(1)$   
    }  
}
```

Ordre de magnitud

La complexitat temporal no ens diu el nombre exacte de vegades que s'executa el codi dintre d'el bucle, sinó l'ordre de magnitud. En els exemples següents, el codi dins del bucle s'executa $3n$, $n + 5$ i $\lceil n/2 \rceil$ vegades, però tots ells tenen complexitat de $O(n)$.

```
for (int i = 1; i <= 3*n; i++) {  
    // codi  
}
```

```
for (int i = 1; i <= n+5; i++) {  
    // codi  
}
```

```
for (int i = 1; i <= n; i += 2) {  
    // codi  
}
```

En aquest altre exemple, la complexitat temporal del codi següent és $O(n^2)$:

```
for (int i = 1; i <= n; i++) {  
    for (int j = i+1; j <= n; j++) {  
        // codi  
    }  
}
```

Fases

Si l'algorisme consta de fases consecutives, la complexitat temporal és la complexitat temporal més gran de les fases. El motiu d'això és que la fase més lenta esdevé el coll d'ampolla del codi.

Per exemple, el següent codi consta de tres fases amb complexitats temporals $O(n)$, $O(n^2)$ i $O(n)$. Així, la complexitat total del temps és $O(n^2)$.

```
for (int i = 1; i <= n; i++) {  
    // codi
```

```

}
for (int i = 1; i <= n; i++) {
    for (int j = 1; j <= n; j++) {
        // codi
    }
}
for (int i = 1; i <= n; i++) {
    // codi
}

```

Diverses variables

De vegades, la complexitat temporal depèn de varis factors. En aquest cas, podem expressar la complexitat temporal com una fórmula de vàries variables.

Per exemple, la complexitat temporal del el codi següent és $O(nm)$:

```

for (int i = 1; i <= n; i++) {
    for (int j = 1; j <= m; j++) {
        // codi
    }
}

```

Recursió

La complexitat temporal d'una funció recursiva depèn del nombre de vegades que es crida la funció multiplicat per la complexitat temporal d'una única crida.

Per exemple, considereu la funció següent:

```

void f(int n) {
    if (n == 1) return;
    f(n-1);
}

```

La crida $f(n)$ provoca n crides a funcions, i la complexitat temporal de cadascuna d'aquestes crides (sense comptar la crida recursiva) és $O(1)$. Així, la complexitat total del temps és $O(n)$.

Com a altre exemple, considereu la funció següent:

```

void g(int n) {
    if (n == 1) return;
    g(n-1);
    g(n-1);
}

```

En aquest cas, cada crida a la funció genera dues altres crides, excepte quan $n = 1$. Vegem què passa quan es crida g amb el paràmetre n . La taula següent mostra el nombre de crides produït per aquesta crida original:

crida de funció	nombre de crides
$g(n)$	1
$g(n-1)$	2
$g(n-2)$	4
...	...
$g(1)$	2^{n-1}

En base a això, la complexitat temporal és

$$1 + 2 + 4 + \dots + 2^{n-1} = 2^n - 1 = O(2^n).$$

2.2 Clases de complexitat

El llistat següent conté les complexitats temporals més comuns dels algorismes:

$O(1)$ El temps d'execució d'un algorisme de **temps constant** no depèn de la mida d'entrada. Un algorisme típic de temps constant és una fórmula que calcula la resposta directament.

$O(\log n)$ Un algorisme **logarítmic** sovint divideix cada entrada per la meitat a cada pas. El seu cost és logarítmic, perquè $\log_2 n$ és igual al nombre de vegades que és necessari dividir n per 2 fins obtenir 1.

$O(\sqrt{n})$ Un **algorisme d'arrel quadrada** és més lent que $O(\log n)$ però més ràpid que $O(n)$. Una propietat especial de les arrels quadrades és que $\sqrt{n} = n/\sqrt{n}$, de manera que l'arrel quadrada \sqrt{n} està, en certa manera, a la meitat (geomètrica) de l'entrada.

$O(n)$ Un algorisme **linial** és aquell que recorre l'entrada un nombre constant de vegades. Sovint aquesta és la millor complexitat temporal possible en els concursos, perquè normalment cal accedir a cada element de l'entrada com a mínim una vegada abans de produir la resposta.

$O(n \log n)$ Aquest cost sovint indica que l'algorisme ordena l'entrada, perquè aquest és el cost dels algorismes d'ordenació eficients. Una altra possibilitat és que l'algorisme utilitzi una estructura de dades on cada operació triga $O(\log n)$ temps.

$O(n^2)$ Un algorisme **quadràtic** sovint conté dos bucles niats un dintre de l'altre. És possible passar per totes les parelles d'elements de l'entrada en temps $O(n^2)$.

$O(n^3)$ Un algorisme **cúbic** sovint conté tres bucles niats. És possible passar per totes les tripletes d'elements de l'entrada en temps $O(n^3)$.

$O(2^n)$ Aquesta complexitat de temps sovint indica que l'algorisme itera per tot els subconjunts dels elements d'entrada. Per exemple, els subconjunts de $\{1, 2, 3\}$ són \emptyset , $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$ i $\{1, 2, 3\}$.

$O(n!)$ Aquesta complexitat de temps sovint indica l'algorisme itera per totes les permutacions dels elements d'entrada. Per exemple, les permutacions de $\{1, 2, 3\}$ són $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$ i $(3, 2, 1)$.

Un algorisme és **polinòmic** quan el seu cost és $O(n^k)$ per alguna constant k . Tots els costos anteriors, exceptuant $O(2^n)$ i $O(n!)$, són polinòmics. A la pràctica, la constant k sol ser petita, i sovint s'associen costos polinòmics amb que l'algorisme és *eficient*.

La majoria dels algorismes d'aquest llibre són polinòmics. Tot i així, hi ha molts problemes importants per als quals no es coneix cap algorisme polinòmic, és a dir, ningú sap com resoldre'ls de manera eficient. Els problemes **NP-hard** són un conjunt important de problemes, per als quals no hi ha cap algorisme polinòmic conegut².

2.3 Estimació de l'eficiència

Quan calculem la complexitat temporal d'un algorisme podem comprovar, abans d'implementar-lo, si és prou eficient per al problema. El punt de partida per les estimacions és el fet que un ordinador modern pot realitzar alguns centenars de milions d'operacions en un segon.

Per exemple, considerem un problema on el límit de temps és d'un segon i la mida d'entrada és $n = 10^5$. Si la complexitat temporal és $O(n^2)$, l'algorisme realitzarà unes $(10^5)^2 = 10^{10}$ operacions. Això hauria de trigar almenys unes desenes de segons, de manera que l'algorisme probablement sigui massa lent per resoldre el problema.

D'altra banda, donada la mida de l'entrada, podem intentar *endevinar* la complexitat de l'algorisme que hem de programar per a resoldre el problema. La taula següent conté algunes estimacions útils suposant un límit de temps d'un segon.³

mida d'entrada	complexitat esperada
$n \leq 10$	$O(n!)$
$n \leq 20$	$O(2^n)$
$n \leq 500$	$O(n^3)$
$n \leq 5000$	$O(n^2)$
$n \leq 10^6$	$O(n \log n)$ o $O(n)$
n és gran	$O(1)$ o $O(\log n)$

Per exemple, si la mida de l'entrada és $n = 10^5$, segurament la solució del problema sigui un algorisme $O(n)$ o $O(n \log n)$. Aquesta informació facilita el disseny de l'algorisme, perquè descarta enfocaments que donarien un algorisme amb un cost massa gran.

²Un llibre clàssic sobre el tema és de M. R. Garey i D. S. Johnson *Informàtica i intractabilitat: una guia per a la teoria de NP-Compleitud* [28].

³No totes les operacions costen el mateix: per exemple, les operacions d'entrada i sortida són més costoses que les operacions aritmètiques.

Tot i així, és important recordar que la complexitat temporal és només una estimació de l'eficiència, perquè amaga els *factors constants*. Per exemple, un algorisme que s'executa en $O(n)$ temps potser fa $n/2$ o $5n$ operacions. Això és un factor important de cara al temps real que farà servir l'algorisme.

2.4 Suma màxima d'un subvector

Sovint hi ha diversos algorismes que poden resoldre un problema i que tenen distintes complexitats temporals. Aquesta secció tracta d'un problema clàssic que té una solució $O(n^3)$ senzilla. Tanmateix, dissenyant un algorisme millor, és possible resoldre el problema en temps $O(n^2)$ temps o fins i tot en temps $O(n)$.

Donat un vector de n nombres, la nostra tasca és calcular el **suma màxima d'un subvector**, és a dir, la suma més gran possible d'una seqüència de valors consecutius del vector⁴. El problema és interessant quan hi ha valors negatius en el vector. Per exemple, en el vector

-1	2	4	-3	5	2	-5	2
----	---	---	----	---	---	----	---

el subvector següent té suma màxima 10:

-1	2	4	-3	5	2	-5	2
----	---	---	----	---	---	----	---

Suposem que es permet triar un subvector buit, de manera que la suma màxima és sempre com a mínim 0.

Algorisme 1

Una manera senzilla de resoldre el problema és passar per tots els subvectors possibles, calcular la suma de valors de cada subvector i mantenir la suma màxima. El codi següent implementa aquest algorisme:

```
int millor = 0;
for (int a = 0; a < n; a++) {
    for (int b = a; b < n; b++) {
        int suma = 0;
        for (int k = a; k <= b; k++) {
            suma += v[k];
        }
        millor = max(millor, suma);
    }
}
cout << millor << "\n";
```

Les variables a i b contenen el primer i el darrer índex del subvector, i la suma de valors es calcula a la variable $suma$. La variable $millor$ conté la suma màxima trobada durant la cerca.

⁴El llibre *Programming Pearls* [8] de J. Bentley va popularitzar aquest problema.

La complexitat temporal de l'algorisme és $O(n^3)$, perquè consta de tres bucles niats que recorren l'entrada.

Algorisme 2

És fàcil fer que l'algorisme 1 sigui més eficient eliminant-ne un bucle. Això és possible si calculem la suma a la vegada que moguem el darrer índex del subvector. El resultat és el codi següent:

```
int millor = 0;
for (int a = 0; a < n; a++) {
    int suma = 0;
    for (int b = a; b < n; b++) {
        suma += v[b];
        millor = max(millor, suma);
    }
}
cout << millor << "\n";
```

Després d'aquest canvi, la complexitat temporal és $O(n^2)$.

Algorisme 3

Sorprenentment, és possible⁵ resoldre el problema en temps $O(n)$, que significa que n'hi ha prou amb un sol bucle. La idea és calcular, per a cada posició del vector, la suma màxima d'un subvector que acaba en aquesta posició. Després d'això, la resposta al problema és el màxim d'aquestes sumes.

Considereu el subproblema de trobar la suma màxima del subvector que acaba a la posició k . Hi ha dues possibilitats:

1. El subvector només conté l'element a la posició k .
2. El subvector consisteix en un subvector que acaba a la posició $k - 1$, seguit de l'element a la posició k .

En aquest darrer cas, ja que volem trobar un subvector amb suma màxima, el subvector que acaba a la posició $k - 1$ també ha de tenir suma màxima. Així, podem resoldre el problema de manera eficient calculant la suma màxima del subvector per a cada posició final d'esquerra a dreta.

El codi següent implementa l'algorisme:

```
int millor = 0, suma = 0;
for (int k = 0; k < n; k++) {
    suma = max(v[k], suma+v[k]);
    millor = max(millor, suma);
}
cout << millor << "\n";
```

⁵En [8], aquest algorisme de temps lineal s'atribueix a J. B. Kadane, i l'algorisme de vegades s'anomena **algorisme de Kadane**.

L'algorisme només conté un bucle que passa per l'entrada, per tant, la complexitat temporal és $O(n)$. Aquesta és també la millor complexitat temporal possible, perquè qualsevol algorisme per aquest problema ha d'examinar tots els elements del vector almenys una vegada.

Comparació d'eficiència

És interessant estudiar com d'eficients són els algorismes a la pràctica. La taula següent mostra els temps de funcionament dels algorismes anteriors per a diferents valors de n en un ordinador modern.

En cada prova, vam generar l'entrada aleatòriament, i no vam mesurar el temps necessari per llegir l'entrada.

mida del vector n	algorisme 1	algorisme 2	algorisme 3
10^2	0.0 s	0.0 s	0.0 s
10^3	0.1 s	0.0 s	0.0 s
10^4	> 10.0 s	0.1 s	0.0 s
10^5	> 10.0 s	5.3 s	0.0 s
10^6	> 10.0 s	> 10.0 s	0.0 s
10^7	> 10.0 s	> 10.0 s	0.0 s

La comparació mostra que tots els algorismes són eficients quan la mida de l'entrada és petita, però les entrades més grans mostren diferències notables entre els temps d'execució dels algorismes. L'algorisme 1 es torna lent quan $n = 10^4$, i l'algorisme 2 es torna lent quan $n = 10^5$. Només l'algorisme 3 pot processar fins i tot les entrades més grans instantàniament.

Capítol 3

Ordenació

El problema de l'**ordenació** és un problema fonamental del disseny d'algorismes. Molts algorismes eficients fan servir l'ordenació com a subrutina, perquè sovint és més fàcil de processar dades si els elements estan ordenats.

Per exemple, el problema de “té un vector dos elements iguals?” és fàcil de resoldre mitjançant l'ordenació. Si el vector conté dos elements iguals, després d'ordenar-los estaran l'un al costat de l'altre, de manera que és fàcil trobar-los. El problema “quin és l'element més freqüent d'un vector?” es pot resoldre de la mateixa manera.

Hi ha molts algorismes per ordenar, i aquests són també bons exemples de com aplicar diferents tècniques del disseny d'algorismes. Els algorismes d'ordenació general eficients treballen en temps $O(n \log n)$, i molts algorismes que utilitzen l'ordenació com a subrutina tenen també aquesta complexitat temporal.

3.1 Teoria de l'ordenació

El problema bàsic de l'ordenació és el següent:

Donat un vector que conté n elements, ordena els elements en ordre creixent.

Per exemple, el vector

1	3	8	2	9	2	5	6
---	---	---	---	---	---	---	---

queda de la manera següent després d'ordenar-lo:

1	2	2	3	5	6	8	9
---	---	---	---	---	---	---	---

Algorismes $O(n^2)$

Alguns algorismes senzills per ordenar un vector treballen en temps $O(n^2)$. Aquests algorismes són curts i generalment consten de dos bucles niats. Un

algorisme famós amb complexitat $O(n^2)$ és **l'ordenació amb bombolla (bubble sort)** on els elements del vector es mouen com si fossin “bombolles”.

L'ordenació amb bombolla consta de n rondes. A cada ronda, l'algorisme itera els elements del vector. Sempre que es troben dos elements consecutius que no estan en ordre correcte, l'algorisme els intercanvia. L'algorisme es pot implementar de la següent manera:

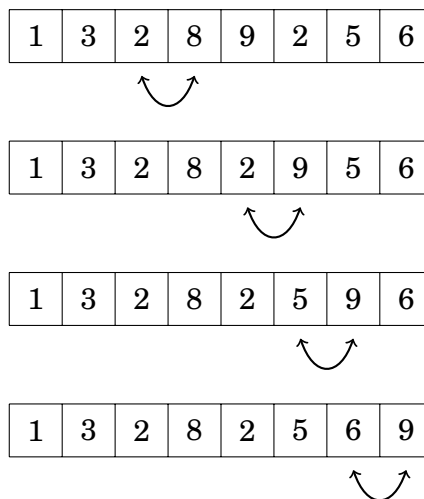
```
for (int i = 0; i < n; i++) {  
    for (int j = 0; j < n-1; j++) {  
        if (v[j] > v[j+1]) {  
            swap(v[j], v[j+1]);  
        }  
    }  
}
```

Després de la primera ronda de l'algorisme, l'element més gran estarà en la posició correcta, i en general, després de k rondes, els k elements més grans estaran en les posicions correctes. Així, després de n rondes, el vector quedarà ordenat.

Per exemple, en el vector

1	3	8	2	9	2	5	6
---	---	---	---	---	---	---	---

la primera ronda d'ordenació amb bombolla intercanvia els elements com segueix:



Inversions

L'ordenació amb bombolla és un exemple d'algorisme d'ordenació que sempre intercanvia elements *consecutius* del vector. La complexitat temporal d'aquest algorisme és $O(n^2)$, donat que aquest és el nombre de comparacions que es fan. (I, en el pitjor cas, és el nombre de swaps).

Un concepte útil a l'hora d'analitzar l'ordenació algorismes és el concepte d'**inversió**, que són els parells d'elements situats en l'ordre incorrecte, és a dir, els parells $(v[a], v[b])$ tal que $a < b$ i $v[a] > v[b]$. Per exemple, el vector

1	2	2	6	3	5	9	8
---	---	---	---	---	---	---	---

té tres inversions: (6, 3), (6, 5) i (9, 8). El nombre d'inversions indica quanta feina és necessària per a ordenar el vector. Un vector està completament ordenat quan no hi ha inversions. D'altra banda, si tots els elements del vector estan en ordre invers, el nombre d'inversions és el més gran possible:

$$1 + 2 + \dots + (n - 1) = \frac{n(n - 1)}{2} = O(n^2)$$

Intercanviar un parell d'elements consecutius en ordre incorrecte elimina exactament una inversió del vector. Per tant, si un algorisme d'ordenació només intercanvia elements consecutius, cada intercanvi elimina com a màxim una inversió i la complexitat temporal de l'algorisme és almenys $O(n^2)$.

Algorismes $O(n \log n)$

És possible ordenar un vector de manera eficient en temps $O(n \log n)$ fent servir algorismes que no es limiten a intercanviar elements consecutius. Un d'aquests algorismes és **merge sort**¹, que es basa en la recursivitat.

El merge sort ordena un subvector $v[a \dots b]$ de la següent manera:

1. Si $a = b$, no feu res, perquè el subvector ja està ordenat.
2. Calcula la posició de l'element central: $k = \lfloor (a + b)/2 \rfloor$.
3. Ordena recursivament el subvector $v[a \dots k]$.
4. Ordena recursivament el subvector $v[k + 1 \dots b]$.
5. Fusiona (*merge*) els subvectors ordenats $v[a \dots k]$ i $v[k + 1 \dots b]$ en un subvector ordenat $v[a \dots b]$.

El merge sort és un algorisme eficient, perquè cada pas redueix a la meitat la mida del subvector. La recursivitat consisteix en $O(\log n)$ nivells, i cada nivell triga temps $O(n)$. És possible fusionar i ordenar els subvectors $v[a \dots k]$ i $v[k + 1 \dots b]$ en temps lineal perquè ja estan ordenats.

Per exemple, considerem el vector següent:

1	3	6	2	8	2	5	9
---	---	---	---	---	---	---	---

Dividim el vector en dos subvectors:

1	3	6	2
8	2	5	9

Ordenem els subvectors de manera recursiva:

1	2	3	6
2	5	8	9

Finalment, l'algorisme fusiona els subvectors ordenats i crea el vector final:

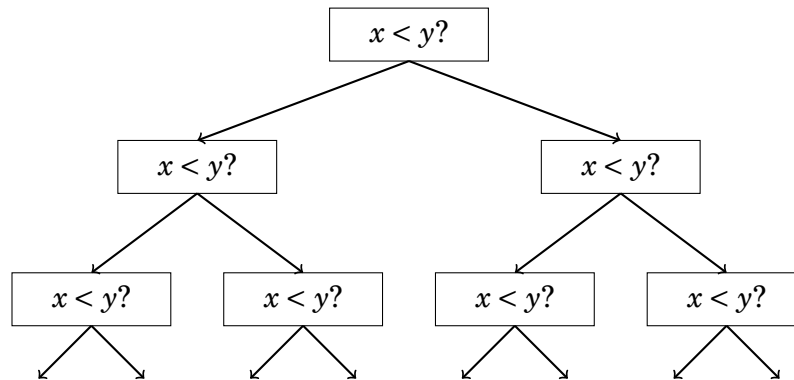
1	2	2	3	5	6	8	9
---	---	---	---	---	---	---	---

¹Segons [47], el merge sort va ser inventat per J. von Neumann el 1945.

Límit inferior per l'ordenació

És possible ordenar més ràpidament que en temps $O(n \log n)$? Resulta que això *no* és possible quan fem servir algorismes d'ordenació que es basen en comparar els elements d'un vector.

Aquest límit inferior $O(n \log n)$ es demostra si hom considera l'ordenació com un procés on cada comparació de dos elements proporciona més informació sobre el contingut del vector original. El procés crea el següent arbre:



Aquí “ $x < y$?” significa que comparem alguns elements x i y . Si $x < y$, el procés continua cap a l'esquerra, i en cas contrari cap a la dreta. Els resultats d'aquest procés són les $n!$ maneres en que els n elements poden aparèixer en el vector. Per aquest motiu, l'alçada de l'arbre ha de ser almenys

$$\log_2(n!) = \log_2(1) + \log_2(2) + \dots + \log_2(n).$$

Obtenim un límit inferior per a aquesta suma escollint els últims $n/2$ elements i canviant el valor de cada element per $\log_2(n/2)$. Això dóna una estimació

$$\log_2(n!) \geq (n/2) \cdot \log_2(n/2),$$

de manera que l'alçada de l'arbre i el mínim nombre possible de passos en un algorisme d'ordenació és $O(n \log n)$.

Ordenació per comptatge

El límit inferior $n \log n$ no s'aplica a algorismes que no comparen els elements del vector però que utilitzen altra informació. Un exemple d'aquests algorismes és l'ordenació per comptatge (**counting sort**) que ordena un vector en temps $O(n)$ assumint que tots els elements del vector són enters entre $0 \dots c$ i $c = O(n)$.

L'algorisme crea un vector addicional de mida c , els índexs del qual són els elements del vector original. L'algorisme itera a través del vector original i calcula quantes vegades cada element apareix al vector.

Per exemple, el vector

1	3	6	9	9	3	5	9
---	---	---	---	---	---	---	---

li correspon el següent vector de comptatge:

1	2	3	4	5	6	7	8	9
1	0	2	0	1	1	0	0	3

Per exemple, el valor a la posició 3 del vector de comptatge és 2, perquè l'element 3 apareix 2 vegades al vector original.

La construcció del vector de comptatge triga temps $O(n)$. Després d'això, el vector ordenat es crea en temps $O(n)$ perquè recuperem el nombre d'ocurrències de cada element en el vector de comptatge. Per tant, la complexitat temporal total és $O(n)$.

L'ordenació per comptatge és un algorisme molt eficient però només es pot fer servir quan el valor c és prou petit, ja que en cas contrari no podem fer servir els elements del vector com a índexs del vector de comptatge.

3.2 Ordenació en C++

Gairebé mai és bona idea fer servir algorismes d'ordenació casolans en un concurs, perquè els llenguatges de programació ja venen amb molt bones implementacions. Per exemple, la biblioteca estàndard de C++ conté la funció `sort` que es pot fer servir per a ordenar vectors i altres estructures de dades.

Hi ha molts avantatges en fer servir les funcions de les biblioteques. Primer, estalvia temps perquè no cal implementar la funció. Segon, la implementació de la biblioteca és certament correcta i eficient: no és probable que la teva funció de classificació casolana sigui millor.

En aquesta secció veurem com fer servir la funció C++ `sort`. El codi següent ordena un vector en ordre creixent:

```
vector<int> v = {4,2,5,3,5,8,3};  
sort(v.begin(),v.end());
```

Després de l'ordenació, el contingut del vector és `[2,3,3,4,5,5,8]`. Els elements s'ordenen per defecte en ordre creixent, però també és possible ordenar-los en ordre invers:

```
sort(v.rbegin(),v.rend());
```

També és possible ordenar un array ordinari de C:

```
int n = 7; // mida del array  
int a[] = {4,2,5,3,5,8,3};  
sort(a,a+n);
```

El codi següent ordena la cadena `s`:

```
string s = "monkey";  
sort(s.begin(), s.end());
```

Ordenar una cadena significa que ordenem tots els seus caràcters. Per exemple, la cadena “monkey” es converteix en “ekmnoy”.

Operadors de comparació

La funció `sort` requereix que es defineix un **operador de comparació** per al tipus de dades dels elements a ordenar. Quan ordenem, es farà servir aquest operador sempre que sigui necessari esbrinar l'ordre de dos elements.

La majoria dels tipus de dades C++ tenen un operador de comparació definit per defecte, i els elements d'aquests tipus es poden ordenar automàticament. Per exemple, els nombres s'ordenen segons els seus valors i les cadenes estan ordenades per ordre alfabètic.

Els parells (`pair`) s'ordenen en funció dels seus primers elements (`first`). Si els primers elements dels dos parells són iguals, aleshores s'ordenen segons els segons elements (`second`):

```
vector<pair<int,int>> v;  
v.push_back({1,5});  
v.push_back({2,3});  
v.push_back({1,2});  
sort(v.begin(), v.end());
```

Després d'això, l'ordre de les parelles és (1,2), (1,5) i (2,3).

De manera semblant, les tuples (`tuple`) s'ordenen pel primer element, seguit pel segon element en cas d'empat, el tercer, etc.²:

```
vector<tuple<int,int,int>> v;  
v.push_back({2,1,4});  
v.push_back({1,5,3});  
v.push_back({2,1,3});  
sort(v.begin(), v.end());
```

Després d'això, l'ordre de les tuples és (1,5,3), (2,1,3) i (2,1,4).

Aquesta idea de les tuples també s'estén als vectors.

Estructures definides per l'usuari

Les estructures definides per l'usuari no tenen un operador de comparació definit per defecte. Una manera de fer-ho és definir l'operador a dintre de la estructura de dades com a funció `operator<`, el paràmetre del qual és un altre element del mateix tipus. L'operador hauria de retornar `true` si l'element és més petit que el paràmetre, i `false` en cas contrari.

Per exemple, l'estructura `P` següent conté les coordenades x i y d'un punt. L'operador de comparació es defineix de manera que els punts s'ordenen per la coordenada x i, en cas d'empat, per la coordenada y .

²Tingueu en compte que en alguns compiladors antics, és necessari fer servir la funció `make_tuple` en lloc de les claus (per exemple, `make_tuple(2,1,4)` en lloc de `{2,1,4}`).

```
struct P {
    int x, y;
    operador bool<(const P &p) {
        if (x != p.x) return x < p.x;
        else return y < p.y;
    }
};
```

Funcions de comparació

També és possible donar una **funció de comparació** externa (callback) a la funció `sort`. Per exemple, la següent funció de comparació `comp` ordena les cadenes primer per longitud i segon per ordre alfabètic:

```
bool comp(const string& a, const string& b) {
    if (a.size() != b.size()) return a.size() < b.size();
    return a < b;
}
```

Ara podem ordenar un vector de cadenes així:

```
sort(v.begin(), v.end(), comp);
```

N. del T.: També podem fer servir expressions lambda:

```
sort(v.begin(), v.end(), [](const string& a, const string& b) {
    return (a.size() != b.size()) ? (a.size() < b.size()) : (a < b);
});
```

3.3 Cerca binària

Un mètode general per cercar un element en un vector és fer servir un bucle `for` que iteri els elements del vector. Per exemple, el codi següent cerca un element x en el vector v :

```
for (int i = 0; i < n; i++) {
    if (v[i] == x) {
        // trobem x a la posicio i
    }
}
```

La complexitat temporal d'aquest codi és $O(n)$, perquè en el pitjor dels casos, cal comprovar tots els elements del vector. Si l'ordre dels elements és arbitrari, aquesta és també la millor solució, perquè no tenim informació addicional per saber en quin lloc del vector hem de cercar l'element x .

Tanmateix, si el vector està *ordenat*, la situació és diferent. En aquest cas és possible realitzar el cerca molt més ràpid, perquè l'ordre dels elements del vector guia la cerca. El següent algorisme de **cerca binària** cerca eficientment un element en un vector ordenat en temps $O(\log n)$.

Mètode 1

La forma habitual d'implementar la cerca binària s'assembla a buscar una paraula en un diccionari. La cerca manté una regió activa al vector, que inicialment conté tots els elements del vector. Aleshores, es fa una sèrie de passos, cadascun dels quals redueix a la meitat la mida de la regió activa.

A cada pas, la cerca comprova l'element central de la regió activa. Si l'element central és l'element objectiu, la cerca acaba. En cas contrari, la cerca continua de forma recursiva a la meitat esquerra o dreta de la regió, en funció del valor de l'element mitjà.

La idea anterior es pot implementar de la següent manera:

```
int a = 0, b = n-1;
while (a <= b) {
    int k = (a+b)/2;
    if (v[k] == x) {
        // trobem x a la posicio k
    }
    if (v[k] > x) b = k-1;
    else a = k+1;
}
```

En aquesta implementació, la regió activa és $a \dots b$, i la regió inicial és $0 \dots n-1$. L'algorisme redueix a la meitat la mida de la regió a cada pas, per tant, la complexitat temporal és $O(\log n)$.

Mètode 2

Un mètode alternatiu per implementar la cerca binària es basa en una manera eficient d'iterar els elements del vector. La idea és fer salts i reduir la velocitat quan ens acostem a l'element objectiu.

La cerca passa per el vector d'esquerra a dreta, i la longitud inicial del salt és $n/2$. A cada pas, la longitud del salt es reduirà a la meitat: primer $n/4$, després $n/8$, $n/16$, etc., fins que finalment la longitud és 1. Després dels salts, l'element objectiu o bé s'ha trobat o bé sabem que no apareix al vector.

El codi següent implementa la idea anterior:

```
int k = 0;
for (int b = n/2; b >= 1; b /= 2) {
    while (k+b < n && v[k+b] <= x) k += b;
}
if (v[k] == x) {
    // trobem x a la posicio x
}
```

```
}
```

Durant la cerca, la variable b conté la longitud actual del salt. La complexitat temporal és també $O(\log n)$, perquè el codi del bucle `while` es realitza com a màxim dues vegades per a cada llargada de salt.

Funcions C++

La biblioteca estàndard de C++ conté les funcions següents que es basen en cerca binària i triguen temps logarítmic:

- `lower_bound` retorna un punter al primer element del vector el valor del qual és almenys x .
- `upper_bound` retorna un punter a primer element del vector el valor del qual és més gran que x .
- `equal_range` retorna els dos punters anteriors.

Les funcions assumeixen que el vector està ordenat. Si aquest primer element no existeix, les funcions retornen un punter a una posició més enllà de l'últim element del vector³. Per exemple, el codi següent esbrina si un vector ordenat conté un element amb el valor x :

```
int k = lower_bound(v.begin(), v.end(), x) - v.begin();
if (k < n && v[k] == x) {
    // trobem x a la posicio k
}
```

El codi següent compta el nombre d'elements amb valor x :

```
auto a = lower_bound(v.begin(), v.end(), x);
auto b = upper_bound(v.begin(), v.end(), x);
cout << b-a << "\n";
```

Utilitzant `equal_range`, el codi queda més curt:

```
auto r = equal_range(v.begin(), v.end(), x);
cout << r.second-r.first << "\n";
```

Trobar la solució més petita

Un ús important de la cerca binària és trobar la posició on canvia el valor d'una funció. Suposem que volem trobar el valor més petit k que és una solució vàlida per a un problema. Ens donen una funció `ok(x)` que retorna `true` si x és vàlida i `false` en cas contrari. A més, sabem que `ok(x)` és `false` quan $x < k$ i `true` quan $x \geq k$. És a dir:

³N. del T: Dit d'altra manera: l'interval semi-obert $[lower_bound(x), upper_bound(x))$ assenya-la en quin lloc són, o haurien de ser, els elements amb valor x .

x	0	1	...	$k-1$	k	$k+1$...
$ok(x)$	f	f	...	f	t	t	...

Ara, el valor de k es pot trobar mitjançant la cerca binària:

```
int x = -1;
for (int b = z; b >= 1; b /= 2) {
    while (!ok(x+b)) x += b;
}
int k = x+1;
```

< La cerca troba el valor més gran de x per al qual $ok(x)$ és false. Així, el següent valor $k = x + 1$ és necessàriament el valor més petit possible per al qual $ok(k)$ és true. La longitud inicial del salt z ha de ser prou gran, per exemple algun valor per al qual sabem per endavant que $ok(z)$ és true.

L'algorisme crida a la funció ok $O(\log z)$ vegades, per tant la complexitat total depèn de la funció ok . Per exemple, si aquesta funció triga temps $O(n)$, la complexitat total és $O(n \log z)$.

Trobar el valor màxim

La cerca binària també es pot fer servir per trobar el valor màxim d'una funció que primer creix i després decreix. La nostra tasca és trobar una posició k tal que

- $f(x) < f(x+1)$ quan $x < k$, i
- $f(x) > f(x+1)$ quan $x \geq k$.

La idea és fer servir la cerca binària per trobar el valor més gran de x tal que $f(x) < f(x+1)$. Això implica que $k = x + 1$ perquè $f(x+1) > f(x+2)$. El codi següent implementa la cerca:

```
int x = -1;
for (int b = z; b >= 1; b /= 2) {
    while (f(x+b) < f(x+b+1)) x += b;
}
int k = x+1;
```

Tingueu en compte que, a diferència de la cerca binària ordinària, en aquest cas no permetem que els valors consecutius de la funció siguin iguals, donat que no sabríem en quina direcció continuar la cerca.

Capítol 4

Estructures de dades

Una **estructura de dades** és una manera d'emmagatzemar dades a la memòria d'un ordinador. És important escollir l'estructura de dades adequada per a un problema, perquè cada estructura de dades té els seus avantatges i inconvenients. La pregunta crucial és: quines operacions són eficients en l'estructura de dades escollida?

Aquest capítol presenta les estructures de dades més importants a la biblioteca estàndard de C++. És molt bona idea fer servir la biblioteca estàndard sempre que sigui possible, perquè ens estalviarà molt de temps. Més endavant parlarem d'estructures de dades més sofisticades no estan disponibles a la biblioteca estàndard.

4.1 Vectors dinàmics

Un **vector dinàmic** és un vector la mida del qual pot canviar durant l'execució del programa. El vector dinàmic més popular en C++ és l'estructura `vector`, que també pot fer-se servir gairebé com un array de C normal.

El codi següent crea un vector buit i hi afegeix tres elements:

```
vector<int> v;  
v.push_back(3); // [3]  
v.push_back(2); // [3,2]  
v.push_back(5); // [3,2,5]
```

Després d'això, podem accedir als elements com si fos un array de C normal:

```
cout << v[0] << "\n"; // 3  
cout << v[1] << "\n"; // 2  
cout << v[2] << "\n"; // 5
```

La funció `size` retorna el nombre d'elements del vector. El codi següent itera el vector i escriu tots els elements:

```
for (int i = 0; i < v.size(); i++) {  
    cout << v[i] << "\n";  
}
```

```
}
```

Una manera més concisa d'iterar un vector és la següent:

```
for (auto x : v) {  
    cout << x << "\n";  
}
```

La funció `back` retorna l'últim element en el vector, i la funció `pop_back` elimina l'últim element:

```
vector<int> v;  
v.push_back(5);  
v.push_back(2);  
cout << v.back() << "\n"; // 2  
v.pop_back();  
cout << v.back() << "\n"; // 5
```

El codi següent crea un vector de cinc elements:

```
vector<int> v = {2,4,2,5,1};
```

Una altra manera de crear un vector és donar el nombre d'elements i el valor inicial de cada element:

```
// mida 10, valor inicial 0  
vector<int> v(10);
```

```
// mida 10, valor inicial 5  
vector<int> v(10, 5);
```

La implementació interna d'un vector fa servir un array ordinari. Si la mida del vector augmenta i l'espai reservat és massa petit, es crea un nou array i tot els elements es mouen al nou espai. Però això no passa sovint, i la complexitat temporal promig de fer un `push_back` és $O(1)$.

L'estructura cadena (`string`) és un vector dinàmic de caràcters. A més, hi ha una sintaxi especial per a les cadenes que no està disponible en les altres estructures de dades. Les cadenes es poden concatenar fent servir el símbol `+`. La funció `substr(k,x)` retorna la subcadena que comença a la posició *k* i té longitud *x*, i la funció `find(t)` troba la posició de la primera ocurrència d'una subcadena *t*.

El codi següent presenta algunes operacions de cadenes:

```
string a = "hatti";  
string b = a+a;  
cout << b << "\n"; // hattihatti  
b[5] = 'v';  
cout << b << "\n"; // hattivatti  
string c = b.substr(3,4);
```



```
cout << c << "\n"; // tiva
```

4.2 Estructures conjunt

Un **conjunt** és una estructura de dades que manté una col·lecció d'elements. Les operacions bàsiques dels conjunts són inserir, cercar i eliminar.

La biblioteca estàndard de C++ conté dues implementacions: L'estructura `set` es basa en arbres binari equilibrats i les seves operacions triguen $O(\log n)$ en cas pitjor. L'estructura `unordered_set` fa serving hashing, i les seves operacions triguen $O(1)$ en mitjana.

Quina implementació triar és sovint qüestió de gustos. El benefici de l'estructura `set` és que manté l'ordre dels elements i ofereix funcions que no estan disponibles a `unordered_set`. D'altra banda, `unordered_set` pot ser més eficient.

El codi següent crea un conjunt que conté nombres enters, i mostra algunes de les operacions. La funció `insert` afegeix un element al conjunt, la funció `count` retorna el nombre d'ocurrències d'un element del conjunt, i la funció `erase` elimina un element del conjunt.

```
set<int> s;  
s.insert(3);  
s.insert(2);  
s.insert(5);  
cout << s.count(3) << "\n"; // 1  
cout << s.count(4) << "\n"; // 0  
s.erase(3);  
s.insert(4);  
cout << s.count(3) << "\n"; // 0  
cout << s.count(4) << "\n"; // 1
```

Un conjunt es pot fer servir com un vector, però no es pot accedir als elements fent servir la notació `[]`. El codi següent crea un conjunt, escriu el nombre d'elements, i itera per tots els elements:

```
set<int> s = {2,5,6,8};  
cout << s.size() << "\n"; // 4  
for (auto x : s) {  
    cout << x << "\n";  
}
```

Una propietat important dels conjunts és que tots els seus elements són *diferents*. Així, la funció `count` sempre retorna 0 (l'element no està al conjunt) o 1 (l'element és al conjunt), i la funció `insert` no s'afegeix mai un element al conjunt si ja hi és. Això es pot veure al codi següent:

```
set<int> s;  
s.insert(5);  
s.insert(5);
```

```
s.insert(5);
cout << s.count(5) << "\n"; // 1
```

C++ també conté les estructures `multiset` i `unordered_multiset`, que funcionen com `set` i `unordered_set` excepte que poden contenir vàries instàncies d'un element. Per exemple, al codi següent afegeix tres instàncies del número 5 a un multiconjunt:

```
multiset<int> s;
s.insert(5);
s.insert(5);
s.insert(5);
cout << s.count(5) << "\n"; // 3
```

La funció `erase` elimina totes les instàncies d'un element d'un multiconjunt:

```
s.erase(5);
cout << s.count(5) << "\n"; // 0
```

Sovint només volem eliminar una instància, que podem fer de la següent manera:

```
s.erase(s.find(5));
cout << s.count(5) << "\n"; // 2
```

4.3 Estructures mapa

Un **mapa** és un vector que consisteix de parells clau-valor. Mentre que les claus d'un vector normal sempre són els nombres enters consecutius $0, 1, \dots, n-1$, on n és la mida del vector, les claus d'un mapa poden ser de qualsevol tipus de dades i no tenen perquè ser consecutius.

La biblioteca estàndard de C++ conté dues implementacions de mapa que es corresponen amb les implementacions de conjunts: l'estructura `map` es basa en un arbre binari equilibrat i accedir als elements triga $O(\log n)$, mentre que l'estructura `unordered_map` fa servir hashing i accedir als elements triga $O(1)$ de mitjana.

El codi següent crea un mapa on les claus són cadenes i els valors són nombres enters:

```
map<string,int> m;
m["monkey"] = 4;
m["banana"] = 3;
m["harpsichord"] = 9;
cout << m["banana"] << "\n"; // 3
```

Si es demana el valor d'una clau però el mapa no la conté, la clau s'afegeix automàticament al mapa amb un valor per defecte. Per exemple, en el codi següent, la clau "aybaltu" s'afegeix al mapa amb valor 0.

```
map<string,int> m;  
cout << m["aybabbu"] << "\n"; // 0
```

La funció count comprova si el mapa conté una clau:

```
if (m.count("aybabbu")) {  
    // la clau existeix  
}
```

El codi següent escriu totes les claus i valors d'un mapa:

```
for (auto x : m) {  
    cout << x.first << " " << x.second << "\n";  
}
```

4.4 Iteradors i intervals

Moltes funcions de la biblioteca estàndard de C++ operen amb iteradors. Un **iterador** (iterator) és una variable que apunta a (assenyala) un element d'una estructura de dades.

Els iteradors begin i end defineixen un interval que conté tots els elements d'una estructura de dades. L'iterador begin apunta al primer element de l'estructura de dades, i l'iterador end apunta a la posició *després de* l'últim element. Per exemple:

```
    { 3, 4, 6, 8, 12, 13, 14, 17 }  
      ↑                               ↑  
    s.begin()                       s.end()
```

Observeu l'asimetria dels iteradors: s.begin() apunta a un element de l'estructura de dades, mentre que s.end() apunta fora de l'estructura de dades. És a dir, l'interval definit pels iteradors és *semi-obert*.

Treballar amb intervals

Els iteradors es fan servir en la biblioteca estàndard de C++ per a passar intervals d'elements en una estructura de dades. Normalment, volem processar tots els elements d'una estructura de dades, i per tant fem servir els iteradors begin i end.

Per exemple, el codi següent ordena un vector fent servir la funció sort, després inverteix l'ordre dels elements fent servir la funció reverse, i finalment barreja l'ordre de els elements fent servir la funció random_shuffle.

```
sort(v.begin(), v.end());  
reverse(v.begin(), v.end());  
random_shuffle(v.begin(), v.end());
```

Aquestes funcions també es poden fer servir amb arrays ordinaris de C. En aquest cas, les funcions reben punters a l'array en lloc d'iteradors:

```
sort(a, a+n);
reverse(a, a+n);
random_shuffle(a, a+n);
```

Iteradors de conjunts

Els iteradors es fan servir sovint per a accedir als elements d'un conjunt. El codi següent crea un iterador `it` que assenyalava a l'element més petit d'un conjunt:

```
set<int>::iterator it = s.begin();
```

També es pot escriure així:

```
auto it = s.begin();
```

Per a accedir a l'element que l'iterador assenyalava es fa servir el símbol `*`. Per exemple, el següent codi imprimeix el primer element del conjunt:

```
auto it = s.begin();
cout << *it << "\n";
```

Els iteradors es poden moure mitjançant els operadors `++` (endavant) i `--` (enrere), és a dir, moure l'iterador a l'element següent o l'element anterior.

El codi següent escriu tots els elements en ordre creixent:

```
for (auto it = s.begin(); it != s.end(); it++) {
    cout << *it << "\n";
}
```

El codi següent escriu l'element més gran del conjunt:

```
auto it = s.end(); it--;
cout << *it << "\n";
```

La funció `find(x)` retorna un iterador que apunta a un element el valor del qual és `x`. Si el conjunt no conté `x`, l'iterador retornat serà `end`.

```
auto it = s.find(x);
if (it == s.end()) {
    // no trobem x
}
```

La funció `lower_bound(x)` retorna un iterador al primer element del conjunt el valor del qual és *almenys* `x`, i la funció `upper_bound(x)` retorna un iterador al primer element del conjunt el valor del qual és *més gran que* `x`. En ambdós

casos, si aquest primer element del conjunt no existeix, retornen end.¹ Aquestes funcions no són compatibles amb l'estructura `unordered_set` perquè aquesta no manté l'ordre dels elements.

Per exemple, el codi següent troba l'element més proper a x :

```
auto it = s.lower_bound(x);
if (it == s.begin()) {
    cout << *it << "\n";
} else if (it == s.end()) {
    it--;
    cout << *it << "\n";
} else {
    int a = *it; it--;
    int b = *it;
    if (x-b < a-x) cout << b << "\n";
    else cout << a << "\n";
}
```

El codi suposa que el conjunt no està buit, i repassa tots els casos possibles utilitzant un iterador `it`. L'iterador apunta al valor més petit el valor del qual és almenys x . Si `it` és igual a `begin`, l'element corresponent és el més proper a x . Si `it` és igual a `end`, l'element més gran del conjunt és el més proper a x . Si no es compleix cap dels dos casos anteriors, l'element més proper a x és o bé l'element assenyalat per `it` o bé l'element anterior.

4.5 Altres estructures

Conjunt de bits

Un conjunt de bits, o **bitset**, és un vector on cada valor és 0 o 1. El codi següent crea un conjunt de bits amb 10 elements:

```
bitset<10> s;
s[1] = 1;
s[3] = 1;
s[4] = 1;
s[7] = 1;
cout << s[4] << "\n"; // 1
cout << s[5] << "\n"; // 0
```

L'avantatge de fer servir conjunts de bits és que requereixen menys memòria que els vectors normals, perquè cada element només fa servir un únic bit de memòria. Per exemple, si emmagatzemèssim n bits en un vector de tipus `int`, necessitaríem $32n$ bits de memòria, però en un conjunt de bits només fan falta n bits de memòria. A més, els valors d'un conjunt de bits es poden manipular de

¹N. del T.: És a dir, `[lower_bound(x), upper_bound(x))` és l'interval semiobert que assenyalava els elements amb valor x .

manera eficient fent servir operadors de bits, amb la qual cosa encara és possible optimitzar més alguns algorismes.

El codi següent mostra una altra manera de crear el conjunt de bits anterior:

```
bitset<10> s(string("0010011010")); // de dreta a esquerra
cout << s[4] << "\n"; // 1
cout << s[5] << "\n"; // 0
```

La funció count retorna el nombre d'uns en el conjunt de bits:

```
bitset<10> s(string("0010011010"));
cout << s.count() << "\n"; // 4
```

El codi següent mostra exemples d'operacions de bits:

```
bitset<10> a(string("0010110110"));
bitset<10> b(string("1011011000"));
cout << (a&b) << "\n"; // 0010010000
cout << (a|b) << "\n"; // 1011111110
cout << (a^b) << "\n"; // 1001101110
```

Deque

Una **deque** és un vector dinàmic la mida del qual es pot canviar eficientment a ambdós extrems del vector. Al igual que els vectors, una deque proporciona les funcions push_back i pop_back, però també inclou les funcions push_front i pop_front que no estan disponibles en un vector normal.

Una deque es pot fer servir de la manera següent:

```
deque<int> d;
d.push_back(5); // [5]
d.push_back(2); // [5,2]
d.push_front(3); // [3,5,2]
d.pop_back(); // [3,5]
d.pop_front(); // [5]
```

La implementació interna d'una deque és més complexa que el d'un vector, i per aquest motiu, una deque és més lenta que un vector. Tot i així, afegir o eliminar elements per qualsevol dels dos extrems triga $O(1)$ de mitjana.

Pila

Una **pila** (stack) és una estructura de dades que proporciona dos operacions que triguen $O(1)$: afegint un element a la part superior, i retirar l'element de la part superior.

El codi següent mostra com fer servir una pila:

```
stack<int> s;
```

```
s.push(3);
s.push(2);
s.push(5);
cout << s.top(); // 5
s.pop();
cout << s.top(); // 2
```

Queue

Una **queue** també ofereix dos operacions de temps $O(1)$: afegir un element al final de la cua, i treure el primer element de la cua. Només és possible accedir al primer i al darrer element de la cua.

El codi següent mostra com fer servir una cua:

```
queue<int> q;
q.push(3);
q.push(2);
q.push(5);
cout << q.front(); // 3
q.pop();
cout << q.front(); // 2
```

Cua de prioritats

Una **cua de prioritats** manté un conjunt d'elements. S'ofereixen les operacions d'inserció i, depenent del tipus de cua, recuperar o eliminar l'element mínim o màxim (només un dels dos). Inserir i eliminar elements triga $O(\log n)$, mentre que obtenir l'element mínim o màxim triga $O(1)$.

En principi, es podria fer servir un conjunt ordenat per a implementar una cua de prioritats, però l'avantatge de fer servir una cua de prioritats és que els factors constants són més petits. Les cues de prioritats s'implementen fent servir una estructura de heap que és molt més senzilla que els arbres binaris balancejats dels conjunts ordenats.

Per defecte, els elements en una cua de prioritats de C++ s'ordenen en ordre decreixent, i només podem trobar i eliminar l'element més gran de la cua. Per exemple:

```
priority_queue<int> q;
q.push(3);
q.push(5);
q.push(7);
q.push(2);
cout << q.top() << "\n"; // 7
q.pop();
cout << q.top() << "\n"; // 5
q.pop();
```

```
q.push(6);
cout << q.top() << "\n"; // 6
q.pop();
```

Si volem crear una cua de prioritat que permeti trobar i eliminar l'element més petit, podem fer el següent:

```
priority_queue<int, vector<int>, greater<int>> q;
```

Estructures de dades “policy-based”

El compilador g++ també ofereix algunes estructures de dades que no formen part de la biblioteca estàndard C++. Aquestes estructures s'anomenen *policy-based*. Per a fer-les servir, hem d'afegir les línies següents:

```
#include <ext/pb_ds/assoc_container.hpp>
using namespace __gnu_pbds;
```

Després d'això, podem fer servir una estructura de dades `indexed_set` que és un set que es pot indexar com un vector. La definició per a valors de tipus `int` és la següent:

```
typedef tree<int, null_type, less<int>, rb_tree_tag,
            tree_order_statistics_node_update> indexed_set;
```

Ara podem crear un conjunt de la manera següent:

```
indexed_set s;
s.insert(2);
s.insert(3);
s.insert(7);
s.insert(9);
```

L'especialitat d'aquest conjunt és que tenim accés als índexs que tindrien els elements en un vector ordenat. La funció `find_by_order` retorna un iterador a l'element que ocupa una posició determinada:

```
auto x = s.find_by_order(2);
cout << *x << "\n"; // 7
```

I la funció `order_of_key` retorna la posició d'un element donat:

```
cout << s.order_of_key(7) << "\n"; // 2
```

Si l'element no apareix al conjunt, obtenim la posició que tindria l'element al conjunt:

```
cout << s.order_of_key(6) << "\n"; // 2
cout << s.order_of_key(8) << "\n"; // 3
```

Ambdues funcions funcionen en temps logarítmic.

4.6 Comparació amb l'ordenació

Sovint és possible resoldre un problema fent servir estructures de dades o ordenació. De vegades hi ha diferències notables en l'eficiència real d'aquests enfocaments, que poden quedar amagats en les seves complexitats temporals.

Considerem un problema on se'ns donen dues llistes A i B , ambdues de n elements. La nostra tasca és calcular el nombre d'elements que pertanyen a les dues llistes. Per exemple, per a les llistes

$$A = [5, 2, 8, 9] \quad \text{i} \quad B = [3, 2, 9, 5],$$

la resposta és 3 perquè els números 2, 5 i 9 pertanyen a les dues llistes.

La solució directa al problema és recórrer tots els parells d'elements en temps $O(n^2)$, però a continuació ens centrarem en els algorismes més eficients.

Algorisme 1

Construïm un conjunt amb els elements que apareixen a A , i després d'això, iterem a través dels elements de B i comprovem si cadascun dels elements també pertanyen a A . Això és eficient perquè els elements de A estan en un conjunt. Utilitzant l'estructura set, la complexitat temporal de l'algorisme és $O(n \log n)$.

Algorisme 2

No necessitem mantenir un conjunt ordenat per a A , per tant, en lloc d'un set fem servir un `unordered_set`. Aquesta és una manera fàcil de fer l'algorisme més eficient, perquè només hem de canviar l'estructura de dades, sense canviar l'algorisme. La complexitat temporal del nou algorisme és $O(n)$.

Algorisme 3

Ordenem en lloc de fer servir estructures de dades. Primer, ordenem les dues llistes A i B . Després d'això, recorrem les dues llistes alhora i trobem els elements comuns. El cost de l'ordenació és $O(n \log n)$, i la resta de l'algorisme triga $O(n)$, per tant, la complexitat total és $O(n \log n)$.

Comparació d'eficiència

La taula següent mostra l'eficiència els algorismes anteriors quan n varia i els elements de les llistes són nombres enters aleatoris entre $1 \dots 10^9$:

n	Algorisme 1	Algorisme 2	Algorisme 3
10^6	1,5 s	0,3 s	0,2 s
$2 \cdot 10^6$	3,7 s	0,8 s	0,3 s
$3 \cdot 10^6$	5,7 s	1,3 s	0,5 s
$4 \cdot 10^6$	7,7 s	1,7 s	0,7 s
$5 \cdot 10^6$	10.0 s	2.3 s	0.9 s

Els Algorismes 1 i 2 són iguals excepte que fem servir diferents estructures de conjunt. En aquest problema, aquesta elecció té un efecte important en el temps d'execució, perquè l'Algorisme 2 és de 4 a 5 vegades més ràpid que l'Algorisme 1.

Tanmateix, l'algorisme més eficient és el que fa servir l'ordenació, l'Algorisme 3. Només triga la meitat del temps que triga l'Algorisme 2. Curiosament, la complexitat temporal tant de l'Algorisme 1 com de l'Algorisme 3 és $O(n \log n)$ però, malgrat això, l'Algorisme 3 és deu vegades més ràpid. Això es pot explicar pel fet que ordenar és un procediment senzill i només s'executa una vegada al començament de l'Algorisme 3, i la resta de l'algorisme triga temps lineal. Per altra banda, l'Algorisme 1 manté un arbre binari equilibrat complex durant tot l'algorisme.²

²N. del T.: L'Algorisme 2 triga $O(n)$, però també és més lent a la pràctica que l'Algorisme 3, que triga temps $O(n \log n)$. En principi, valors cada cops més grans de n haurien d'afavorir l'Algorisme 2, fins a ser un nombre arbitràriament gran de vegades més ràpid que l'Algorisme 3. A la pràctica, $O(\log n)$ creix molt lentament; a més a més, algunes operacions, com ara fer servir memòria addicional, poden ser més costoses com més gran sigui la n i això pot negar l'avantatge teòric de ser $O(\log n)$ cops més ràpid.

Capítol 5

Cerca completa

Cerca completa és un mètode general que es pot fer servir per a resoldre gairebé qualsevol problema algorísmic. La idea és generar totes les possibles solucions del problema amb força bruta i, a continuació, seleccionar la millor solució, o comptar el nombre de solucions, segons el problema.

La cerca completa és una bona tècnica si hi ha prou temps per a generar totes les solucions, ja que la cerca sol ser fàcil d'implementar i sempre dona la resposta correcta. Si la cerca completa és massa lenta haurem de fer servir altres tècniques, com els algorismes *greedy* o la programació dinàmica.

5.1 Generar subconjunts

Considerem el problema de generar tots els subconjunts d'un conjunt de n elements. Per exemple, els subconjunts de $\{0, 1, 2\}$ són \emptyset , $\{0\}$, $\{1\}$, $\{2\}$, $\{0, 1\}$, $\{0, 2\}$, $\{1, 2\}$ i $\{0, 1, 2\}$. Hi ha dos mètodes comuns per a generar subconjunts: fer una cerca recursiva o aprofitar la representació binària dels nombres enters.

Mètode 1

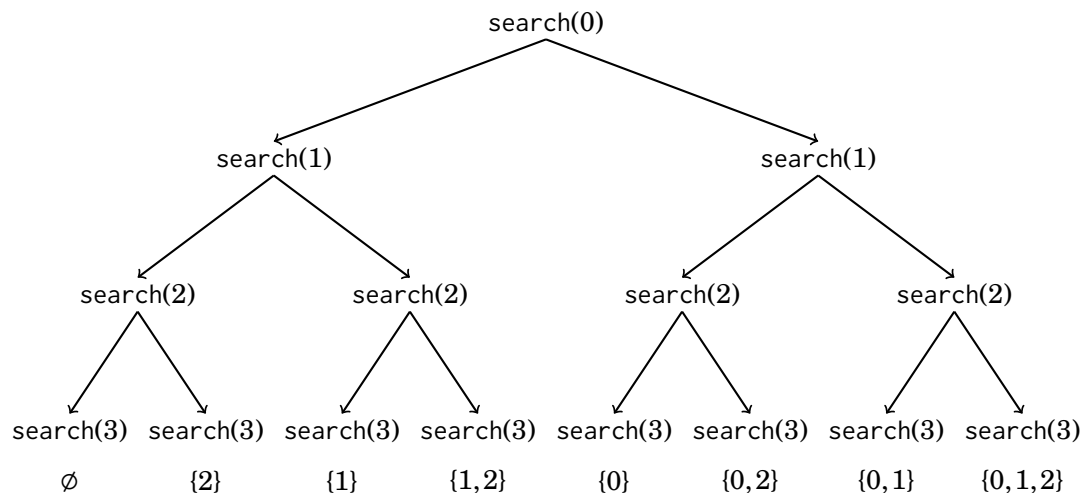
La recursivitat és una manera elegant de passar per tots els subconjunts d'un conjunt. La següent funció `search` genera els subconjunts de $\{0, 1, \dots, n-1\}$. La funció manté un vector global `v` que anirà contenint els elements de cada subconjunt. La cerca comença quan es crida la funció amb el paràmetre 0.

```
vector<int> v;  
  
void cerca(int k) {  
    if (k == n) {  
        // tracta el subconjunt v;  
    } altrament {  
        cerca(k+1);  
        v.push_back(k);  
        cerca(k+1);  
        v.pop_back();  
    }  
}
```

}

Quan la funció cerca es crida amb el paràmetre k , aquesta decideix si inclou o no l'element k al subconjunt i , en ambdós casos, es crida recursivament a ella mateixa amb paràmetre $k + 1$. Tanmateix, si $k = n$, la funció s'adona que ha processat tots els elements i , per tant, s'ha generat un nou subconjunt v .

L'arbre següent il·lustra les crides de funció quan $n = 3$. Sempre podem triar la branca esquerra (k no s'inclou al subconjunt) o la branca dreta (k s'inclou al subconjunt).



Mètode 2

Una altra manera de generar subconjunts es fer servir la representació en binari dels nombres enters. Cada subconjunt d'un conjunt de n elements es pot representar com una seqüència de n bits, que correspon a un nombre enter entre $0 \dots 2^n - 1$. Els uns en la seqüència de bits indiquen quins elements formen part del subconjunt.

La convenció habitual és que l'últim bit correspon a l'element 0, el penúltim bit és l'element 1, etcètera. Per exemple, la representació binària de 25 és 11001, que correspon al subconjunt $\{0, 3, 4\}$.

El codi següent¹ tracta tots els subconjunts d'un conjunt de n elements

```

for (int b = 0; b < (1<<n); b++) {
    // tracta el subconjunt b
}

```

El codi següent mostra com trobar el subconjunt que conté els elements corresponents a una seqüència de bits. Quan tractem cada subconjunt b , el codi construeix el vector v amb els elements del subconjunt.

```

for (int b = 0; b < (1<<n); b++) {

```

¹N. del T.: La notació $1 \ll n$ és l'operador *shift*, i indica que el nombre 1 es desplaça n posicions a l'esquerra, és a dir, 2^n

```

vector<int> v;
for (int i = 0; i < n; i++) {
    if (b&(1<<i)) v.push_back(i);
}
}

```

5.2 Generar permutacions

A continuació considerem el problema de generar totes les permutacions d'un conjunt de n elements. Per exemple, les permutacions de $\{0, 1, 2\}$ són $(0, 1, 2)$, $(0, 2, 1)$, $(1, 0, 2)$, $(1, 2, 0)$, $(2, 0, 1)$ i $(2, 1, 0)$. De nou, hi ha dos enfocaments: fer servir la recursivitat o recórrer les permutacions iterativament.

Mètode 1

A l'igual que amb els subconjunts, podem generar permutacions fent servir recursivitat. La següent funció cerca recorre les permutacions del conjunt $\{0, 1, \dots, n-1\}$. La funció construeix un vector global `perm` que conté la permutació, i la cerca comença quan la funció és crida sense paràmetres.

```

vector<int> perm;
vector<bool> triat(n);

void cerca() {
    if (permutacio.size() == n) {
        // tracta la permutacio perm
    } else {
        for (int i = 0; i < n; i++) {
            if (triat[i]) continue;
            triat[i] = true;
            perm.push_back(i);
            cerca();
            triat[i] = false;
            perm.pop_back();
        }
    }
}

```

Cada crida a la funció afegeix un nou element al vector `perm`. El vector `triat` indica quins elements ja han estat inclosos a la permutació. Cada cop que la mida del vector `perm` coincideix amb la mida del conjunt vol dir que hem generat una permutació nova.

Mètode 2

Un altre mètode per generar permutacions és començar amb la permutació $\{0, 1, \dots, n-1\}$ i fer servir repetidament una funció que construeix la següent

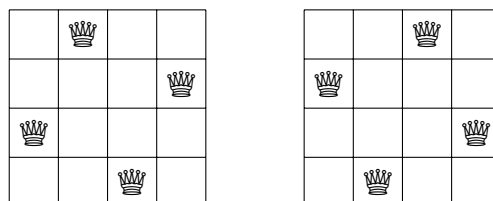
permutació en ordre creixent. La biblioteca estàndard de C++ conté la funció `next_permutation` que es pot fer servir per a això:

```
vector<int> perm;
per (int i = 0; i < n; i++) {
    perm.push_back(i);
}
do {
    // tractar la permutacio perm
} while (next_permutation(perm.begin(), perm.end()));
```

5.3 Backtracking

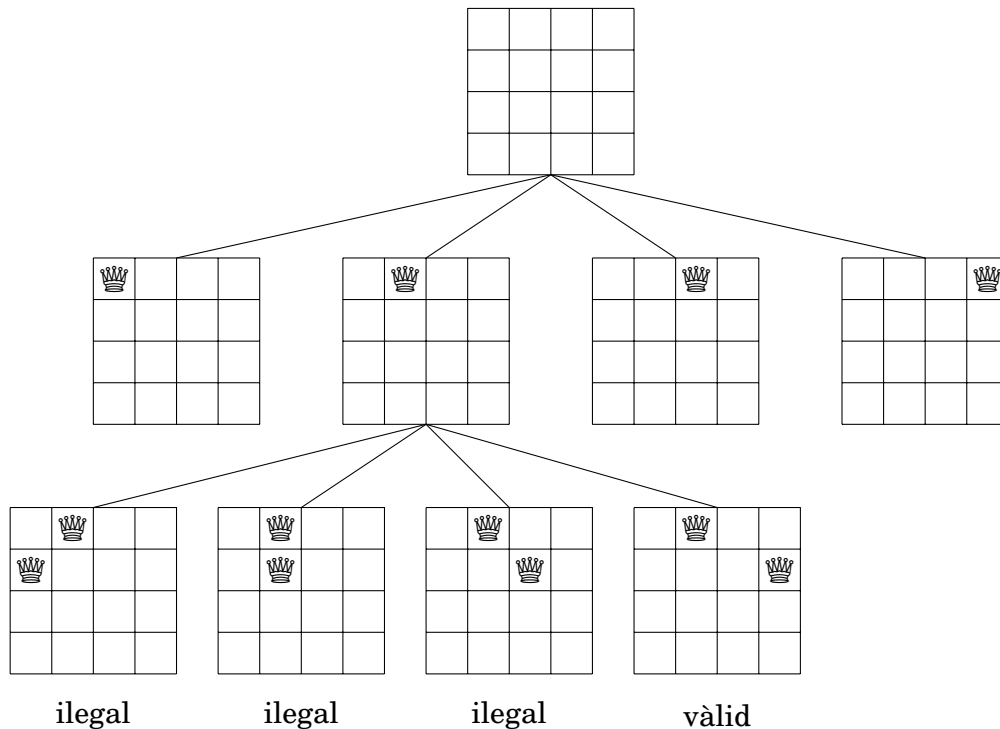
Un algorisme de **backtracking** comença amb una solució buida i amplia la solució pas a pas. La cerca passa recursivament per les diferents maneres en que es pot construir una solució.

Com a exemple, considerem el problema de calcular el nombre de maneres en què es poden col·locar n reines en un taulell d'escacs de mida $n \times n$ sense que dues reines s'amenacin. Per exemple, quan $n = 4$, hi ha dues solucions possibles:



El problema es pot resoldre fent servir backtracking, col·locant les reines al taulell fila per fila. Més precisament, per a cada fila, col·locarem una sola reina i de manera que cap de les reines anteriors l'ataqui. Quan col·loquem n reines haurem trobat una solució al problem.

Per exemple, quan $n = 4$, aquestes són algunes solucions parcials generades per l'algorisme de backtracking:



Al nivell inferior, els tres primers taulells són il·legals, perquè les reines s'ataquen entre elles. Tanmateix, el quart taulell és vàlid, i es pot estendre a una solució completa posant dues reines més al tauler. Només hi ha una manera de col·locar aquestes dues reines.

L'algorisme es pot implementar de la següent manera:

```
int compte = 0;
vector<bool> columna(n), diag1(2*n), diag2(2*n);

void cerca(int y) {
    if (y == n) {
        compte++;
        return;
    }
    for (int x = 0; x < n; x++) {
        if (columna[x] || diag1[x+y] || diag2[x-y+n-1]) continue;
        columna[x] = diag1[x+y] = diag2[x-y+n-1] = 1;
        cerca(y+1);
        columna[x] = diag1[x+y] = diag2[x-y+n-1] = 0;
    }
}
```

La cerca comença cridant `cerca(0)`. La mida del tauler és $n \times n$, i el codi calcula el nombre de solucions a compte.

El codi assumeix que les files i columnes del tauler estan numerats de 0 a $n - 1$. Quan la funció `cerca` és crida amb el paràmetre y , col·loca una dama a la fila y i després es crida recursivament amb el paràmetre $y + 1$. Quan $y = n$ s'ha trobat una solució i la variable `compte` s'incrementa en un.

El vector columna porta el compte de les columnes que tenen una reina, i els vectors diag1 i diag2 porten el compte de les diagonals amb reina. No està permès afegir una altra reina a una columna o diagonal que ja en conté una. Per exemple, les columnes i diagonals de el tauler 4×4 es numeren de la manera següent:

0	1	2	3
0	1	2	3
0	1	2	3
0	1	2	3

column

0	1	2	3
1	2	3	4
2	3	4	5
3	4	5	6

diag1

3	4	5	6
2	3	4	5
1	2	3	4
0	1	2	3

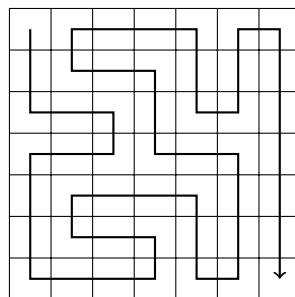
diag2

Sigui $q(n)$ el nombre de maneres per posar n reines en un tauler d'escacs $n \times n$. El procés de backtracking anterior ens diu que, per exemple, $q(8) = 92$. Quan n augmenta, la cerca es torna lenta ràpidament, perquè el nombre de solucions augmenta exponencialment. Per exemple, calcular $q(16) = 14772512$ fent servir l'algorisme anterior triga aproximadament un minut en un ordinador modern².

5.4 Podar la cerca

Sovint podem optimitzar el backtracking podant l'arbre de cerca. La idea és afegir “intel·ligència” a l'algorisme perquè es doni compte com més aviat possible que una solució parcial no es pot ampliar a una solució completa. Aquestes optimitzacions poden tenir un gran impacte sobre l'eficiència de la cerca.

Considerem el problema de calcular el nombre de camins en un taulell $n \times n$ des de la cantonada superior esquerra a la cantonada inferior dreta de manera que el camí visiti cada casella exactament una vegada. Per exemple, en un taulell 7×7 , hi ha 111712 camins d'aquest tipus. Un dels camins és el següent:



Ens centrem en el cas de 7×7 , perquè el seu nivell de dificultat és adequat a les nostres necessitats. Començarem amb un algorisme de backtracking senzill, i després l'optimitzarem pas a pas fent servir observacions de com podem podar la cerca. Després de cada optimització, mesurarem el temps d'execució de l'algorisme i el nombre de crides recursives, per a veure clarament l'efecte de cadascuna d'aquestes optimitzacions.

²No es coneix cap manera de calcular eficientment valors més grans de $q(n)$. El rècord actual és $q(27) = 234907967154122528$, calculat el 2016 [55].

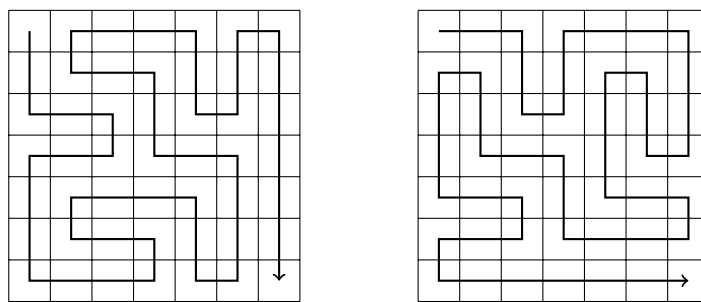
Algorisme bàsic

La primera versió de l'algorisme no conté cap optimització. Simplement fem servir el backtracking per generar tots els camins possibles des de la cantonada superior esquerra fins a la cantonada inferior dreta i comptar el nombre d'aquests camins.

- temps de funcionament: 483 segons
- nombre de crides recursives: 76 mil milions

Optimització 1

En tota solució, primer avancem un pas cap avall o cap a la dreta. Sempre hi ha dos camins que són simètrics sobre la diagonal del taulell que passa pel primer pas. Per exemple, els camins següents són simètrics:

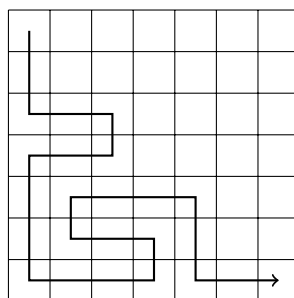


Per tant, podem decidir que el primer pas és sempre cap avall (o cap a la dreta), i multiplicar per dos el nombre de solucions.

- temps de funcionament: 244 segons
- nombre de crides recursives: 38 mil milions

Optimització 2

Si el camí arriba al quadrat inferior dret abans d'haver visitat tots els altres quadrats de la quadrícula, està clar que no serà possible completar la solució. Un exemple d'això és el camí següent:

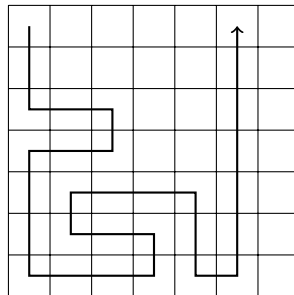


Amb aquesta observació, podem acabar la cerca immediatament si arribem massa aviat a la casella inferior dreta.

- temps de funcionament: 119 segons
- nombre de crides recursives: 20 mil milions

Optimització 3

Si el camí toca una paret i pot girar a l'esquerra o a la dreta, la graella es divideix en dues parts que contenen caselles no visitades. Per exemple, en la situació següent, el camí pot girar a l'esquerra o a la dreta:

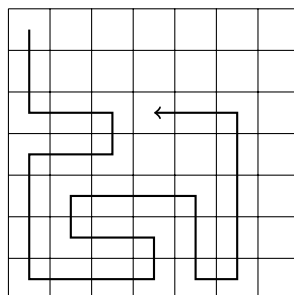


En aquest cas, ja no podem visitar totes les caselles, i podem acabar la cerca. Aquesta optimització és molt útil:

- temps de funcionament: 1.8 segons
- nombre de crides recursives: 221 milions

Optimització 4

La idea de l'optimització 3 es pot generalitzar: si el camí no pot continuar endavant però pot girar a l'esquerra o a la dreta, el taulell es divideix en dues parts que contenen caselles no visitades. Per exemple, considerem el camí següent:



Està clar que ja no podem visitar totes les caselles, de manera que acabem la cerca. Després d'aquesta optimització, la cerca és molt eficient:

- temps de funcionament: 0.6 segons
- nombre de crides recursives: 69 milions

Ara és un bon moment per deixar d'optimitzar l'algorisme i veure què hem aconseguit. El temps d'execució de l'algorisme original era 483 segons i, després de les optimitzacions, el temps ha baixat a només 0.6 segons. Les optimitzacions han fet que l'algorisme sigui gairebé 1000 vegades més ràpid.

Aquest és un fenomen habitual en el backtracking, perquè els arbres de cerca solen ser molt grans i fins i tot observacions simples poden podar eficaçment la cerca. Les optimitzacions que es produeixen durant els primers passos de l'algorisme, és a dir, a la part superior de l'arbre de cerca, són especialment útils.

5.5 Trobar-se al mig

Trobar-se al mig és una tècnica on es divideix l'espai de cerca en dues parts d'aproximadament la mateixa mida. Es realitzen cerques independents per a cada part, i finalment es combinen els resultats de les cerques.

La tècnica es pot fer servir si hi ha una manera eficient de combinar el resultat de les cerques. En aquesta situació, les dues cerques poden requerir menys temps que una cerca gran. Típicament, fent servir la tècnica de trobar-se al mig podem transformar un factor 2^n en un factor $2^{n/2}$.

Com a exemple, considerem el problema on se'ns dona una llista de n nombres i un nombre x , i volem saber si és possible triar alguns números de la llista de manera que la seva suma sigui x . Per exemple, donada la llista $[2, 4, 5, 9]$ i $x = 15$, podem triar els números $[2, 4, 9]$ per obtenir $2 + 4 + 9 = 15$. Tanmateix, fent servir la mateixa llista i $x = 10$, ja no és possible obtenir la suma.

Un algorisme senzill que resol el problema és iterar tots els subconjunts dels elements i comprovar si la suma d'algun dels subconjunts és x . El temps d'execució d'aquest algorisme és $O(2^n)$, perquè hi ha 2^n subconjunts. No obstant això, fent servir la tècnica de trobar-se al mig, podem aconseguir un algorisme de temps $O(2^{n/2})$ més eficient³. Tingueu en compte que $O(2^n)$ i $O(2^{n/2})$ són diferents complexitats perquè $2^{n/2}$ és igual a $\sqrt{2^n}$.

La idea és dividir la llista en dues llistes A i B de manera que cada llista contingui aproximadament la meitat dels números. La primera cerca genera tots els subconjunts de A i emmagatzema les seves sumes en una llista S_A . De manera semblant, la segona cerca crea una llista S_B a partir de B . Després d'això, n'hi ha prou comprovant si és possible triar un element de S_A i un altre de S_B de manera que la seva suma sigui x . Això només és possible si hi ha alguna manera de formar la suma x amb els números de la llista original.

Per exemple, suposem que la llista és $[2, 4, 5, 9]$ i $x = 15$. Primer, dividim la llista en $A = [2, 4]$ i $B = [5, 9]$. Després d'això, creem llistes $S_A = [0, 2, 4, 6]$ i $S_B = [0, 5, 9, 14]$. En aquest cas, podem formar la suma $x = 15$, perquè S_A conté la suma 6, S_B conté la suma 9 i $6 + 9 = 15$. Això correspon a la solució $[2, 4, 9]$.

Podem implementar l'algorisme de manera que la seva complexitat temporal és $O(2^{n/2})$. Primer, generem llistes ordenades S_A i S_B , que es pot fer en $O(2^{n/2})$, fent servir una tècnica semblant a la fusió (merge). Després, donat que les llistes

³Aquesta idea va ser introduïda l'any 1974 per E. Horowitz i S. Sahni [39].

estan ordenades, podem comprovar en temps $O(2^{n/2})$ si la suma x es pot crear a partir de S_A i S_B .

Capítol 6

Algorismes greedy

Un **algorisme greedy** (cobdiciós) es aquell que construeix una solució al problema fent sempre la tria que sembla millor en aquell moment. Un algorisme greedy mai desfà una opció ja triada, sinó que construeix la solució final directament. Per aquest motiu, els algorismes greedy solen ser molt eficients.

La dificultat de dissenyar algorismes greedy és trobar una estratègia que sempre produeixi una solució òptima al problema. En un algorisme greedy ha de passar que les eleccions que són localment òptimes siguin també globalment òptimes. Sovint no és fàcil d'argumentar perquè un algorisme greedy concret funciona.

6.1 Problema de les monedes

Com a primer exemple, considerem un problema on se'ns dóna un conjunt de monedes i la nostra feina és formar una quantitat de diners n fent servir les monedes. Els valors de les monedes són $\text{monedes} = \{c_1, c_2, \dots, c_k\}$, i cada moneda es pot utilitzar tantes vegades com vulguem. Quin és el nombre mínim de monedes necessàries?

Per exemple, si les monedes són les monedes d'euro (en cèntims)

$$\{1, 2, 5, 10, 20, 50, 100, 200\}$$

i $n = 520$, necessitem almenys quatre monedes. La solució òptima és seleccionar monedes $200 + 200 + 100 + 20$ la suma dels quals és 520.

Algorisme greedy

Un algorisme greedy senzill que resol el problema consisteix en seleccionar sempre la moneda més gran possible, fins que s'hagi construït la suma de diners requerida. Aquest algorisme funciona en el cas d'exemple, perquè primer seleccionem dues monedes de 200 cèntims, després una moneda de 100 cèntims i finalment una moneda de 20 cèntims. Però, com sabem que aquest algorisme sempre funciona?

Resulta que si les monedes són les monedes d'euro, l'algorisme greedy *sempre* funciona, és a dir, sempre produeix una solució amb el mínim nombre possible de monedes. Això es pot argumentar de la manera següent:

En primer lloc, cada moneda 1, 5, 10, 50 i 100 pot aparèixer com a màxim una vegada en una solució òptima, perquè si la solució contingués dues d'aquestes monedes, podríem canviar-les per una sola moneda i obtenir una solució millor. Per exemple, si la solució contingués les monedes $5 + 5$, podríem substituir-les per una moneda 10.

De la mateixa manera, les monedes 2 i 20 només poden aparèixer com a màxim dues vegades en una solució òptima, perquè d'altra forma podrem reemplaçar les monedes $2 + 2 + 2$ per monedes $5 + 1$ i les monedes $20 + 20 + 20$ per monedes $50 + 10$. A més, una solució òptima no pot contenir les monedes $2 + 2 + 1$ o $20 + 20 + 10$, perquè podríem substituir-les per monedes 5 i 50.

Fent servir aquestes observacions, hem de veure que per cada moneda x no és possible construir de manera òptima una suma x o qualsevol suma més gran utilitzant només monedes que són més petits que x . Per exemple, si $x = 100$, la suma òptima més gran fent servir només monedes més petites és $50 + 20 + 20 + 5 + 2 + 2 = 99$. Així, l'algorisme greedy que sempre selecciona la moneda més gran produeix la solució òptima.

Aquest exemple mostra que pot ser difícil argumentar perquè un algorisme greedy funciona, fins i tot si l'algorisme mateix és simple.

Cas general

En el cas general, el conjunt de monedes pot contenir qualsevol moneda i l'algorisme greedy ja *no* produeix necessàriament una solució òptima.

Podem demostrar que un algorisme greedy no funciona mostrant un contraexemple on l'algorisme ens dona una resposta incorrecta. En aquest problema és fàcil trobar-ne un: si les monedes són $\{1, 3, 4\}$ i la suma objectiu és 6, l'algorisme greedy produeix la solució $4 + 1 + 1$ mentre que la solució òptima és $3 + 3$.

No se sap si el problema generalitzat de la moneda es pot resoldre fent servir algun algorisme greedy¹. Tanmateix, com veurem al capítol 7, en alguns casos el problema generalitzat pot ser resolt eficientment fent servir un algorisme de programació dinàmica que sempre dona la resposta correcta.

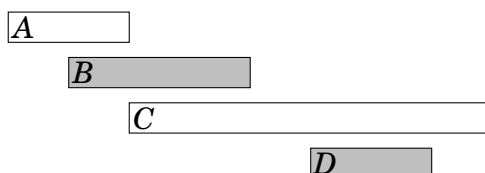
6.2 Scheduling

Molts problemes de *scheduling* (planificació horària) es poden resoldre fent servir algorismes greedy. Un problema clàssic és el següent: donats n esdeveniments amb els corresponents temps d'inici i de final, troba un scheduling que inclogui tants esdeveniments com sigui possible. No és permet seleccionar un esdeveniment parcialment. Per exemple, considerem els esdeveniments següents:

¹No obstant això, és possible *comprovar* en temps polinòmic si l'algorisme greedy que s'ha presentat en aquest capítol funciona amb un conjunt determinat de monedes [53].

esdeveniment	hora d'inici	hora final
<i>A</i>	1	3
<i>B</i>	2	5
<i>C</i>	3	9
<i>D</i>	6	8

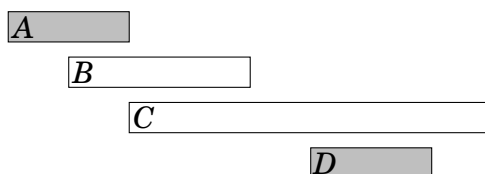
En aquest cas, el nombre màxim d'esdeveniments és dos. Per exemple, podem seleccionar els esdeveniments *B* i *D* com segueix:



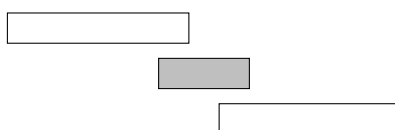
És possible inventar diversos algorismes greedy per al problema, però quin d'ells funciona en tots els casos?

Algorisme 1

La primera idea és començar seleccionant els esdeveniments més *curts* possibles. En l'exemple anterior l'algorisme selecciona els següent esdeveniments:



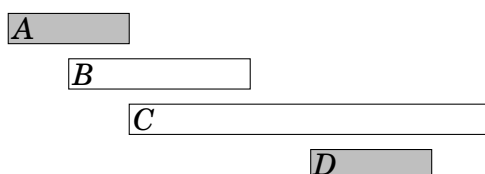
Tanmateix, seleccionar esdeveniments curts no sempre és una estratègia correcta. Per exemple, l'algorisme falla en el cas següent:



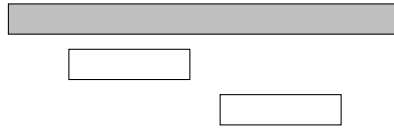
Si seleccionem l'esdeveniment més curt, només podem seleccionar un, quan en realitat seria possible seleccionar-ne dos.

Algorisme 2

Una altra idea és seleccionar sempre l'esdeveniment que *comença* tan d'hora com sigui possible. Aquest algorisme selecciona els esdeveniments següents:



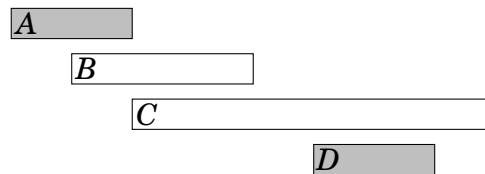
Tanmateix, també podem trobar un contraexemple per a aquest algorisme. Per exemple, en el cas següent, l'algorisme només selecciona un esdeveniment:



Si seleccionem el primer esdeveniment, ja no és possible seleccionar-ne cap d'altre, quan en realitat hauríem pogut seleccionar dos esdeveniments

Algorisme 3

La tercera idea és seleccionar l'esdeveniment que *acabi* tan *d'hora* com sigui possible. Aquest algorisme selecciona els esdeveniments següents:



Resulta que aquest algorisme *sempre* produeix una solució òptima.

Perquè l'algorisme funciona? Sigui X l'esdeveniment que acaba primer, i considerem una solució òptima que no tingui X . Com que X és l'esdeveniment que acaba primer, podem intercanviar el primer element de la solució òptima per X sense causar cap conflicte. Per tant, per a qualsevol solució òptima sense X , n'existeix una altra solució òptima amb X . Això demostra que la primera tria de l'algorisme no és errònia. El mateix argument s'aplica a les següents tries.

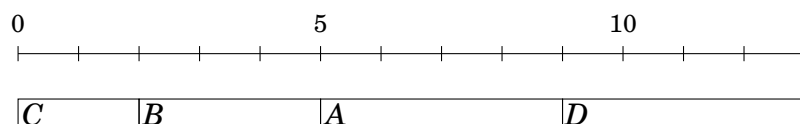
6.3 Tasques i terminis

Considerem ara un problema on se'ns dona n tasques amb durades i terminis i la nostra feina és triar en quin ordre s'han de fer les tasques. Per a cada tasca, guanyem $d - x$ punts on d és la data límit de la tasca i x és el moment en què acabem la tasca. Quina és la puntuació total més gran que podem obtenir?

Per exemple, suposem que les tasques són les següents:

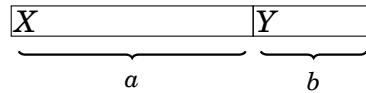
tasca	durada	termini
A	4	2
B	3	5
C	2	7
D	4	5

En aquest cas, aquesta és una assignació òptima:

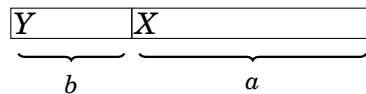


En aquesta solució, C ens dóna 5 punts, B ens dóna 0 punts, A ens dóna -7 punts i D ens dóna -8 punts, per tant, la puntuació total és de -10.

Sorprenentment, la solució òptima a aquest problema no depèn en absolut dels terminis. Una estratègia greedy correcta és simplement realitzar les tasques ordenades per la seva durada en ordre creixent. La raó d'això és que si mai fem dues tasques una darrere l'altra de manera que la primera tasca triga més que la segona tasca, podem millorar la solució intercanviant les tasques. Per exemple, considerem l'assignació següent:



Com que es dóna $a > b$, hauríem d'intercanviar les tasques:



Ara X ens dóna b punts menys però Y ens dóna a punts més, de manera que la puntuació total augmenta en $a - b > 0$. En una solució òptima, per cada dues tasques consecutives qualsevol, la tasca més curta s'ha de fer abans que la tasca més llarga. Per tant, les tasques s'han de fer ordenades en funció de la seva durada.

6.4 Minimitzar sumes

Considerem el problema on se'ns donen n nombres a_1, a_2, \dots, a_n i la nostra tasca és trobar un valor x que minimitzi la suma

$$|a_1 - x|^c + |a_2 - x|^c + \dots + |a_n - x|^c.$$

Ens centrem en els casos $c = 1$ i $c = 2$.

Cas $c = 1$

En aquest cas hem de minimitzar la suma

$$|a_1 - x| + |a_2 - x| + \dots + |a_n - x|.$$

Per exemple, si els números són $[1, 2, 9, 2, 6]$, la millor solució és seleccionar $x = 2$ que produeix la suma

$$|1 - 2| + |2 - 2| + |9 - 2| + |2 - 2| + |6 - 2| = 12.$$

En el cas general, la millor opció per x és la *mediana* dels nombres, és a dir, el nombre del mig després d'ordenar-los. Per exemple, la llista $[1, 2, 9, 2, 6]$ es converteix en $[1, 2, 2, 6, 9]$ després d'ordenar, i la mediana és 2.

La mediana és una opció òptima, perquè si x és més petit que la mediana, la suma es fa més petita en augmentar x , i si x és més gran que la mediana, la suma es fa més petita en disminuir x . Per tant, la solució òptima és que x sigui la mediana. Si n és parell i hi ha dues medianes, qualsevol de les medianes o els valors entre les dues són òptimes.

Cas $c = 2$

En aquest cas, hem de minimitzar la suma

$$(a_1 - x)^2 + (a_2 - x)^2 + \cdots + (a_n - x)^2.$$

Per exemple, si els nombres són $[1, 2, 9, 2, 6]$, la millor solució és seleccionar $x = 4$ que dona lloc a la suma

$$(1 - 4)^2 + (2 - 4)^2 + (9 - 4)^2 + (2 - 4)^2 + (6 - 4)^2 = 46.$$

En el cas general, la millor opció per x és la *mitjana* dels nombres. A l'exemple, la mitjana és $(1 + 2 + 9 + 2 + 6)/5 = 4$. Aquest resultat s'obté presentant la suma de la següent manera:

$$nx^2 - 2x(a_1 + a_2 + \cdots + a_n) + (a_1^2 + a_2^2 + \cdots + a_n^2)$$

Podem ignorar l'última part perquè no depèn de x . Les parts restants formen una funció $nx^2 - 2xs$ on $s = a_1 + a_2 + \cdots + a_n$. Aquesta és una paràbola que s'obre cap amunt amb arrels $x = 0$ i $x = 2s/n$. El valor mínim és la mitjana de les arrels $x = s/n$, és a dir, la mitjana dels nombres a_1, a_2, \dots, a_n .

6.5 Comprensió de dades

Una **codificació binària** assigna a cada caràcter d'una cadena un **codi** format per bits. Podem *comprimir* la cadena fent servir la codificació que reemplaça cada caràcter pel seu codi corresponent. Per exemple, la següent codificació assigna aquests codis als caràcters: A–D:

caràcter	codi
A	00
B	01
C	10
D	11

Aquesta codificació té **longitud constant** perquè cada codi té la mateixa mida. Fent servir aquesta codificació, la cadena AABACDACA es transforma en els 18 bits

000001001011001000

. Tanmateix, podem comprimir encara més la cadena si fem servir codificacions de **longitud variable**, on cada codi pot tenir longituds diferents. D'aquesta manera podem fer servir codis curts pels caràcters que apareixen sovint i codis llargs pels caràcters infreqüents. Resulta que la següent codificació és **òptima** per a la cadena anterior:

caràcter	codi
A	0
B	110
C	10
D	111

Una codificació òptima comprimeix la cadena en el mínim espai possible. En aquest cas, la cadena comprimida

001100101110100,

només necessita 15 bits en lloc de 18. Fer servir la millor codificació ens estalvia 3 bits.

Exigim també que no hi hagi cap codi que sigui prefix d'un altre codi. Per exemple, està prohibit que la codificació contingui els codis 10 i 1011. El motiu és que volem poder recuperar la cadena original a partir de la cadena comprimida. Si un codi pogués ser prefix d'un altre això no sempre seria possible. Per exemple, la codificació següent *no* és vàlida:

caràcter	codi
A	10
B	11
C	1011
D	111

Amb aquesta codificació, no sabem si 1011 és la compressió de AB o de C.

Codificació de Huffman

La **codificació de Huffman**² és un algorisme greedy que construeix una codificació òptima per a comprimir una cadena determinada. L'algorisme construeix un arbre binari en funció de les freqüències dels caràcters de la cadena, i trobem el codi que assignem a cada caràcter recurrent un camí des de l'arrel fins al node corresponent. Quan cop que ens movem a l'esquerra escrivim un bit 0, i quan ens movem a la dreta escrivim un bit 1.

Inicialment, cada caràcter és un node el pes del qual és el nombre de vegades que el caràcter apareix a la cadena. A continuació, combinem els dos nodes de pes mínim per a crear un nou node el pes del qual és la suma dels pesos dels nodes originals. El procés continua fins que combinem tots els nodes.

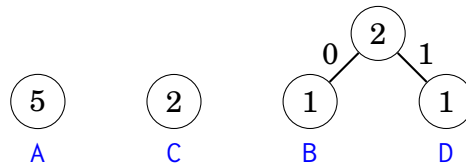
A continuació mostrem com es crea el codi de Huffman per a la cadena AABACDAC. Al principi, hi ha quatre nodes, un per cada caràcter de la cadena:



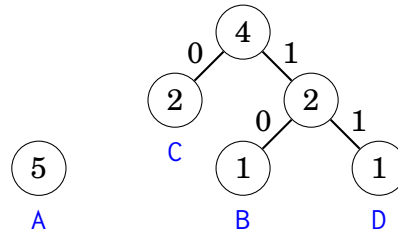
El node que representa el caràcter A té pes 5 perquè el caràcter A apareix 5 vegades a la cadena, i el mateix per la resta de pesos.

El primer pas és combinar els nodes dels caràcters B i D, tots dos amb pes 1. El resultat és:

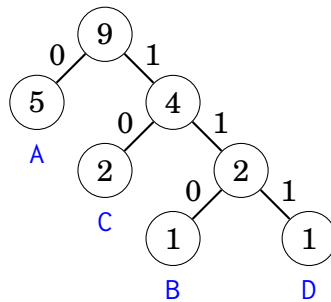
²D. A. Huffman va descobrir aquest mètode en un treball de final de curs a l'universitat i va publicar l'algorisme el 1952 [40].



Després d'això, combinen els nodes amb pes 2:



Finalment, combinen els dos nodes restants:



Ara tots els nodes estan connectats en un arbre, i podem recuperar la codificació llegint l'arbre:

caràcter	codi
A	0
B	110
C	10
D	111

³(N. del T.) Donem una intuïció de perquè la codificació de Huffman és òptima. Primer, en tota solució òptima passa que com més freqüent és un caràcter, més curt és el seu codi. Altrament, intercanviariem els codis de dos caràcters que no complissin aquesta propietat i obtindríem una codificació millor. Per tant, els dos caràcters menys freqüents són els que tenen codis més llargs. Aquests dos codis tenen la mateixa mida. Altrament, eliminariem l'últim bit del codi més llarg i obtindríem una codificació millor. Això és vàlid perquè el nou codi escurçat no és part de la codificació perquè era un prefix, i no és prefix de cap altre codi perquè els dos codis originals eren els més llargs. L'algorisme greedy descrit sorgeix d'aplicar repetidament aquesta propietat, i de fer servir l'arbre resultant per a generar una codificació òptima.

Capítol 7

Programació dinàmica

La **programació dinàmica** (*dynamic programming* o *DP*) és una tècnica que combina la correcció de la cerca completa amb l'eficiència dels algorismes greedy. La programació dinàmica es pot aplicar si el problema es pot dividir en subproblemes superposats però que es poden resoldre independentment.

La programació dinàmica té dos usos:

- **Trobar una solució òptima:** Volem trobar una solució que sigui tan gran (o tan petita) com sigui possible.
- **Comptar el nombre de solucions:** Volem calcular el nombre total de solucions.

Primer veurem com la programació dinàmica es pot fer servir per trobar una solució òptima, i després la farem servir per comptar les solucions.

Entendre la programació dinàmica és una fita en la carrera de qualsevol programador competitiu. Tot i que la idea bàsica és senzilla, el repte és com aplicar programació dinàmica als diferents problemes. Aquest capítol presenta un conjunt de problemes clàssics que són un bon punt de partida.

7.1 Problema de les monedes

Primer ens centrem en un problema que ja vam veure al capítol 6: Donat un conjunt de valors de monedes $\text{monedes} = \{c_1, c_2, \dots, c_k\}$ i una suma objectiu de diners n , la nostra feina és obtenir suma n fent servir el menor nombre de monedes possible.

Al capítol 6, vam resoldre el problema amb un algorisme greedy que sempre triava la moneda amb valor més gran possible. L'algorisme greedy funciona, per exemple, quan les monedes són les monedes d'euro, però en el cas general l'algorisme greedy no produeix necessàriament una solució òptima.

Ara és el moment de resoldre el problema de manera eficient fent servir la programació dinàmica, i fer que l'algorisme funcioni per qualsevol conjunt de monedes. L'algorisme de programació dinàmica es basa en una funció recursiva que passa per totes les possibles maneres de formar la suma, com si fos un algorisme de força bruta. No obstant això, l'algorisme de programació dinàmica perquè fa

servir *memoization* (escriure notes) i calcula la resposta a cada subproblema un sol cop.

Formulació recursiva

La idea en la programació dinàmica és formular el problema de manera recursiva de manera que la solució al problema pugui ser calculada a partir de solucions a subproblemes més petits. En el problema de la moneda, un problema natural recursiu és el següent: quin és el menor nombre de monedes necessari per a obtenir una suma x qualsevol?

Sigui $\text{resol}(x)$ el mínim nombre de monedes necessàries per a obtenir x . El resultat de la funció depèn dels valors de les monedes. Per exemple, si $\text{monedes} = \{1, 3, 4\}$, els primers valors de la funció són els següents:

$\text{resol}(0)$	$=$	0
$\text{resol}(1)$	$=$	1
$\text{resol}(2)$	$=$	2
$\text{resol}(3)$	$=$	1
$\text{resol}(4)$	$=$	1
$\text{resol}(5)$	$=$	2
$\text{resol}(6)$	$=$	2
$\text{resol}(7)$	$=$	2
$\text{resol}(8)$	$=$	2
$\text{resol}(9)$	$=$	3
$\text{resol}(10)$	$=$	3

Per exemple, $\text{resol}(10) = 3$, perquè calen almenys 3 monedes per formar la suma 10. La solució òptima és $3 + 3 + 4 = 10$.

La propietat essencial de resol és que els seus valors poden ser calculada recursivament a partir dels seus valors més petits. La idea és centrar-se en la *primera* moneda que triem per la suma. Per exemple, en l'escenari anterior, la primera moneda pot ser 1, 3 o 4. Si primer triem la moneda 1, la feina restant és obtenir la suma 9 fent servir el nombre mínim de monedes, que és un subproblema del problema original. Per descomptat, el mateix s'aplica si triem les monedes 3 o 4. Per tant, podem calcular el nombre mínim de monedes amb la següent fórmula recursiva:

$$\begin{aligned}\text{resol}(x) = \min(&\text{resol}(x - 1) + 1, \\ &\text{resol}(x - 3) + 1, \\ &\text{resol}(x - 4) + 1).\end{aligned}$$

El cas base de la recursivitat és $\text{resol}(0) = 0$, perquè no calen monedes per formar una suma buida. Per exemple,

$$\text{resol}(10) = \text{resol}(7) + 1 = \text{resol}(4) + 2 = \text{resol}(0) + 3 = 3.$$

Ara estem llestos per donar una fórmula recursiva general que calcula el

nombre mínim de monedes necessàries per obtenir una suma x :

$$\text{resol}(x) = \begin{cases} \infty & x < 0 \\ 0 & x = 0 \\ \min_{c \in \text{monedes}} \text{resol}(x - c) + 1 & x > 0 \end{cases}$$

Primer, si $x < 0$, el valor és ∞ , perquè és impossible formar una quantitat negativa de diners. Després, si $x = 0$, el valor és 0, perquè no fan falta monedes per obtenir una suma buida. Finalment, si $x > 0$, la variable c recorre totes les maneres de triar la primera moneda de la suma.

Una vegada hem trobat una fórmula recursiva, podem implementar directament la solució en C++ (la constant INF denota infinit):

```
int resol(int x) {
    if (x < 0) return INF;
    if (x == 0) return 0;
    int millor = INF;
    for (auto c : monedes) {
        millor = min(millor, resol(x-c)+1);
    }
    return millor;
}
```

Amb tot i això, aquesta funció no és eficient, perquè hi ha un nombre exponencial de maneres de construir la suma. A continuació, veurem com podem fer que aquesta funció sigui eficient fent servir una tècnica anomenada memoization.

Memoization

La idea de la programació dinàmica és fer servir **memoization** per calcular de manera eficient els valors d'una funció recursiva. Això vol dir que els valors de la funció s'emmagatzemen en un vector un cop calculats. Per a cada paràmetre, el valor de la funció es calcula recursivament només una vegada, i després d'això, el valor es pot obtenir directament consultant el vector.

En aquest problema, fem servir un vector

```
vector<int> valor(N, 0);
```

on $\text{valor}[x]$ indica el valor de $\text{resol}(x)$ o 0 si encara no l'hem calculat. La constant N ha estat triada per a que tots els valors necessaris càpiguen als vectors.

Ara podem implementar la funció eficientment de la manera següent:

```
int resol(int x) {
    if (x < 0) return INF;
    if (x == 0) return 0;
    if (valor[x]) return valor[x];
    int millor = INF;
    for (auto c : monedes) {
```

```

        millor = min(millor, resol(x-c)+1);
    }
    valor[x] = millor;
    return millor;
}

```

La funció gestiona els casos bàsics $x < 0$ i $x = 0$ com abans. A continuació, la funció comprova si el valor ja s'ha desat anteriorment a `valor[x]` i, si és així, el la retorna directament. En cas contrari, la funció calcula el valor `resol(x)` recursivament i l'emmagatzema en `valor[x]`.

Aquest codi és eficient perquè la resposta per a cada entrada x només es calcula una vegada. Un cop emmagatzemat el valor `resol(x)` a `valor[x]` el podem recuperar eficientment quan la funció es torna a cridar amb el paràmetre x . La complexitat de l'algorisme és $O(nk)$, on n és la suma objectiu i k és el nombre de monedes.

Tingueu en compte que també podem construir el vector `valor` de manera *iterativa* amb un bucle que calculi tots els valors de `resol` per als paràmetres $0 \dots n$:

```

valor[0] = 0;
for (int x = 1; x <= n; x++) {
    valor[x] = INF;
    for (auto c : monedes) {
        if (x-c >= 0) {
            valor[x] = min(valor[x], valor[x-c]+1);
        }
    }
}

```

De fet, la majoria dels programadors competitius prefereixen aquesta implementació, perquè és més curta i té factors constants més petits. A partir d'ara, també farem servir implementacions iteratives en els nostres exemples. Tot i això, sovint és més fàcil pensar en les solucions de programació dinàmica en termes de funcions recursives.

Construir una solució

De vegades se'ns demana trobar tant el valor d'una solució òptima com donar un exemple un exemple de solució. En el problema de les monedes, per exemple, podem declarar un altre vector que es guardi per cada suma de diners la primera moneda d'una solució òptima:

```

vector<int> primera(N);

```

Podem modificar l'algorisme de la següent manera:

```

valor[0] = 0;
for (int x = 1; x <= n; x++) {

```



```

    valor[x] = INF;
    per (auto c : monedes) {
        if (x-c >= 0 && valor[x-c]+1 < valor[x]) {
            valor[x] = valor[x-c]+1;
            primera[x] = c;
        }
    }
}

```

Després d'això, podem fer servir el codi següent per a imprimir les monedes que apareixen en una solució òptima amb suma n :

```

while (n > 0) {
    cout << primera[n] << "\n";
    n -= primera[n];
}

```

Comptar el nombre de solucions

Considerem ara una altra versió del problema de les monedes on la nostra feina és calcular el nombre de maneres de produir la suma x fent servir les monedes. Per exemple, si $\text{monedes} = \{1, 3, 4\}$ i $x = 5$, hi ha un total de 6 maneres:

- $1 + 1 + 1 + 1 + 1$
- $1 + 1 + 3$
- $1 + 3 + 1$
- $3 + 1 + 1$
- $1 + 4$
- $4 + 1$

De nou, podem resoldre el problema de manera recursiva. Sigui $\text{solve}(x)$ el nombre de maneres d'obtenir formar la suma x . Per exemple, si $\text{monedes} = \{1, 3, 4\}$, aleshores $\text{solve}(5) = 6$ i la fórmula recursiva és

$$\begin{aligned} \text{resol}(x) = & \text{resol}(x-1) + \\ & \text{resol}(x-3) + \\ & \text{resol}(x-4). \end{aligned} \quad (7.1)$$

La funció recursiva general és la següent:

$$\text{resol}(x) = \begin{cases} 0 & x < 0 \\ 1 & x = 0 \\ \sum_{c \in \text{monedes}} \text{resol}(x-c) & x > 0 \end{cases} \quad (7.2)$$

Si $x < 0$, el valor és 0, perquè no hi ha solucions. Si $x = 0$, el valor és 1, perquè només hi ha manera d'obtenir la suma buida. En cas contrari calculem la suma de tots els valors de la forma $\text{resol}(x-c)$ on c pertany a monedes .

El codi següent construeix un vector num tal que $\text{num}[x]$ és igual el valor de $\text{solve}(x)$ per a $0 \leq x \leq n$:

```

num[0] = 1;
for (int x = 1; x <= n; x++) {
    for (auto c : monedes) {
        if (x-c >= 0) {
            num[x] += num[x-c];
        }
    }
}

```

Sovint el nombre de solucions és tan gran que no se'ns demana calcular el nombre exacte sinó la resposta mòdul m on, per exemple, $m = 10^9 + 7$. Això es pot fer fent que tots els càlculs es fàcil mòdul m . En el codi anterior, n'hi ha prou afegint la línia

```
num[x] %= m;
```

després de

```
num[x] += num[x-c];
```

Amb això ja hem discutit totes les idees bàsiques de la programació dinàmica. Com que la programació dinàmica es pot fer servir en moltes situacions diferents, presentarem ara un conjunt de problemes que mostren més exemples sobre les possibilitats de la programació dinàmica.


7.2 Subseqüència creixent més llarga

El nostre primer problema és trobar la **subseqüència creixent més llarga** en un vector v de n elements. Aquesta és una longitud màxima d'una seqüència d'elements del vector, triats d'esquerra a dreta, i que cada element de la seqüència és més gran que l'element anterior. Per exemple, al vector

0	1	2	3	4	5	6	7
6	2	5	1	7	4	8	3

es té que la subseqüència creixent més llarga conté 4 elements:

0	1	2	3	4	5	6	7
6	2	5	1	7	4	8	3



Sigui $\text{longitud}(k)$ la longitud de la subseqüència creixent més llarga que acaba a la posició k . Si calculèssim tots els valors de $\text{longitud}(k)$ on $0 \leq k \leq n-1$, descobriríem quina és la longitud de la subseqüència creixent més llarga. Per

exemple, els valors de la funció per al vector anterior són els següents:

```
longitud(0) = 1
longitud(1) = 1
longitud(2) = 2
longitud(3) = 1
longitud(4) = 3
longitud(5) = 2
longitud(6) = 4
longitud(7) = 2
```

Per exemple, $\text{longitud}(6) = 4$, perquè la subseqüència creixent més llarga que acaba a la posició 6 consta de 4 elements.

Per calcular un valor de $\text{longitud}(k)$, hauríem de trobar una posició $i < k$ per a la qual $v[i] < v[k]$ i $\text{longitud}(i)$ és tan gran com sigui possible. D'aquí deduïm que $\text{longitud}(k) = \text{longitud}(i) + 1$, perquè aquesta és una manera òptima d'afegir $v[k]$ a una subseqüència. Tanmateix, si no existeix aquesta posició i , llavors $\text{longitud}(k) = 1$, és a dir, l'única subseqüència possible es aquella que només conté l'element $v[k]$.

Podem fer servir programació dinàmica perquè tots els valors de la funció es poden calcular a partir de valors més petits. En el codi següent emmagatzemem els valors de la funció en el vector `longitud`.

```
for (int k = 0; k < n; k++) {
    longitud[k] = 1;
    for (int i = 0; i < k; i++) {
        if (v[i] < v[k]) {
            longitud[k] = max(longitud[k], longitud[i]+1);
        }
    }
}
```

Aquest codi funciona en temps $O(n^2)$ perquè consta de dos bucles niats. Tanmateix, també és possible implementar el càlcul anterior de programació dinàmica de manera més eficient en temps $O(n \log n)$. Pots trobar una manera de fer-ho?

7.3 Camins en una quadrícula

El problema següent és trobar un camí que vagi de la cantonada superior esquerra a la cantonada inferior dreta d'una quadrícula $n \times n$, movent-nos únicament cap avall i cap a la dreta. Cada quadrat conté un nombre enter positiu, i el camí s'ha de construir de tal manera que la suma dels valors al llarg el camí sigui la més gran possible.

La imatge següent mostra un camí òptim en una quadrícula:

3	7	9	2	7
9	8	3	5	5
1	7	9	8	5
3	8	6	4	10
6	3	9	7	8

La suma dels valors del camí és 67, i aquesta és la suma més gran possible per a qualsevol camí des del cantonada superior esquerra a la cantonada inferior dreta.

Suposem que les files i columnes de la quadrícula estan numerades de l'1 al n , i que $\text{valor}[y][x]$ és el valor de la casella (y, x) . Sigui $\text{suma}(y, x)$ la suma màxima de camins que van des de la cantonada superior esquerra a la casella (y, x) . Aleshores, $\text{suma}(n, n)$ ens diu la suma màxima de la cantonada superior esquerra a la cantonada inferior dreta. Per exemple, a la quadrícula anterior, $\text{suma}(5, 5) = 67$.

Les sumes màximes es poden calcular de manera recursiva com segueix:

$$\text{suma}(y, x) = \max(\text{suma}(y, x - 1), \text{suma}(y - 1, x)) + \text{valor}[y][x]$$

La fórmula recursiva es basa en l'observació que un camí que acaba a la casella (y, x) ha de passar per la casella $(y, x - 1)$ o la casella $(y - 1, x)$:



Per tant, només hem de triar la direcció que maximitzi la suma. Si definim $\text{suma}(y, x) = 0$ per $y = 0$ o $x = 0$ (els camins no poden sortir de la quadrícula), tenim que la fórmula recursiva també funciona quan $y = 1$ o $x = 1$.

Com que la funció suma té dos paràmetres, el vector de la programació dinàmica també té dues dimensions. Per exemple, podem fer servir la matriu

```
vector<vector<int>> suma(N, vector<int>(N));
```

i calcula les sumes de la següent manera:

```
for (int y = 1; y <= n; y++) {
    for (int x = 1; x <= n; x++) {
        suma[y][x] = max(suma[y][x-1], suma[y-1][x]) + valor[y][x];
    }
}
```

La complexitat temporal de l'algorisme és $O(n^2)$.

7.4 Problemes de motxilla

El terme **motxilla** (*knapsack*) es refereix a problemes on es dóna un conjunt d'objectes, i volem buscar subconjunts amb algunes propietats. Els problemes de motxilla sovint es poden resoldre fent servir programació dinàmica.

En aquest apartat, ens centrem en el següent problema: donada una llista de pesos $[w_1, w_2, \dots, w_n]$, determinar totes les sumes que es poden construir fent servir els pesos. Per exemple, donats els pesos $[1, 3, 3, 5]$, podem obtenir les següents sumes:

0	1	2	3	4	5	6	7	8	9	10	11	12
X	X		X	X	X	X	X	X	X		X	X

En aquest cas, totes les sumes entre $0 \dots 12$ són possibles, excepte 2 i 10. Per exemple, la suma 7 és possible perquè podem seleccionar els pesos $[1, 3, 3]$.

Per resoldre el problema, ens centrem en els subproblemes on només fem servir els primers k pesos per construir sumes. Sigui $\text{possible}(x, k) = \text{true}$ si podem construir una suma x fent servir els primers k pesos, i $\text{possible}(x, k) = \text{false}$ en cas contrari. Els valors de la funció es poden calcular recursivament de la següent manera:

$$\text{possible}(x, k) = \text{possible}(x - w_k, k - 1) \vee \text{possible}(x, k - 1)$$

La fórmula es basa en el fet que podem fer servir o no fer servir el pes w_k a la suma. Si fem servir w_k , la tasca restant és trobar la suma $x - w_k$ fent servir els primers $k - 1$ pesos, i si no fem servir w_k , la tasca restant és formar la suma x fent servir els primers pesos $k - 1$. Els casos base són

$$\text{possible}(x, 0) = \begin{cases} \text{true} & x = 0 \\ \text{false} & x \neq 0 \end{cases}$$

ja que sense pesos només podem formar la suma 0.

La taula següent mostra tots els valors de la funció per als pesos $[1, 3, 3, 5]$ (el símbol "X" indica els valors reals):

$k \backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12
0	X												
1	X	X											
2	X	X		X	X								
3	X	X		X	X		X	X					
4	X	X		X	X	X	X	X	X	X		X	X

Després de calcular aquests valors, $\text{possible}(x, n)$ ens diu si podem construir la suma x fent servir *tots* pesos.

Sigui W la suma total dels pesos. El codi següent es correspon a la funció recursiva anterior i troba la solució en temps $O(nW)$.

```
possible[0][0] = true;
for (int k = 1; k <= n; k++) {
    for (int x = 0; x <= W; x++) {
        if (x-w[k] >= 0) possible[x][k] |= possible[x-w[k]][k-1];
        possible[x][k] |= possible[x][k-1];
    }
}
```

No obstant això, en aquest cas hi ha una millor implementació que només fa servir un vector unidimensional `possible[x]` que indica si podem construir un subconjunt amb la suma x . El truc és actualitzar el vector de dreta a esquerra per cada nou pes¹:

```
possible[0] = cert;
for (int k = 1; k <= n; k++) {
    for (int x = W; x >= 0; x--) {
        if (possible[x]) possible[x+w[k]] = true;
    }
}
```

Tingueu en compte que la idea presentada aquí es pot fer servir per molts problemes de motxilla. Per exemple, si ens donen objectes amb pesos i valors, podem determinar quin subconjunt d'objectes donaria valor màxim per a cadascun dels pesos possibles.

7.5 Distància d'edició

La **distància d'edició** o **distància de Levenshtein**² és el nombre mínim d'operacions d'edició que fan falta per transformar una cadena en una altra cadena. Les operacions d'edició permeses són les següents:

- inserir un caràcter (per exemple, $ABC \rightarrow ABCA$)
- eliminar un caràcter (per exemple, $ABC \rightarrow AC$)
- modificar un caràcter (per exemple, $ABC \rightarrow ADC$)

Per exemple, la distància d'edició entre LOVE i MOVIE és 2, perquè primer podem realitzar l'operació $LOVE \rightarrow MOVE$ (modificar) i després l'operació $MOVE \rightarrow MOVIE$ (inserir). Aquest és el nombre més petit possible d'operacions, perquè és evident que una sola operació no és suficient.

Suposem que se'ns dona una cadena x de longitud n i una cadena y de longitud m , i volem calcular la distància d'edició entre x i y . Per resoldre el problema, definim una funció $\text{dist}(a, b)$ que dona la distància d'edició entre prefixos $x[0 \dots a]$

¹N. del T.: Una solució alternativa i general consisteix en fer servir dos vectors unidimensionals, un per la fila actual (k) i un per la fila anterior ($k-1$).

²La distància rep el nom de V. I. Levenshtein, que la va estudiar per la seva relació amb les codificacions binàries [49].

i $y[0 \dots b]$. Així, utilitzant aquesta funció, la distància d'edició entre x i y és igual a $\text{dist}(n-1, m-1)$.

Podem calcular valors de dist com segueix:

$$\begin{aligned} \text{dist}(a, b) = \min(&\text{dist}(a, b-1) + 1, \\ &\text{dist}(a-1, b) + 1, \\ &\text{dist}(a-1, b-1) + \text{cost}(a, b)). \end{aligned}$$

Aquí $\text{cost}(a, b) = 0$ si $x[a] = y[b]$, i en cas contrari $\text{cost}(a, b) = 1$. La fórmula considera les següents maneres d'editar la cadena x :

- $\text{dist}(a, b-1)$: inserir un caràcter al final de x
- $\text{dist}(a-1, b)$: eliminar l'últim caràcter de x
- $\text{dist}(a-1, b-1)$: modificar, si és necessari, l'últim caràcter de x

En els dos primers casos, fa falta una sola operació d'edició (inserir o eliminar). En el darrer cas, si $x[a] = y[b]$, no fa falta gastar cap operació, però en cas contrari necessitem una operació d'edició (modificar).

La taula següent mostra els valors de dist en el cas exemple:

		M	O	V	I	E
	0	1	2	3	4	5
L	1	1	2	3	4	5
O	2	2	1	2	3	4
V	3	3	2	1	2	3
E	4	4	3	2	2	2

La cantonada inferior esquerra de la taula ens diu que la distància d'edició entre LOVE i MOVIE és 2. La taula també mostra com construir una seqüència mínima d'operacions d'edició. En aquest trobem el camí següent:

		M	O	V	I	E
	0	1	2	3	4	5
L	1	1	2	3	4	5
O	2	2	1	2	3	4
V	3	3	2	1	2	3
E	4	4	3	2	2	2

Els últims caràcters de AMOR i MOVIE són iguals, de manera que la distància d'edició entre ells és igual a la distància d'edició entre LOV i MOVI. Podem fer servir una operació d'edició per eliminar l'últim caràcter I de MOVI. Per tant, la distància d'edició és un més gran que la distància d'edició entre LOV i MOV, etc.

7.6 Comptar rajoles

De vegades, els estats d'una solució de programació dinàmica són més complexes que les combinacions fixes de nombres. Per exemple, considerem el problema de calcular el nombre de maneres diferents d'omplir una quadrícula $n \times m$ fent servir fitxes de mida 1×2 i 2×1 . Una solució vàlida per a la quadrícula 4×7 és



i el nombre total de solucions és 781.

El problema es pot resoldre amb programació dinàmica si passem per la quadrícula fila per fila. Cada fila d'una solució es pot representar com una cadena que conté m caràcters del conjunt $\{\sqcap, \sqcup, \sqsubset, \sqsupset\}$. Per exemple, la solució anterior consta de quatre files que es corresponen amb les següents cadenes:

- $\sqcap \sqsubset \sqcap \sqsubset \sqcap \sqsubset \sqcap$
- $\sqcup \sqsubset \sqcup \sqcap \sqcap \sqcup$
- $\sqsubset \sqsubset \sqsubset \sqcup \sqcup \sqcap$
- $\sqsubset \sqsubset \sqsubset \sqsubset \sqcup$

Sigui $\text{count}(k, x)$ el nombre de maneres de construir una solució per a les files $1 \dots k$ de la graella de manera que la cadena x correspon amb la fila k . Aquí és possible utilitzar la programació dinàmica, perquè l'estat d'una fila està restringit només per l'estat de la fila anterior.

Una solució és vàlida si la fila 1 no conté el caràcter \sqcup , la fila n no conté el caràcter \sqcap , i totes les files consecutives són *compatibles*. Per exemple, les files $\sqcup \sqsubset \sqcup \sqcap \sqcap \sqcup$ i $\sqsubset \sqsubset \sqsubset \sqcup \sqcup \sqcap$ són compatibles, mentre que les files $\sqcap \sqsubset \sqcap \sqsubset \sqcap$ i $\sqsubset \sqsubset \sqsubset \sqsubset \sqcup$ no són compatibles.

Com que una fila consta de m caràcters i n'hi ha quatre opcions per a cada caràcter, el nombre de files diferents és com a màxim 4^m . Per tant, la complexitat temporal de la solució és $O(n4^{2m})$ perquè podem passar pels $O(4^m)$ estats possibles de cada fila i, per a cada estat, hi ha $O(4^m)$ estats possibles de la fila anterior. A la pràctica, és una bona idea girar la quadrícula per a que el costat més curt tingui longitud m , donat que el factor 4^{2m} domina la complexitat temporal.

És possible millorar la solució amb una representació més compacta de les files. Resulta que n'hi ha prou amb saber quines de les columnes de la fila anterior contenen el quadrat superior d'una rajola vertical. Així, podem representar una fila utilitzant només caràcters \sqcap i \square , on \square és una combinació de caràcters \sqcup , \sqsubset i \sqsupset . Fent servir aquesta representació, només n'hi ha 2^m files diferents, i la complexitat temporal és $O(n2^{2m})$.

Com a nota final, també hi ha una fórmula directa sorprenent per calcular el nombre de rajoles³:

$$\prod_{a=1}^{\lceil n/2 \rceil} \prod_{b=1}^{\lceil m/2 \rceil} 4 \cdot \left(\cos^2 \frac{\pi a}{n+1} + \cos^2 \frac{\pi b}{m+1} \right)$$

Aquesta fórmula és molt eficient, perquè calcula el nombre de mosaics en $O(nm)$ temps, però com que la resposta és un producte de nombres reals, fer servir la fórmula requereix emmagatzemar els resultats intermedis amb precisió.

³Sorprenentment, aquesta fórmula va ser descoberta l'any 1961 per dos equips de recerca [43, 67] que funcionaven de manera independent.

Capítol 8

Anàlisi amortitzada

La complexitat temporal d'un algorisme sovint és fàcil d'analitzar simplement mirant l'estructura de l'algorisme: quins bucles conté l'algorisme i quantes vegades s'executen. Tanmateix, de vegades una anàlisi directa no dona una imatge real de l'eficiència de l'algorisme.

L'anàlisi amortitzada es pot fer servir per analitzar algorismes que contenen operacions la complexitat temporal dels quals varia. La idea és estimar el temps total utilitzat per totes aquestes operacions durant l'execució de l'algorisme, en lloc de centrar-se en les operacions individuals.

8.1 Mètode dels dos punters

En el **mètode dels dos punters**, es fan servir dos punters per iterar pels valors d'un vector. Els punters només es poden moure en una direcció, cosa que garanteix que l'algorisme funciona de manera eficient. A continuació discutim dos problemes que es poden resoldre mitjançant el mètode dels dos punters.

Suma de subvector

Com a primer exemple, considerem un problema en què se'ns dona un vector de n enters positius i una suma objectiu x , i volem trobar un subvector la suma del qual és x o informar que no hi ha aquesta subvector.

Per exemple, el vector

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

conté un subvector la suma del qual és 8:

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

Aquest problema es pot resoldre en temps $O(n)$ fent servir el mètode dels dos punters. La idea és mantenir punters que assenyalin el primer i l'últim valor d'un subvector. En cada gir, el punter esquerre es mou un pas cap a la dreta i el punter dret es mou cap a la dreta sempre que la suma del subvector resultant

sigui com a màxim x . Si la suma es converteix exactament en x , s'ha trobat una solució.

Com a exemple, considereu el vector següent i una suma objectiu $x = 8$:

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

El subvector inicial conté els valors 1, 3 i 2 la suma dels quals és 6:

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

Aleshores, el punter esquerre es mou un pas cap a la dreta. El punter dret no es mou, perquè, en cas contrari, la suma del subvector superaria x .

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

De nou, el punter esquerre es mou un pas cap a la dreta, i aquesta vegada el punter dret es mou tres passos cap a la dreta. La suma del subvector és $2 + 5 + 1 = 8$, de manera que s'ha trobat un subvector la suma del qual és x .

1	3	2	5	1	1	2	3
---	---	---	---	---	---	---	---

El temps d'execució de l'algorisme depèn del nombre de passos que es mou el punter dret. Tot i que no hi ha cap límit superior útil sobre quants passos pot moure el punter en un *única* gir, sabem que el punter es mou *un total de* $O(n)$ passos durant l'algorisme, perquè només es mou cap a la dreta.

Com que tant el punter esquerre com el dret es mouen $O(n)$ passos durant l'algorisme, l'algorisme funciona en el temps $O(n)$.

Problema 2SUMA

Un altre problema que es pot resoldre mitjançant el mètode dels dos punters és el següent problema, també conegut com a **problema 2SUMA**: donat un vector de n nombres i una suma objectiu x , trobeu, si existeixen, dos valors del vector de manera que suma és x .

Per resoldre el problema, primer ordenem els valors del vector en ordre creixent. Després d'això, iterem el vector fent servir dos punters. El punter esquerre comença al primer valor i es mou un pas cap a la dreta en cada torn. El punter dret comença a l'últim valor i sempre es mou cap a l'esquerra fins que la suma del valor esquerre i dret és com a màxim x . Si la suma és exactament x , s'ha trobat una solució.

Per exemple, considereu el vector següent i una suma objectiu $x = 12$:

1	4	5	6	7	9	9	10
---	---	---	---	---	---	---	----

Les posicions inicials dels punters són les següents. La suma dels valors és $1 + 10 = 11$ que és més petit que x .



A continuació, el punter esquerre es mou un pas cap a la dreta. El punter dret es mou tres passos cap a l'esquerra i la suma es converteix en $4 + 7 = 11$.



Després d'això, el punter esquerre torna a moure's un pas cap a la dreta. El punter dret no es mou i s'ha trobat una solució $5 + 7 = 12$.



El temps d'execució de l'algorisme és $O(n \log n)$, perquè primer ordena el vector en temps $O(n \log n)$, i després els dos punters mouen $O(n)$ passos.

Tingueu en compte que és possible resoldre el problema d'una altra manera en temps $O(n \log n)$ fent servir la cerca binària. En aquesta solució, iterem a través del vector i per a cada valor del vector intentem trobar un altre valor que produeixi la suma x . Això es pot fer fent n cerques binàries, cadascuna de les quals triga temps $O(\log n)$.

Un problema més difícil és el **problema 3SUMA**, on es demana trobar *tres* valors del vector la suma dels quals és x . Utilitzant la idea de l'algorisme anterior, aquest problema es pot resoldre en temps $O(n^2)$ ¹. Veus com?

8.2 Element menor més propers

L'anàlisi amortitzada s'utilitza sovint per estimar el nombre d'operacions realitzades en una estructura de dades. Les operacions poden estar distribuïdes de manera desigual, de tal forma que la majoria d'operacions tenen lloc en una fase determinada de l'algorisme, però el nombre total d'operacions és limitat.

Per exemple, considereu el problema de trobar per a cada element d'un vector l'**element menor més proper**, és a dir, el primer element menor que l'element original i que el precedeix en el vector. És possible que no existeixi aquest element, i en aquest cas l'algorisme hauria d'informar-ho. A continuació veurem com es pot resoldre el problema de manera eficient mitjançant una estructura de pila.

Recorrem el vector d'esquerra a dreta i mantenim una pila d'elements del vector. Per cada posició del vector, traiem elements de la pila fins que l'element superior sigui més petit que l'element actual o la pila estigui buida. Aleshores,

¹Durant molt de temps, es va pensar que resoldre el problema 3SUMA de manera més eficient que en temps $O(n^2)$ no seria possible. Tanmateix, el 2014, es va veure en [30] que no era així.

informem que l'element superior és l'element menor més proper a l'element actual, o si la pila està buida, l'element no existeix. Finalment, afegim l'element actual a la pila.

Com a exemple, considereu el vector següent:

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

En primer lloc, els elements 1, 3 i 4 s'afegeixen a la pila, perquè cada element és més gran que l'element anterior. Així, l'element menor més proper de 4 és 3 i l'element menor més proper de 3 és 1.

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

1 → 3 → 4

El següent element 2 és menor que els dos elements superiors de la pila. Així, els elements 3 i 4 s'eliminen de la pila i, a continuació, s'afegeix l'element 2 a la pila. El seu element menor més proper és 1:

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

1 → 2

Aleshores, l'element 5 és més gran que l'element 2, de manera que s'afegirà a la pila i el seu element menor més proper és 2:

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

1 → 2 → 5

Després d'això, l'element 5 s'elimina de la pila i els elements 3 i 4 s'afegeixen a la pila:

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

1 → 2 → 3 → 4

Finalment, tots els elements excepte l'1 s'eliminen de la pila i l'últim element 2 s'afegeix a la pila:

1	3	4	2	5	3	4	2
---	---	---	---	---	---	---	---

1 → 2

L'eficiència de l'algorisme depèn del nombre total d'operacions fets a la pila. Si l'element actual és més gran que l'element superior de la pila, s'afegeix directament a la pila, la qual cosa és eficient. Tanmateix, de vegades la pila pot contenir diversos elements més grans i es necessita temps per eliminar-los. Tot i així, cada element s'afegeix *exactament una vegada* a la pila i s'elimina *com a màxim una vegada* de la pila. Així, cada element provoca $O(1)$ operacions de pila, i l'algorisme funciona en temps $O(n)$.

8.3 Mínim de finestra lliscant

Una finestra lliscant (**sliding window**) és un subvector de mida constant que es mou d'esquerra a dreta a través del vector. A cada posició de la finestra, volem calcular una certa informació sobre els elements dins de la finestra. En aquesta secció, ens centrem en el problema de mantenir el **mínim de la finestra**, és a dir, el valor més petit dins de cada finestra.

El mínim de la finestra lliscant es pot calcular fent servir una idea semblant a la dels elements menors més propers. Mantenim una cua on cada element és més gran que l'element anterior, i el primer element sempre es correspon amb l'element mínim de la finestra. Després de cada moviment de la finestra, treiem elements del final de la cua fins que l'últim element de la cua sigui més petit que l'element de la nova finestra, o la cua quedi buida. També eliminem el primer element de la cua si ja no està dins de la finestra. Finalment, afegim l'element de la nova finestra al final de la cua.

Com a exemple, considereu el vector següent:

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---

Suposem que la mida de la finestra lliscant és 4. A la primera posició de la finestra, el valor més petit és 1:

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---

1	→	4	→	5
---	---	---	---	---

Aleshores, la finestra es mou un pas cap a la dreta. El nou element 3 és més petit que els elements 4 i 5 de la cua, de manera que els elements 4 i 5 s'eliminen de la cua i l'element 3 s'afegeix a la cua. El valor més petit segueix sent 1.

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---

1	→	3
---	---	---

Després d'això, la finestra es mou de nou i l'element més petit 1 ja no pertany a la finestra. Així, s'elimina de la cua i el valor més petit és ara 3. També s'afegeix el nou element 4 a la cua.

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---

3	→	4
---	---	---

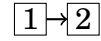
El següent element nou 1 és més petit que tots els elements de la cua. Així, tots els elements s'eliminen de la cua i aquesta només contindrà l'element 1:

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---

1

Finalment la finestra arriba a la seva última posició. L'element 2 s'afegeix a la cua, però el valor més petit dins de la finestra segueix sent 1.

2	1	4	5	3	4	1	2
---	---	---	---	---	---	---	---



Com que cada element del vector s'afegeix a la cua exactament una vegada i s'elimina de la cua com a màxim una vegada, l'algorisme funciona en temps $O(n)$.

Capítol 9

Consultes d'interval

En aquest capítol parlem de les estructures de dades que ens permeten processar de manera eficient les consultes d'interval. En una **consulta d'interval**, la nostra tasca és calcular un valor basat en un subvector d'un vector. Les consultes d'interval típiques són:

- $\text{sum}_q(a, b)$: calculate the sum of values in range $[a, b]$
- $\text{min}_q(a, b)$: find the minimum value in range $[a, b]$
- $\text{max}_q(a, b)$: find the maximum value in range $[a, b]$

Per exemple, considereu l'interval $[3, 6]$ al vector següent:

0	1	2	3	4	5	6	7
1	3	8	4	6	1	3	4

En aquest cas, $\text{sum}_q(3, 6) = 14$, $\text{min}_q(3, 6) = 1$ i $\text{max}_q(3, 6) = 6$.

Una manera senzilla de processar les consultes d'interval és fer servir un bucle que recorre tots els valors de matriu de l'interval. Per exemple, la funció següent es pot fer servir per processar consultes de suma en un vector:

```
int sum(int a, int b) {
    int s = 0;
    for (int i = a; i <= b; i++) {
        s += array[i];
    }
    return s;
}
```

Aquesta funció funciona en temps $O(n)$, on n és la mida del vector. Així, podem processar q consultes en temps $O(nq)$ fent servir la funció. Tanmateix, si n i q són grans, aquest enfocament és lent. Afortunadament, resulta que hi ha maneres de processar les consultes d'interval de manera molt més eficient.

9.1 Consultes de vector estàtiques

Primer ens centrem en una situació en què el vector és *estàtic*, és a dir, els valors del vector mai canvien entre consultes. En aquest cas, n'hi ha prou amb construir una estructura de dades estàtica que ens indiqui la resposta a qualsevol consulta possible.

Consultes de suma

Podem processar fàcilment les consultes de suma en un vector estàtic mitjançant la construcció d'un vector suma de prefixos (**prefix sum array**). Cada valor del vector suma de prefixos és igual a la suma de valors del vector original fins a aquesta posició, és a dir, el valor a la posició k és $\text{sum}_q(0, k)$. El vector suma de prefixos es pot construir en temps $O(n)$.

Per exemple, considereu el vector següent:

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

El vector suma de prefixos corresponent és:

0	1	2	3	4	5	6	7
1	4	8	16	22	23	27	29

Com que el vector suma de prefixos conté tots els valors de $\text{sum}_q(0, k)$, podem calcular qualsevol valor de $\text{sum}_q(a, b)$ en temps $O(1)$ com segueix:

$$\text{sum}_q(a, b) = \text{sum}_q(0, b) - \text{sum}_q(0, a - 1)$$

Definit $\text{sum}_q(0, -1) = 0$, la fórmula anterior també es compleix quan $a = 0$.

Per exemple, considereu l'interval $[3, 6]$:

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

En aquest cas $\text{sum}_q(3, 6) = 8 + 6 + 1 + 4 = 19$. Aquesta suma es pot calcular a partir de dos valors de la matriu de suma de prefix:

0	1	2	3	4	5	6	7
1	4	8	16	22	23	27	29

Així, $\text{sum}_q(3, 6) = \text{sum}_q(0, 6) - \text{sum}_q(0, 2) = 27 - 8 = 19$.

També és possible generalitzar aquesta idea a dimensions superiors. Per exemple, podem construir una matriu de suma de prefixos bidimensional que es pot utilitzar per calcular la suma de qualsevol submatriu rectangular en temps $O(1)$. Cada suma d'aquesta matriu correspon a una submatriu que comença a la cantonada superior esquerra de la matriu.

La figura següent mostra aquesta idea:

		<i>D</i>				<i>C</i>			
		<i>B</i>				<i>A</i>			

La suma de la submatriu gris es pot calcular mitjançant la fórmula

$$S(A) - S(B) - S(C) + S(D),$$

on $S(X)$ indica la suma de valors d'una submatriu des de la cantonada superior esquerra fins a la posició de X .

Consultes mínimes

Les consultes de mínim són més difícils de resoldre que les consultes de suma. Tot i així, hi ha un mètode de preprocessament de temps $O(n \log n)$ força senzill després del qual podem respondre qualsevol consulta mínima en $temps O(1)$ ¹. Tingueu en compte que com que les consultes mínimes i màximes es poden processar de manera similar, ens podem centrar en les consultes mínimes.

La idea és precalcular tots els valors de $\min_q(a, b)$ on $b - a + 1$ (la longitud de l'interval) és una potència de dos. Per exemple, per al vector

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

es calculen els valors següents:

a	b	$\min_q(a, b)$	a	b	$\min_q(a, b)$	a	b	$\min_q(a, b)$
0	0	1	0	1	1	0	3	1
1	1	3	1	2	3	1	4	3
2	2	4	2	3	4	2	5	1
3	3	8	3	4	6	3	6	1
4	4	6	4	5	1	4	7	1
5	5	1	5	6	1	0	7	1
6	6	4	6	7	2			
7	7	2						

¹Aquesta tècnica es va introduir a [7] i de vegades s'anomena mètode del vector dispersa (**sparse array**). També hi ha tècniques més sofisticades [22] on el temps de preprocessament és només $O(n)$, però aquests algorismes no són necessaris en la programació competitiva.

El nombre de valors precalculats és $O(n \log n)$, perquè hi ha $O(\log n)$ longituds d'interval que són potències de dos. Els valors es poden calcular de manera eficient mitjançant la fórmula recursiva

$$\min_q(a, b) = \min(\min_q(a, a + w - 1), \min_q(a + w, b)),$$

on $b - a + 1$ és una potència de dos i $w = (b - a + 1)/2$. Calcular tots aquests valors requereix temps $O(n \log n)$.

Després d'això, qualsevol valor de $\min_q(a, b)$ es pot calcular en temps $O(1)$ com el mínim de dos valors precalculats. Sigui k la potència de dos més gran que no superi $b - a + 1$. Podem calcular el valor de $\min_q(a, b)$ mitjançant la fórmula

$$\min_q(a, b) = \min(\min_q(a, a + k - 1), \min_q(b - k + 1, b)).$$

A la fórmula anterior, l'interval $[a, b]$ es representa com la unió dels intervals $[a, a + k - 1]$ i $[b - k + 1, b]$, tots dos de longitud k .

Com a exemple, considereu l'interval $[1, 6]$:

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

La longitud de l'interval és 6, i la potència més gran de dos que no supera 6 és 4. Així, el rang $[1, 6]$ és la unió dels intervals $[1, 4]$ i $[3, 6]$:

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

0	1	2	3	4	5	6	7
1	3	4	8	6	1	4	2

Com que $\min_q(1, 4) = 3$ i $\min_q(3, 6) = 1$, concloem que $\min_q(1, 6) = 1$.

9.2 Arbre binari indexat

Un **arbre binari indexat** o un **arbre de Fenwick**² es pot veure com una variant dinàmica d'un vector suma de prefixos. Aquest admet dues operacions de temps $O(\log n)$: processar una consulta de suma d'interval i actualitzar un valor.

L'avantatge d'un arbre binari indexat és que ens permet actualitzar de manera eficient els valors del vector entre les consultes de suma. Això no és possible si fem servir un vector suma de prefix, perquè després de cada actualització, caldria tornar a reconstruir tot el vector suma de prefix en temps $O(n)$.

²L'estructura d'arbre binari indexat va ser presentada per P.M. Fenwick el 1994 [21].

Estructura

Encara que el nom de l'estructura sigui *arbre* binari indexat, normalment es representa com un vector. En aquesta secció suposem que tots els vectors comencen amb index 1 (en lloc de 0), perquè dóna lloc a una implementació més senzilla.

Sigui $p(k)$ la potència de dos més gran que divideix k . Emmagatzemem un arbre binari indexat com un vector arbre de manera que

$$\text{tree}[k] = \text{sum}_q(k - p(k) + 1, k),$$

és a dir, cada posició k conté la suma de valors en un interval del vector original la longitud del qual és $p(k)$ i que acaba a la posició k . Per exemple, com que $p(6) = 2$, $\text{tree}[6]$ conté el valor de $\text{sum}_q(5, 6)$.

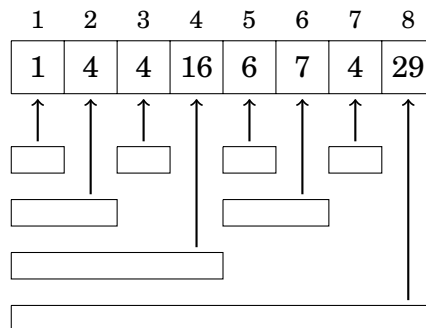
Per exemple, considereu el vector següent:

1	2	3	4	5	6	7	8
1	3	4	8	6	1	4	2

L'arbre binari indexat corresponent és el següent:

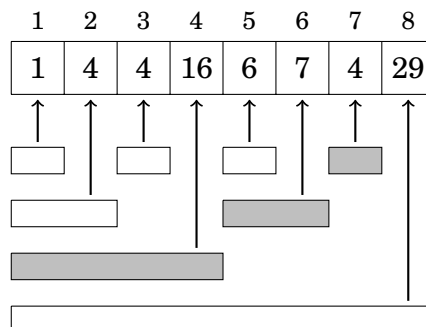
1	2	3	4	5	6	7	8
1	4	4	16	6	7	4	29

La imatge següent mostra més clarament com cada valor de l'arbre binari indexat correspon a un interval del vector original:



Fent servir un arbre binari indexat, qualsevol valor de $\text{sum}_q(1, k)$ es pot calcular en temps $O(\log n)$, perquè un interval $[1, k]$ sempre es pot dividir en $O(\log n)$ intervals les sumes dels quals estan emmagatzemades a l'arbre.

Per exemple, l'interval $[1, 7]$ consta dels intervals següents:



Així, podem calcular la suma corresponent de la següent manera:

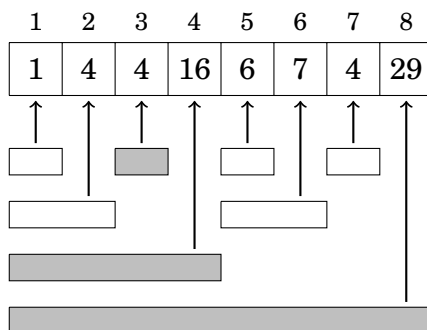
$$\text{sum}_q(1, 7) = \text{sum}_q(1, 4) + \text{sum}_q(5, 6) + \text{sum}_q(7, 7) = 16 + 7 + 4 = 27$$

Per calcular el valor de $\text{sum}_q(a, b)$ on $a > 1$, podem fer servir el mateix truc que hem fet servir amb els vectors suma de prefixos:

$$\text{sum}_q(a, b) = \text{sum}_q(1, b) - \text{sum}_q(1, a - 1).$$

Com que podem calcular tant $\text{sum}_q(1, b)$ com $\text{sum}_q(1, a - 1)$ en temps $O(\log n)$, la complexitat total és $O(\log n)$.

Quan actualitzem un valor en el vector original, hem d'actualitzar diversos valors de l'arbre binari indexat. Per exemple, si el valor a la posició 3 canvia, les sumes dels intervals següents canvien:



Com que cada element del vector pertany a $O(\log n)$ intervals de l'arbre binari indexat, n'hi ha prou amb actualitzar $O(\log n)$ valors de l'arbre.

Implementació

Les operacions d'un arbre binari indexat es poden implementar de manera eficient mitjançant operacions de bits. El fet clau és que podem calcular qualsevol valor de $p(k)$ mitjançant la fórmula

$$p(k) = k \& -k.$$

La funció següent calcula el valor de $\text{sum}_q(1, k)$:

```
int sum(int k) {
    int s = 0;
    while (k >= 1) {
        s += tree[k];
        k -= k&-k;
    }
    return s;
}
```

La funció següent augmenta en x unitats el valor del vector a la posició k (x pot ser positiu o negatiu):

```

void add(int k, int x) {
    while (k <= n) {
        tree[k] += x;
        k += k&-k;
    }
}

```

La complexitat temporal d'ambdues funcions és $O(\log n)$, perquè les funcions accedeixen als $O(\log n)$ valors de l'arbre binari indexat, i cada moviment a la següent posició triga temps $O(1)$.

9.3 Arbre de segments

Un **arbre de segments**³ (*segment tree*) és una estructura de dades que admet dues operacions: processar una consulta d'interval i actualitzar un valor del vector. Els arbres de segments poden suportar consultes de suma, consultes de mínim i màxim i moltes altres consultes perquè ambdues operacions funcionen en $O(\log n)$ temps.

En comparació amb un arbre binari indexat, l'avantatge d'un arbre de segments és que és una estructura de dades més general. Tot i que els arbres indexats binaris només admeten consultes de suma⁴, Els arbres de segments també admeten altres consultes. D'altra banda, un arbre de segments requereix més memòria i és una mica més difícil d'implementar.

Estructura

Un arbre de segments és un arbre binari on els nodes del nivell inferior de l'arbre corresponen als elements del vector i els altres nodes contenen la informació necessària per a processar les consultes d'interval.

En aquesta secció, suposem que la mida del vector és una potència de dos i utilitzem una indexació basada en zero, perquè resulta més convenient. Si la mida del vector no és una potència de dos, sempre podem afegir-hi elements addicionals.

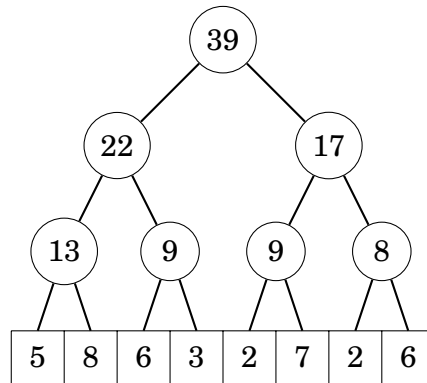
Primer parlarem dels arbres de segments que admeten consultes de suma. Com a exemple, considereu el vector següent:

0	1	2	3	4	5	6	7
5	8	6	3	2	7	2	6

L'arbre de segments corresponent és el següent:

³La implementació de baix a dalt d'aquest capítol correspon a la de [62]. Estructures similars es van fer servir a finals dels 70 per a resoldre problemes geomètrics [9].

⁴De fet, utilitzant *dos* arbres indexats binaris és possible suportar consultes de mínim [16], però és més complicat que utilitzar un arbre de segments.

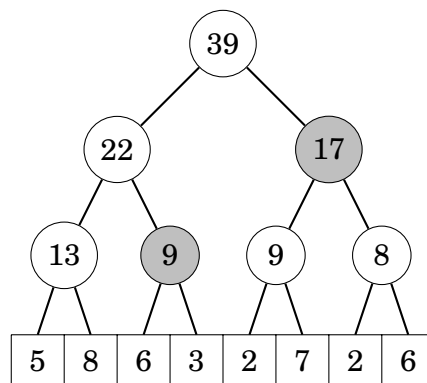


Cada node intern de l'arbre es correspon a un interval del vector la mida del qual és una potència de dos. En l'arbre anterior, el valor de cada node intern és la suma dels valors corresponents del vector i es pot calcular com la suma dels valors del fill esquerre i dret.

Resulta que qualsevol rang $[a, b]$ es pot dividir en $O(\log n)$ els valors dels quals s'emmagatzemen als nodes de l'arbre. Per exemple, considereu l'interval $[2, 7]$:

0	1	2	3	4	5	6	7
5	8	6	3	2	7	2	6

Aquí $\text{sum}_q(2, 7) = 6 + 3 + 2 + 7 + 2 + 6 = 26$. En aquest cas, els dos nodes següents es corresponen amb l'interval:

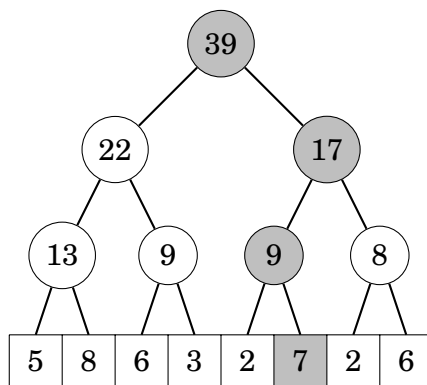


Així, una altra manera de calcular la suma és $9 + 17 = 26$.

Quan la suma es calcula fent servir nodes situats el més alt possible a l'arbre, fan falta com a màxim dos nodes a cada nivell de l'arbre. Per tant, el nombre total de nodes és $O(\log n)$.

Quan actualitzem un element del vector, hem d'actualitzar tots els nodes el valor dels quals depèn de l'element actualitzat. Això es pot fer travessant el camí des de l'element actualitzat fins al node superior i actualitzant els nodes al llarg del camí.

La imatge següent mostra quins nodes d'arbre canvien si l'element 7 del vector canvia:

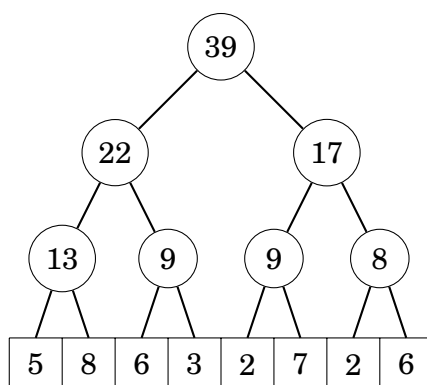


El camí de baix a dalt sempre consta de $O(\log n)$ nodes, de manera que cada actualització té aquest cost.

Implementació

Emmagatzemem un arbre de segments com un vector de $2n$ elements on n és la mida potència de dos del vector original. Els nodes de l'arbre s'emmagatzemen de dalt a baix: $\text{tree}[1]$ és el node superior, $\text{tree}[2]$ i $\text{tree}[3]$ són els seus fills, etcètera. Finalment, els valors de $\text{tree}[n]$ a $\text{tree}[2n - 1]$ corresponen als valors del vector original al nivell inferior de l'arbre.

Per exemple, l'arbre de segments



s'emmagatzema de la manera següent:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
39	22	17	13	9	9	8	5	8	6	3	2	7	2	6

Fent servir aquesta representació, el pare de $\text{tree}[k]$ és $\text{tree}[\lfloor k/2 \rfloor]$, i els seus fills són $\text{tree}[2k]$ i $\text{tree}[2k + 1]$. Tingueu en compte que això implica que la posició d'un node és parell si és fill esquerre i imparell si és fill dret.

La funció següent calcula el valor de $\text{sum}_q(a, b)$:

```
int sum(int a, int b) {
    a += n; b += n;
    int s = 0;
    while (a <= b) {
```

```

    if (a%2 == 1) s += tree[a++];
    if (b%2 == 0) s += tree[b--];
    a /= 2; b /= 2;
}
return s;
}

```

La funció manté un interval que inicialment és $[a + n, b + n]$. Aleshores, a cada pas, l'interval es mou al nivell superior en l'arbre, i abans d'això, els valors dels nodes que no pertanyen a l'interval superior s'afegeixen a la suma.

La funció següent incrementa en x unitats l'element a la posició k del vector:

```

void add(int k, int x) {
    k += n;
    tree[k] += x;
    for (k /= 2; k >= 1; k /= 2) {
        tree[k] = tree[2*k] + tree[2*k+1];
    }
}

```

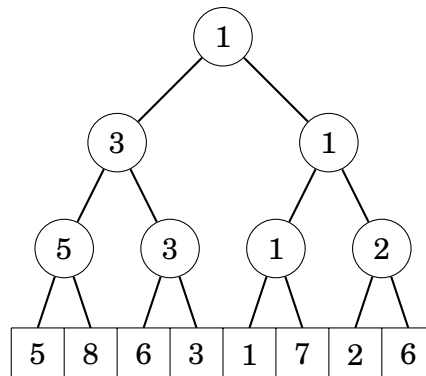
Primer, la funció actualitza el valor al nivell inferior de l'arbre. Després d'això, la funció actualitza els valors de tots els nodes interns de l'arbre, fins que arriba al node superior de l'arbre.

Les dues funcions anteriors funcionen en temps $O(\log n)$, perquè un arbre de segments de n elements consta de $O(\log n)$ nivells, i les funcions puguen l'arbre un nivell a cada pas.

Altres consultes

Els arbres de segments poden suportar totes les consultes d'interval on és possible dividir un interval en dues parts, calcular la resposta per separat per a ambdues parts i després combinar les respostes de manera eficient. Exemples d'aquestes consultes són el mínim i el màxim, el màxim comú divisor i les operacions de bits *and*, *or* i *xor*.

Per exemple, l'arbre de segment següent admet consultes de mínim:

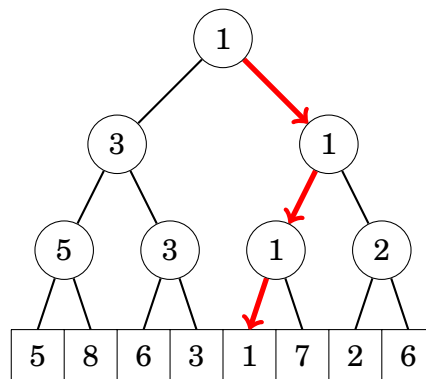


En aquest cas, cada node d'arbre conté el valor més petit de l'interval del vector corresponent. El node superior de l'arbre conté el valor més petit de tot

el vector. Les operacions es poden implementar com abans, però en comptes de sumes, es calculen mínims.

L'estructura d'arbre de segments també ens permet fer servir la cerca binària per a localitzar elements del vector. Per exemple, si l'arbre admet consultes de mínims, podem trobar la posició de l'element més petit en temps $O(\log n)$.

Per exemple, a l'arbre anterior, es pot trobar l'element amb el valor mínim 1 travessant un camí cap avall des del node superior:



9.4 Tècniques addicionals

Compressió de l'índex

Una limitació de les estructures de dades que es construeixen sobre un vector és que els elements s'indexen mitjançant nombres enters consecutius. Les dificultats sorgeixen quan es necessiten índexs grans. Per exemple, si volem fer servir l'índex 10^9 , el vector hauria de contenir 10^9 elements que requeririen massa memòria.

No obstant això, sovint podem ignorar aquesta limitació fent servir compressió de l'índex (**index compression**), on els índexs originals es substitueixen per índexs 1,2,3,, etc. Això es pot fer si coneixem prèviament tots els índexs necessaris durant l'algorisme.

La idea és substituir cada índex original x per $c(x)$ on c és una funció que comprimeix els índexs. Necessitem que l'ordre dels índexs no canviï, de manera que si $a < b$, llavors $c(a) < c(b)$. Això ens permet realitzar consultes còmodament encara que els índexs estiguin comprimits.

Per exemple, si els índexs originals són 555, 10^9 i 8, els nous índexs són:

$$\begin{aligned} c(8) &= 1 \\ c(555) &= 2 \\ c(10^9) &= 3 \end{aligned}$$

Actualitzacions d'interval

Fins ara, hem implementat estructures de dades que admeten consultes d'interval i actualitzacions de valors únics. Considerem ara una situació oposada, on hem d'actualitzar intervals i recuperar valors únics. Ens centrem en una operació que augmenta tots els elements d'un interval $[a, b]$ en x .

Sorprenentment, podem utilitzar les estructures de dades presentades en aquest capítol també en aquesta situació. Per a fer-ho, construïm una **vector de diferències** els valors del qual indiquen les diferències entre valors consecutius del vector original. Així, el vector original és el vector suma de prefixos del vector de diferències. Per exemple, considereu el vector següent:

0	1	2	3	4	5	6	7
3	3	1	1	1	5	2	2

El vector de diferències per al vector anterior és el següent:

0	1	2	3	4	5	6	7
3	0	-2	0	0	4	-3	0

Per exemple, el valor 2 a la posició 6 del vector original correspon a la suma $3 - 2 + 4 - 3 = 2$ al vector de diferències.

L'avantatge del vector de diferències és que podem actualitzar un interval del vector original canviant només dos elements del vector de diferències. Per exemple, si volem augmentar en 5 els valors del vector original entre les posicions 1 i 4, n'hi ha prou amb augmentar en 5 el valor del vector de diferències a la posició 1 i disminuir en 5 el valor a la posició 5. El resultat és el següent:

0	1	2	3	4	5	6	7
3	5	-2	0	0	-1	-3	0

De manera més general, per augmentar en x els valors de l'interval $[a, b]$, augmentem en x el valor a la posició a i reduïm en x el valor a la posició $b + 1$. Per tant, només cal actualitzar valors únics i processar consultes de suma, de manera que podem utilitzar un arbre binari indexat o un arbre de segments.

Un problema més difícil és donar tant suport a les consultes d'interval com a les actualitzacions d'interval. Al capítol 28 veurem que fins i tot això és possible.

Manipulació de bits

10.1 Representació binària

Aquí teniu la representació binària del nombre de tipus int 43:

$$b_k 2^k + \dots + b_2 2^2 + b_1 2^1 + b_0 2^0.$$
$$1 \cdot 2^5 + 1 \cdot 2^3 + 1 \cdot 2^1 + 1 \cdot 2^0 = 43.$$

99

Per exemple, la representació binària del nombre int `-43` és

1111111111111111111111111010101.

En una representació sense signe, només es poden fer servir nombres no negatius, però el límit superior dels valors és més gran. Una variable sense signe de n bits pot contenir qualsevol nombre enter entre 0 i $2^n - 1$. Per exemple, en C++, una variable `unsigned int` pot contenir qualsevol nombre enter entre 0 i $2^{32} - 1$.

Hi ha una connexió entre les representacions: un nombre amb signe $-x$ és igual a un nombre sense signe $2^n - x$. Per exemple, el codi següent mostra que el número amb signe $x = -43$ és igual al nombre sense signe $y = 2^{32} - 43$:

```
int x = -43;
unsigned int y = x;
cout << x << "\n"; // -43
cout << y << "\n"; // 4294967253
```

Si un nombre és més gran que el límit superior de la representació de bits, el nombre es desbordarà. En una representació amb signe, el nombre que segueix $2^{n-1} - 1$ és -2^{n-1} , i en una representació sense signe, el nombre que segueix $2^n - 1$ és 0¹. Per exemple, considereu el codi següent:

```
int x = 2147483647
cout << x << "\n"; // 2147483647
x++;
cout << x << "\n"; // -2147483648
```

Inicialment, el valor de x és $2^{31} - 1$. Aquest és el valor més gran que es pot emmagatzemar en una variable int, de manera que el nombre següent després de $2^{31} - 1$ és -2^{31} .

10.2 Operacions de bits

Operació *and*

L'operació **and** x & y produeix un nombre que té un bit en les posicions on tant x com y tenen un bit. Per exemple, $22 \& 26 = 18$, perquè

$$\begin{array}{rcl} & 10110 & (22) \\ \& & 11010 & (26) \\ \hline = & 10010 & (18) \end{array}$$

¹(N. del T.) En C++, està permès fer operacions aritmètiques que causin *overflow* en tipus sense signe, però fer-ho amb tipus amb signe és comportament no definit (*undefined behavior*). No és un problema dels processadors, que poden fer *overflow* de nombres amb signe sense problema, sinó dels compiladors de C++, que optimitzen el codi agressivament sota el supòsit que el vostre programa mai comet *undefined behavior*. No ho feu.

Mitjançant l'operació (*and*), podem comprovar si un nombre x és parell perquè $x \& 1 = 0$ si x és parell, i $x \& 1 = 1$ si x és senar. De manera més general, x és divisible per 2^k exactament quan $x \& (2^k - 1) = 0$.

Operació *or*

L'operació **OR** $x \mid y$ produeix un nombre que té un bit en les posicions on x o y tenen un bit. Per exemple, $22 \mid 26 = 30$, perquè

$$\begin{array}{r} 10110 \quad (22) \\ \mid 11010 \quad (26) \\ \hline = 11110 \quad (30) \end{array}$$

Operació *xor*

L'operació **xor** $x \wedge y$ produeix un nombre que té un bit en les posicions on exactament un de x i y tenen un bit. Per exemple, $22 \wedge 26 = 12$, perquè

$$\begin{array}{r} 10110 \quad (22) \\ \wedge 11010 \quad (26) \\ \hline = 01100 \quad (12) \end{array}$$

Operació *not*

L'operació **not** $\sim x$ produeix un nombre on tots els bits de x s'han invertit. En complement a 2 es compleix que $\sim x = -x - 1$. Per exemple, $\sim 29 = -30$.

El resultat de l'operació *not* a nivell de bits depèn de la longitud de la representació binària, perquè l'operació inverteix tots els bits. Per exemple, si els nombres són *ints* de 32 bits, el resultat és el següent:

$$\begin{array}{rcl} x & = & 29 \quad 000000000000000000000000011101 \\ \sim x & = & -30 \quad 111111111111111111111111110010 \end{array}$$

Desplaçament de bits

El desplaçament de bits (*bit shift*) a l'esquerra $x \ll k$ afegeix k bits zero al nombre, i el desplaçament de bits a la dreta $x \gg k$ elimina els darrers k bits del nombre. Per exemple, $14 \ll 2 = 56$, perquè 14 i 56 són 1110 i 111000 en binari. De la mateixa manera, $49 \gg 3 = 6$, perquè 49 i 6 són 110001 i 110.

Tingueu en compte que $x \ll k$ correspon a multiplicar x per 2^k , i $x \gg k$ correspon a dividir x per 2^k arrodonint a un nombre enter.

Aplicacions

Un nombre de la forma $1 \ll k$ té un bit a la posició k i tots els altres bits són zero, de manera que podem fer servir aquests nombres per accedir als bits individuals d'un nombre donat. En particular, el k -èsim bit d'un nombre és 1 exactament quan $x \& (1 \ll k)$ no és zero. El codi següent imprimeix la representació binària d'un nombre x de tipus *int*:

```
for (int i = 31; i >= 0; i--) {
    if (x & (1 << i)) cout << "1";
    else cout << "0";
}
```

També és possible modificar bits individuals fent servir idees similars. Per exemple, la fórmula $x \mid (1 \ll k)$ posa el k -èsim bit de x a 1, la fórmula $x \& \sim(1 \ll k)$ posa el k -èsim bit de x a 0, i la fórmula $x \wedge (1 \ll k)$ inverteix el k -èsim bit de x .

La fórmula $x \& (x - 1)$ posa l'últim bit de x a zero, i la fórmula $x \& -x$ posa tots els bits a zero, excepte l'últim. La fórmula $x \mid (x - 1)$ inverteix tots els bits després de l'últim bit. Tingueu en compte també que un nombre positiu x és una potència de dos exactament quan $x \& (x - 1) = 0$.

Funcions addicionals

El compilador g++ proporciona les funcions següents per comptar bits:

- `__builtin_clz(x)`: el nombre de zeros al principi del nombre
- `__builtin_ctz(x)`: el nombre de zeros al final del nombre
- `__builtin_popcount(x)`: el nombre d'uns en el nombre
- `__builtin_parity(x)`: la paritat (parell o senar) del nombre d'uns

Aquestes funcions es poden fer servir de la manera següent:

```
int x = 5328; // 000000000000000000001010011010000
cout << __builtin_clz(x) << "\n"; // 19
cout << __builtin_ctz(x) << "\n"; // 4
cout << __builtin_popcount(x) << "\n"; // 5
cout << __builtin_parity(x) << "\n"; // 1
```

Tot i que les funcions anteriors només admeten nombres de tipus `int`, també hi ha versions `longlong` de les funcions amb el sufix `ll`.

10.3 Representació de conjunts

Cada subconjunt d'un conjunt $\{0, 1, 2, \dots, n - 1\}$ es pot representar com un nombre enter de n bits els bits del qual indiquen quins elements pertanyen al subconjunt. Aquesta és una manera eficient de representar conjunts, perquè cada element només requereix un bit de memòria i les operacions de conjunt es poden implementar com a operacions de bits.

Per exemple, com que `int` és un tipus de 32 bits, un nombre `int` pot representar qualsevol subconjunt del conjunt $\{0, 1, 2, \dots, 31\}$. La representació binària del conjunt $\{1, 3, 4, 8\}$ és

000000000000000000000000100011010,

que correspon al nombre $2^8 + 2^4 + 2^3 + 2^1 = 282$.

Implementació de conjunts

El codi següent declara una variable `int x` que pot contenir un subconjunt de $\{0, 1, 2, \dots, 31\}$. Després d'això, el codi afegeix els elements 1, 3, 4 i 8 al conjunt i imprimeix la seva mida.

```
int x = 0;
x |= (1<<1);
x |= (1<<3);
x |= (1<<4);
x |= (1<<8);
cout << __builtin_popcount(x) << "\n"; // 4
```

El codi següent imprimeix tots els elements que pertanyen al conjunt:

```
for (int i = 0; i < 32; i++) {
    if (x&(1<<i)) cout << i << " ";
}
// output: 1 3 4 8
```

Operacions de conjunts

Les operacions de conjunt es poden implementar de la següent manera com a operacions de bits:

	set syntax	bit syntax
intersection	$a \cap b$	$a \& b$
union	$a \cup b$	$a \mid b$
complement	\bar{a}	$\sim a$
difference	$a \setminus b$	$a \& (\sim b)$

Per exemple, el codi següent construeix primer els conjunts $x = \{1, 3, 4, 8\}$ i $y = \{3, 6, 8, 9\}$, i després construeix el conjunt $z = x \cup y = \{1, 3, 4, 6, 8, 9\}$:

```
int x = (1<<1)|(1<<3)|(1<<4)|(1<<8);
int y = (1<<3)|(1<<6)|(1<<8)|(1<<9);
int z = x|y;
cout << __builtin_popcount(z) << "\n"; // 6
```

Iteració de subconjunts

El codi següent passa per tots els subconjunts de $\{0, 1, \dots, n-1\}$:

```
for (int b = 0; b < (1<<n); b++) {
    // process subset b
}
```

El codi següent passa pels subconjunts amb exactament k elements:

```
for (int b = 0; b < (1<<n); b++) {
    if (__builtin_popcount(b) == k) {
        // process subset b
    }
}
```

El codi següent passa pels subconjunts d'un conjunt x :

```
int b = 0;
do {
    // process subset b
} while (b=(b-x)&x);
```

10.4 Optimitzacions de bits

Molts algorismes es poden optimitzar mitjançant operacions de bits. Aquestes optimitzacions no canvien la complexitat temporal de l'algorisme, però poden tenir un gran impacte en el temps real d'execució del codi. En aquesta secció comentem exemples d'aquestes situacions.

Distàncies de Hamming

La **distància de Hamming** $\text{hamming}(a, b)$ entre dues cadenes a i b d'igual longitud és el nombre de posicions on les cadenes difereixen. Per exemple,

$$\text{hamming}(01101, 11001) = 2.$$

Considereu el problema següent: donada una llista de n cadenes de bits, cadascuna de longitud k , calculeu la distància de Hamming mínima entre dues cadenes de la llista. Per exemple, la resposta per a $[00111, 01101, 11110]$ és 2, perquè

- $\text{hamming}(00111, 01101) = 2$,
- $\text{hamming}(00111, 11110) = 3$ i
- $\text{hamming}(01101, 11110) = 3$.

Una manera senzilla de resoldre el problema és passar per tots els parells de cordes i calcular les seves distàncies de Hamming, que dóna lloc a un algorisme de temps $O(n^2k)$. La funció següent es pot utilitzar per calcular distàncies:

```
int hamming(string a, string b) {
    int d = 0;
    for (int i = 0; i < k; i++) {
        if (a[i] != b[i]) d++;
    }
    return d;
}
```

Tanmateix, si k és petit, podem optimitzar el codi emmagatzemant les cadenes binàries com a nombres enters i calculant les distàncies de Hamming mitjançant operacions de bits. En particular, si $k \leq 32$, podem simplement emmagatzemar les cadenes com a valors `int` i utilitzar la funció següent per calcular distàncies:

```
int hamming(int a, int b) {
    return __builtin_popcount(a^b);
}
```

A la funció anterior, l'operació *xor* construeix una cadena binària que té un bit en posicions on a i b difereixen, i a continuació calculem el nombre de bits mitjançant la funció `__builtin_popcount`.

Per a comparar les implementacions, generem una llista de 10.000 cadenes binàries aleatòries de longitud 30. Amb el primer enfocament, la cerca triga 13.5 segons, però si fem servir l'optimització de bits, només triga 0.5 segons. El codi optimitzat per bits és gairebé 30 vegades més ràpid que el codi original.

Comptar subquadrícules

Com a altre exemple, considereu el problema següent: Donada una quadrícula $n \times n$ cada quadrat de la qual és negre (1) o blanc (0), calculeu el nombre de subquadrícules que tenen totes les cantonades negres. Per exemple, la quadrícula



conté dues subquadrícules d'aquest tipus:



Hi ha un algorisme de temps $O(n^3)$ que resol el problema: passeu per tots els parells de files $O(n^2)$ i per a cada parell (a, b) calculeu el nombre de columnes que contenen un quadrat negre a les dues files en $O(n)$ temps. El codi següent assumeix que `color[y][x]` denota el color de la fila y i la columna x :

```
int count = 0;
for (int i = 0; i < n; i++) {
    if (color[a][i] == 1 && color[b][i] == 1) count++;
}
```

Aleshores, aquestes columnes representen subquadrícules $\text{count}(\text{count} - 1)/2$ amb cantonades negres, perquè podem triar-ne dues per formar una subquadrícula.

Per optimitzar aquest algorisme, dividim la quadrícula en blocs de columnes de manera que cada bloc consta de N columnes consecutives. Aleshores, cada fila s'emmagatzema com una llista de números de N bits que descriuen els colors dels quadrats. Ara podem processar N columnes simultàniament gràcies a les operacions de bits. Al codi següent, `color[y][k]` representa un bloc de N colors com a bits.

```
int count = 0;
for (int i = 0; i <= n/N; i++) {
    count += __builtin_popcount(color[a][i]&color[b][i]);
}
```

L'algorisme resultant funciona en temps $O(n^3/N)$.

Generem una graella aleatòria de 2500×2500 i comparem la implementació original i la implementació optimitzada per bits. El codi original triga 29.6 segons, mentre que la versió optimitzada per bits només triga 3.1 segons amb $N = 32$ (nombres int) i 1.7 segons amb $N = 64$ (nombres longlong).

10.5 Programació dinàmica

Les operacions de bits proporcionen una manera eficient i còmoda d'implementar algorismes de programació dinàmica els estats dels quals contenen subconjunts d'elements, perquè aquests estats es poden emmagatzemar com a nombres enters. A continuació discutim exemples de combinació d'operacions de bits i programació dinàmica.

Selecció òptima

Com a primer exemple, considereu el problema següent: Ens donen els preus de k productes durant n dies i volem comprar cada producte exactament una vegada. Tanmateix, podem comprar com a màxim un producte al dia. Quin és el preu total mínim? Per exemple, considereu l'escenari següent ($k = 3$ i $n = 8$):

	0	1	2	3	4	5	6	7
product 0	6	9	5	2	8	9	1	6
product 1	8	2	6	2	7	5	7	2
product 2	5	3	9	7	3	5	1	4

En aquest escenari, el preu total mínim és de 5:

	0	1	2	3	4	5	6	7
product 0	6	9	5	2	8	9	1	6
product 1	8	2	6	2	7	5	7	2
product 2	5	3	9	7	3	5	1	4

Sigui $\text{price}[x][d]$ el preu del producte x el dia d . Per exemple, a l'escenari anterior $\text{price}[2][3] = 7$. Sigui $\text{total}(S, d)$ el preu total mínim per comprar un subconjunt S de productes en els primers d dies. Amb aquesta funció, la solució al problema és $\text{total}(\{0 \dots k-1\}, n-1)$.

Primer, $\text{total}(\emptyset, d) = 0$, perquè no costa res comprar un conjunt buit, i $\text{total}(\{x\}, 0) = \text{price}[x][0]$, perquè hi ha una manera de comprar un producte el primer dia. Aleshores, es pot utilitzar la recurrència següent:

$$\text{total}(S, d) = \min(\text{total}(S, d-1), \min_{x \in S} (\text{total}(S \setminus x, d-1) + \text{price}[x][d]))$$

Això vol dir que o bé no comprem cap producte el dia d o bé comprem un producte x que pertany a S . En aquest últim cas, eliminem x de S i afegim el preu de x al preu total.

El següent pas és calcular els valors de la funció mitjançant la programació dinàmica. Per emmagatzemar els valors de la funció, declarem una matriu

```
int total[1<<K][N];
```

on K i N són constants adequadament grans. La primera dimensió de la matriu correspon a una representació de bits d'un subconjunt.

En primer lloc, els casos en què $d = 0$ es poden processar de la següent manera:

```
for (int x = 0; x < k; x++) {
    total[1<<x][0] = price[x][0];
}
```

Aleshores, la recurrència es tradueix al codi següent:

```
for (int d = 1; d < n; d++) {
    for (int s = 0; s < (1<<k); s++) {
        total[s][d] = total[s][d-1];
        for (int x = 0; x < k; x++) {
            if (s & (1<<x)) {
                total[s][d] = min(total[s][d],
                                   total[s^(1<<x)][d-1] + price[x][d]);
            }
        }
    }
}
```

La complexitat temporal de l'algorisme és $O(n2^k k)$.

De les permutacions als subconjunts

Utilitzant la programació dinàmica, sovint és possible canviar una iteració sobre permutacions en una iteració sobre subconjunts². El benefici d'això és que $n!$, el

²Aquesta tècnica va ser introduïda el 1962 per M. Held i R. M. Karp [34].

nombre de permutacions, és molt més gran que 2^n , el nombre de subconjunts. Per exemple, si $n = 20$, llavors $n! \approx 2,4 \cdot 10^{18}$ i $2^n \approx 10^6$. Per tant, per a certs valors de n , podem passar de manera eficient pels subconjunts però no per les permutacions.

Com a exemple, considereu el problema següent: hi ha un ascensor amb un pes màxim x i n persones amb pes conegut que volen anar de la planta baixa a la planta superior. Quin és el nombre mínim de trajectes necessaris si les persones entren a l'ascensor en un ordre òptim?

Per exemple, suposem que $x = 10$, $n = 5$ i els pesos són els següents:

person	weight
0	2
1	3
2	3
3	5
4	6

En aquest cas, el nombre mínim de viatges és 2. Un ordre òptim és $\{0, 2, 3, 1, 4\}$, que divideix les persones en dos viatges: primer $\{0, 2, 3\}$ (pes total 10) i després $\{1, 4\}$ (pes total 9).

El problema es pot resoldre fàcilment en temps $O(n!n)$ provant totes les permutacions possibles de n persones. Tanmateix, podem utilitzar la programació dinàmica per obtenir un algorisme de temps $O(2^n n)$ més eficient. La idea és calcular per a cada subconjunt de persones dos valors: el nombre mínim de viatges necessaris i el pes mínim de les persones que viatgen en l'últim grup.

Sigui $\text{weight}[p]$ el pes de la persona p . Definim dues funcions: $\text{rides}(S)$ és el nombre mínim de viatges per a un subconjunt S , i $\text{last}(S)$ és el pes mínim de l'últim viatge. Per exemple, en l'escenari anterior

$$\text{rides}(\{1, 3, 4\}) = 2 \quad \text{i} \quad \text{last}(\{1, 3, 4\}) = 5,$$

perquè els viatges òptims són $\{1, 4\}$ i $\{3\}$, i el segon viatge té un pes de 5. Per descomptat, el nostre objectiu final és calcular el valor de $\text{rides}(\{0 \dots n-1\})$.

Podem calcular els valors de les funcions recurrents de manera directa i després aplicar la programació dinàmica. La idea és passar per totes les persones que pertanyen a S i triar de manera òptima l'última persona p que entra a l'ascensor. Cadascuna d'aquestes opcions genera un subproblema per a un subconjunt de persones més petit. Si es dona $\text{last}(S \setminus p) + \text{weight}[p] \leq x$, podem afegir p a l'últim viatge. En cas contrari, haurem de reservar un viatge nou que inicialment només contingui p .

Per implementar la programació dinàmica, declarem un vector

```
pair<int,int> best[1<<N];
```

que conté per a cada subconjunt S un parell $(\text{rides}(S), \text{últim}(S))$. Establim el valor per al grup buit de la següent manera:

```
best[0] = {1,0};
```

Aleshores, podem omplir el vector de la següent manera:

```
for (int s = 1; s < (1<<n); s++) {  
    // initial value: n+1 rides are needed  
    best[s] = {n+1,0};  
    for (int p = 0; p < n; p++) {  
        if (s&(1<<p)) {  
            auto option = best[s^(1<<p)];  
            if (option.second+weight[p] <= x) {  
                // add p to an existing ride  
                option.second += weight[p];  
            } else {  
                // reserve a new ride for p  
                option.first++;  
                option.second = weight[p];  
            }  
            best[s] = min(best[s], option);  
        }  
    }  
}
```

Observeu que el bucle anterior garanteix que, per a dos subconjunts S_1 i S_2 tals que $S_1 \subset S_2$, tractem S_1 abans que S_2 . És a dir, els valors de programació dinàmica es calculen en l'ordre correcte.

Comptar subconjunts

El nostre darrer problema en aquest capítol és el següent: sigui $X = \{0 \dots n-1\}$, i assignem a cada subconjunt $S \subset X$ un enter $\text{value}[S]$. La nostra tasca és calcular per cada S la suma dels valors dels subconjunts de S , és a dir,

$$\text{sum}(S) = \sum_{A \subset S} \text{value}[A],$$

Per exemple, suposem que $n = 3$ i els valors són els següents:

- $\text{value}[\emptyset] = 3$
- $\text{value}[\{0\}] = 1$
- $\text{value}[\{1\}] = 4$
- $\text{value}[\{0,1\}] = 5$
- $\text{value}[\{2\}] = 5$
- $\text{value}[\{0,2\}] = 1$
- $\text{value}[\{1,2\}] = 3$
- $\text{value}[\{0,1,2\}] = 3$

En aquest cas, per exemple,

$$\begin{aligned} \text{sum}(\{0,2\}) &= \text{value}[\emptyset] + \text{value}[\{0\}] + \text{value}[\{2\}] + \text{value}[\{0,2\}] \\ &= 3 + 1 + 5 + 1 = 10. \end{aligned}$$

Com que hi ha un total de 2^n subconjunts, una solució possible és passar per tots els parells de subconjunts en $O(2^{2n})$ temps. Tanmateix, utilitzant la

programació dinàmica, podem resoldre el problema en temps $O(2^n n)$. La idea és centrar-se en les sumes on els elements que es poden eliminar de S estan restringits.

Sigui $\text{partial}(S, k)$ la suma de valors dels subconjunts de S amb la restricció que només els elements $0 \dots k$ es poden eliminar de S . Per exemple,

$$\text{partial}(\{0, 2\}, 1) = \text{value}[\{2\}] + \text{value}[\{0, 2\}],$$

perquè només podem eliminar els elements $0 \dots 1$. Podem calcular valors de sum fent servir valors de partial , perquè

$$\text{sum}(S) = \text{partial}(S, n - 1).$$

Els casos base de la funció són

$$\text{partial}(S, -1) = \text{value}[S],$$

perquè en aquest cas no es pot eliminar cap element de S . Per al cas general podem fer servir la recurrència següent:

$$\text{partial}(S, k) = \begin{cases} \text{partial}(S, k - 1) & k \notin S \\ \text{partial}(S, k - 1) + \text{partial}(S \setminus \{k\}, k - 1) & k \in S \end{cases}$$

Aquí ens centrem en l'element k . Si $k \in S$, tenim dues opcions: podem mantenir k a S o eliminar-lo de S .

Hi ha una manera especialment intel·ligent d'implementar el càlcul de sumes. Podem declarar un vector

```
int sum[1<<N];
```

que conté la suma de cada subconjunt. El vector s'inicia de la següent manera:

```
for (int s = 0; s < (1<<n); s++) {
    sum[s] = value[s];
}
```

Aleshores, podem omplir el vector de la següent manera:

```
for (int k = 0; k < n; k++) {
    for (int s = 0; s < (1<<n); s++) {
        if (s & (1<<k)) sum[s] += sum[s ^ (1<<k)];
    }
}
```

Aquest codi calcula els valors de $\text{partial}(S, k)$ per a $k = 0 \dots n - 1$ al vector sum . Com que $\text{partial}(S, k)$ sempre es basa en $\text{partial}(S, k - 1)$, podem reutilitzar el vector sum , la qual cosa dóna una implementació molt eficient.

Part II

Algorismes de grafs

Capítol 11

Introducció als grafs

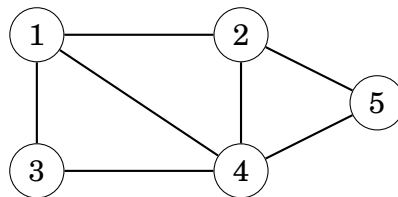
Molts problemes de programació es poden resoldre modelant el problema com un problema de grafs i fent servir l'algorisme de grafs adequat. Un exemple típic de grafs és la xarxa de carreteres i ciutats d'un país. De vegades, però, el graf està amagat dins del problema i pot ser difícil detectar-lo.

Aquesta part del llibre tracta els algorismes de grafs, especialment centrant-se en temes importants en la programació competitiva. En aquest capítol, repassem conceptes relacionats amb els grafs i estudiem diferents maneres de representar els grafs en algorismes.

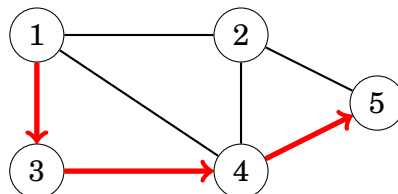
11.1 Vocabulari de grafs

Un **graf** consta de **nodes** i **arestes**. En aquest llibre, la variable n indica el nombre de nodes en un graf i la variable m indica el nombre d'arestes. Els nodes es numeren fent servir nombres $1, 2, \dots, n$.

Per exemple, el graf següent consta de 5 nodes i 7 arestes:



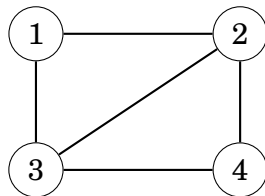
Un **camí** condueix des del node a al node b a través de les arestes del graf. La **longitud** d'un camí és el nombre d'arestes que té. Per exemple, el graf anterior conté un camí $1 \rightarrow 3 \rightarrow 4 \rightarrow 5$ de longitud 3 des del node 1 fins al node 5:



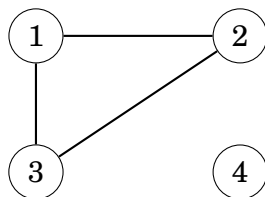
Un camí és un **cicle** si el primer i l'últim node són el mateix. Per exemple, el graf anterior conté un cicle $1 \rightarrow 3 \rightarrow 4 \rightarrow 1$. Un camí és **simple** si cada node apareix com a màxim una vegada al camí.

Connectivitat

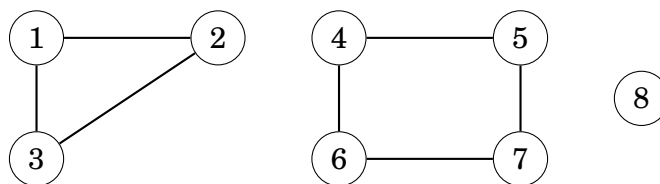
Un graf és **connex** si hi ha un camí entre dos nodes qualsevol. Per exemple, el graf següent és connex:



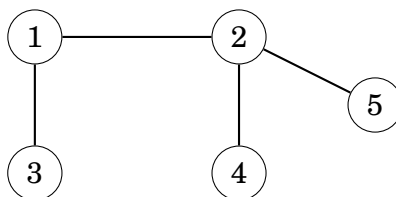
El graf següent no és connex, perquè no és possible anar del node 4 a cap altre node:



Les parts connexes d'un graf s'anomenen **components connexes**. Per exemple, el graf següent conté tres components connexes: {1, 2, 3}, {4, 5, 6, 7} i {8}.

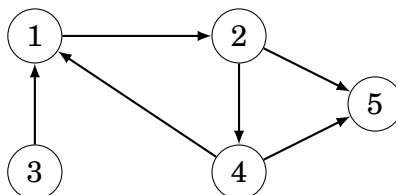


Un **arbre** és un graf connex que consta de n nodes i $n - 1$ arestes. Entre dos nodes qualsevols de l'arbre hi ha un camí únic. Per exemple, el graf següent és un arbre:



Arestes dirigides

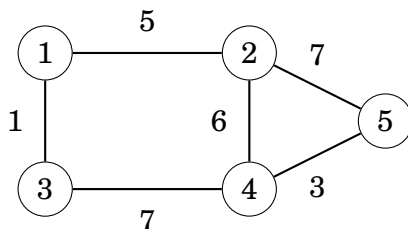
Un graf és **dirigit** si les arestes només es poden recórrer en una direcció. Per exemple, el graf següent és un graf dirigit:



El graf anterior conté un camí $3 \rightarrow 1 \rightarrow 2 \rightarrow 5$ des del node 3 fins al node 5, però no hi ha cap camí des del node 5 fins al node 3.

Arestes amb pesos

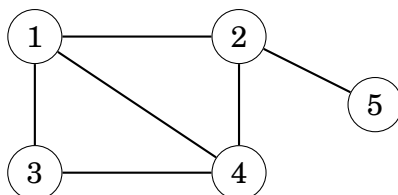
Un graf té **pesos** si a cada aresta se li assigna un **pes**. Els pesos s'interpreten sovint com a longituds de l'aresta. Per exemple, el graf següent és un graf amb pesos:



La longitud d'un camí d'un graf amb pesos és la suma dels pesos de les arestes del camí. Per exemple, al graf anterior, la longitud del camí $1 \rightarrow 2 \rightarrow 5$ és 12 i la longitud del camí $1 \rightarrow 3 \rightarrow 4 \rightarrow 5$ és 11. Aquest darrer camí és el camí **més curt** des del node 1 fins al node 5.

Veïns i graus

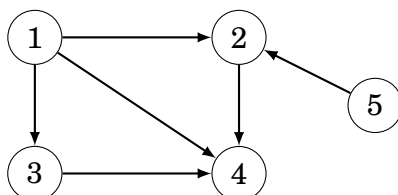
Dos nodes són **veïns** o **adjacents** si hi ha una aresta entre ells. El **grau** d'un node és el nombre d'arestes incidents al node, que coincideix amb el nombre de nodes veïns si el graf és simple. Per exemple, en el graf següent, els veïns del node 2 són 1, 4 i 5, i el seu grau és 3.



La suma de graus d'un graf és sempre $2m$, on m és el nombre d'arestes, perquè cada aresta té dos extrems, i per tant incrementa el grau de dos nodes en una unitat. Per aquest motiu, la suma dels graus sempre és un nombre parell.

Un graf és **regular** si el grau de cada node és una constant d . Un graf és **complet** si el graf conté una aresta per cada parell de nodes i , per tant, cada node té grau $n - 1$.

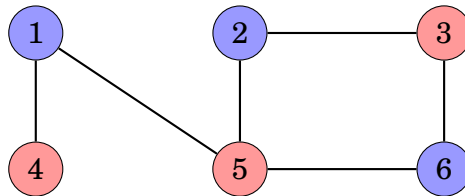
En un graf dirigit, el **grau d'entrada** d'un node és el nombre d'arestes que apunten al node, i el **grau de sortida** d'un node és el nombre d'arestes que surten del node. Per exemple, al graf següent, el grau d'entrada del node 2 és 2 i el grau de sortida del node 2 és 1.



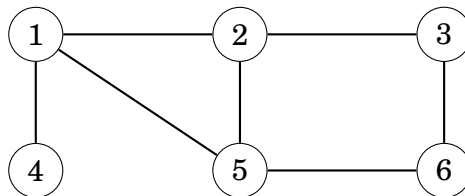
Coloracions

Una **coloració** d'un graf és una assignació de nodes a colors de manera que no hi hagi dos nodes adjacents amb el mateix color.

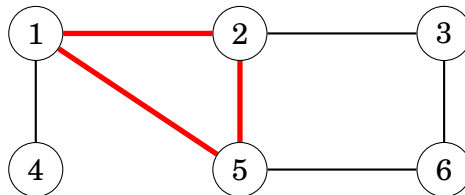
Un graf és **bipartit** si és possible acolorir-lo amb dos colors. Es pot demostrar que un graf és bipartit exactament quan no conté cap cicle amb un nombre senar d'arestes. Per exemple, el graf $[[[11]]]$ és bipartit, perquè es pot acolorir de la següent manera:



Tanmateix, el graf

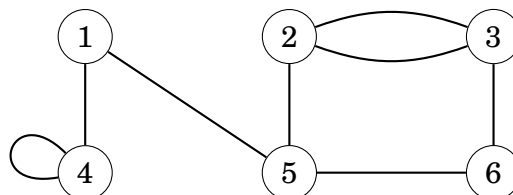


no és bipartit, perquè no és possible acolorir el següent cicle de tres nodes amb dos colors:



Grafs simples

Un graf és **simple** si cap aresta comença i acaba al mateix node, i no hi ha múltiples arestes entre dos nodes. Sovint assumim que els grafs són simples. Per exemple, el graf següent *no* és simple:



11.2 Representació de grafs

Hi ha diverses maneres de representar grafs en algorismes. L'elecció d'una estructura de dades depèn de la mida del graf i de la forma en què l'algorisme el processa. A continuació mostrarem tres representacions comunes.

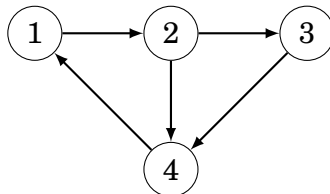
Llista d'adjacència

Per a representar un graf com a llista d'adjacència, assignem a cada node x del graf una **llista d'adjacència** que consta dels nodes als quals hi ha una aresta des de x . Les llistes d'adjacència són la manera més popular de representar grafs, i la majoria dels algorismes es poden implementar eficientment amb llistes d'adjacència.

Una manera convenient d'emmagatzemar les llistes d'adjacència és declarar un *array* de vectors de la manera següent:

```
vector<int> adj[N];
```

La constant N s'escull de manera que es pugui emmagatzemar totes les llistes d'adjacència. Per exemple, el graf



es pot emmagatzemar de la següent manera:

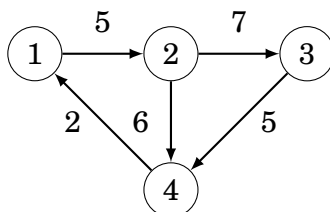
```
adj[1].push_back(2);  
adj[2].push_back(3);  
adj[2].push_back(4);  
adj[3].push_back(4);  
adj[4].push_back(1);
```

Si el graf no està dirigit, podem fer servir una representació semblant, però afegim cada aresta en ambdues direccions.

Per a un graf amb pesos, l'estructura es pot ampliar de la següent manera:

```
vector<pair<int,int>> adj[N];
```

En aquest cas, la llista d'adjacència del node a conté el parell (b, w) sempre quan hi ha una aresta des del node a fins al node b amb pes w . Per exemple, el graf



es pot emmagatzemar de la següent manera:

```
adj[1].push_back({2,5});
adj[2].push_back({3,7});
adj[2].push_back({4,6});
adj[3].push_back({4,5});
adj[4].push_back({1,2});
```

L'avantatge d'utilitzar llistes d'adjacència és que podem trobar de manera eficient els nodes als quals ens podem moure des d'un node determinat a través d'una aresta. Per exemple, el següent bucle passa per tots els nodes als quals ens podem moure des del node s :

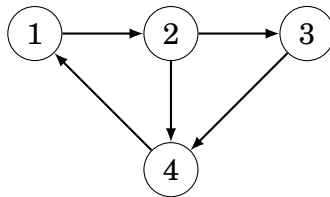
```
for (auto u : adj[s]) {
    // fer coses amb el node 'u'
}
```

Matriu d'adjacència

Una **matriu d'adjacència** és una matriu bidimensional que indica quines arestes pertanyen al graf. Amb una matriu d'adjacència podem comprovar de manera eficient si hi ha una aresta entre dos nodes. La matriu es pot emmagatzemar com un *array* de vectors

```
int adj[N][N];
```

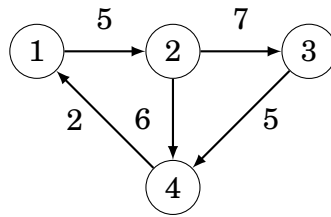
on cada valor $adj[a][b]$ indica si el graf conté una aresta des del node a fins al node b . Si l'aresta s'inclou al graf, llavors $adj[a][b] = 1$, i en cas contrari $adj[a][b] = 0$. Per exemple, el graf



es pot representar de la següent manera:

	1	2	3	4
1	0	1	0	0
2	0	0	1	1
3	0	0	0	1
4	1	0	0	0

Si el graf té pesos, la representació amb matriu d'adjacència s'estén de manera que la matriu contingui el pes a l'aresta si l'aresta existeix. Utilitzant aquesta representació, el graf



es correspon amb la matriu següent:

	1	2	3	4
1	0	5	0	0
2	0	0	7	6
3	0	0	0	5
4	2	0	0	0

L'inconvenient de la representació amb matrius d'adjacència és que la matriu conté n^2 elements, i normalment la majoria d'ells són zero. Per aquest motiu, la representació no es pot fer servir si el graf és gran.

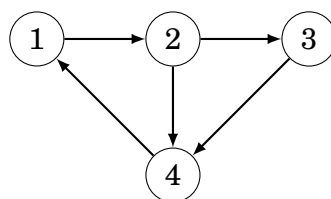
Llista d'arestes

Una **llista d'arestes** conté totes les arestes d'un graf en un cert ordre. Aquesta és una manera convenient de representar un graf si l'algorisme processa totes les arestes del graf i no cal trobar les arestes que comencen en un node determinat.

La llista d'arestes es pot emmagatzemar en un vector

```
vector<pair<int,int>> edges;
```

on cada parell (a,b) indica que hi ha una aresta des del node a fins al node b . Així, el graf



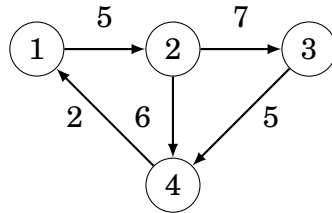
es pot representar de la següent manera:

```
edges.push_back({1,2});
edges.push_back({2,3});
edges.push_back({2,4});
edges.push_back({3,4});
edges.push_back({4,1});
```

Si el graf té pesos, l'estructura es pot ampliar de la següent manera:

```
vector<tuple<int,int,int>> edges;
```

Cada element d'aquesta llista té la forma (a,b,w) , el que significa que hi ha una aresta des del node a fins al node b amb pes w . Per exemple, el graf



es pot representar com segueix¹:

```
edges.push_back({1,2,5});  
edges.push_back({2,3,7});  
edges.push_back({2,4,6});  
edges.push_back({3,4,5});  
edges.push_back({4,1,2});
```

¹En alguns compiladors antics, és necessari fer servir la funció `make_tuple` en lloc de les claus (per exemple, `make_tuple(1,2,5)` en lloc de `{1,2,5}`).

Capítol 12

Recorreguts en grafs

Aquest capítol tracta dos algorismes fonamentals de grafs: la cerca en profunditat i la cerca en amplada. Ambdós algorismes reben un node inicial del graf i visiten tots els nodes als quals es pot accedir des del node inicial. La diferència entre els algorismes és l'ordre en què visiten els nodes.

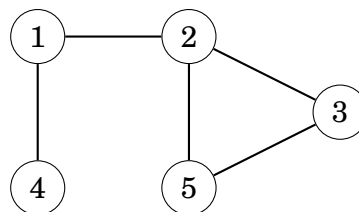
12.1 Cerca en profunditat

Cerca en profunditat (*depth-first search*, DFS) és una tècnica senzilla per a recórrer grafs. L'algorisme comença en un node inicial, i continua per tots els altres nodes als quals es pot accedir des del node inicial fent servir les arestes del graf.

La cerca en profunditat sempre recorre un únic camí del graf mentre trobi nodes nous per explorar. Després d'això, torna als nodes anteriors i comença a explorar altres parts del graf. L'algoritme té control dels nodes que ja han estat visitats, de manera que processa cada node només una vegada.

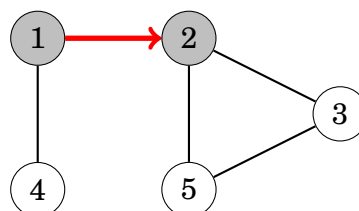
Exemple

Posem com a exemple la cerca en profunditat del graf següent:

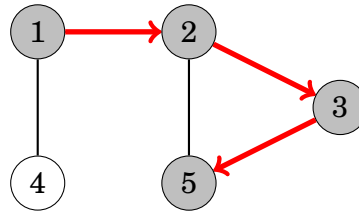


Podem començar la cerca en qualsevol node del graf, per exemple, el node 1.

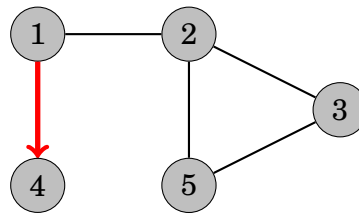
La cerca passa primer pel node 2:



Després d'això, visita els nodes 3 i 5:



Els veïns del node 5 són el 2 i el 3, però la cerca ja ha visitat ambdós, de manera que tornem als nodes anteriors. També hem vist els veïns dels nodes 3 i 2, per la qual cosa el següent moviment és del node 1 al node 4:



Després d'això, la cerca finalitza perquè ha visitat tots els nodes.

La complexitat temporal de la cerca en profunditat és $O(n + m)$ on n és el nombre de nodes i m és el nombre d'arestes, perquè l'algorisme processa cada node i aresta una sola vegada.

Implementació

La cerca en profunditat es pot implementar còmodament mitjançant recursivitat. La següent funció `dfs` comença una cerca en profunditat en un node determinat. La funció suposa que el graf s'emmagatzema com a llistes d'adjacència en un vector

```
vector<int> adj[N];
```

i també manté un vector

```
bool visited[N];
```

per conèixer els nodes visitats. Inicialment, cada valor del vector és `false`, i quan la cerca arriba al node s , el valor de `visited[s]` es converteix en `true`. La funció es pot implementar de la següent manera:

```
void dfs(int s) {
    if (visited[s]) return;
    visited[s] = true;
    // process node s
    for (auto u: adj[s]) {
        dfs(u);
    }
}
```

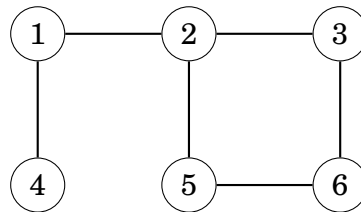
12.2 Cerca en amplada

La **cerca en amplada** (*breadth-first search*, BFS) visita els nodes en ordre creixent a la seva distància des del node inicial. Així, podem calcular la distància des del node inicial fins a la resta de nodes mitjançant la cerca en amplada. Tanmateix, la cerca en amplada és més difícil d'implementar que la cerca en profunditat.

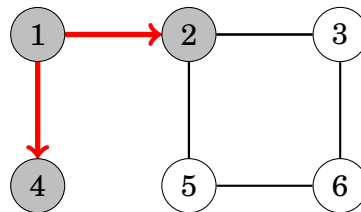
La cerca en amplada passa per tots els nodes d'un nivell abans de considerar els nodes del nivell següent. Primer, la cerca explora tots els nodes a distància 1 del node inicial, després els nodes a distància 2, i així successivament. Aquest procés continua fins que s'han visitat tots els nodes.

Exemple

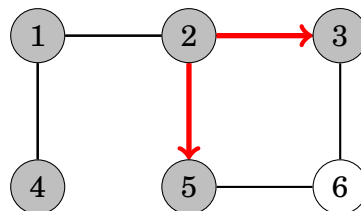
Considerem com la cerca en amplada processa el graf següent:



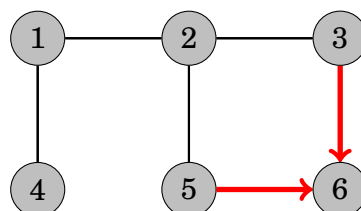
Suposem que la cerca comença al node 1. En primer lloc, processem tots els nodes als quals es pot arribar des del node 1 mitjançant una única aresta:



Després d'això, passem als nodes 3 i 5:



Finalment, visitem el node 6:



Ara hem calculat les distàncies des del node inicial fins a tots els nodes del graf. Les distàncies són les següents:

node	distance
1	0
2	1
3	2
4	1
5	2
6	3

A l'igual que en la cerca en profunditat, la complexitat temporal de la cerca en amplada és $O(n + m)$, on n és el nombre de nodes i m és el nombre d'arestes.

Implementació

La cerca en profunditat és més difícil d'implementar que la cerca en profunditat, perquè l'algoritme visita els nodes de parts diferents del graf. Una implementació típica es basa en una cua que conté els nodes. En cada pas processem un node de la cua.

El següent codi suposa que el graf s'emmagatzema com a llistes d'adjacència i manté les estructures de dades següents:

```
queue<int> q;
bool visited[N];
int distance[N];
```

La cua q conté els nodes que s'han de processar en ordre de distància creixent. Els nous nodes sempre s'afegeixen al final de la cua, i el node al començament de la cua és el següent node a processar. El vector `visited` indica quins nodes ja s'han visitat, i el vector `distance` contindrà les distàncies des del node inicial a tots els nodes del graf.

La cerca es pot implementar de la següent manera, començant pel node x :

```
visited[x] = true;
distance[x] = 0;
q.push(x);
while (!q.empty()) {
    int s = q.front(); q.pop();
    // process node s
    for (auto u : adj[s]) {
        if (visited[u]) continue;
        visited[u] = true;
        distance[u] = distance[s]+1;
        q.push(u);
    }
}
```

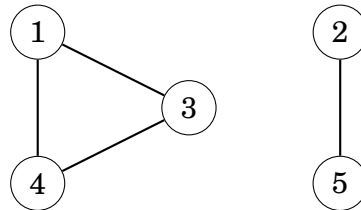
12.3 Aplicacions

Fent servir els algorismes de recorreguts en grafs podem comprovar moltes propietats dels mateixos. Podem fer servir tant la cerca en profunditat com la cerca en amplada, però a la pràctica, la cerca en profunditat és una millor opció perquè és més fàcil d'implementar. En les aplicacions següents assumirem que el graf no està orientat.

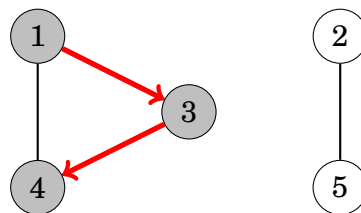
Comprovació de connectivitat

Un graf és connex si hi ha un camí entre dos nodes qualsevol del graf. Així, podem comprovar si un graf és connex començant per un node arbitrari i esbrinant si podem arribar a tots els altres nodes.

Per exemple, en el graf



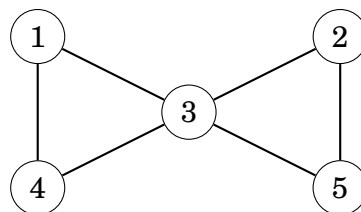
es pot veure que una cerca en profunditat des del node 1 visita els nodes següents:



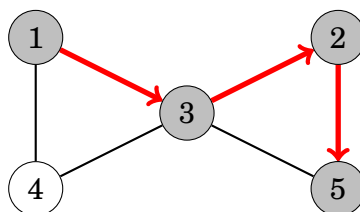
Com que la cerca no ha visitat tots els nodes, en conclouem que el graf no és connex. De manera semblant, podem trobar totes les components connexes d'un graf iterant pels nodes i començant una nova cerca en profunditat cada cop que trobem un node que no pertany a cap component connexa.

Trobar cicles

Un graf conté un cicle si al recórrer el graf trobem un node amb un veí que ja ha estat visitat, i que no sigui el node anterior en el camí actual. Per exemple, el graf



conté dos cicles i podem trobar un d'ells de la següent manera:



Després de passar del node 2 al node 5 observem que el veí 3 del node 5 ja ha estat visitat. Així, el graf conté un cicle que passa pel node 3, per exemple, $3 \rightarrow 2 \rightarrow 5 \rightarrow 3$.

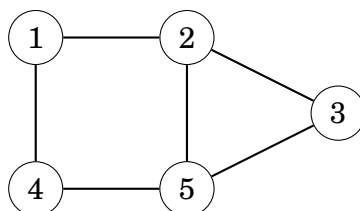
Una altra manera d'esbrinar si un graf conté un cicle és calcular el nombre de nodes i arestes de cada component connexa. Si un component conté c nodes i cap cicle, ha de contenir exactament $c - 1$ arestes (ha de ser un arbre). Si hi ha c o més arestes, la component sens dubte conté un cicle.

Comprovar si un graf és bipartit

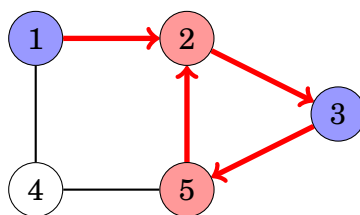
Un graf és bipartit si els seus nodes es poden acolorir amb dos colors de manera que no hi hagi nodes adjacents amb el mateix color. Comprovar si un graf és bipartit mitjançant algorismes de recorregut de grafs és sorprenentment fàcil.

La idea és pintar el node inicial de blau, tots els seus veïns de vermell, tots els seus veïns de blau, etc. Si en algun moment de la cerca observem que dos nodes adjacents tenen el mateix color, això vol dir que el graf no és bipartit. En cas contrari, el graf és bipartit i hem trobat una coloració.

Per exemple, el graf



no és bipartit, perquè si fem una cerca des del node 1 veiem el següent:



Observem que el color dels nodes 2 i 5 és vermell, tot i que són adjacents. Per tant, el graf no pot ser bipartit.

Aquest algorisme sempre funciona, perquè quan només hi ha dos colors disponibles, el color del node inicial d'un component determina els colors de tots els altres nodes del component, sense importar si acolorim el node inicial vermell o blau.

Tingueu en compte que en el cas general, és difícil esbrinar si els nodes d'un graf es poden acolorir amb k colors de manera que cap node adjacent tingui el

mateix color. Fins i tot quan $k = 3$, és sap que el problema és NP-difícil i no es coneix cap algorisme eficient.

Capítol 13

Camins més curts

Trobar el camí més curt entre dos nodes d'un graf és un problema important que té moltes aplicacions pràctiques. Per exemple, donada una xarxa de carreteres, calcula la longitud més curta possible d'una ruta entre dues ciutats, tenint en compte les longituds de les carreteres.

En un graf sense pesos, la longitud d'un camí és igual al nombre d'arestes, i podem fer servir la cerca en amplada per trobar el camí més curt. Tanmateix, en aquest capítol ens centrem en els grafs amb pesos on necessitem algorismes més sofisticats per trobar els camins mínims.

13.1 Algorisme de Bellman–Ford

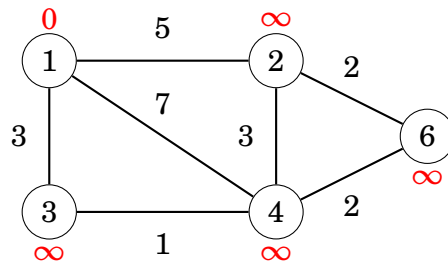
L'**algorisme de Bellman–Ford**¹ troba els camins més curts des d'un node inicial a tots els nodes del graf. L'algorisme pot processar tot tipus de grafs, sempre que el graf no contingui un cicle de longitud negativa. L'algorisme és capaç de detectar si el graf té cicles de longitud negativa.

L'algorisme manté un vector amb les millors distàncies conegudes des del node inicial fins a tots els nodes del graf. Inicialment, la distància al node inicial és 0 i la distància a tots els altres nodes és infinita. L'algorisme redueix les distàncies trobant arestes que escurcen els camins fins que no és possible reduir cap distància.

Exemple

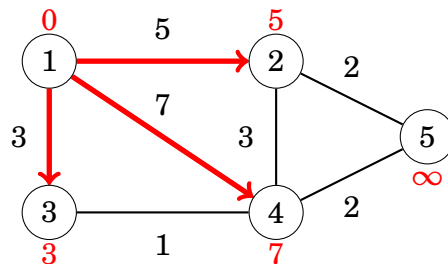
Mostrem com funciona l'algorisme de Bellman–Ford en el graf següent:

¹L'algorisme rep el nom de R.E. Bellman i L.R. Ford que el van publicar de manera independentment els anys 1958 i el 1956 [5, 24].

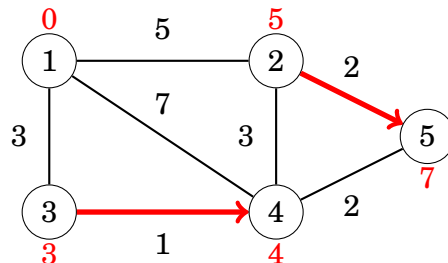


A cada node del graf li assignem una distància. Inicialment, la distància al node inicial és 0 i la distància a tots els altres nodes és infinita.

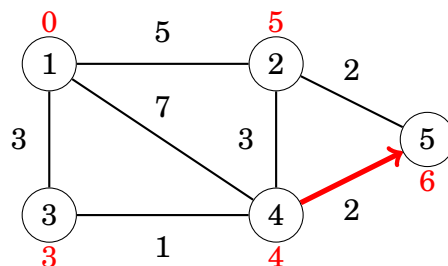
L'algorisme cerca arestes que redueixin distàncies. En primer lloc, totes les arestes del node 1 redueixen les distàncies:



Després d'això, les arestes $2 \rightarrow 5$ i $3 \rightarrow 4$ redueixen les distàncies:

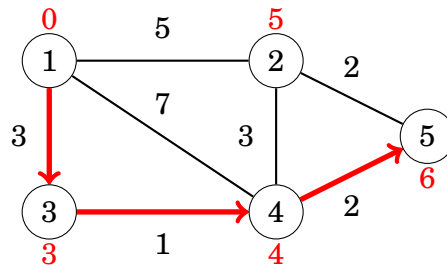


Finalment, hi ha un canvi més:



Després d'això, cap aresta pot reduir cap distància. Això vol dir que les distàncies són definitives i hem calculat correctament les distàncies més curtes des del node inicial fins a tots els nodes del graf.

Per exemple, la distància més curta del node 1 al node 5 és 3 i es correspon amb el camí següent:



Implementació

La següent implementació de l'algorisme de Bellman-Ford determina les distàncies més curtes des d'un node x a tots els nodes del graf. El codi suposa que el graf s'emmagatzema com una llista de arestes `edges` que consta de tuples de la forma (a, b, w) , el que significa que hi ha una aresta des del node a fins al node b amb pes w .

L'algorisme consta de $n - 1$ rondes, i a cada ronda l'algoritme passa per totes les arestes del graf i intenta reduir les distàncies. L'algorisme construeix un vector `distance` que conté les millors distàncies conegudes de x a tots els nodes del graf. La constant `INF` denota una distància infinita.

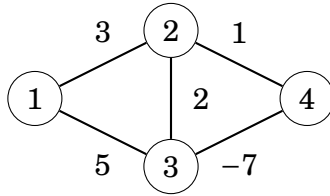
```
for (int i = 1; i <= n; i++) distance[i] = INF;
distance[x] = 0;
for (int i = 1; i <= n-1; i++) {
    for (auto e : edges) {
        int a, b, w;
        tie(a, b, w) = e;
        distance[b] = min(distance[b], distance[a]+w);
    }
}
```

La complexitat temporal de l'algorisme és $O(nm)$, perquè l'algoritme consta de $n - 1$ rondes i itera per totes les m arestes durant una ronda. Si no hi ha cicles negatius al graf, totes les distàncies són finals després de $n - 1$ rondes, perquè els camins mínims no poden repetir nodes, i per tant contenen com a molt $n - 1$ arestes.

A la pràctica, sovint no és necessari fer $n - 1$ rondes per a trobar les distàncies definitives. Podem fer l'algorisme més eficient si l'aturem quan no reduïm cap distància durant una ronda.

Cicles negatius

L'algorisme de Bellman-Ford també es pot fer servir per comprovar si el graf conté un cicle de longitud negativa. Per exemple, el graf



conté un cicle negatiu $2 \rightarrow 3 \rightarrow 4 \rightarrow 2$ de longitud -4 .

Si el graf conté un cicle negatiu, podem escurçar infinites vegades qualsevol camí que contingui el cicle repetint el cicle una i altra vegada. Per tant, el concepte de camí més curt no té sentit en aquesta situació.

Podem detectar cicles negatius executant l'algorisme de Bellman-Ford exactament n rondes. Si l'última ronda redueix alguna distància, el graf conté un cicle negatiu. Observeu que aquest algorisme es pot fer servir per cercar un cicle negatiu a tot el graf independentment del node inicial.

Algorisme SPFA

L'algorisme **SPFA** ("Shortest Path Faster Algorithm") [20] és una variant de l'algorisme de Bellman-Ford que sovint és més eficient que l'algorisme original. L'algoritme SPFA no passa per totes les arestes a cada ronda, sinó que tria les arestes que s'han d'examinar d'una manera més intel·ligent.

L'algorisme manté una cua de nodes que es poden fer servir per reduir les distàncies. En primer lloc, l'algoritme afegeix el node inicial x a la cua. A continuació, l'algorisme considera el primer node a de la cua, i comprova si les arestes de la forma $a \rightarrow b$ redueixen la distància, i si és el cas, afegeix el node b a la cua.

L'eficiència de l'algorisme SPFA depèn de l'estructura del graf: l'algorisme és sovint eficient, però la seva complexitat temporal en el pitjor dels casos segueix sent $O(nm)$ i és possible crear entrades que fan que l'algorisme sigui tan lent com l'algorisme original de Bellman-Ford.

13.2 Algorisme de Dijkstra

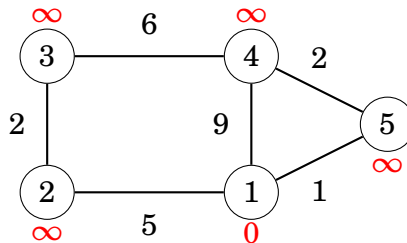
Algorisme de Dijkstra² troba els camins més curts des del node inicial fins a tots els nodes del graf, a l'igual que l'algorisme de Bellman-Ford. L'avantatge de l'algorisme de Dijkstra és que és més eficient i es pot fer servir per processar grafs grans. Tanmateix, l'algorisme requereix que no hi hagi arestes de pes negatius al graf.

Igual que l'algorisme de Bellman-Ford, l'algoritme de Dijkstra manté les distàncies als nodes i les redueix durant la cerca. L'algorisme de Dijkstra és eficient, perquè només processa cada aresta del graf una vegada, fent servir el fet que no hi ha arestes negatives.

²E. W. Dijkstra va publicar l'algorisme el 1959 [14]; tanmateix, el seu article original no esmenta com implementar l'algorisme de manera eficient.

Exemple

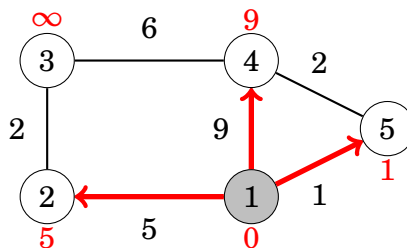
Considerem com funciona l'algorisme de Dijkstra al graf següent quan el node inicial és el node 1:



A l'igual que en l'algorisme de Bellman-Ford, inicialment la distància al node inicial és 0 i la distància a tots els altres nodes és infinita.

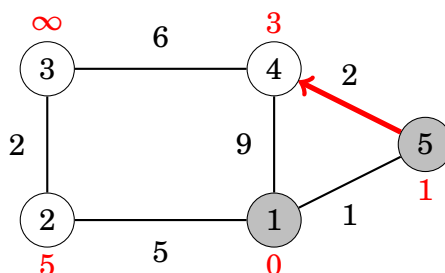
A cada pas, l'algorisme de Dijkstra selecciona aquell node que encara no s'ha processat i que està a distància mínima. El primer d'aquests nodes és el node 1 a distància 0.

Quan es selecciona un node, l'algorisme passa per totes les arestes que surten del node i redueix les distàncies:

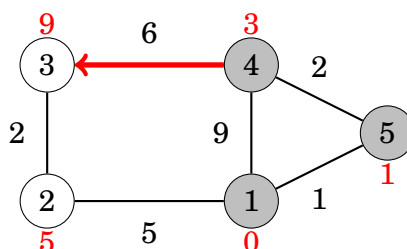


En aquest cas, les arestes del node 1 redueixen les distàncies als nodes 2, 4 i 5, que passen a ser ara 5, 9 i 1.

El següent node a processar és el node 5 a distància 1. Això redueix la distància al node 4 de 9 a 3:

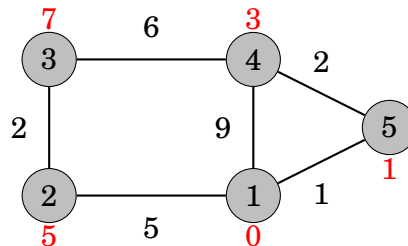


Després d'això, el següent node és el node 4, i fa que la distància al node 3 sigui 9:



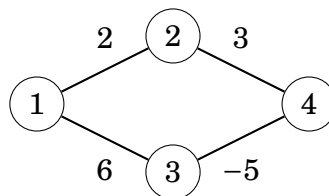
Una propietat notable de l'algorisme de Dijkstra és que sempre que treballem amb un node trobem la seva distància final. Per exemple, en aquest punt de l'algorisme, les distàncies 0, 1 i 3 són les distàncies finals als nodes 1, 5 i 4.

Després d'això, l'algorisme processa els dos nodes restants i les distàncies resultants són les següents:



Arestes negatives

L'eficiència de l'algorisme de Dijkstra es basa en el fet que el graf no conté arestes negatives. Si hi ha una aresta negativa l'algorisme pot donar resultats incorrectes. Per exemple:



El camí més curt des del node 1 al node 4 és $1 \rightarrow 3 \rightarrow 4$ i la seva longitud és 1. Tanmateix, l'algorisme de Dijkstra troba el camí $1 \rightarrow 2 \rightarrow 4$ seguint les arestes de pes mínim. L'algorisme no té en compte que en l'altre camí el pes -5 serveix per compensar el pes 6.

Implementació

La següent implementació de l'algorisme de Dijkstra calcula les distàncies mínimes des d'un node x a els altres nodes del graf. El graf s'emmagatzema com a llistes d'adjacència de manera que $\text{adj}[a]$ conté un parell (b, w) quan hi ha una aresta del node a fins al node b amb pes w .

Una implementació eficient de l'algorisme de Dijkstra requereix trobar de manera eficient el node de distància mínima que no s'ha processat encara. L'estructura de dades adequada és una cua de prioritats que conté els nodes ordenats per distància. Amb una cua de prioritats trobem el següent node a processar en temps logarítmic.

Al codi següent, la cua de prioritats q conté els parells de la forma $(-d, x)$ per a indicar que la distància al node x és d . El vector `distance` conté la distància a cada node, i el vector `processed` indica si s'ha processat un node. Inicialment la distància és 0 a x i ∞ a tots els altres nodes.


```

for (int i = 1; i <= n; i++) distance[i] = INF;
distance[x] = 0;
q.push({0,x});
while (!q.empty()) {
    int a = q.top().second; q.pop();
    if (processed[a]) continue;
    processed[a] = true;
    for (auto u : adj[a]) {
        int b = u.first, w = u.second;
        if (distance[a]+w < distance[b]) {
            distance[b] = distance[a]+w;
            q.push({-distance[b],b});
        }
    }
}
}

```

Tingueu en compte que la cua de prioritat conté distàncies *negatives* als nodes. La raó d'això és que la versió predeterminada de la cua de prioritats C++ troba els elements màxims, mentre que volem trobar els elements mínims. Mitjançant l'ús de distàncies negatives, podem utilitzar directament la cua de prioritat per defecte³. Tingueu en compte també que pot haver-hi diverses instàncies del mateix node a la cua de prioritats; tanmateix, només processarem la instància amb distància mínima.

La complexitat temporal de la implementació anterior és $O(n + m \log m)$, perquè l'algoritme passa per tots els nodes del graf i afegeix, per cada aresta, com a màxim una distància a la cua de prioritats.

13.3 Algorisme de Floyd-Warshall

L'algorisme de **Floyd-Warshall**⁴ proporciona una manera alternativa d'abordar el problema de trobar els camins més curts. Aquest, a diferència dels altres algorismes d'aquest capítol, troba tots els camins més curts entre tots els nodes en una sola execució.

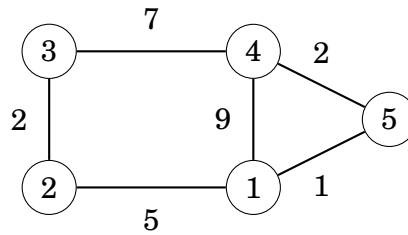
L'algorisme manté una matriu que conté les distàncies entre els nodes. Al principi, les distàncies es calculen únicament amb arestes directes entre els nodes, i a continuació, l'algorisme redueix les distàncies utilitzant nodes intermedis en camins.

Exemple

Considerem com funciona l'algorisme de Floyd-Warshall al graf següent:

³Per descomptat, també podem declarar la cua de prioritats com al capítol 4.5 i fer servir distàncies positives, però la implementació seria una mica més llarga.

⁴L'algorisme rep el nom de R.W. Floyd i S. Warshall que el van publicar de manera independent el 1962 [23, 70].



Inicialment, la distància de cada node a si mateix és 0, i la distància entre els nodes a i b és x si hi ha una aresta entre els nodes a i b amb pes x . Totes les altres distàncies són infinites.

En aquest graf, la matriu inicial és la següent:

	1	2	3	4	5
1	0	5	∞	9	1
2	5	0	2	∞	∞
3	∞	2	0	7	∞
4	9	∞	7	0	2
5	1	∞	∞	2	0

L'algorisme consta de rondes consecutives. A cada ronda, l'algoritme selecciona un nou node que pot actuar com a node intermedi en els camins a partir d'ara, i les distàncies es redueixen mitjançant aquest node.

A la primera ronda, el node 1 és el nou node intermedi. Hi ha un nou camí entre els nodes 2 i 4 de longitud 14, perquè el node 1 els connecta. També hi ha un nou camí entre els nodes 2 i 5 amb longitud 6.

	1	2	3	4	5
1	0	5	∞	9	1
2	5	0	2	14	6
3	∞	2	0	7	∞
4	9	14	7	0	2
5	1	6	∞	2	0

A la segona ronda, el node 2 és el nou node intermedi. Això crea nous camins entre els nodes 1 i 3 i entre els nodes 3 i 5:

	1	2	3	4	5
1	0	5	7	9	1
2	5	0	2	14	6
3	7	2	0	7	8
4	9	14	7	0	2
5	1	6	8	2	0

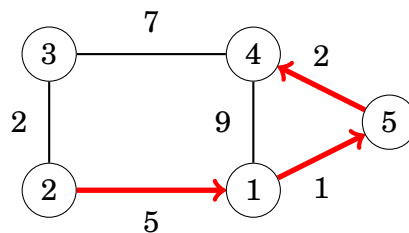
A la tercera ronda, el node 3 és el nou node intermedi. Hi ha un nou camí entre els nodes 2 i 4:

	1	2	3	4	5
1	0	5	7	9	1
2	5	0	2	9	6
3	7	2	0	7	8
4	9	9	7	0	2
5	1	6	8	2	0

L'algorisme continua així, fins que tots els nodes han estat designats nodes intermedis. Un cop acabat l'algorisme, la matriu conté les distàncies mínimes entre dos nodes qualsevol:

	1	2	3	4	5
1	0	5	7	3	1
2	5	0	2	8	6
3	7	2	0	7	8
4	3	8	7	0	2
5	1	6	8	2	0

Per exemple, la matriu ens indica que la distància més curta entre els nodes 2 i 4 és 8. Això correspon al camí següent:



Implementació

L'avantatge de l'algorisme de Floyd-Warshall és que és fàcil d'implementar. El codi següent construeix una matriu de distàncies on $distance[a][b]$ és la distància més curta entre els nodes a i b . Primer, l'algorisme inicialitza $distance$ fent servir la matriu d'adjacència adj del graf:

```
for (int i = 1; i <= n; i++) {
    for (int j = 1; j <= n; j++) {
        if (i == j) distance[i][j] = 0;
        else if (adj[i][j]) distance[i][j] = adj[i][j];
        else distance[i][j] = INF;
    }
}
```

Després d'això, les distàncies més curtes es poden trobar de la següent manera:

```
for (int k = 1; k <= n; k++) {
    for (int i = 1; i <= n; i++) {
```

```
    for (int j = 1; j <= n; j++) {  
        distance[i][j] = min(distance[i][j],  
                             distance[i][k]+distance[k][j]);  
    }  
}  
}
```

La complexitat temporal de l'algorisme és $O(n^3)$, perquè conté tres bucles niats que passen per tots els nodes del graf.

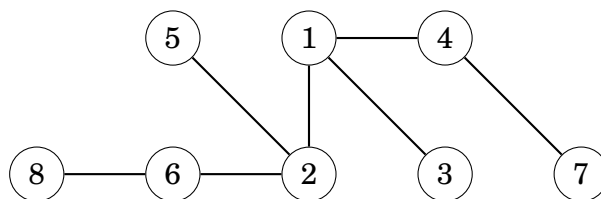
Com que la implementació de l'algorisme de Floyd-Warshall és senzilla, aquest algoritme pot ser una bona opció encara que només sigui necessari trobar un únic camí mínim del graf. Tanmateix, l'algorisme només pot fer-se servir quan el graf és tan petit que la complexitat cúbica resultant és acceptable.

Capítol 14

Algorismes d'arbres

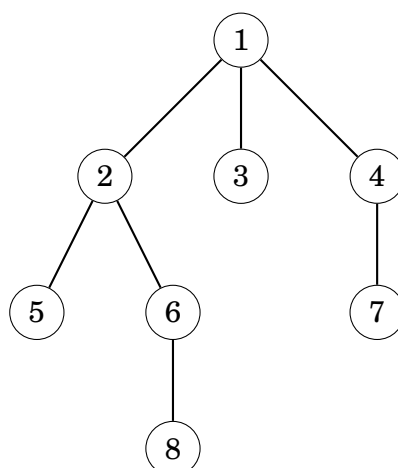
Un **arbre** és un graf connex i acíclic que consta de n nodes i $n - 1$ arestes. Quan eliminem qualsevol aresta d'un arbre el dividim en dos components connexes, i quan afegim qualsevol aresta a un arbre creem un cicle. A més, sempre hi ha un camí únic entre dos nodes qualsevol d'un arbre.

Per exemple, l'arbre següent consta de 8 nodes i 7 arestes:



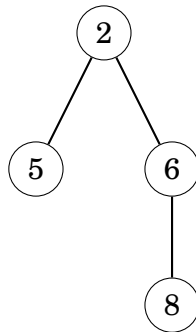
Les **fulles** d'un arbre són els nodes de grau 1, és a dir, amb només un veí. Per exemple, les fulles de l'arbre anterior són els nodes 3, 5, 7 i 8.

En un arbre **arrelat**, un dels nodes s'anomena l'**arrel** de l'arbre, i tots els altres nodes es col·loquen sota l'arrel. Per exemple, a l'arbre següent, el node 1 és el node arrel.



En un arbre arrelat, els **fills** d'un node són els seus veïns inferiors, i el **pare** d'un node és el seu veí superior. Cada node té exactament un pare, excepte l'arrel que no té cap pare. Per exemple, a l'arbre anterior, els fills del node 2 són els nodes 5 i 6, i el seu pare és el node 1.

L'estructura d'un arbre arrelat és *recursiva*: cada node de l'arbre actua com a arrel d'un **subarbre** que conté el node en si i tots els nodes que es troben als subarbres dels seus fills. Per exemple, a l'arbre anterior, el subarbre del node 2 consta dels nodes 2, 5, 6 i 8:



14.1 Recorregut d'arbres

Els algorismes generals de recorreguts de grafs es poden utilitzar per recórrer els nodes d'un arbre. Tanmateix, el recorregut d'un arbre és més fàcil d'implementar que el d'un graf general, perquè no hi ha cicles a l'arbre i no és possible arribar a un node des de múltiples direccions.

La manera típica de recórrer un arbre és iniciar una cerca en profunditat en un node arbitrari. Es pot fer servir la funció recursiva següent:

```
void dfs(int s, int e) {  
    // process node s  
    for (auto u : adj[s]) {  
        if (u != e) dfs(u, s);  
    }  
}
```

La funció té dos paràmetres: el node actual s i el node anterior e . El propòsit del paràmetre e és assegurar-se que la cerca només es mou als nodes que encara no s'han visitat.

El següent codi inicia la cerca al node x :

```
dfs(x, 0);
```

En la primera crida $e = 0$, perquè no hi ha cap node anterior, i es permet avançar en qualsevol direcció de l'arbre.

Programació dinàmica

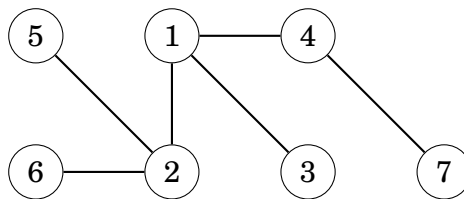
La programació dinàmica es pot fer servir per calcular informació quan recorrem un arbre. Per exemple, podem calcular en temps $O(n)$ la mida del subarbre de cada node, o la longitud del camí més llarg des de cada node a una fulla.

Com a exemple, calculem per a cada node s un valor $\text{count}[s]$: la mida del seu subarbre. El subarbre conté el node i tots els nodes dels subarbres dels seus fills, de manera que podem calcular aquesta mida recursivament amb el codi següent:

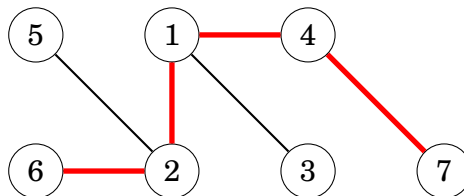
```
void dfs(int s, int e) {
    count[s] = 1;
    for (auto u : adj[s]) {
        if (u == e) continue;
        dfs(u, s);
        count[s] += count[u];
    }
}
```

14.2 Diàmetre

El **diàmetre** d'un arbre és la longitud màxima d'un camí entre dos nodes. Per exemple, considereu l'arbre següent:



El diàmetre d'aquest arbre és 4, i es correspon al següent camí:



Tingueu en compte que poden haver-hi diversos camins de longitud màxima. Al camí anterior, podem substituir el node 6 pel node 5 per obtenir un altre camí de longitud 4.

A continuació mostrem dos algorismes de temps $O(n)$ que calculen el diàmetre d'un arbre. El primer algorisme es basa en la programació dinàmica, i el segon algorisme utilitza dues cerques en profunditat.

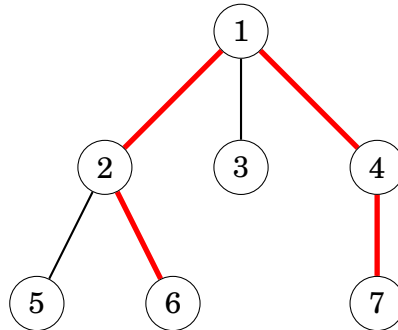
Algorisme 1

Una manera general d'abordar molts problemes dels arbres és arrelar primer l'arbre de manera arbitrària. Després d'això, podem intentar resoldre el problema per separat per a cada subarbre. El nostre primer algorisme per calcular el diàmetre es basa en aquesta idea.

Una observació important és que cada camí d'un arbre arrelat té un *punt més alt*: el node més alt que pertany al camí. Així, podem trobar per a cada node x

el camí més llarg que té x com a node més alt del camí. Un d'aquests camins correspon al diàmetre de l'arbre.

Per exemple, a l'arbre següent, el node 1 és el punt més alt del camí que correspon al diàmetre:



Calculem per a cada node x dos valors:

- $\text{toLeaf}(x)$: the maximum length of a path from x to any leaf
- $\text{maxLength}(x)$: the maximum length of a path whose highest point is x

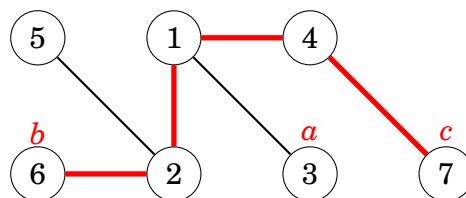
Per exemple, a l'arbre anterior, $\text{toLeaf}(1) = 2$, perquè hi ha un camí $1 \rightarrow 2 \rightarrow 6$, i $\text{maxLength}(1) = 4$, perquè hi ha un camí $6 \rightarrow 2 \rightarrow 1 \rightarrow 4 \rightarrow 7$. En aquest cas, $\text{maxLength}(1)$ és igual al diàmetre.

La programació dinàmica es pot utilitzar per calcular els valors anteriors per a tots els nodes en temps $O(n)$. Primer, per calcular $\text{toLeaf}(x)$, considerem els fills c de x i afegim un al màxim dels valors $\text{toLeaf}(c)$. A continuació, per a calcular $\text{maxLength}(x)$ escollim dos fills diferents a i b de manera que la suma $\text{toLeaf}(a) + \text{toLeaf}(b)$ sigui màxima, i afegim dos a aquesta suma.

Algorisme 2

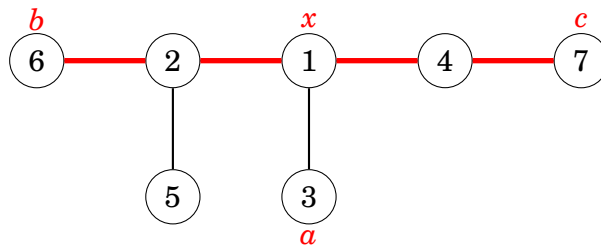
Una altra manera eficient de calcular el diàmetre d'un arbre es fer dues cerques en profunditat. Primer, triem un node arbitrari a de l'arbre i trobem el node b més llunyà de a . Aleshores, trobem el node més llunyà c de b . El diàmetre de l'arbre és la distància entre b i c .

Al graf següent, a , b i c podrien ser:



Aquest és un mètode elegant, però per què funciona?

És útil dibuixar l'arbre de manera que el camí que correspon al diàmetre sigui horitzontal i tots els altres nodes en penguin:

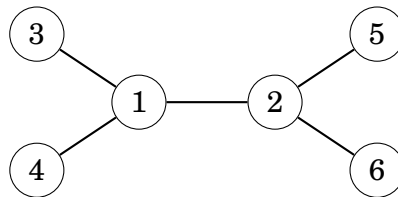


El node x indica el lloc on el camí des del node a s'uneix al camí que correspon al diàmetre. El node més allunyat de a és el node b , el node c o algun altre node que estigui almenys tan lluny del node x , i que, per tant, hagués pogut reemplaçar a b o c com a punt final del camí que es correspon al diàmetre.

14.3 Tots els camins més llargs

El problema següent és calcular per a cada node de l'arbre la longitud màxima d'un camí que comença al node. Això és una generalització del problema del diàmetre de l'arbre, perquè la longitud màxima és el diàmetre de l'arbre. Aquest problema també es pot resoldre en temps $O(n)$.

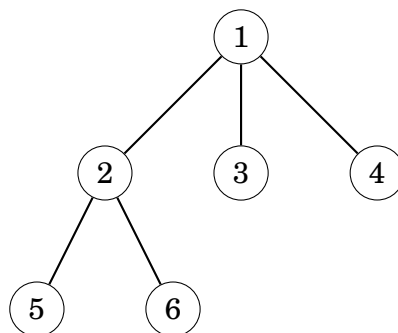
Com a exemple, considereu l'arbre següent:



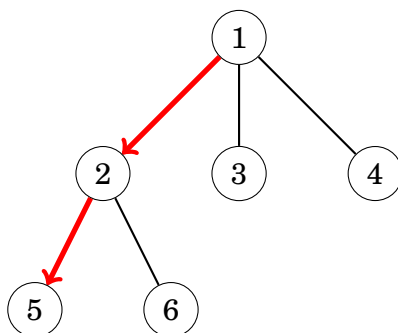
Sigui $\text{maxLength}(x)$ la longitud màxima d'un camí que comença al node x . Per exemple, a l'arbre anterior, $\text{maxLength}(4) = 3$, perquè hi ha un camí $4 \rightarrow 1 \rightarrow 2 \rightarrow 6$. Aquí teniu una taula completa dels valors:

node x	1	2	3	4	5	6
$\text{maxLength}(x)$	2	2	3	3	3	3

A l'igual que abans, un bon punt de partida per resoldre el problema és arrelar l'arbre de manera arbitrària:

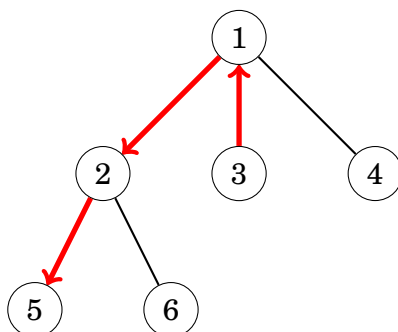


La primera part del problema és calcular per a cada node x la longitud màxima d'un camí que passa per un fill de x . Per exemple, el camí més llarg des del node 1 passa pel seu fill 2:

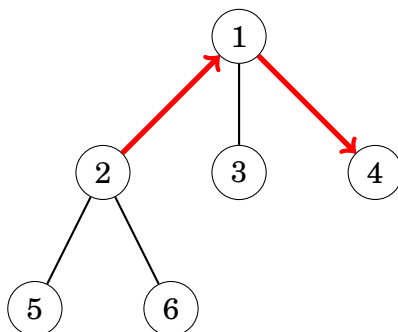


Aquesta part és fàcil de resoldre en temps $O(n)$ fent servir programació dinàmica, com abans.

La segona part del problema és calcular per a cada node x la longitud màxima d'un camí a través del seu pare p . Per exemple, el camí més llarg des del node 3 passa pel seu pare 1:



A primera vista, sembla que hauríem de triar el camí més llarg que surt de p . Tanmateix, això *no* sempre funciona, perquè el camí més llarg des de p pot passar per x . Aquí teniu un exemple d'aquesta situació:



Tot i així, podem resoldre la segona part en temps $O(n)$ emmagatzemant *dues* longituds màximes per a cada node x :

- $\text{maxLength}_1(x)$: the maximum length of a path from x
- $\text{maxLength}_2(x)$ the maximum length of a path from x in another direction than the first path

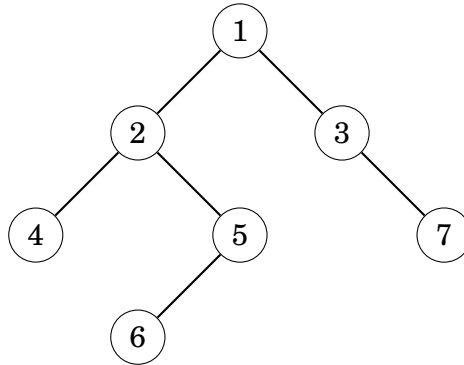
Per exemple, al graf anterior, $\text{maxLength}_1(1) = 2$ utilitzant el camí $1 \rightarrow 2 \rightarrow 5$, i $\text{maxLength}_2(1) = 1$ utilitzant el camí $1 \rightarrow 3$.

Finalment, si el camí que correspon a $\text{maxLength}_1(p)$ passa per x , arribem a la conclusió que la longitud màxima és $\text{maxLength}_2(p) + 1$, i en cas contrari la longitud màxima és $\text{maxLength}_1(p) + 1$.

14.4 Arbres binaris

Un **arbre binari** és un arbre arrelat on cada node té un subarbre fill esquerre i un subarbre fill dret. Permetem que els subarbres siguin buits. Per tant, cada node en un arbre binari té zero, un o dos fills.

Per exemple, l'arbre següent és un arbre binari:



Els nodes d'un arbre binari tenen tres ordenacions naturals que es corresponen a maneres distintes de recórrer l'arbre recursivament:

- **pre-order**: first process the root, then traverse the left subtree, then traverse the right subtree
- **in-order**: first traverse the left subtree, then process the root, then traverse the right subtree
- **post-order**: first traverse the left subtree, then traverse the right subtree, then process the root

Per a l'arbre anterior, l'ordenació dels nodes de l'arbre en pre-ordre és [1, 2, 4, 5, 6, 3, 7], en in-ordre és [4, 2, 6, 5, 1, 3, 7] i en post-ordre és [4, 6, 5, 2, 7, 3, 1].

Si coneixem l'ordenació en pre-ordre i in-ordre d'un arbre, podem reconstruir l'estructura exacta de l'arbre. Per exemple, l'arbre anterior és l'únic arbre possible amb pre-ordre [1, 2, 4, 5, 6, 3, 7] i in-ordre [4, 2, 6, 5, 1, 3, 7]. De manera similar, si tenim l'ordenació en post-ordre i in-ordre també podem determinar l'estructura d'un arbre.

Tanmateix, la situació és diferent si només coneixem l'ordenació en pre-ordre i post-ordre d'un arbre. En aquest cas, pot haver-hi més d'un arbre que coincideix amb les ordenacions. Per exemple, els dos arbres



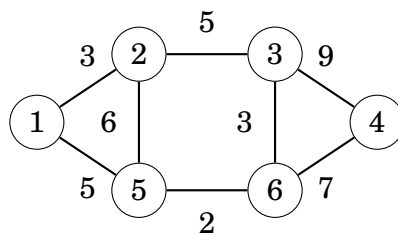
tenen pre-ordre [1, 2] i post-ordre [2, 1], però les estructures dels arbres són diferents.

Capítol 15

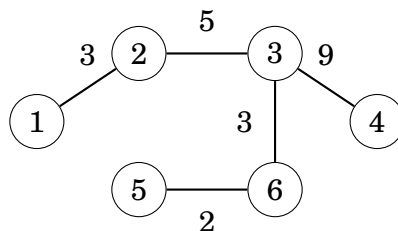
Arbres d'expansió

Un **arbre d'expansió** (*spanning tree*) d'un graf és un arbre que conté tots els nodes del graf i algunes de les arestes. Com que és un arbre, els arbres d'expansió són connexos, sense cicles i existeix un sol camí entre dos nodes qualsevol. Normalment hi ha diverses maneres de construir un arbre d'expansió.

Per exemple, considereu el graf següent:

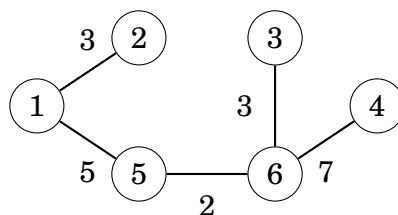


Un possible arbre d'expansió és el següent:

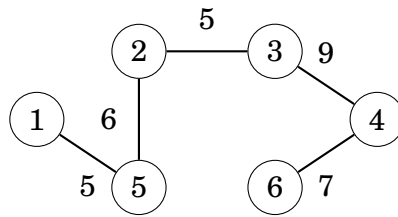


El pes d'un arbre d'expansió és la suma dels pesos de les seves arestes. Per exemple, el pes de l'arbre d'expansió anterior és $3 + 5 + 9 + 3 + 2 = 22$.

Un **arbre d'expansió mínim** és aquell que té pes mínim. En l'exemple anterior, el pes mínim és 20, i aquest arbre es pot construir com segueix:



De la mateixa manera, un **arbre d'expansió màxim** és aquell que té pes màxim. En l'exemple anterior és 32:



Tingueu en compte que un graf pot tenir diversos arbres d'expansió mínims i màxims, de manera que els arbres no són únics.

Resulta que es poden fer servir diversos algorismes greedy per construir arbres d'expansió mínim i màxim. En aquest capítol, discutim dos algorismes que processen les arestes del graf ordenades segons els seus pesos. Ens centrem en trobar arbres d'expansió mínim, però els mateixos algorismes poden trobar arbres d'expansió màxim processant les arestes en ordre invers.

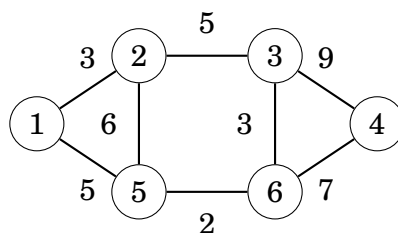
15.1 Algorisme de Kruskal

En l'**algorisme de Kruskal**¹, omplirem un arbre d'expansió aresta a aresta. L'algorisme recorre totes les arestes ordenades pels seus pesos, i afegeix l'aresta si no crea cap cycle amb les arestes ja seleccionades.

L'algorisme manté les components connexes de l'arbre d'expansió que estem omplint. Inicialment, cadascun dels n nodes del graf pertany a una component connexa distinta. Només afegim arestes quan unim dues components distintes, ja que altrament tindrem un cycle. L'algorisme acaba quan, després d'afegir $n - 1$ arestes, tenim una sola component connexa resultant. Quan això passa hem trobat un arbre d'expansió mínim.

Exemple

Considerem ara com l'algorisme de Kruskal processa el graf següent:



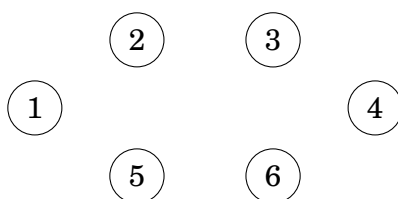
El primer pas de l'algorisme és ordenar les arestes en ordre creixent dels seus pesos. El resultat és el següent:

¹L'algorisme va ser publicat l'any 1956 per J. B. Kruskal [48].

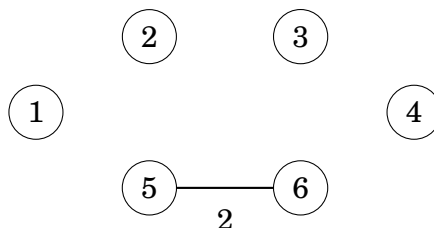
aresta	pes
5-6	2
1-2	3
3-6	3
1-5	5
2-3	5
2-5	6
4-6	7
3-4	9

Després d'això, l'algorisme recorre la llista i afegeix cada aresta a l'arbre si aquesta uneix dues components separades.

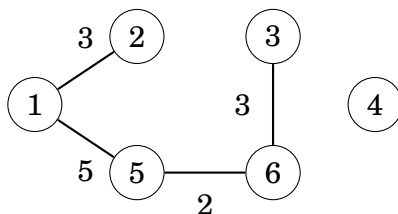
Inicialment, cada node té la seva pròpia component connexa:



La primera aresta que s'afegeix a l'arbre és la aresta 5-6 que crea la component {5,6} resultant d'unir les components {5} i {6}:



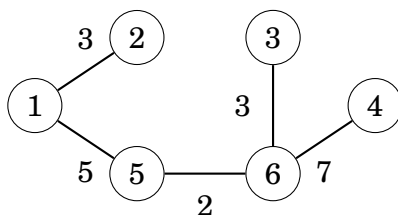
Després d'això, les arestes 1-2, 3-6 i 1-5 s'afegeixen de la mateixa manera:



Després d'aquests passos, la majoria de components s'han unit i hi ha dues components a l'arbre: {1,2,3,5,6} i {4}.

L'aresta següent de la llista és l'aresta 2-3, però no s'inclou a l'arbre perquè els nodes 2 i 3 ja estan a la mateixa component. El mateix passa amb l'aresta 2-5.

Finalment, l'aresta 4-6 s'afegeix a l'arbre:

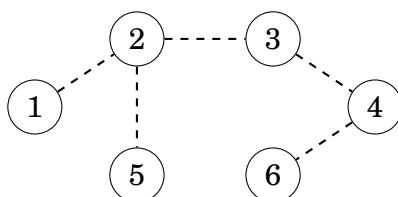


Després d'això, l'algorisme no afegeix cap aresta nova, perquè el graf està connectat i hi ha un camí entre dos nodes qualsevol. El graf resultant és un arbre d'expansió mínim amb pes $2 + 3 + 3 + 5 + 7 = 20$.

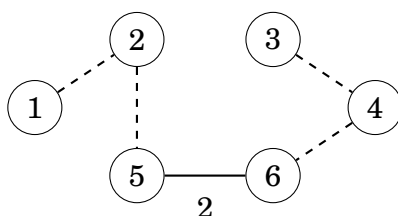
Per què funciona això?

És bo demanar-se per què l'estratègia greedy de l'algorisme de Kruskal sempre funciona.

Vegem què passa si l'aresta de mínim pes del graf *no* s'inclogués a l'arbre d'expansió. Per exemple, suposem que l'arbre d'expansió del graf anterior no inclou l'aresta de pes mínim 5–6. No coneixem l'estructura exacta de l'arbre d'expansió, però sens dubte contindrà algunes arestes. Imaginem-nos un arbre com aquest:



No és possible que l'arbre anterior sigui un arbre d'expansió mínim. La raó d'això és que si afegim a aquest arbre l'aresta de pes mínim 5–6 estem creant un cicle. Si treiem qualsevol altra aresta del cicle trobarem un nou arbre d'expansió amb pes *menor*:



Per aquesta raó, sabem que sempre és òptim incloure qualsevol aresta de pes mínim a l'arbre d'expansió mínim. Amb un argument similar, podem demostrar que afegir l'aresta següent en ordre de pes també és òptim, i així successivament. Per tant, l'algorisme de Kruskal funciona correctament i sempre troba un arbre d'expansió mínim.

Implementació

Quan s'implementa l'algorisme de Kruskal, és convenient fer servir la representació de llista d'arestes del graf. La primera fase de l'algorisme ordena les arestes de la llista en temps $O(m \log m)$ temps. La segona fase de l'algorisme construeix l'arbre d'abast mínim de la següent manera:

```
for (...) {  
    if (!same(a,b)) unite(a,b);  
}
```

El bucle passa per les arestes de la llista i considera arestes de la forma $a-b$. Es necessiten dues funcions: la funció `same` determina si a i b estan en la mateixa component connexa, i la funció `unite` uneix les components connexes que contenen a i b .

El problema és com implementar de manera eficient les funcions `same` i `unite`. Una possibilitat és implementar la funció `same` com a recorregut de grafs i comprovar si podem passar del node a al node b . Tanmateix, la complexitat temporal d'aquesta funció seria $O(n + m)$ i l'algorisme resultant seria lent, perquè es cridarà la funció `same` per a cada aresta del graf.

Resoldrem el problema fent servir una estructura *union-find* (o també coneguda com *merge-find-set* o *MF set*) que implementa ambdues funcions en temps $O(\log n)^2$. Així, la complexitat temporal de l'algorisme de Kruskal és $O(m \log n)$ després d'ordenar la llista de arestes.

15.2 Estructura *union-find*

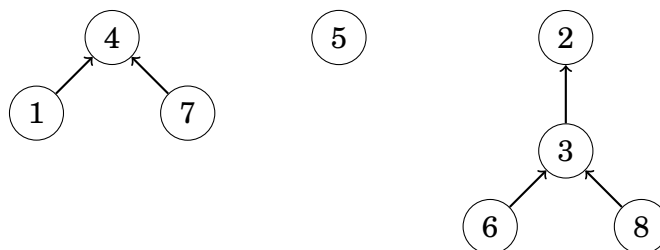
Una **estructura *union-find*** manté una col·lecció de conjunts. Els conjunts són disjunts, de manera que cap element pertany a més d'un conjunt. Aquesta estructura admet dues operacions de temps $O(\log n)$: l'operació `unite` uneix dos conjunts, i l'operació `find` troba el representant del conjunt que conté un element donat³.

Estructura

En una estructura *union-find*, un element de cada conjunt és el representant del conjunt, i hi ha una cadena des de qualsevol altre element del conjunt fins al seu representant. Per exemple, suposem que els conjunts són $\{1, 4, 7\}$, $\{5\}$ i $\{2, 3, 6, 8\}$:

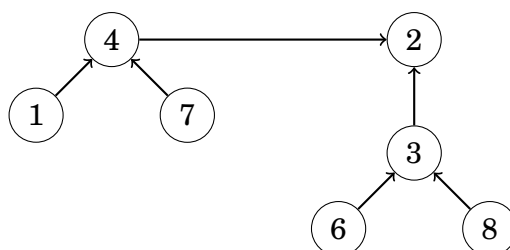
²(N. del T.): De fet, s'implementa en temps $O(\alpha(n))$, on $\alpha(n)$ és la funció inversa d'Ackermann. Aquesta funció creix encara més lentament que el logaritme, i a efectes pràctics es pot considerar simplement una constant petita.

³L'estructura presentada aquí va ser introduït el 1971 per JD Hopcroft i JD Ullman [38]. Més tard, el 1975, R. E. Tarjan va estudiar una variant més sofisticada de l'estructura [64] que es discuteix en molts llibres d'algorísmica.



En aquest cas els representants dels conjunts són 4, 5 i 2. Podem trobar el representant d'un element seguint la cadena que comença en l'element donat. Per exemple, l'element 2 és el representant del conjunt que conté l'element 6, perquè seguim la cadena $6 \rightarrow 3 \rightarrow 2$. Dos elements pertanyen al mateix conjunt exactament quan els seus representants són els mateixos.

Podem unir dos conjunts connectant el representant d'un conjunt amb el representant de l'altre conjunt. Per exemple, els conjunts $\{1, 4, 7\}$ i $\{2, 3, 6, 8\}$ es poden unir de la següent manera:



El conjunt resultant conté els elements $\{1, 2, 3, 4, 6, 7, 8\}$. A partir d'aquí, l'element 2 és el representant de tot el conjunt i l'antic representant 4 apunta a l'element 2.

L'eficiència de l'estructura d'unió-troba depèn de com s'uneixen els conjunts. Resulta que podem seguir una estratègia senzilla: fent la connexió des del representant del conjunt *més petit* al representant del conjunt *més gran* (o, si els conjunts són de la mateixa mida, podem fer una elecció arbitrària). Amb aquesta estratègia, es pot veure que la longitud de qualsevol cadena és $O(\log n)$, de manera que es pot trobar el representant de qualsevol element de manera eficient seguint la cadena corresponent.

Implementació

L'estructura *union-find* s'implementa amb vectors. En la següent implementació, el vector `link` conté per a cada element l'element següent de la cadena, o el propi element si és un representant del conjunt, i el vector `size` indica per a cada representant la mida del seu conjunt corresponent.

Inicialment, cada element pertany a un conjunt distint:

```

for (int i = 1; i <= n; i++) link[i] = i;
for (int i = 1; i <= n; i++) size[i] = 1;

```

La funció `find` retorna el representant d'un element x , que es troba seguint la cadena que comença a x .

```
int find(int x) {
    while (x != link[x]) x = link[x];
    return x;
}
```

La funció `same` verifica si els elements a i b pertanyen al mateix conjunt. Això es pot fer fàcilment utilitzant la funció `find`:

```
bool same(int a, int b) {
    return find(a) == find(b);
}
```

La funció `unite` uneix els conjunts que contenen els elements a and b (els elements han de pertànyer a conjunts distints). La funció primer troba els representants dels conjunts i després connecta el conjunt més petit al conjunt més gran.

```
void unite(int a, int b) {
    a = find(a);
    b = find(b);
    if (size[a] < size[b]) swap(a,b);
    size[a] += size[b];
    link[b] = a;
}
```

La complexitat temporal de la funció `find` és $O(\log n)$ perquè la longitud de cada cadena és $O(\log n)$. En aquest cas, les funcions `same` i `unite` també funcionen en temps $O(\log n)$. La funció `unite` assegura que la longitud de cada cadena és $O(\log n)$ perquè connecta el conjunt més petit al conjunt més gran.

15.3 Algorisme de Prim

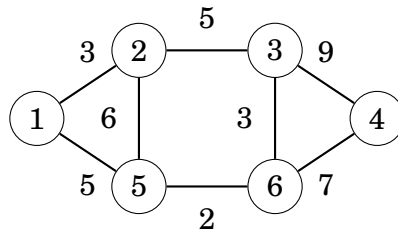
L'**algorisme de Prim**⁴ és un mètode alternatiu per trobar un arbre d'expansió mínim. L'algorisme afegeix primer un node arbitrari a l'arbre. Després d'això, l'algorisme sempre tria una aresta de pes mínim que afegeix un nou node a l'arbre. Al final, s'han afegit tots els nodes a l'arbre i s'ha trobat un arbre d'expansió mínim.

L'algorisme de Prim s'assembla a l'algorisme de Dijkstra. La diferència és que l'algorisme de Dijkstra sempre selecciona una aresta la distància de la qual des del node inicial és mínima, però l'algorisme de Prim simplement selecciona la aresta de pes mínim que afegeix un nou node a l'arbre.

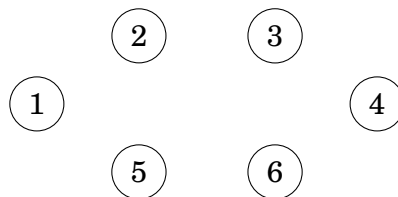
⁴L'algorisme rep el nom de R. C. Prim que el va publicar l'any 1957 [54]. No obstant això, el mateix algorisme ja va ser descobert l'any 1930 per V. Jarník.

Exemple

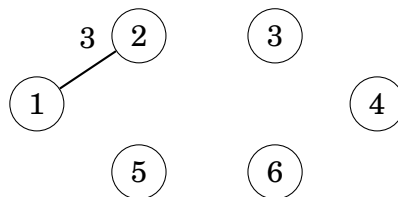
Considerem com funciona l'algorisme de Prim en el graf següent:



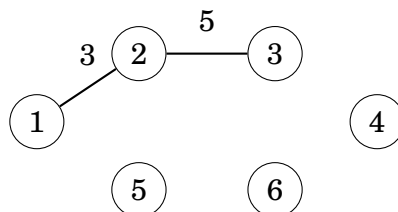
Inicialment, no hi ha arestes entre els nodes:



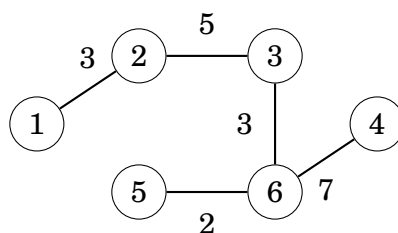
El node inicial és un node arbitrari, així que escollim el node 1. Primer, afegim el node 2 que està connectat per una aresta de pes 3:



Després d'això, hi ha dos arestes amb pes 5, de manera que podem afegir el node 3 o el node 5 a l'arbre. Afegim primer el node 3:



El procés continua fins que tots els nodes s'han inclòs en l'arbre:



Implementació

A l'igual que l'algorisme de Dijkstra, l'algorisme de Prim s'implementa de manera eficient amb una cua de prioritats. La cua de prioritats ha de contenir tots els nodes que es poden connectar a la component actual mitjançant una única aresta, en ordre creixent dels pesos de les arestes corresponents.

La complexitat temporal de l'algorisme de Prim és $O(n + m \log m)$ que és igual a la complexitat temporal de l'algorisme de Dijkstra. A la pràctica, els algorismes de Prim i Kruskal són tots dos eficients, i l'elecció de l'algorisme és qüestió de gustos. La majoria dels programadors competitius fan servir l'algorisme de Kruskal.

Capítol 16

Grafs dirigits

En aquest capítol, ens centrem en dues classes de grafs dirigits:

- **Acyclic graphs:** There are no cycles in the graph, so there is no path from any node to itself¹.
- **Successor graphs:** The outdegree of each node is 1, so each node has a unique successor.

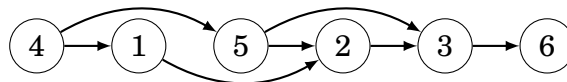
Resulta que en ambdós casos podem dissenyar algorismes eficients que es basen en propietats especials dels grafs.

16.1 Ordenació topològica

Una **ordenació topològica** és una ordenació dels nodes d'un graf dirigit de manera que quan hi ha un camí des del node a fins al node b es compleix que el node a apareix abans que el node b en l'ordenació. Per exemple, per al graf



es té que $[4, 1, 5, 2, 3, 6]$ és una possible ordenació topològica:



Un graf acíclic sempre té una ordenació topològica. Tanmateix, si el graf conté un cicle, no és possible formar una ordenació topològica, perquè cap node del cicle pot aparèixer abans que els altres nodes del cicle en l'ordenació. Resulta que la cerca en profunditat es pot fer servir tant per comprovar si un graf dirigit conté un cicle com, si no conté un cicle, per construir una ordenació topològica.

¹Directed acyclic graphs are sometimes called DAGs.

Algorisme

La idea és iterar tots els nodes del graf, iniciant una nova cerca en profunditat cada cop que trobem un node sense processar. Durant el recorregut, els nodes poden tenir tres estats possibles:

- state 0: the node has not been processed (white)
- state 1: the node is under processing (light gray)
- state 2: the node has been processed (dark gray)

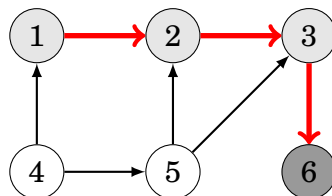
Inicialment, l'estat de cada node és 0. Quan una cerca arriba a un node per primera vegada, el seu estat passa a ser 1. Un cop hem processat tots els successors del node, el seu estat passa a ser 2.

Si el graf conté un cicle, ho descobrirem durant la cerca, perquè tard o d'hora arribarem a un node l'estat del qual és 1. En aquest cas, no és possible construir un ordenament topològic.

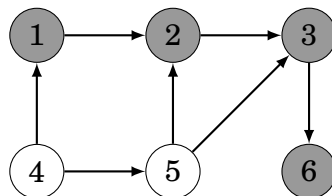
Si el graf no conté cap cicle, podem construir una ordenació topològica afegint cada node a una llista quan l'estat del node esdevé 2. Aquesta llista conté una ordenació topològica en ordre invers.

Exemple 1

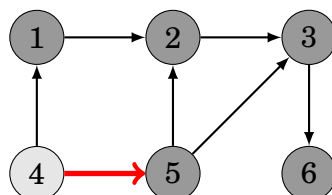
Al graf d'exemple, la cerca passa del node 1 al node 6:



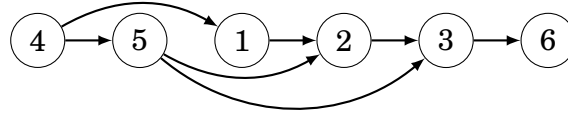
Amb això hem processat el node 6, de manera que l'afegim a la llista. Després d'això, també afegim els nodes 3, 2 i 1 a la llista:



En aquest punt, la llista és [6,3,2,1]. La cerca següent comença al node 4:



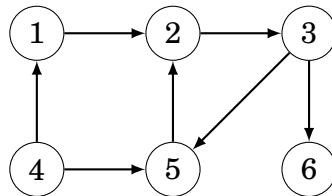
Així, la llista final és [6,3,2,1,5,4]. Hem processat tots els nodes, de manera que hem trobat una ordenació classificació topològica, la llista inversa [4,5,1,2,3,6]:



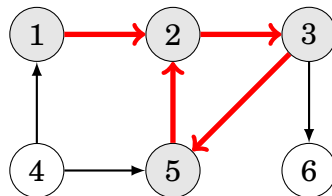
Tingueu en compte que les ordenacions topològiques no són úniques, i que per tant, un graf donat pot tenir moltes ordenacions topològiques distintes.

Exemple 2

Considerem ara un graf per al qual no podem construir cap ordenació topològica, perquè el graf conté un cicle:



La cerca continua de la següent manera:



La cerca arriba al node 2 l'estat del qual és 1, el que significa que el graf conté un cicle. En aquest exemple, hi ha un cicle $2 \rightarrow 3 \rightarrow 5 \rightarrow 2$.

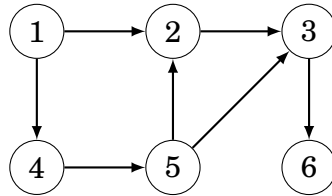
16.2 Programació dinàmica

Quan un graf dirigit és acíclic podem aplicar-hi programació dinàmica. Per exemple, podem resoldre de manera eficient els problemes següents relacionats amb els camins des d'un node inicial fins a un node final:

- how many different paths are there?
- what is the shortest/longest path?
- what is the minimum/maximum number of edges in a path?
- which nodes certainly appear in any path?

Comptar el nombre de camins

Per exemple, calculem el nombre de camins del node 1 al node 6 en el graf següent:



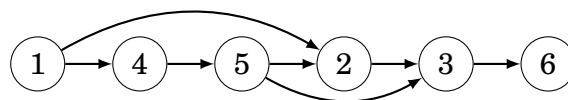
Hi ha un total de tres camins d'aquest tipus:

- $1 \rightarrow 2 \rightarrow 3 \rightarrow 6$
- $1 \rightarrow 4 \rightarrow 5 \rightarrow 2 \rightarrow 3 \rightarrow 6$
- $1 \rightarrow 4 \rightarrow 5 \rightarrow 3 \rightarrow 6$

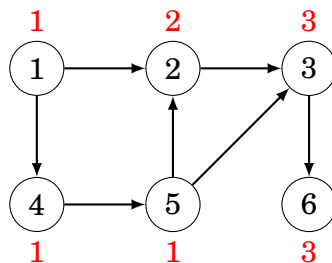
Sigui $\text{paths}(x)$ el nombre de camins des del node 1 fins al node x . Com a cas base, $\text{camins}(1) = 1$. Podem calcular els valors de $\text{camins}(x)$ fent servir la recursió

$$\text{paths}(x) = \text{paths}(a_1) + \text{paths}(a_2) + \dots + \text{paths}(a_k)$$

on a_1, a_2, \dots, a_k són els nodes dels quals surt una aresta cap a x . Com que el graf és acíclic, una ordenació topològica ens permet calcular els valors de $\text{paths}(x)$ amb programació dinàmica. Per exemple, aquesta és una ordenació topològica del graf:



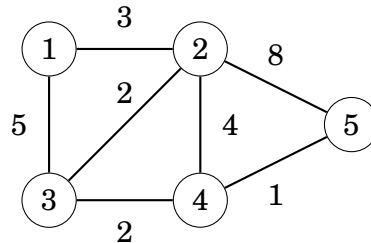
Per tant, els nombres de camins són:



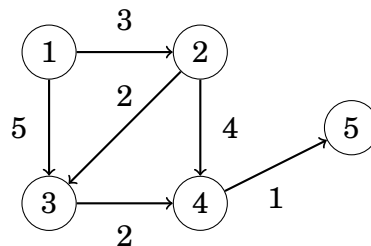
Per exemple, per calcular el valor de $\text{paths}(3)$, fem servir la fórmula $\text{paths}(2) + \text{paths}(5)$, perquè hi ha arestes que surten dels nodes 2 i 5 al node 3. Com que $\text{paths}(2) = 2$ i $\text{paths}(5) = 1$, concloem que $\text{paths}(3) = 3$.

Extensió de l'algorisme de Dijkstra

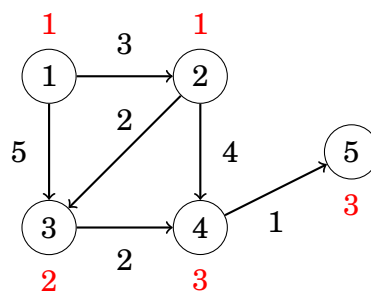
Un subproducte de l'algorisme de Dijkstra és un graf acíclic dirigit que indica per a cada node del graf original les possibles maneres d'arribar al node des del node inicial fent servir camins mínims. Podem aplicar programació dinàmica en aquest graf. Per exemple, en el graf



els camins mínims des del node 1 poden fer servir les arestes següents:



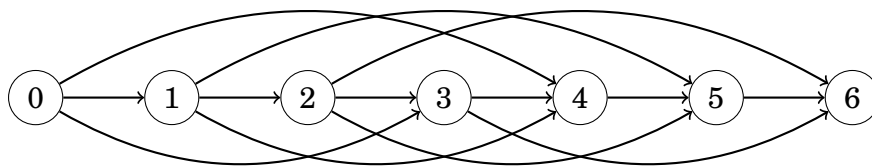
Ara podem, per exemple, calcular el nombre de camins mínims del node 1 al node 5 mitjançant la programació dinàmica:



Representació problemes com a graf

De fet, qualsevol problema de programació dinàmica es pot representar com un graf acíclic dirigit. En aquest graf, cada node correspon a un estat de la programació dinàmica, i les arestes indiquen que un estat depèn d'un altre.

Per exemple, considereu el problema de formar una suma de diners n fent servir monedes $\{c_1, c_2, \dots, c_k\}$. En aquest problema, podem construir un graf on cada node es correspon amb una suma de diners, i les arestes mostren com triar una moneda. Per exemple, per a les monedes $\{1, 3, 4\}$ i $n = 6$, el graf és el següent:



Utilitzant aquesta representació, el camí mínim del node 0 al node n es correspon amb una solució que fa servir el nombre mínim de monedes, i el nombre total de camins des del node 0 al node n és igual al nombre total de solucions.

16.3 Camins de successió

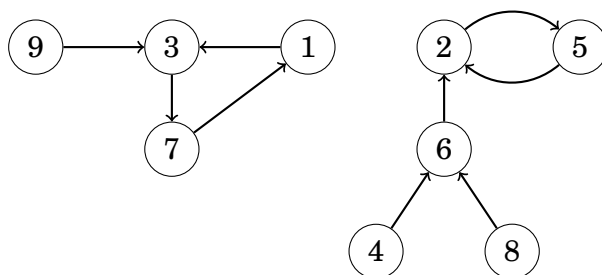
Per a la resta del capítol, ens centrarem en **grafs de successió**. En aquests grafs, el grau de sortida de cada node és 1, és a dir, exactament una aresta comença en cada node. Un graf successor consta d'un o més components, cadascun dels quals conté un cicle i alguns camins que van a parar al cicle.

Els grafs de successió de vegades s'anomenen **grafs funcionals**. La raó d'això és que qualsevol graf de successió es correspon amb una funció que defineix les arestes del graf. El paràmetre de la funció és un node del graf i la funció retorna el successor d'aquest node.

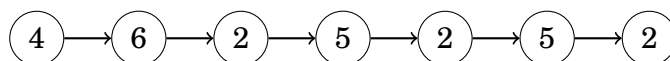
Per exemple, la funció

x	1	2	3	4	5	6	7	8	9
$\text{succ}(x)$	3	5	7	6	2	2	1	6	3

defineix el següent graf:



Com que cada node d'un graf de successió té un únic successor, també podem definir una funció $\text{succ}(x, k)$ que ens retorna el node al qual arribarem si comencem pel node x i avancem k passos endavant. Per exemple, al graf anterior $\text{succ}(4, 6) = 2$, perquè arribarem al node 2 avançant 6 passos des del node 4:



Una manera senzilla de calcular el valor de $\text{succ}(x, k)$ és començar al node x i avançar k passos endavant, i triga temps $O(k)$. Tanmateix, fent servir preporcessament, qualsevol valor de $\text{succ}(x, k)$ es pot calcular en temps $O(\log k)$.

La idea és precalcular tots els valors de $\text{succ}(x, k)$ on k és una potència de dos i com a màxim u , on u és el nombre màxim de passos que permetem avançar. Això es pot fer de manera eficient fent servir la següent recursió:

$$\text{succ}(x, k) = \begin{cases} \text{succ}(x) & k = 1 \\ \text{succ}(\text{succ}(x, k/2), k/2) & k > 1 \end{cases}$$

Precalculer els valors requereix temps $O(n \log u)$, perquè hem de calcular $O(\log u)$ valors per cada node. En el graf anterior, els primers valors són els següents:

x	1	2	3	4	5	6	7	8	9
$\text{succ}(x, 1)$	3	5	7	6	2	2	1	6	3
$\text{succ}(x, 2)$	7	2	1	2	5	5	3	2	7
$\text{succ}(x, 4)$	3	2	7	2	5	5	1	2	3
$\text{succ}(x, 8)$	7	2	1	2	5	5	3	2	7
...									

Després d'això, podem calcular qualsevol valor de $\text{succ}(x, k)$ representant el nombre de passos k com una suma de potències de dos. Per exemple, si volem calcular el valor de $\text{succ}(x, 11)$, primer formem la representació $11 = 8 + 2 + 1$. Amb això tenim

$$\text{succ}(x, 11) = \text{succ}(\text{succ}(\text{succ}(x, 8), 2), 1).$$

Per exemple, al graf anterior

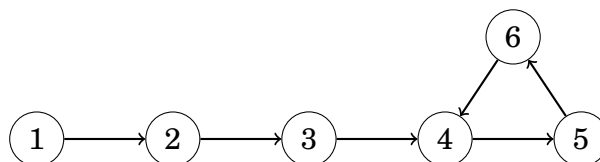
$$\text{succ}(4, 11) = \text{succ}(\text{succ}(\text{succ}(4, 8), 2), 1) = 5.$$

Aquesta representació sempre consta de $O(\log k)$ parts, de manera que calcular un valor de $\text{succ}(x, k)$ triga temps $O(\log k)$.

16.4 Detecció de cicles

Considerem un graf de successió que només conté un camí que acaba en un cicle. Podem fer-nos les preguntes següents: si comencem la nostra cerca al node inicial, quin és el primer node del cicle i quants nodes conté el cicle?

Per exemple, en el graf



comencem el nostre camí al node 1, el primer node que pertany al cicle és el node 4, i el cicle consta de tres nodes (4, 5 i 6).

Una manera senzilla de detectar el cicle és caminar pel graf i fer un seguiment de tots els nodes que s'han visitat. Un cop visitat un node per segona vegada,

podem concloure que el node és el primer node del cicle. Aquest mètode funciona en temps $O(n)$ i fa servir memòria $O(n)$.

Tanmateix, hi ha millors algorismes per a la detecció de cicles. La complexitat temporal d'aquests algorismes segueix sent $O(n)$, però només fan servir memòria $O(1)$. Aquesta és una millora important si n és gran. A continuació parlarem de l'algorisme de Floyd que compleix aquesta propietat.

Algorisme de Floyd

L'**algorisme de Floyd**² avança pel graf fent servir dos punters a i b . Tots dos punters comencen en un node x que és el node inicial del graf. Aleshores, a cada pas, el punter a fa un pas endavant mentre que el punter b fa dos passos endavant. El procés continua fins que els punters es troben entre ells:

```
a = succ(x);
b = succ(succ(x));
while (a != b) {
    a = succ(a);
    b = succ(succ(b));
}
```

A continuació, fem que el punter a torni a apuntar a x , i avancem ambdós punters pas a pas fins que es tornin a trobar. El punt on es tornen a trobar és el primer node del cicle.

```
a = x;
while (a != b) {
    a = succ(a);
    b = succ(b);
}
first = a;
```

La durada del cicle es calcula de la següent manera:

```
b = succ(a);
length = 1;
while (a != b) {
    b = succ(b);
    length++;
}
```

Per què funciona l'algorisme de Floyd? Quan els punters a i b es troben per primer cop, el punter a ha fet k passos i el punter b ha fet $2k$ passos. El punter b ha avançat $2k - k = k$ passos més que el punter a , i com que aquests k passos addicionals han estat donant voltes de més al cicle, deduïm que la mida c del cicle divideix k . Sigui $k = v + w$, on v és el nombre de passos que el punter a ha

²La idea de l'algorisme s'esmenta a [46] i s'atribueix a R. W. Floyd; tanmateix, no se sap si Floyd va descobrir realment l'algorisme.

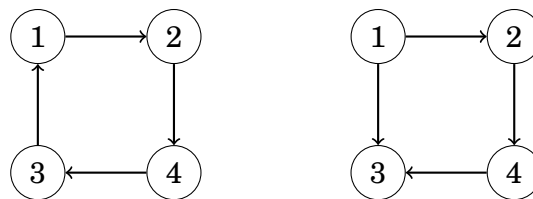
trigat a arribar al primer node del cicle i w el nombre de passos fets després. Amb aquesta notació, en la segona fase de l'algorisme els punters a i b arribaran al primer node del cicle en v i $c - w$ passos respectivament. Però com que c divideix $k = v + w$, es compleix que $v + w = 0$ mòdul c o, equivalentment, $v = c - w$ mòdul c . Aquesta congruència garanteix que en la segona fase de l'algorisme els punters a i b es tornaran a trobar, i això només pot passar en el primer node del cicle. Quan això passa, podem calcular la mida c del cicle fent que b doni una volta de més, fins retrobar a .

Capítol 17

Grafs fortament connexos

En un graf dirigit, les arestes només es poden recórrer en una direcció, de manera que encara que el graf sigui connex, això no garanteix que sempre hi hagi un camí dirigit entre dos nodes. Per aquest motiu, definim un nou concepte més fort que el de connectivitat.

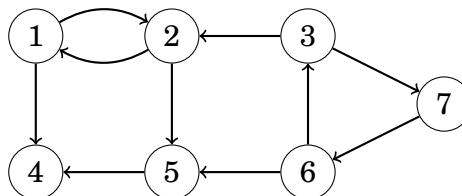
Un graf és **fortament connex** si hi ha un camí des de qualsevol node a tots els altres nodes del graf. Per exemple, a la imatge següent, el graf esquerre és fortament connex mentre que el graf dret no.



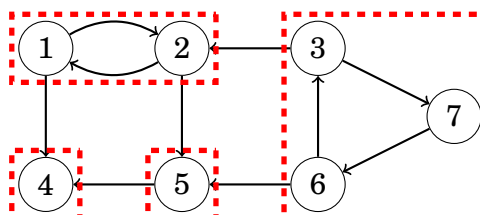
El graf dret no és fortament connex perquè, per exemple, no hi ha cap camí des del node 2 fins al node 1.

Les **components fortament connexes** d'un graf divideixen el graf en parts fortament connexes tan grans com sigui possible. Les components fortament connexes d'un graf formen un **graf de components**, que és un graf acíclic i dirigit que representa l'estructura del graf original.

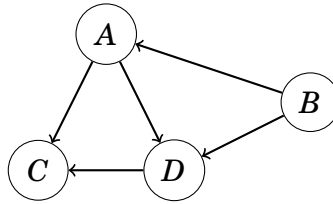
Per exemple, el graf



té les següents components fortament connexes:



El graf de components corresponent és el següent:



Les components són $A = \{1, 2\}$, $B = \{3, 6, 7\}$, $C = \{4\}$ i $D = \{5\}$.

Com que el graf de components és acíclic i dirigit, és més fàcil de processar que el graf original. Com que el graf no conté cicles, sempre podem construir una ordenació topològica i utilitzar tècniques de programació dinàmica del capítol 16.

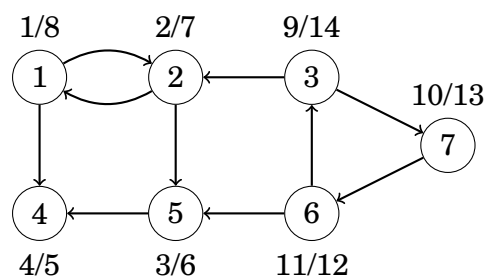
17.1 Algorisme de Kosaraju

L'algorisme de Kosaraju¹ és un mètode eficient per trobar les components fortament connexes d'un graf dirigit. L'algorisme realitza dues cerques en profunditat: la primera cerca construeix una llista de nodes segons l'estructura del graf, i la segona cerca forma les components fortament connectats.

Cerca 1

La primera fase de l'algorisme de Kosaraju construeix una llista de nodes en l'ordre en el que apareixen en una cerca en profunditat. L'algorisme itera pels nodes i comença una cerca en profunditat per cada node no processat. Cada node s'afegeix a la llista un cop s'ha processat.

En aquest exemple els nodes es processen en l'ordre següent:



Fem servir la notació x/y per a indicar que el node s'ha començat a processar en el pas x i s'ha acabat de processar en el pas y . La llista resultant és la següent:

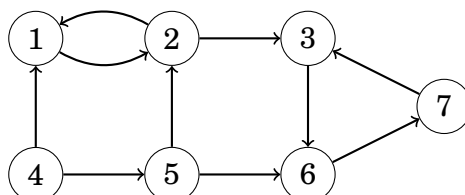
¹Segons [1], S. R. Kosaraju va inventar aquest algorisme l'any 1978 però no el va publicar. El 1981, el mateix algorisme va ser redescobert i publicat per M. Sharir [57].

node	pas de final de processament
4	5
5	6
2	7
1	8
6	12
7	13
3	14

2

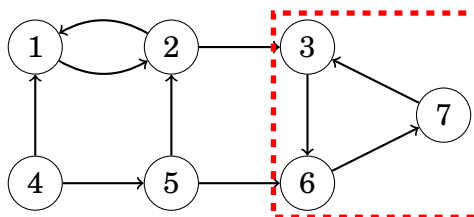
Cerca 2

En la segona fase de l'algorisme formem les components fortament connexes del graf. En primer lloc, l'algorisme inverteix totes les arestes del graf. Després d'invertir les arestes, el graf d'exemple és el següent:



Ara l'algorisme recorre en ordre *invers* la llista de nodes creats per la primera cerca. Per cada node que encara no pertany a cap component, l'algorisme crea una nova component i inicia una cerca en profunditat que afegeix tots els nodes nous trobats a la nova component.

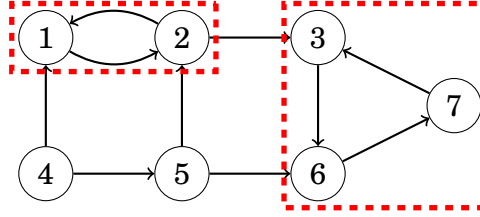
Al graf d'exemple, la primera component comença al node 3:



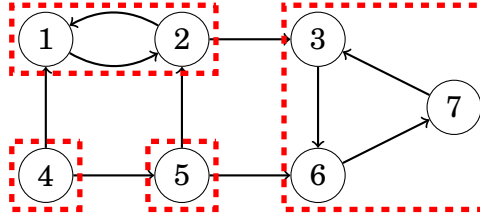
Tingueu en compte que, com que totes les arestes estan invertides, no es possible “escapar-se” a altres parts del graf. (N. del T.) Com que tractem els nodes en ordre topològic de les components fortament connexes, quan tractem una component sabem que ja hem acabat de tractar totes aquelles altres components des de les quals es podia arribar a la nostra en el graf original. Això garanteix que la segona cerca no pot sortir de la component fortament connexa.

Els nodes següents a la llista són els nodes 7 i 6, però ja pertanyen a una component, de manera que la nova component comença al node 1:

²(N. del T.) La propietat clau d'aquesta llista és que si, es pot anar de u a v però no és pot anar de v a u , es té que v apareix abans que u en la llista. Dit d'altra manera: la llista ordena els nodes en *ordre topològic invers* de les components fortament connexes respectives en el graf de components.



Finalment, l'algorisme processa els nodes 5 i 4 que creen les components fortament connexes que falten:



La complexitat temporal de l'algorisme és $O(n+m)$, perquè l'algorisme realitza dues cerques en profunditat.

17.2 Problema 2SAT

Els grafs fortament connexos també estan relacionats amb el problema **2SAT**³. En aquest problema, se'ns dona una fórmula lògica

$$(a_1 \vee b_1) \wedge (a_2 \vee b_2) \wedge \cdots \wedge (a_m \vee b_m),$$

on cada a_i i b_i és una variable lògica (x_1, x_2, \dots, x_n) o una negació d'una variable lògica ($\neg x_1, \neg x_2, \dots, \neg x_n$). Els símbols " \wedge " i " \vee " denoten operadors lògics "AND" i "OR". La nostra tasca és assignar un valor a cada variable perquè la fórmula sigui certa, o bé afirmar que això no és possible.

Per exemple, la fórmula

$$L_1 = (x_2 \vee \neg x_1) \wedge (\neg x_1 \vee \neg x_2) \wedge (x_1 \vee x_3) \wedge (\neg x_2 \vee \neg x_3) \wedge (x_1 \vee x_4)$$

és certa quan les variables s'assignen de la següent manera:

$$\begin{cases} x_1 = \text{false} \\ x_2 = \text{false} \\ x_3 = \text{true} \\ x_4 = \text{true} \end{cases}$$

Tanmateix, la fórmula

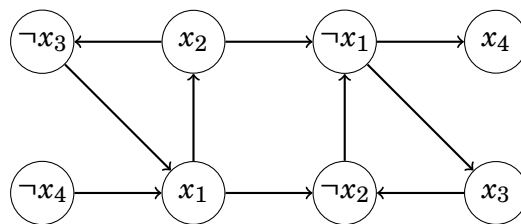
$$L_2 = (x_1 \vee x_2) \wedge (x_1 \vee \neg x_2) \wedge (\neg x_1 \vee x_3) \wedge (\neg x_1 \vee \neg x_3)$$

³L'algorisme que es presenta aquí es va introduir a [4]. També hi ha un altre algorisme de temps lineal conegut [19] que es basa en el retrocés (backtracking).

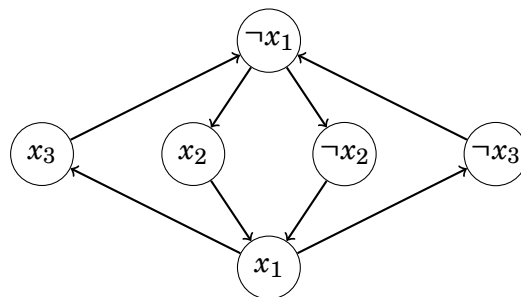
sempre és falsa, independentment de com assignem els valors. El motiu és que no podem triar un valor per a x_1 sense crear una contradicció. Si x_1 és fals, tant x_2 com $\neg x_2$ han de ser certs, cosa que és impossible; i si x_1 és cert, tant x_3 com $\neg x_3$ han de ser certs, cosa que també és impossible.

El problema 2SAT es pot representar com un graf els nodes del qual es corresponen amb variables x_i i negacions $\neg x_i$, i les arestes determinen les connexions entre les variables. Cada parell $(a_i \vee b_i)$ genera dues arestes: $\neg a_i \rightarrow b_i$ i $\neg b_i \rightarrow a_i$. Això vol dir que si a_i no es certa, b_i ha de ser certa, i viceversa.⁴

El graf de la fórmula L_1 és:



I el graf de la fórmula L_2 és:



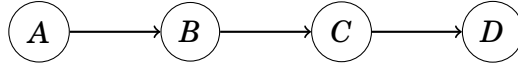
L'estructura del graf ens indica si és possible assignar valors a les variables de manera que la fórmula sigui certa. Resulta que això es pot fer exactament quan no hi ha nodes x_i i $\neg x_i$ que pertanyin a la mateixa component fortament connexa. Si existeixen aquests nodes, aleshores el graf conté un camí de x_i a $\neg x_i$ i també un camí de $\neg x_i$ a x_i , de manera que x_i implica $\neg x_i$ i $\neg x_i$ implica x_i , cosa que no és possible.

Al graf de la fórmula L_1 no hi ha nodes x_i i $\neg x_i$ que pertanyin a la mateixa component fortament connexa, de manera que existeix una solució. Al graf de la fórmula L_2 tots els nodes pertanyen a la mateixa component fortament connexa, i no hi ha cap solució.

Si existeix una solució, els valors de les variables es poden trobar passant pels nodes del graf de components en un ordre topològic invers. A cada pas, processem una component que no conté arestes que condueixin a una component sense processar. Si no s'han assignat valors a les variables de la component, busquem una assignació de valors compatible, i si ja tenen valors, no els canviem. El procés continua fins que a cada variable se li ha assignat un valor.

El graf de components per a la fórmula L_1 és el següent:

⁴(N. del T.) És útil interpretar una aresta dirigida (x, y) com “ x implica y ”, i un camí dirigit com una cadena d'implícacions lògiques.



Les components són $A = \{\neg x_4\}$, $B = \{x_1, x_2, \neg x_3\}$, $C = \{\neg x_1, \neg x_2, x_3\}$ i $D = \{x_4\}$. Quan construïm la solució, primer processem la component D , i x_4 esdevé cert. Després d'això, processem la component C on x_1 i x_2 esdevenen falsos i x_3 esdevé cert. Totes les variables tenen valors assignats, de manera que les components restants A i B no canvien les variables.

Tingueu en compte que aquest mètode funciona perquè el graf té una estructura especial: si hi ha camins des del node x_i al node x_j i des del node x_j al node $\neg x_j$, aleshores el node x_i mai no pot ser cert. El motiu es que també ha d'haver-hi un camí des del node $\neg x_j$ fins al node $\neg x_i$, i tant x_i com x_j esdevenen falsos.

Un problema més difícil és el **problema 3SAT**, on cada part de la fórmula és de la forma $(a_i \vee b_i \vee c_i)$, on a_i, b_i i c_i són també variables lògiques o les seves negacions. Aquest problema és NP-difícil, de manera que no es coneix cap algorisme eficient que el resolgui.

Capítol 18

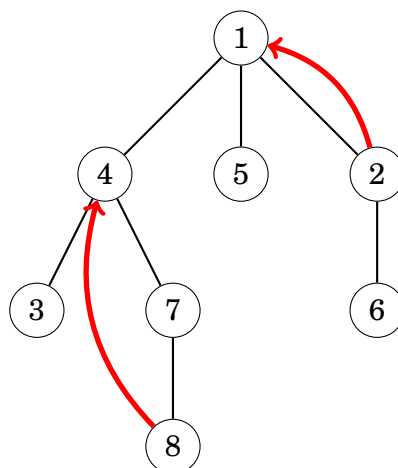
Consultes d'arbres

En aquest capítol es tracta de tècniques per processar consultes sobre subarbres i camins d'un arbre arrelat, com ara les següents:

- quin és el k -èssim avantpassat d'un node?
- quina és la suma dels valors del subarbre d'un node?
- quina és la suma dels valors en un camí entre dos nodes?
- quin és l'avantpassat comú més baix a dos nodes?

18.1 Trobar avantpassats

El k -èssim **avantpassat** d'un node x en un arbre arrelat és el node al qual arribarem si ens movem k nivells amunt d'un node x . Sigui $\text{ancestor}(x, k)$ el k -èssim avantpassat d'un node x (o 0 si no existeix). Per exemple, a l'arbre següent, $\text{ancestor}(2, 1) = 1$ i $\text{ancestor}(8, 2) = 4$.



Una manera fàcil de calcular $\text{ancestor}(x, k)$ és fer una seqüència de k moviments a l'arbre. Tanmateix, la complexitat temporal d'aquest mètode és $O(k)$, i això pot ser molt lent, perquè un arbre de n nodes pot tenir una cadena de n descendents.

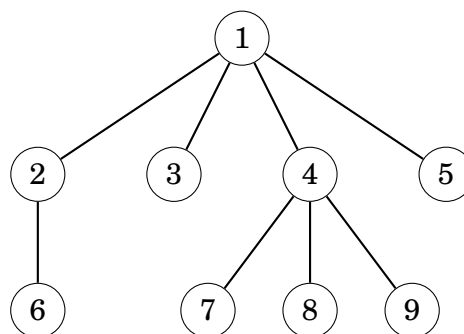
Afortunadament, amb una tècnica semblant a la del capítol 16.3, qualsevol valor de $\text{ancestor}(x, k)$ pot calcular-se eficientment en temps $O(\log k)$ un cop acabat el preprocessament. La idea és precalcular tots els valors $\text{ancestor}(x, k)$ on $k \leq n$ és una potència de dos. Per exemple, els valors de l'arbre anterior són els següents:

x	1	2	3	4	5	6	7	8
$\text{ancestor}(x, 1)$	0	1	4	1	1	2	4	7
$\text{ancestor}(x, 2)$	0	0	1	0	0	1	1	4
$\text{ancestor}(x, 4)$	0	0	0	0	0	0	0	0
...								

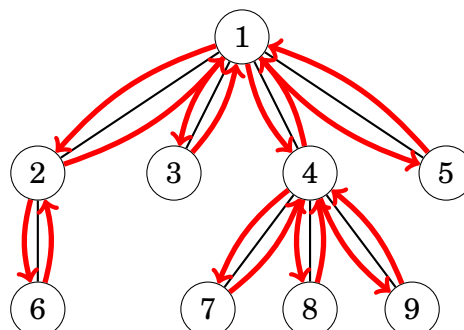
El preprocessament triga temps $O(n \log n)$, perquè es calculen $O(\log n)$ per cada node. Després d'això, qualsevol valor de $\text{ancestor}(x, k)$ es pot calcular en temps $O(\log k)$ representant k com a suma de potències de dos.

18.2 Subarbres i camins

Un **vector recorregut d'arbre** () conté els nodes d'un arbre arrelat en l'ordre en què una cerca en profunditat des del node arrel els visitaria. Per exemple, a l'arbre



una cerca en profunditat es fa com segueix:



Per tant, el vector recorregut corresponent és:

1	2	6	3	4	7	8	9	5
---	---	---	---	---	---	---	---	---

Consultes de subarbre

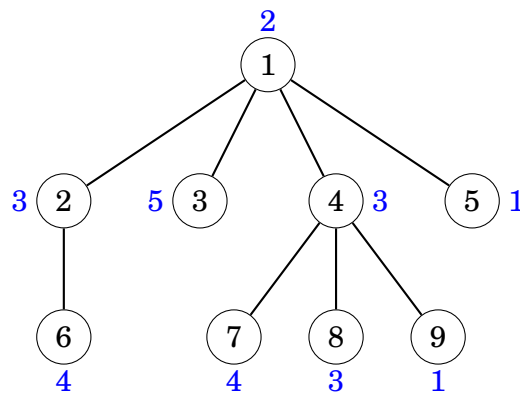
Cada subarbre d'un arbre es correspon a un subvector del vector recorregut d'arbre, on el primer element del subvector conté el node arrel del subarbre. Per exemple, el subvector següent conté els nodes del subarbre arrelat al node 4:

1	2	6	3	4	7	8	9	5
---	---	---	---	---	---	---	---	---

Amb aquest coneixement podem processar eficientment les consultes relacionades amb subarbres d'un arbre. Per exemple, considereu un problema on assignem un valor a cada node i la nostra feina és donar suport a les consultes següents:

- actualitzar el valor d'un node,
- calcular la suma de valors del subarbre arrelat a un node.

Considereu l'arbre següent on els números blaus són els valors dels nodes. Per exemple, la suma del subarbre arrelat al node 4 és $3 + 4 + 3 + 1 = 11$.



La idea és construir un vector recorregut d'arbre que contingui tres valors per cada node: l'identificador del node, la mida del subarbre i el valor del node. Per exemple, el vector associat a l'arbre anterior és:

node id	1	2	6	3	4	7	8	9	5
subtree size	9	2	1	1	4	1	1	1	1
node value	2	3	4	5	3	4	3	1	1

Amb aquesta vector, podem calcular la suma de valors de qualsevol subarbre esbrinant primer la mida del subarbre i després els valors dels nodes corresponents. Per exemple, els valors del subarbre arrelat al node 4 es poden trobar de la manera següent:

node id	1	2	6	3	4	7	8	9	5
subtree size	9	2	1	1	4	1	1	1	1
node value	2	3	4	5	3	4	3	1	1

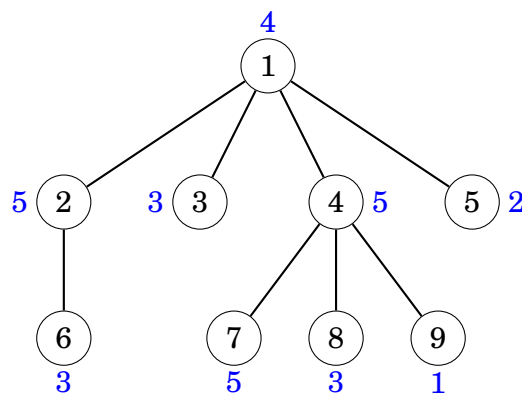
Per respondre les consultes de manera eficient, n'hi ha prou amb emmagatzemar els valors dels nodes en un arbre binari indexat o segmentat. Després d'això, podem actualitzar un valor i calcular la suma de valors en temps $O(\log n)$.

Consultes de camí

Amb un vector recorregut d'arbre també podem calcular eficientment les sumes de valors dels camins del node arrel a qualsevol altre node de l'arbre. Considereu el problema on la nostra tasca és donar suport a les consultes següents:

- canviar el valor d'un node,
- calcular la suma de valors en el camí de l'arrel a un node.

Per exemple, en l'arbre següent, la suma de valors del node arrel al node 7 és $4 + 5 + 5 = 14$:



Podem resoldre aquest problema com abans, però ara guardem en el nostre vector de tuples les sumes dels camins corresponents. Per exemple, el vector següent es correspon amb l'arbre anterior:

node id	1	2	6	3	4	7	8	9	5
subtree size	9	2	1	1	4	1	1	1	1
path sum	4	9	12	7	9	14	12	10	6

Quan el valor d'un node augmenta en x , les sumes dels camins de tots els nodes del seu subarbre augmenten en x . Per exemple, si el valor del node 4 augmenta en 1, el vector canvia de la següent manera:

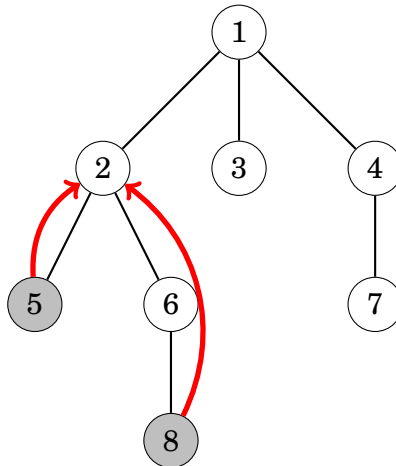
node id	1	2	6	3	4	7	8	9	5
subtree size	9	2	1	1	4	1	1	1	1
path sum	4	9	12	7	10	15	13	11	6

Així, podem implementar les dues operacions si som capaços d'augmentar tots els valors d'un interval i recuperar un sol valor. Això es pot fer en temps $O(\log n)$ fent servir arbre binari indexat o segmentat (vegeu el capítol 9.4).

18.3 Avantpassat comú més baix

La **avantpassat comú més baix** de dos nodes d'un arbre arrelat és el node més baix que conté ambdós nodes en el seu subarbre. Un problema típic és respondre eficientment consultes d'aquest tipus.

Per exemple, a l'arbre següent, l'avantpassat comú més baix dels nodes 5 i 8 és el node 2:



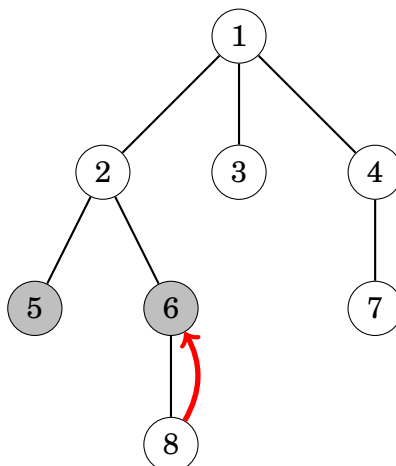
A continuació, mostrem dues tècniques eficients per trobar l'avantpassat comú més baix de dos nodes.

Mètode 1

Una manera de resoldre el problema és fer servir que podem trobar de manera eficient el k -èssim avantpassat de qualsevol node de l'arbre. Amb això, podem dividir el problema de trobar l'avantpassat comú més baix en dues parts.

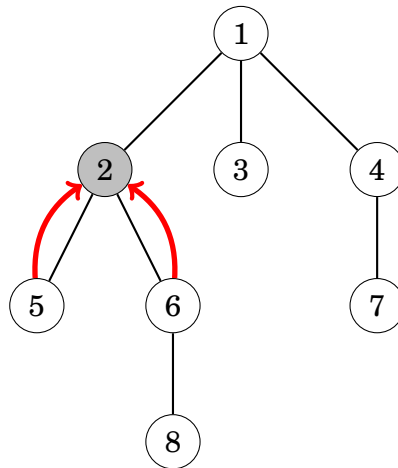
Fem servir dos punters que apunten inicialment als dos nodes. Primer, movem un dels punters cap amunt fins que ambdós punters estiguin a la mateixa alçada.

En l'escenari d'exemple, movem el segon punter un nivell cap amunt. Ara apunta al node 6, que es troba al mateix nivell que el node 5:



Després d'això, trobem el mínim nombre de passos necessaris cap a amunt per a que els dos punters apuntin al mateix node. Aquest node és l'avantpassat comú més baix.

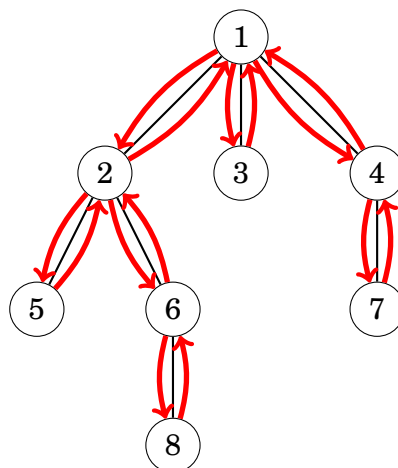
En l'escenari d'exemple, n'hi ha prou amb moure els dos punters un pas cap amunt fins al node 2, que és l'avantpassat comú més baix:



Les dos parts de l'algorisme es poden realitzar en temps $O(\log n)$ amb informació precalculada, de manera que podem trobar l'avantpassat comú més baix de dos nodes qualsevols en temps $O(\log n)$.

Mètode 2

Una altra manera de resoldre el problema es basa en un vector recorregut d'arbre¹. Una vegada més, la idea és d'avançar pels nodes amb una cerca en profunditat:



Ara, però, fem servir un vector recorregut d'arbre diferent de l'anterior: afegim cada node al vector *sempre* que la cerca en profunditat passa pel node, i no només

¹Aquest algorisme d'avantpassat comú més baix es va presentar a [7]. Aquesta tècnica de vegades s'anomena **tècnica de recorregut d'Euler** [66].

la primera vegada que el veiem. Per tant, un node amb k fills apareix $k + 1$ vegades al vector, i el vector conté $2n - 1$ elements.

Emmagatzemem dos valors al vector: l'identificador del node i la profunditat del node a l'arbre. El vector següent es correspon amb l'arbre anterior:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
node id	1	2	5	2	6	8	6	2	1	3	1	4	7	4	1
depth	1	2	3	2	3	4	3	2	1	2	1	2	3	2	1

Ara podem trobar l'avantpassat comú més baix dels nodes a i b trobant el node amb profunditat *mínima* entre els nodes a i b del vector. Per exemple, l'avantpassat comú més baix dels nodes 5 i 8 es pot trobar de la següent manera:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
node id	1	2	5	2	6	8	6	2	1	3	1	4	7	4	1
depth	1	2	3	2	3	4	3	2	1	2	1	2	3	2	1

↑

El node 5 es troba a la posició 2, el node 8 es troba a la posició 5 i el node de profunditat mínima entre les posicions 2...5 és el node 2 a la posició 3, i té profunditat 2. Així, l'avantpassat comú més baix dels nodes 5 i 8 és el node 2.

Per tant, per trobar l'avantpassat comú més baix de dos nodes n'hi ha prou amb processar una consulta d'interval mínim. Com que el vector és estàtic, podem processar consultes com aquesta en temps $O(1)$ després del preprocessament de temps $O(n \log n)$.

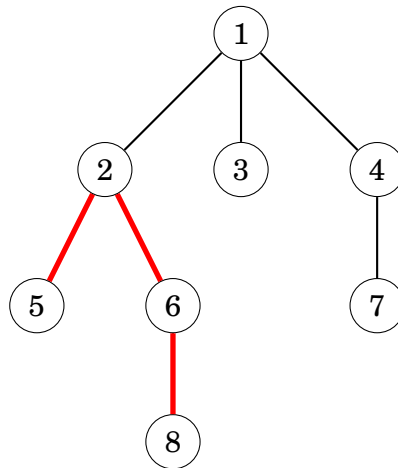
Distàncies dels nodes

La distància entre els nodes a i b és igual a la longitud del camí de a a b . Resulta que el problema de calcular la distància entre dos nodes es redueix a trobar el seu avantpassat comú més baix.

Primer, arrelem l'arbre de manera arbitrària. Després d'això, la distància dels nodes a i b es pot calcular mitjançant la fórmula

$$\text{depth}(a) + \text{depth}(b) - 2 \cdot \text{depth}(c),$$

on c és l'avantpassat comú més baix de a i b i $\text{depth}(s)$ indica la profunditat del node s . Per exemple, considereu la distància entre els nodes 5 i 8:



L'avantpassat comú més baix dels nodes 5 i 8 és el node 2. Les profunditats dels nodes són $\text{depth}(5) = 3$, $\text{depth}(8) = 4$ i $\text{depth}(2) = 2$, de manera que la distància entre els nodes 5 i 8 és $3 + 4 - 2 \cdot 2 = 3$.

18.4 Algorismes *offline*

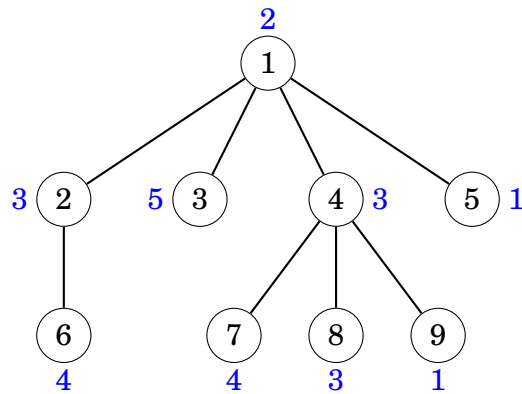
Fins ara, hem tractat els algorismes *online* de consultes d'arbres. Aquests algorismes són capaços de processar les consultes una rera l'altra de manera que cada consulta es respon abans de rebre la següent.

Tanmateix, en molts problemes, no és necessari fer-ho així. En aquesta secció ens centrem en els algorismes *offline*. Aquests algorismes reben un conjunt de consultes que es permés respondre en qualsevol ordre. Sovint és més fàcil dissenyar un algorisme *offline* que no pas un algorisme *online*.

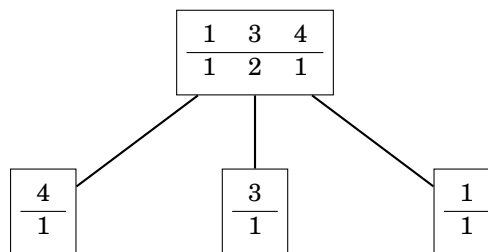
Fusionar estructures de dades

Un mètode per construir algorismes *offline* és fer un recorregut de l'arbre en profunditat i mantenir estructures de dades als nodes. Per cada node s , creem una estructura de dades $d[s]$ que es basa en les estructures de dades dels fills de s . Aleshores, fem servir aquesta estructura de dades per processar totes les consultes relacionades amb s .

Com a exemple, considereu el problema següent: donat un arbre on cada node té algun valor, processa consultes de la forma "troba el nombre de nodes amb valor x al subarbre del node s ". Per exemple, a l'arbre següent, el subarbre arrelat al node 4 conté dos nodes el valor dels quals és 3.



En aquest problema, podem fer servir mapes per respondre les consultes. Per exemple, els mapes per al node 4 i els seus fills són els següents:



Si creem una estructura de dades com aquesta per a cada node, podem processar fàcilment totes les consultes relacionades amb un node donat just després de crear l'estructura de dades pel node en qüestió. Per exemple, el mapa anterior ens indica que el subarbre arrelat al node 4 conté dos nodes el valor dels quals és 3.

Tanmateix, seria massa lent crear totes les estructures de dades des de zero. En canvi, per cada node s , creem una estructura de dades inicial $d[s]$ que només conté el valor de s . Després d'això, recorrem els fills u de s i fusionem $d[s]$ i totes les estructures de dades $d[u]$ ².

Per exemple, en l'arbre anterior, el mapa del node 4 es crea fusionant els mapes següents:



El primer mapa és l'estructura de dades inicial del node 4, i els altres tres mapes corresponen als nodes 7, 8 i 9.

La fusió al node s es pot fer de la següent manera: per cada fill u de s fusionem $d[s]$ i $d[u]$. Fem la fusió copiant el contingut de $d[u]$ sobre $d[s]$, però si la mida de $d[s]$ és més petita que la mida de $d[u]$, intercanviem (*swap*) els continguts dels

²(N. del T.) I, un cop finalitzada la fusió de $d[s]$, responem totes les consultes relatives al node s . Això ens obliga a ordenar les consultes en post-ordre.

mapes abans de fer la fusió. Fent això cada valor només es copia $O(\log n)$ vegades durant el recorregut d'arbre³, la qual cosa garanteix que l'algorisme és eficient.

Podem intercanviar el contingut de dues estructures de dades a i b de manera eficient fent servir:

```
swap(a, b);
```

El codi anterior sempre funciona en temps constant quan a i b són estructures de dades de la biblioteca estàndard de C++.

Avantpassats comuns més baixos

També hi ha un algorisme *offline* per processar un conjunt de consultes d'avantpassats comuns més baixos⁴. L'algorisme es basa en l'estructura *union-find* (vegeu el capítol 15.2), i un benefici és que és més fàcil d'implementar que els algorismes presentats anteriorment.

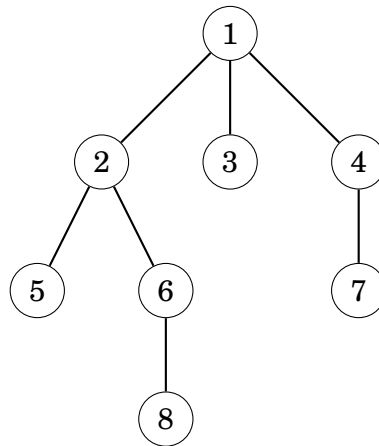
L'algorisme rep com a entrada un conjunt de parells de nodes i determina, per a cada parell, l'avantpassat comú més baix dels nodes. L'algorisme realitza una travessa de l'arbre a la profunditat i manté una estructura *union-find*, on inicialment cada node pertany a un conjunt separat. Per a cada conjunt emmagatzemem també el node més alt que pertany al conjunt.

Quan l'algorisme visita un node x , processa totes les consultes de la forma (x, y) on y ja ha estat visitat amb anterioritat, i respon que l'avantpassat comú més baix de x i y és el node més alt del conjunt associat a y . Un cop hem acabat de processar el node x l'algorisme uneix els conjunts de x i el seu pare.

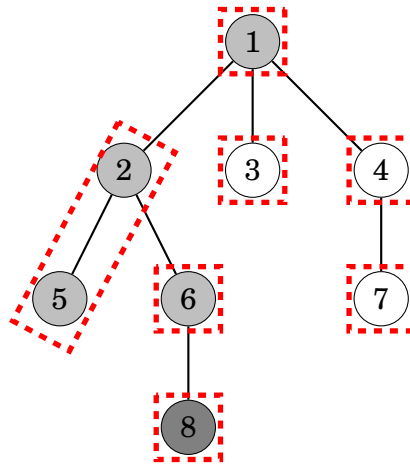
Per exemple, suposem que volem trobar els avantpassats comuns més baixos dels parells de nodes $(5, 8)$ i $(2, 7)$ a l'arbre següent:

³(N. del T.): Això no és cert: és possible construir contra-exemples d'alçada \sqrt{n} on un mapa es copia \sqrt{n} vegades. Tot i això, l'algorisme sí és eficient, pel següent motiu. Considerem un algorisme semblant al nostre però on en lloc de fer *swaps* en funció de la mida dels mapes $d[s]$ i $d[u]$, fem sempre un únic *swap* entre $d[s]$ i $d[u_{max}]$, on u_{max} és un fill de s amb subarbre de mida màxima, i els mapes $d[u]$ de la resta de fills es copien a sobre d'aquest. (Això es coneix com “*small to large*”). Aquest algorisme calcula els mateixos mapes que l'algorisme original, però fa igual o més feina en les fusions, perquè no fa servir el criteri òptim per estalviar-se còpies. Amb aquest canvi es compleix que tots els mapes es copien $O(\log n)$ vegades, ja que només copiem un mapa $d[u]$ sobre $d[s]$ si l'arbre arrelat a s és el doble de gran que l'arbre arrelat a u . Com que comencem amb n mapes de mida 1, el cost total de l'algorisme és $O(n \log n)$. I d'aquí deduïm que l'algorisme original també té cost total $O(n \log n)$.

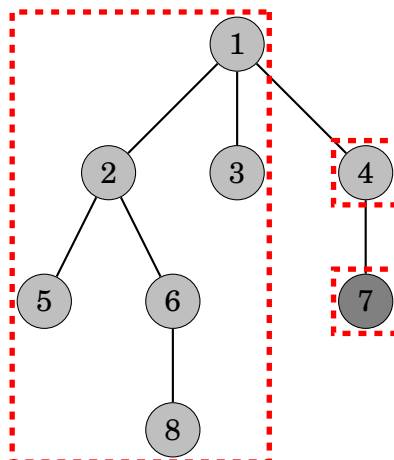
⁴Aquest algorisme va ser publicat per R. E. Tarjan el 1979 [65].



En els arbres següents, els nodes grisos denoten nodes visitats i els grups de nodes amb guions pertanyen al mateix conjunt. Quan l'algorisme visita el node 8, s'adona que el node 5 ja ha estat visitat i que el node més alt del conjunt corresponent és 2. Per tant, l'avantpassat comú més baix dels nodes 5 i 8 és 2:



Més tard, en visitar el node 7, l'algorisme determina que l'avantpassat comú més baix dels nodes 2 i 7 és 1:



Capítol 19

Camins i circuits

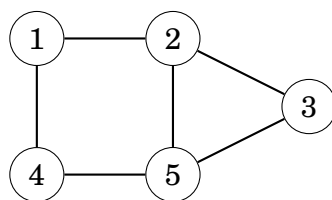
Aquest capítol parla de dos tipus de camins en els grafs:

- Un **Camí eulerià** és un camí que travessa cada aresta exactament una vegada.
- Un **Camí hamiltonià** és un camí que visita cada node exactament una vegada.

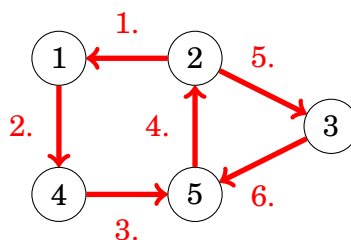
Tot i que els camins eulerians i hamiltonians semblen conceptes similars a primera vista, els problemes computacionals relacionats amb ells són molt diferents. Resulta que hi ha una regla senzilla que determina si un graf conté un camí eulerià, i també hi ha un algorisme eficient per trobar aquest camí si existeix. Al contrari, comprovar l'existència d'un camí hamiltonià és un problema NP-difícil, i no es coneix cap algorisme eficient per resoldre el problema.

19.1 Camins eulerians

Una **camí eulerià**¹ és un camí que passa exactament una vegada per cada aresta del graf. Per exemple, el graf

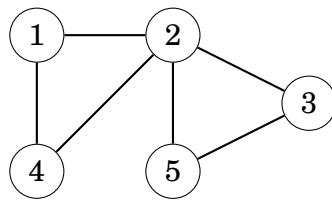


té un camí eulerià des del node 2 fins al node 5:

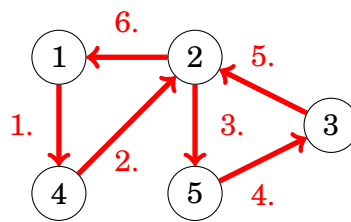


¹L. Euler va estudiar aquests camins el 1736 quan va resoldre el famós problema del pont de Königsberg. Aquest va ser el naixement de la teoria de grafs.

Un **circuit eulerià** és un camí eulerià que comença i acaba al mateix node. Per exemple, el graf



té un circuit eulerià que comença i acaba al node 1:



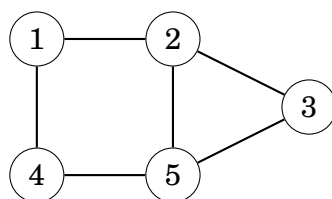
Existència

L'existència de camins i circuits eulerians depèn dels graus dels nodes. Un graf no dirigit té un camí eulerià exactament quan té un sol component connex i

- el grau de cada node és parell o
- el grau de dos nodes és senar, i el grau de tots els altres nodes és parell.

En el primer cas, cada camí eulerià és també un circuit eulerià. En el segon cas, els nodes de grau senar són els nodes inicials i finals d'un camí eulerià que no és un circuit eulerià ².

Per exemple, en el graf



els nodes 1, 3 i 4 tenen grau 2, i els nodes 2 i 5 tenen grau 3. Com que hi ha dos nodes de grau senar, hi ha un camí eulerià entre els nodes 2 i 5, però el graf no conté cap circuit eulerià.

En un graf dirigit ens fixem en els graus d'entrada i sortida dels nodes. Un graf dirigit conté un camí eulerià exactament quan té un sol component connex i

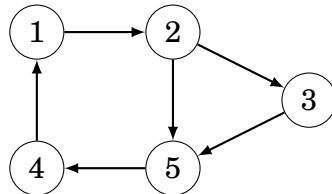
- en tots els nodes el grau d'entrada és igual al grau de sortida, o

²(N. del T.) És clar que, si un graf té un node de grau senar, no pot tenir un circuit eulerià, perquè els circuits eulerians tenen tots els nodes de grau parell. Allò sorprenent és que aquesta condició també sigui suficient.

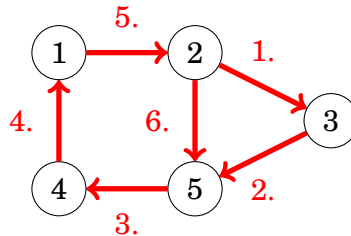
- hi ha un node on el grau d'entrada és un més que el grau de sortida, hi ha un altre node on el grau de sortida és un més que el grau d'entrada, i en tots els altres nodes el grau d'entrada és igual al grau de sortida.

En el primer cas, cada camí eulerià també és un circuit eulerià, i en el segon cas, el graf conté un camí eulerià que comença al node amb excés de grau de sortida i acaba al node amb excés de grau d'entrada.

Per exemple, al graf



els nodes 1, 3 i 4 tenen grau d'entrada i de sortida 1, el node 2 té grau d'entrada 1 i grau de sortida 2, i el node 5 té grau d'entrada 2 i grau de sortida 1. Per tant, el graf conté un camí eulerià des del node 2 fins al node 5:



Algorisme de Hierholzer

Algorisme de Hierholzer³ és un mètode eficient per construir un circuit eulerià. L'algorisme consta de diverses rondes, cadascuna de les quals afegeix noves arestes al circuit. Suposem que el graf conté un circuit eulerià; en cas contrari l'algorisme de Hierholzer no el pot trobar.

En primer lloc, l'algorisme construeix un circuit que conté algunes (no necessàriament totes) les arestes del graf. Després d'això, l'algorisme amplia el circuit pas a pas afegint-hi subcircuit. El procés continua fins que totes les arestes s'han afegit al circuit.

L'algorisme sempre amplia el circuit mitjançant un node x que pertany al circuit però que té una aresta de sortida que no pertany al circuit. L'algorisme construeix un nou camí des del node x només amb arestes que no pertanyen al circuit. Tard o d'hora, el camí tornarà al node x , creant un nou subcircuit que afegim al circuit original.

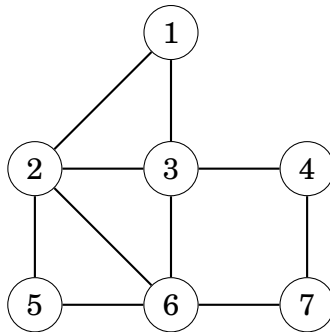
Si el graf només conté camins eulerians, també podem trobar-los amb l'algorisme de Hierholzer, afegint una aresta addicional al graf entre els dos nodes especials i eliminant-la després de construir el circuit. Per exemple, en un graf no dirigit, afegim l'aresta addicional entre els dos nodes de grau senar.

A continuació veurem com l'algorisme de Hierholzer construeix un circuit eulerià per a un graf no dirigit.

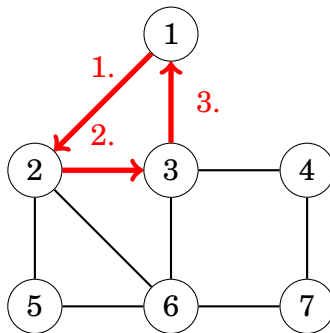
³L'algorisme es va publicar el 1873 després de la mort de Hierholzer [35].

Exemple

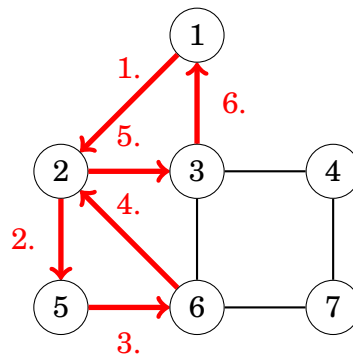
Considerem el graf següent:



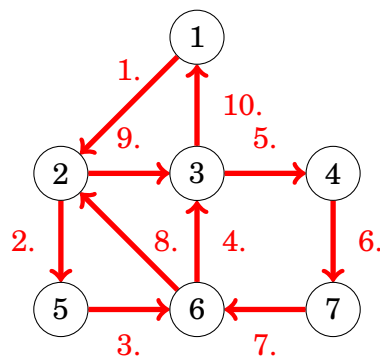
Suposem que l'algorisme primer crea un circuit que comença al node 1. Un circuit possible és $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$:



Després d'això, l'algorisme afegeix el subcircuit $2 \rightarrow 5 \rightarrow 6 \rightarrow 2$ al circuit:



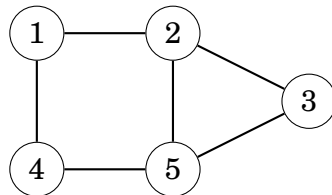
Finalment, l'algorisme afegeix el subcircuit $6 \rightarrow 3 \rightarrow 4 \rightarrow 7 \rightarrow 6$ al circuit:



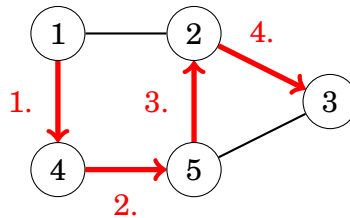
Ara totes les arestes estan incloses al circuit, així que hem construït amb èxit un circuit eulerià.

19.2 Camins de Hamilton

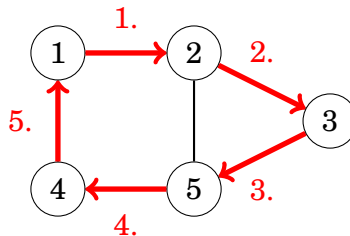
Un **camí hamiltonià** és un camí que visita cada node del graf exactament una vegada. Per exemple, el graf



conté un camí hamiltonià des del node 1 fins al node 3:



Si un camí hamiltonià comença i acaba al mateix node, s'anomena **circuit hamiltonià**. El graf anterior també té un circuit hamiltonià que comença i acaba al node 1:



Existència

No es coneix cap mètode eficient per provar si un graf conté un camí hamiltonià, i el problema és NP-difícil. Tot i així, en alguns casos especials, podem estar segurs que un graf conté un camí hamiltonià.

Una observació senzilla és que si el graf és complet, és a dir, si hi ha una aresta entre tots els parells de nodes, aleshores també conté un camí hamiltonià. També hi ha resultats més forts:

- **Teorema de Dirac:** Si el grau de cada node és almenys $n/2$, el graf conté un camí hamiltonià.
- **Teorema d'Ore:** Si la suma de graus de cada parell de nodes no adjacents és almenys n , el graf conté un camí hamiltonià.

Una propietat comuna en aquests teoremes i altres resultats és que garanteixen l'existència d'un camí hamiltonià si el graf té *un gran nombre* d'arestes. Això té sentit, perquè com més arestes conté el graf, més possibilitats hi ha de construir un camí hamiltonià.

Construcció

Com que no hi ha cap manera eficient de comprovar si existeix un camí hamiltonià, és clar que tampoc no hi ha cap mètode per construir el camí de manera eficient, perquè en cas contrari simplement intentariem construir el camí i mirariem si tenim èxit.

Una manera senzilla de buscar un camí hamiltonià és amb un algorisme de marxa enrera (*backtracking*) que passi per totes les maneres possibles de construir el camí. La complexitat temporal d'aquest algorisme és com a molt $O(n!)$, perquè hi ha $n!$ maneres diferents d'ordenar n nodes.

Una solució més eficient fa servir programació dinàmica (vegeu el capítol 10.5). La idea és calcular els valors d'una funció $\text{possible}(S, x)$, on S és un subconjunt de nodes i x és un dels nodes. La funció indica si hi ha un camí hamiltonià que visita els nodes de S i acaba al node x . És possible implementar aquesta solució en temps $O(2^n n^2)$.

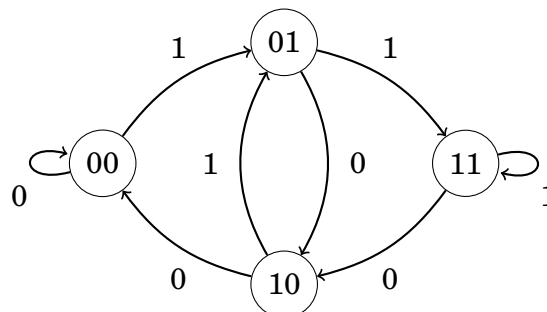
19.3 Seqüències de De Bruijn

Una **seqüència de De Bruijn** és una cadena que conté cada subcadena de longitud n exactament una vegada com a subcadena, per a un alfabet fix de k caràcters. La longitud d'aquesta cadena és de $k^n + n - 1$ caràcters. Per exemple, quan $n = 3$ i $k = 2$, un exemple de seqüència de De Bruijn és

0001011100.

Les subcadenaes d'aquesta cadena són totes les combinacions de tres bits: 000, 001, 010, 011, 100, 101, 110 i 111.

Resulta que cada seqüència de De Bruijn es correspon amb un camí eulerià en un graf. La idea és construir un graf on cada node contingui una cadena de $n - 1$ caràcters i cada aresta afegeix un caràcter a la cadena. El graf següent correspon a l'escenari anterior:



Un camí eulerià en aquest graf correspon a una cadena que conté totes les cadenes de longitud n . La cadena conté els caràcters del node inicial i tots els caràcters de les arestes. El node inicial té $n - 1$ caràcters i hi ha k^n caràcters a les arestes, de manera que la longitud de la cadena és $k^n + n - 1$.

19.4 Ruta del cavall

Una **ruta del cavall** és una seqüència de moviments d'un cavall en un tauler d'escacs $n \times n$, seguint les regles dels escacs, de manera que el cavall visita cada casella exactament una vegada. La ruta del cavall és *tancada* si el cavall torna a la casella inicial i, en cas contrari, és *oberta*.

Per exemple, aquí hi ha una ruta oberta del cavall en un tauler de 5×5 :

1	4	11	16	25
12	17	2	5	10
3	20	7	24	15
18	13	22	9	6
21	8	19	14	23

Una ruta del cavall es correspon a un camí hamiltonià en un graf els nodes del qual representen les caselles del tauler, i dos nodes estan connectats amb una aresta si un cavall pot moure's entre elles segons les regles dels escacs.

Una manera natural de construir una ruta del cavall és amb *backtracking*. La cerca es pot fer més eficient fent servir *heurístiques* que intenten guiar el cavall de manera que trobi ràpidament una ruta sencera.

Regla de Warnsdorf

La regla de Warnsdorf és una heurística senzilla i eficaç per trobar una ruta del cavall⁴. Utilitzant la regla, és possible construir de manera eficient una ruta fins i tot en un tauler gran. La idea és moure el cavall a una casella on el nombre de moviments possibles sigui el més *petit* possible.

Per exemple, en la situació següent, hi ha cinc caselles possibles a les quals es pot moure el cavall (quadrats *a*...*e*):

1				<i>a</i>
		2		
<i>b</i>				<i>e</i>
	<i>c</i>		<i>d</i>	

⁴Aquesta heurística es va proposar al llibre de Warnsdorf [69] l'any 1823. També hi ha algorismes polinomials per trobar les rutes del cavall [52], però són més complicats.

En aquesta situació, la regla de Warnsdorf mou el cavall al quadrat a , perquè després d'aquesta elecció només hi ha un únic moviment possible, mentre que les altres opcions mouen el cavall a caselles on hi ha tres moviments possibles.

Capítol 20

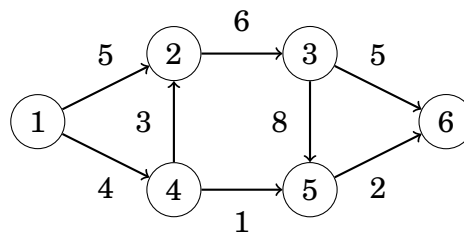
Fluxos i talls

En aquest capítol considerem els dos problemes següents:

- **Trobar un flux màxim:** Quina és la quantitat màxima de flux que podem enviar d'un node a un altre node?
- **Trobar un tall mínim:** Què és el pes mínim d'un conjunt de arestes que separa dos nodes del graf?

L'entrada d'aquests dos problemes és un graf dirigit i amb pesos que conté dos nodes especials: el node origen (font, *source*) és un node sense arestes entrants, i el node destí (aiguera, *sink*) és un node sense arestes sortints.

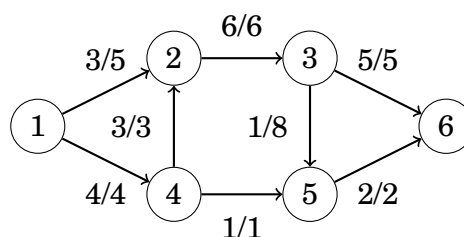
Com a exemple, utilitzarem el graf següent on el node 1 és el node origen i el node 6 és el node destí:



Flux màxim

En el problema del **flux màxim**, la nostra tasca és enviar el màxim flux possible des del node origen al destí. El pes de cada aresta és la màxima capacitat de flux pot passar per l'aresta. Per cada node intermedi s'ha de complir que el flux entrant i sortint és igual.

Per exemple, el flux màxim del graf d'exemple és 7. La imatge següent mostra com podem dirigir el flux:

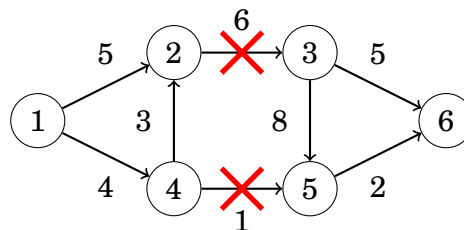


La notació v/k significa que un flux de v unitats s'encamina a través d'una aresta la capacitat de la qual és de k unitats. La mida del flux és de 7, perquè l'origen envia $3 + 4$ unitats de flux i el destí rep $5 + 2$ unitats de flux. És fàcil veure que aquest flux és màxim, perquè la capacitat total de les arestes que condueixen al destí és 7.

Tall mínim

En el problema del **tall mínim**, la nostra tasca és eliminar un conjunt d'arestes del graf de manera que no hi hagi cap camí des de l'origen fins al destí, i que el pes total (o mida) de les arestes eliminades sigui mínim.

La mida mínima d'un tall al graf d'exemple és 7. N'hi ha prou amb eliminar les arestes $2 \rightarrow 3$ i $4 \rightarrow 5$:



Després d'eliminar les arestes, no hi haurà camí des de l'origen al destí. La mida del tall és 7, perquè els pesos de les arestes eliminades són 6 i 1. El tall és mínim, perquè no hi ha cap manera vàlida d'eliminar arestes del graf de manera que el seu pes total sigui inferior a 7.

En l'exemple anterior la mida màxima d'un flux i la mida mínima d'un tall són iguals, però no és cap coincidència. Resulta que un flux màxim i un tall mínim són *sempre* iguals, de manera que els conceptes són dues cares de la mateixa moneda.

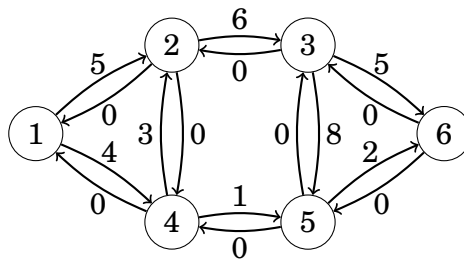
A continuació, parlarem de l'algorisme Ford-Fulkerson que es pot utilitzar per trobar el flux màxim i el tall mínim d'un graf. L'algorisme també ens ajuda a entendre *per què* són igualment grans.

20.1 Algorisme de Ford-Fulkerson

L'algorisme de **Ford-Fulkerson** [25] troba el flux màxim d'un graf. L'algorisme comença amb un flux buit, i a cada pas troba un camí des de l'origen fins al destí que genera més flux. Finalment, quan l'algorisme ja no pot augmentar el flux, ha trobat el flux màxim.

L'algorisme utilitza una representació especial del graf on cada aresta original té una aresta inversa en l'altra direcció. El pes de cada aresta indica quant més flux podríem afegir-hi. Al principi de l'algorisme, el pes de cada aresta original és igual a la capacitat de la aresta i el pes de cada aresta inversa és zero.

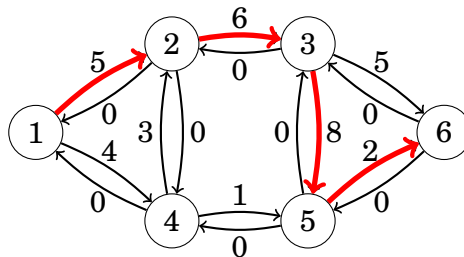
La nova representació del graf d'exemple és la següent:



Descripció de l'algorisme

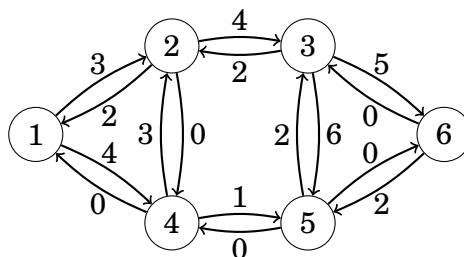
L'algorisme de Ford-Fulkerson consta de diverses rondes. A cada ronda, l'algorisme troba un camí des de l'origen fins al destí de manera que cada aresta del camí tingui un pes positiu. Si hi ha més d'un camí possible disponible, triem un qualsevol.

Per exemple, suposem que hem triat el camí següent:



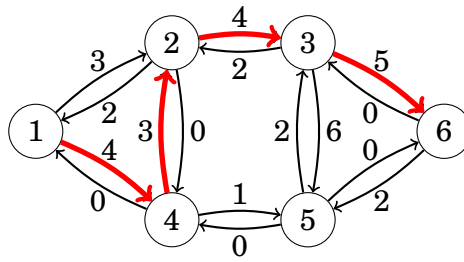
Després de triar el camí, el flux augmenta en x unitats, on x és el pes de l'aresta més petita del camí. A més, el pes de cada aresta del camí disminueix en x i el pes de cada aresta inversa augmenta en x .

En el camí anterior, els pesos de les arestes són 5, 6, 8 i 2. El pes més petit és 2, de manera que el flux augmenta en 2 i el nou graf és el següent:



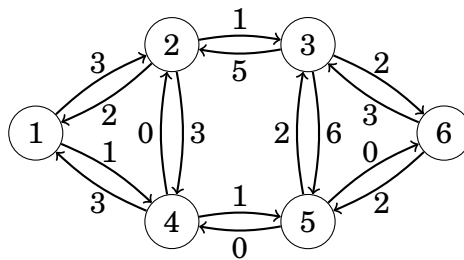
La idea és que augmentar el flux disminueix la quantitat de flux que pot passar per les arestes en el futur. D'altra banda, és possible cancel·lar l'augment fent servir les arestes inverses del graf si resulta que seria beneficiós dirigir el flux d'una altra manera.

L'algorisme augmenta el flux sempre que hi hagi un camí des de l'origen fins al destí a través de arestes de pes positiu. En el nostre exemple, considerem el següent camí:

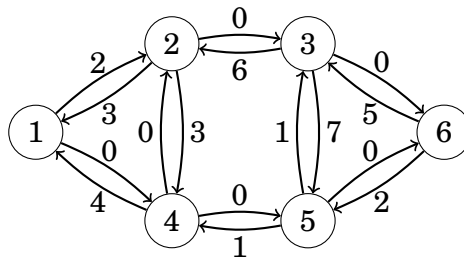


El pes mínim de la aresta d'aquest camí és 3, de manera que el camí augmenta el flux en 3 i el flux total després de processar el camí és de 5.

El nou graf és:



Encara necessitem dues rondes més abans d'arribar al flux màxim. Per exemple, podem triar els camins $1 \rightarrow 2 \rightarrow 3 \rightarrow 6$ i $1 \rightarrow 4 \rightarrow 5 \rightarrow 3 \rightarrow 6$. Ambdós camins augmenten el flux en 1, i el graf final és el següent:



Ja no és possible augmentar el flux, perquè no hi ha cap camí des de l'origen fins al destí amb pesos positius. Per tant, l'algorisme s'acaba i el flux màxim és 7.

Trobar els camins

L'algorisme de Ford-Fulkerson no especifica com hem de triar els camins que augmenten el flux. En qualsevol cas, l'algorisme acabarà tard o d'hora i trobarà correctament el flux màxim. No obstant això, l'eficiència de l'algorisme depèn de la manera com s'escullen els camins.

Una manera senzilla de trobar els camins és fer servir la cerca en profunditat. Normalment, això funciona bé, però en el pitjor dels casos cada camí només augmenta el flux en 1 i l'algorisme és lent. Afortunadament, podem evitar aquesta situació utilitzant una de les tècniques següents:

L'algorisme d'**Edmonds-Karp** [18] tria cada camí de manera que el nombre d'arestes del camí sigui el més petit possible. Això és fàcil de fer amb la cerca en

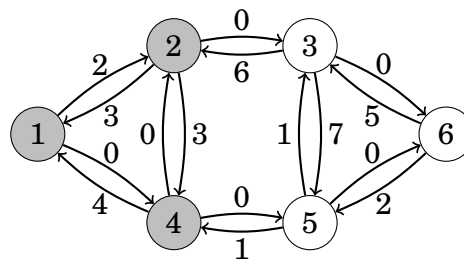
amplada. Es pot demostrar que això garanteix que el flux augmenta ràpidament, i la complexitat temporal de l'algorisme és $O(m^2n)$.

L'algorisme de reescalat (*scaling algorithm*) [2] utilitza la cerca en profunditat per trobar camins on cada pes de la aresta sigui almenys un valor llinar. Inicialment, el valor llinar és un nombre gran, per exemple, la suma de tots els pesos de les arestes del graf. Sempre que no es pot trobar un camí, el valor llinar es divideix per 2. La complexitat temporal de l'algorisme és $O(m^2 \log c)$, on c és el valor llinar inicial.

A la pràctica, l'algorisme de reescalat és més fàcil d'implementar, perquè la cerca en profunditat es pot fer servir per trobar camins. Tots dos algorismes són prou eficients per als problemes que solen aparèixer als concursos de programació.

Talls mínims

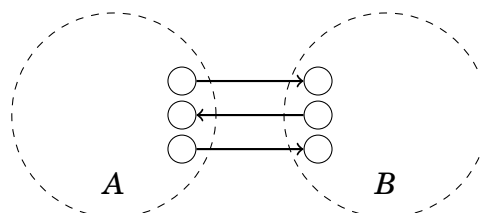
Resulta que un cop l'algorisme Ford-Fulkerson ha trobat un flux màxim, també ha determinat un tall mínim. Sigui A el conjunt de nodes als quals es pot arribar des del node origen fent servir arestes de pes positiu, i B el conjunt restant. Al graf d'exemple, A conté els nodes 1, 2 i 4:



Ara el tall mínim consisteix en les arestes del graf original que comencen en algun node a A , acaben en algun node de B , i la seva capacitat s'utilitza plenament en el flux màxim. Al graf anterior, aquestes arestes són $2 \rightarrow 3$ i $4 \rightarrow 5$, que corresponen al tall mínim $6 + 1 = 7$.

Per què el flux produït per l'algorisme és màxim i per què el tall és mínim? El motiu és que un graf no pot contenir un flux la mida del qual sigui més gran que el pes de qualsevol tall del graf. Per tant, sempre que un flux i un tall són igualment grans, són un flux màxim i un tall mínim.

Considerem ara qualsevol tall del graf de manera que l'origen pertanyi a A , el destí pertanyi a B i hi hagi algunes arestes entre els conjunts:



La mida del tall és la suma de les arestes que van de A a B . Aquest és un límit superior per al flux del graf, perquè el flux ha de procedir de A a B . Així, la mida d'un flux màxim és menor o igual a la mida de qualsevol tall del graf.

D'altra banda, l'algorisme de Ford-Fulkerson produeix un flux la mida del qual és *exactament* tan gran com la mida d'un tall al graf. Així, el flux ha de ser un flux màxim i el tall ha de ser un tall mínim.

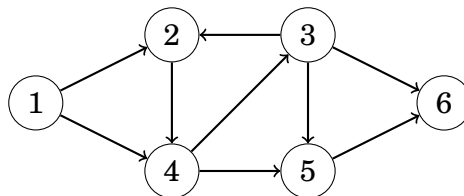
20.2 Camins discontinus

Molts problemes de grafes es poden resoldre reduint-los al problema del flux màxim. Per exemple, considerem el següent problema: donat un graf dirigit amb un origen i un destí, troba el nombre màxim de camins disjunts des de l'origen fins al destí.

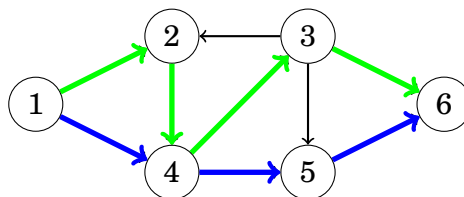
Camins aresta-disjunts

Primer tractem el cas de **camins aresta-disjunts** des de l'origen al destí. Això vol dir que hem de construir conjunts de camins de manera que cada aresta aparegui com a màxim un cop.

Per exemple, considereu el graf següent:



En aquest graf, el nombre màxim de camins aresta-disjunts és 2. Podem triar els camins $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 6$ i $1 \rightarrow 4 \rightarrow 5 \rightarrow 6$ de la següent manera:

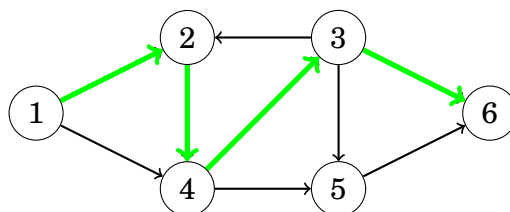


Resulta que el nombre màxim de camins aresta-disjunts és igual al flux màxim del graf, suposant que la capacitat de cada aresta és 1. Un cop trobat el flux màxim, els camins disjunts de aresta es poden trobar de manera *greedy* buscant camins des de l'origen fins al destí.

Camins node-disjunts

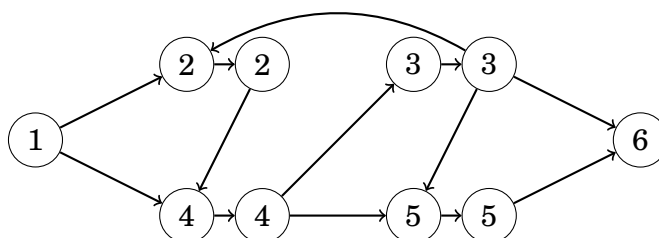
Considerem ara un altre problema: trobar el nombre màxim de **camins node-disjunts** des de l'origen fins al destí. En aquest problema, tots els nodes, excepte l'origen i el destí, poden aparèixer com a màxim un cop al camí. El nombre de camins disjunts entre nodes és menor o igual al nombre de camins aresta-disjunts.

Per exemple, al graf anterior, el nombre màxim de camins node-disjunts és 1:

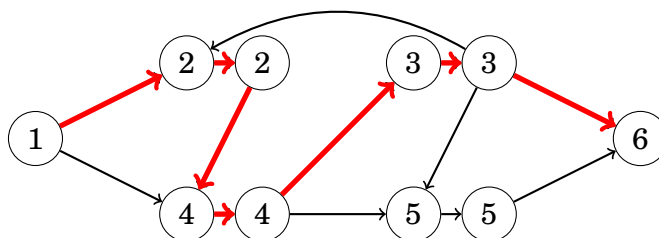


També podem reduir aquest problema al problema de flux màxim. Com que cada node pot aparèixer com a màxim un cop en els camins, hem de limitar el flux que passa pels nodes. Un mètode estàndard és dividir cada node en dos nodes de manera que el primer node rebi les arestes d'entrada del node original, el segon node tingui les arestes de sortida del node original i hi hagi una nova aresta des del primer node fins al segon node.

En el nostre exemple, el graf esdevé:



El flux màxim per al graf és:



Així, el nombre màxim de camins disjunts entre nodes des de l'origen al destí és 1.

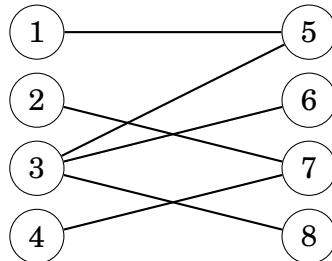
20.3 Emparellaments màxims

El problema dels **emparellaments màxims** (*maximum matchings*) consisteix en trobar un conjunt de mida màxima de parelles de nodes en un graf no dirigit de manera que cada parella estigui connectada amb una aresta i cada node pertanyi com a molt a un parella.

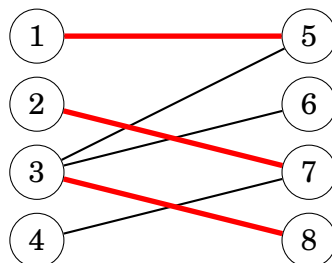
Hi ha algorismes polinomials per trobar emparellaments màxims en grafs generals [17], però aquests algorismes són complexos i rarament es veuen en concursos de programació. Tanmateix, en els grafs bipartits, el problema d'emparellaments màxims és molt més fàcil de resoldre, perquè es redueix a trobar un flux màxim.

Trobar emparellaments màxims

Els nodes d'un graf bipartit sempre es poden dividir en dos grups de manera que totes les arestes del graf van del grup esquerre al grup dret. Per exemple, al graf bipartit següent, els grups són $\{1, 2, 3, 4\}$ i $\{5, 6, 7, 8\}$.

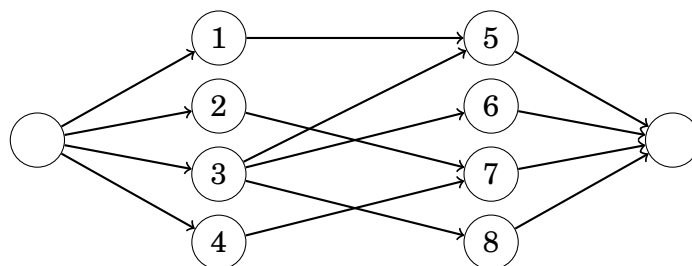


La mida d'un emparellament màxim és 3:

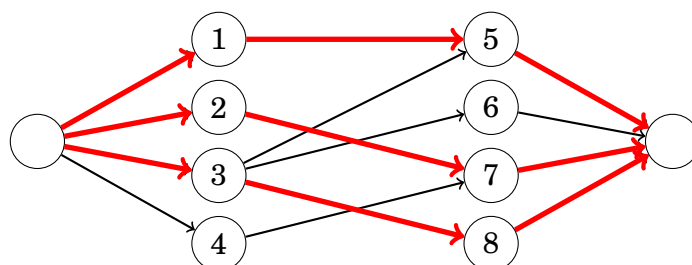


El problema de l'emparellament màxim bipartit es redueix al problema del flux màxim afegint dos nodes nous al graf: un d'origen i un de destí. També afegim arestes de l'origen a cada node esquerre i de cada node dret al destí. Amb això, la mida d'un flux màxim al graf és igual a la mida d'un emparellament màxim al graf original.

Per exemple, la reducció del graf anterior és:



El flux màxim d'aquest graf és:

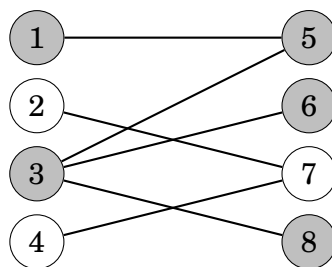


Teorema de Hall

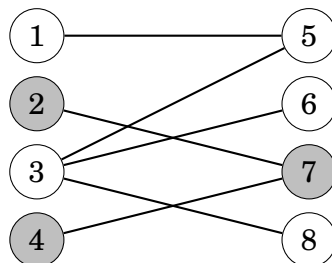
Teorema de Hall es pot fer servir per esbrinar si un graf bipartit té una concordança que conté tots els nodes esquerre o dret. Si el nombre de nodes esquerres i drets és el mateix, el teorema de Hall ens diu si és possible construir un **emparellament perfecte** que contingui tots els nodes del graf.

Suposem que volem trobar un emparellament que contingui tots els nodes esquerre. Sigui X qualsevol conjunt de nodes esquerre i sigui $f(X)$ el conjunt dels seus veïns. Segons el teorema de Hall, un emparellament que conté tots els nodes esquerre existeix si i només si quan, per a cada X , es compleix la condició $|X| \leq |f(X)|$.

Veiem el teorema de Hall al graf d'exemple. Per a $X = \{1, 3\}$ es compleix $f(X) = \{5, 6, 8\}$:



La condició del teorema de Hall es compleix, perquè $|X| = 2$ i $|f(X)| = 3$. Per a $X = \{2, 4\}$, en canvi, es compleix $f(X) = \{7\}$:



En aquest cas, $|X| = 2$ i $|f(X)| = 1$, de manera que la condició del teorema de Hall no es compleix. Això vol dir que no és possible formar un emparellament perfecte per al graf. Aquest resultat no és sorprenent, perquè ja sabíem que l'emparellament màxim del graf era 3 i no 4.

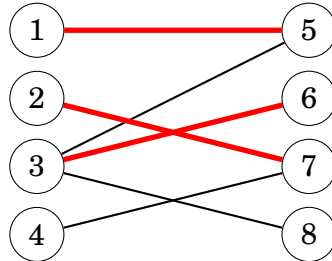
Si la condició del teorema de Hall no es compleix, el conjunt X proporciona una explicació de *per què* no podem formar aquest emparellament. Com que X conté més nodes que $f(X)$, no hi ha prou parelles per a tots els nodes de X . Per exemple, al graf anterior, els dos nodes 2 i 4 haurien d'estar connectats amb el node 7, cosa que no és possible.

Teorema de Kőnig

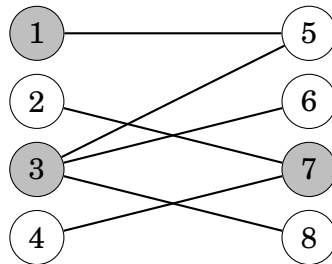
Una **cobertura de nodes mínim** d'un graf és un conjunt mínim de nodes de manera que cada aresta del graf té almenys un extrem al conjunt. En un graf general, trobar una cobertura de nodes mínim és un problema NP-difícil.

Tanmateix, si el graf és bipartit, el **teorema de König** ens diu que la mida d'una cobertura de nodes mínim i la mida d'un emparellament màxim sempre coincideixen. Així, podem calcular la mida d'una cobertura de nodes mínim trobant un flux màxim.

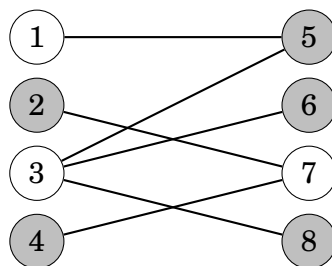
Considerem el graf següent amb un emparellament màxim de mida 3:



El teorema de König ens diu que la mida d'una cobertura de nodes mínim també és 3. Aquesta cobertura es pot construir com segueix:



Els nodes que *no* pertanyen a una cobertura de nodes mínim formen un **conjunt independent màxim**. Aquest és el conjunt de nodes més gran possible de manera que no hi ha dos nodes del conjunt connectats amb una aresta. Una vegada més, trobar un conjunt màxim independent en un graf general és un problema NP-difícil, però en un graf bipartit podem utilitzar el teorema de König per resoldre el problema de manera eficient. En el graf d'exemple, el conjunt independent màxim és el següent:

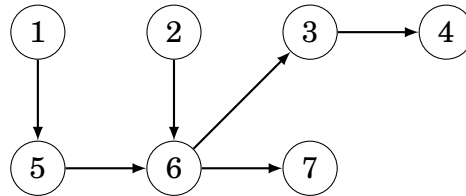


20.4 Cobertura de camins

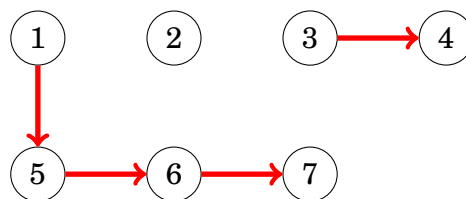
Una **cobertura de camins** és un conjunt de camins en un graf de manera que cada node del graf pertany a almenys un camí. Resulta que, en els grafs acíclics dirigits, podem reduir el problema de trobar una cobertura de camins mínima al problema de trobar un flux màxim en un altre graf.

Cobertura de camins node-disjunts

En una **cobertura de camins node-disjunts**, cada node pertany exactament a un camí. Per exemple, considereu el graf següent:



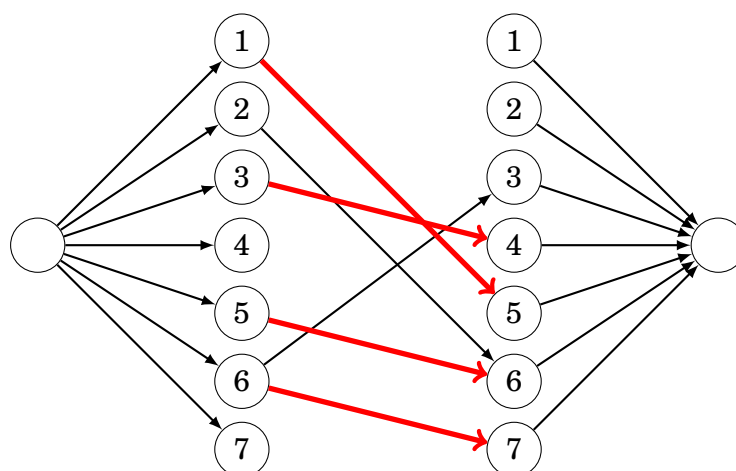
Una cobertura de camins mínima en aquest graf consta de tres camins. Per exemple, podem triar els camins següents:



Tingueu en compte que un dels camins només conté el node 2, perquè acceptem que un camí no contingui cap aresta.

Podem trobar una cobertura mínima de camins node-disjunt construint un *graf d'emparellaments* on cada node del graf original està representat per dos nodes: un node esquerre i un node dret. Hi ha una aresta des d'un node esquerre fins a un node dret si existeix aquesta aresta al graf original. A més, el graf d'emparellaments conté un origen i un destí, i hi ha arestes des de l'origen fins a tots els nodes esquerres i des de tots els nodes drets fins al destí.

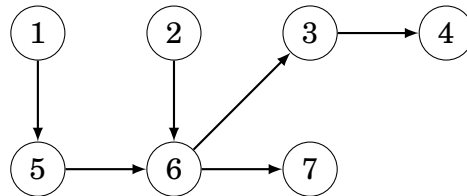
Un emparellament màxim en el graf resultant correspon a una cobertura mínima de camins node-disjunts en el graf original. Per exemple, el graf d'emparellaments següent conté un emparellament màxim de mida 4:



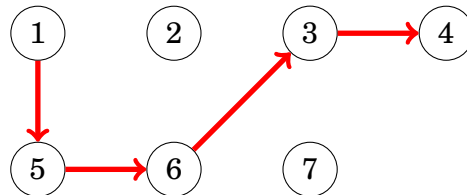
Cada aresta de l'emparellament màxim es correspon amb una aresta en la cobertura mínima de camins node-disjunts entre nodes del graf original. Així, la mida de la cobertura mínima de camins node-disjunt és $n - c$, on n és el nombre de nodes del graf original i c és la mida de l'emparellament màxim.

Cobertura general de camins

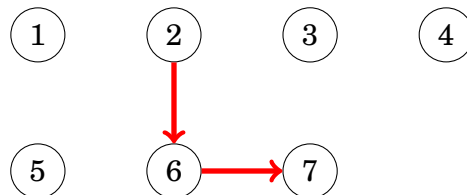
Una **cobertura general de camins** és una cobertura de camins on un node pot pertànyer a més d'un camí. Una cobertura general de camins mínima pot ser més petita que una cobertura de camins node-disjunts mínima, perquè un node es pot fer servir diverses vegades en els camins. Considereu de nou el graf següent:



La cobertura general de camins mínima d'aquest graf consta de dos camins. Per exemple, el primer camí pot ser el següent:

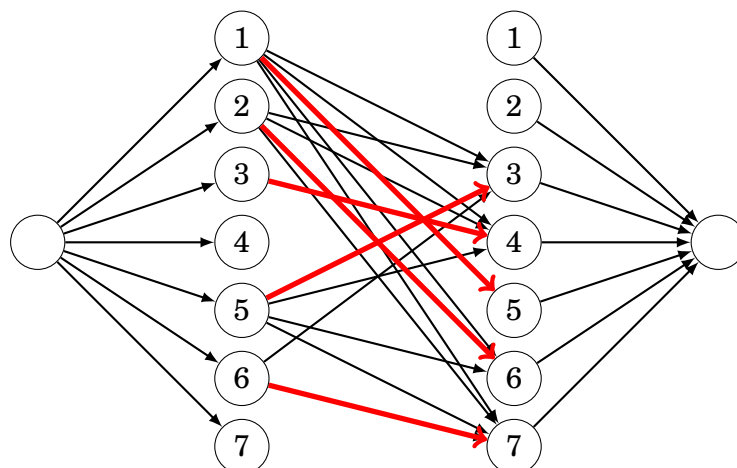


I el segon camí pot ser aquest:



Una cobertura general de camins mínima es pot trobar gairebé com una cobertura de camins node-disjunts mínima. N'hi ha prou amb afegir algunes arestes noves al graf d'emparellaments perquè hi hagi una aresta $a \rightarrow b$ sempre quan hi hagi un camí de a a b al graf original (possiblement a través de diverses arestes).

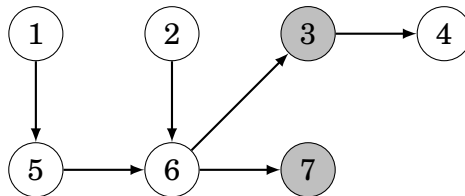
El graf d'emparellaments pel graf anterior és el següent:



Teorema de Dilworth

Una **anticadena** és un conjunt de nodes d'un graf de manera que no hi ha cap camí des de cap node a un altre node fent servir les arestes del graf. El **teorema de Dilworth** estableix que en un graf acíclic dirigit, la mida d'una cobertura general de camins mínima és igual a la mida d'una anticadena màxima.

Per exemple, els nodes 3 i 7 formen una anticadena al graf següent:



Aquesta és una anticadena màxima, perquè no és possible construir cap anticadena que contingui tres nodes. Hem vist abans que la mida d'una cobertura general de camins mínima d'aquest graf consta de dos camins.

Part III

Temes avançats

Capítol 21

Teoria de nombres

La **teoria de nombres** és una branca de les matemàtiques que estudia els nombres enters. La teoria dels nombres és un camp fascinant, perquè moltes qüestions que impliquen nombres enters són molt difícils de resoldre encara que a primera vista semblin senzilles.

Com a exemple, considereu l'equació següent:

$$x^3 + y^3 + z^3 = 33$$

És fàcil trobar tres nombres reals x , y i z que compleixin l'equació. Per exemple, podem triar

$$\begin{aligned}x &= 3, \\y &= \sqrt[3]{3}, \\z &= \sqrt[3]{3}.\end{aligned}$$

Tanmateix, és un problema obert en teoria de nombres si hi ha tres *enters* x , y i z que satisfacin l'equació [6].

En aquest capítol, ens centrarem en els conceptes bàsics i algorismes de la teoria dels nombres. Al llarg del capítol, assumirem que tots els nombres són enters si no s'indica el contrari.

21.1 Nombres primers i factors

Un nombre a s'anomena **factor** o **divisor** d'un nombre b si a divideix b . Si a és un factor de b , escrivim $a \mid b$, i en cas contrari escrivim $a \nmid b$. Per exemple, els factors de 24 són 1, 2, 3, 4, 6, 8, 12 i 24.

Un nombre $n > 1$ és un nombre **primer** si els seus únics factors positius són 1 i n . Per exemple, 7, 19 i 41 són primers, però 35 no és primer, perquè $5 \cdot 7 = 35$. Per a cada nombre $n > 1$, hi ha una **factorització en nombres primers** única

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k},$$

on p_1, p_2, \dots, p_k són nombres primers diferents i $\alpha_1, \alpha_2, \dots, \alpha_k$ són nombres positius. Per exemple, la factorització en nombres primers per a 84 és

$$84 = 2^2 \cdot 3^1 \cdot 7^1.$$

El **nombre de factors** d'un nombre n és

$$\tau(n) = \prod_{i=1}^k (\alpha_i + 1),$$

perquè per a cada p_i primer, hi ha $\alpha_i + 1$ maneres de triar quantes vegades apareix en el factor. Per exemple, el nombre de factors de 84 és $\tau(84) = 3 \cdot 2 \cdot 2 = 12$. Els factors són 1, 2, 3, 4, 6, 7, 12, 14, 21, 28, 42 i 84.

La **suma de factors** de n és

$$\sigma(n) = \prod_{i=1}^k (1 + p_i + \dots + p_i^{\alpha_i}) = \prod_{i=1}^k \frac{p_i^{\alpha_i+1} - 1}{p_i - 1},$$

on aquesta última fórmula es basa en la fórmula de la progressió geomètrica. Per exemple, la suma de factors de 84 és

$$\sigma(84) = \frac{2^3 - 1}{2 - 1} \cdot \frac{3^2 - 1}{3 - 1} \cdot \frac{7^2 - 1}{7 - 1} = 7 \cdot 4 \cdot 8 = 224.$$

El **producte de factors** de n és

$$\mu(n) = n^{\tau(n)/2},$$

perquè podem formar $\tau(n)/2$ parells a partir dels factors, cadascun amb el producte n . Per exemple, els factors de 84 produeixen els parells $1 \cdot 84$, $2 \cdot 42$, $3 \cdot 28$, etc., i el producte dels factors és $\mu(84) = 84^6 = 351298031616$.

Un nombre n s'anomena **perfecte** si $n = \sigma(n) - n$, és a dir, n és igual a la suma dels seus factors entre 1 i $n - 1$. Per exemple, 28 és un nombre perfecte, perquè $28 = 1 + 2 + 4 + 7 + 14$.

Infinitos nombres primers

És fàcil demostrar que hi ha una quantitat infinita de nombres primers. Si el nombre de primers fos finit, podríem construir un conjunt $P = \{p_1, p_2, \dots, p_n\}$ que contingués tots els primers. Per exemple, $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, etc. Tanmateix, utilitzant P , podríem formar un nou nombre primer

$$p_1 p_2 \cdots p_n + 1$$

més gran que tots els elements de P . Això és una contradicció, i per tant hi ha infinits nombres primers.

Densitat dels nombres primers

La densitat de nombres primers és la freqüència amb que els nombres primers apareixen entre tots els nombres. Sigui $\pi(n)$ la quantitat de nombres primers entre 1 i n . Per exemple, $\pi(10) = 4$, perquè hi ha 4 nombres primers entre 1 i 10: 2, 3, 5 i 7.

És possible demostrar que

$$\pi(n) \approx \frac{n}{\ln n},$$

el que vol dir que els nombres primers són força freqüents. Per exemple, la quantitat de nombres primers entre 1 i 10^6 és $\pi(10^6) = 78498$ i $10^6 / \ln 10^6 \approx 72382$.

Conjectures

Hi ha moltes *conjectures* que tenen a veure amb nombres primers. La majoria de la gent pensa que les conjectures són certes, però ningú les ha pogut demostrar. Per exemple, les conjectures següents són famoses:

- **Conjectura de Goldbach:** Cada enter parell $n > 2$ es pot representar com a suma $n = a + b$ de manera que tant a com b són primers.
- **Conjectura dels nombres primers bessons:** Hi ha un nombre infinit de parelles de la forma $\{p, p + 2\}$, on tant p com $p + 2$ són primers.
- **La conjectura de Legendre:** Sempre hi ha un nombre primer entre els nombres n^2 i $(n + 1)^2$, on n és qualsevol nombre enter positiu.

Algorismes bàsics

Si un nombre n no és primer, es pot representar com a producte $a \cdot b$, on $a \leq \sqrt{n}$ o $b \leq \sqrt{n}$, de manera que certament té un factor entre 2 i $\lfloor \sqrt{n} \rfloor$. Amb aquesta observació, podem comprovar si un nombre és primer i trobar la factorització en primers d'un nombre en temps $O(\sqrt{n})$.

La següent funció `prime` verifica si el nombre donat n és primer. La funció intenta dividir n per tots els nombres entre 2 i $\lfloor \sqrt{n} \rfloor$, i si cap d'ells divideix n , llavors n és primer.

```
bool prime(int n) {
    if (n < 2) return false;
    for (int x = 2; x*x <= n; x++) {
        if (n%x == 0) return false;
    }
    return true;
}
```

La funció `factors` següent construeix un vector que conté la factorització en nombres primers de n . La funció divideix n pels seus factors primers i els afegeix al vector. El procés acaba quan el nombre resultant n no té factors entre 2 i $\lfloor \sqrt{n} \rfloor$. Si $n > 1$, és un nombre primer i esdevé l'últim factor.

```
vector<int> factors(int n) {
    vector<int> f;
    for (int x = 2; x*x <= n; x++) {
        while (n%x == 0) {
            f.push_back(x);
            n /= x;
        }
    }
    if (n > 1) f.push_back(n);
    return f;
}
```

Tingueu en compte que cada factor primer apareix al vector tantes vegades com divideix el nombre. Per exemple, $24 = 2^3 \cdot 3$, de manera que el resultat de la funció és $[2, 2, 2, 3]$.

Sedàs d'Eratòstenes

El **sedàs d'Eratòstenes** és un algorisme de preprocessament que construeix un vector mitjançant el qual podem comprovar de manera eficient si un nombre determinat entre $2 \dots n$ és primer i, si no ho és, troba un factor primer del nombre.

L'algorisme crea un vector *sieve*, del qual farem servir les posicions $2, 3, \dots, n$. El valor $\text{sieve}[k] = 0$ vol dir que k és primer, i el valor $\text{sieve}[k] \neq 0$ vol dir que k no és primer i que $\text{sieve}[k]$ és un dels seus factors primers.

L'algorisme itera els nombres $2 \dots n$ un per un. Sempre que troba un nou nombre primer x , l'algorisme enregistra que els múltiples de x ($2x, 3x, 4x, \dots$) no són primers, ja que x és un divisor.

Per exemple, si $n = 20$, el vector resultant és:

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	2	0	3	0	2	3	5	0	3	0	7	5	2	0	3	0	5

El codi següent implementa el sedàs d'Eratòstenes. El codi assumeix que cada element de *sieve* és inicialment zero.

```
for (int x = 2; x <= n; x++) {
    if (sieve[x]) continue;
    for (int u = 2*x; u <= n; u += x) {
        sieve[u] = x;
    }
}
```

El bucle intern de l'algorisme s'executa n/x vegades per cada valor de x . Així, el temps d'execució de l'algorisme està limitat per la suma harmònica

$$\sum_{x=2}^n n/x = n/2 + n/3 + n/4 + \dots + n/n = O(n \log n).$$

De fet, l'algorisme és més eficient, perquè el bucle intern només s'executa si el nombre x és primer. Es pot demostrar que el temps d'execució de l'algorisme és $O(n \log \log n)$, una complexitat molt propera a $O(n)$.

Algorisme d'Euclides

El **màxim comú divisor** (*greatest common divisor*) dels nombres a i b , $\text{gcd}(a, b)$, és el nombre més gran que divideix a i b , i el **mínim comú múltiple** (*least common multiple*) de a i b , $\text{lcm}(a, b)$, és el nombre més petit que és divisible per a i b . Per exemple, $\text{gcd}(24, 36) = 12$ i $\text{lcm}(24, 36) = 72$.

El màxim comú divisor i el mínim comú múltiple estan connectats de la següent manera:

$$\text{lcm}(a, b) = \frac{ab}{\text{gcd}(a, b)}$$

Algorisme d'Euclides¹ proporciona una manera eficient de trobar el màxim comú divisor de dos nombres. L'algorisme es basa en la fórmula següent:

$$\text{gcd}(a, b) = \begin{cases} a & b = 0 \\ \text{gcd}(b, a \bmod b) & b \neq 0 \end{cases}$$

Per exemple,

$$\text{gcd}(24, 36) = \text{gcd}(36, 24) = \text{gcd}(24, 12) = \text{gcd}(12, 0) = 12.$$

L'algorisme es pot implementar de la següent manera:

```
int gcd(int a, int b) {
    if (b == 0) return a;
    return gcd(b, a%b);
}
```

Es pot demostrar que l'algorisme d'Euclides funciona en temps $O(\log n)$, on $n = \min(a, b)$. El pitjor cas per a l'algorisme és quan a i b són nombres de Fibonacci consecutius. Per exemple,

$$\text{gcd}(13, 8) = \text{gcd}(8, 5) = \text{gcd}(5, 3) = \text{gcd}(3, 2) = \text{gcd}(2, 1) = \text{gcd}(1, 0) = 1.$$

Funció φ d'Euler

Els nombres a i b són **coprims** si $\text{gcd}(a, b) = 1$. La **funció φ (fi) d'Euler** $\varphi(n)$ és la quantitat de nombres coprims amb n entre 1 i n . Per exemple, $\varphi(12) = 4$, perquè 1, 5, 7 i 11 són coprims amb 12.

El valor de $\varphi(n)$ es pot calcular a partir de la factorització en primers de n mitjançant la fórmula

$$\varphi(n) = \prod_{i=1}^k p_i^{\alpha_i - 1} (p_i - 1).$$

Per exemple, $\varphi(12) = 2^1 \cdot (2 - 1) \cdot 3^0 \cdot (3 - 1) = 4$. Tingueu en compte que $\varphi(n) = n - 1$ si n és primer.

21.2 Aritmètica modular

En l'**aritmètica modular**, el conjunt de nombres està limitat de manera que només s'utilitzen els nombres $0, 1, 2, \dots, m - 1$, on m és una constant. Cada nombre

¹Euclides va ser un matemàtic grec que va viure cap al 300 aC. Aquest és potser el primer algorisme conegut de la història.

x es representa amb el nombre $x \bmod m$: el residu després de dividir x per m . Per exemple, si $m = 17$, llavors 75 es representa amb $75 \bmod 17 = 7$.

Sovint podem calcular les restes abans de fer càlculs. En particular, es compleixen les següents fórmules:

$$\begin{aligned}(x + y) \bmod m &= (x \bmod m + y \bmod m) \bmod m \\(x - y) \bmod m &= (x \bmod m - y \bmod m) \bmod m \\(x \cdot y) \bmod m &= (x \bmod m \cdot y \bmod m) \bmod m \\x^n \bmod m &= (x \bmod m)^n \bmod m\end{aligned}$$

Exponenciació modular

Sovint cal calcular de manera eficient el valor de $x^n \bmod m$. Això es pot fer en temps $O(\log n)$ amb la recursivitat següent:

$$x^n = \begin{cases} 1 & n = 0 \\ x^{n/2} \cdot x^{n/2} & n \text{ is even} \\ x^{n-1} \cdot x & n \text{ is odd} \end{cases}$$

És important que, quan n és parell, el valor $x^{n/2}$ només es calculi una vegada. Això garanteix que la complexitat temporal de l'algorisme és $O(\log n)$, ja que n sempre es redueix a la meitat quan és parell.

La funció següent calcula el valor de $x^n \bmod m$:

```
int modpow(int x, int n, int m) {
    if (n == 0) return 1%m;
    long long u = modpow(x, n/2, m);
    u = (u*u)%m;
    if (n%2 == 1) u = (u*x)%m;
    return u;
}
```

Teorema de Fermat i Teorema d'Euler

El **teorema de Fermat** afirma que

$$x^{m-1} \bmod m = 1$$

quan m és primer i x i m són coprimers. Això també implica

$$x^k \bmod m = x^{k \bmod (m-1)} \bmod m.$$

De manera més general, **teorema d'Euler** afirma que

$$x^{\varphi(m)} \bmod m = 1$$

quan x i m són coprimers. El teorema de Fermat és una conseqüència del teorema d'Euler, perquè si m és primer, aleshores $\varphi(m) = m - 1$.

Invers modular

L'invers de x mòdul m és un nombre x^{-1} tal que

$$xx^{-1} \bmod m = 1.$$

Per exemple, si $x = 6$ i $m = 17$, aleshores $x^{-1} = 3$, ja que $6 \cdot 3 \bmod 17 = 1$.

Els inversos modulars es fan servir per dividir nombres mòdul m , ja que dividir per x és el mateix que multiplicar per x^{-1} . Per exemple, per avaluar el valor de $36/6 \bmod 17$, podem fer servir la fórmula $2 \cdot 3 \bmod 17$, ja que $36 \bmod 17 = 2$ i $6^{-1} \bmod 17 = 3$.

Tanmateix, no sempre existeix un invers modular. Per exemple, si $x = 2$ i $m = 4$, l'equació

$$xx^{-1} \bmod m = 1$$

no es pot resoldre, perquè tots els múltiples de 2 són parells i el residu no pot ser mai 1 quan $m = 4$. Resulta que el valor de $x^{-1} \bmod m$ es pot calcular exactament quan x i m són coprimers.

Si existeix un invers modular, es pot calcular mitjançant la fórmula

$$x^{-1} = x^{\varphi(m)-1}.$$

Si m és primer, la fórmula esdevé

$$x^{-1} = x^{m-2}.$$

Per exemple,

$$6^{-1} \bmod 17 = 6^{17-2} \bmod 17 = 3.$$

Aquesta fórmula ens permet calcular eficaçment inversos modulars mitjançant l'algorisme d'exponentació modular. La fórmula es deriva fent servir el teorema d'Euler. En primer lloc, l'invers modular satisfà l'equació següent:

$$xx^{-1} \bmod m = 1.$$

D'altra banda, segons el teorema d'Euler,

$$x^{\varphi(m)} \bmod m = xx^{\varphi(m)-1} \bmod m = 1,$$

per tant, els nombres x^{-1} i $x^{\varphi(m)-1}$ són iguals.

Aritmètica i ordinadors

En programació, els nombres enters sense signe es representen mòdul 2^k , on k és el nombre de bits del tipus de dades. Una conseqüència habitual d'això és que els nombres tornen a començar de 0 si es fan massa gran.

Per exemple, en C++, els nombres del tipus `unsigned int` es representen mòdul 2^{32} . El codi següent declara una variable `unsigned int` el valor de la qual és 123456789. Després d'això, el valor es multiplicarà per si mateix i el resultat és $123456789^2 \bmod 2^{32} = 2537071545$.

```
unsigned int x = 123456789;
cout << x*x << "\n"; // 2537071545
```

21.3 Resolució d'equacions

Equacions diofàntiques

Una **equació diofàntica** és una equació de la forma

$$ax + by = c,$$

on a , b i c són constants i s'han de trobar els valors de x i y . Cada nombre de l'equació ha de ser un nombre enter. Per exemple, una solució per a l'equació $5x + 2y = 11$ és $x = 3$ i $y = -2$.

Les equacions diofàniques es poden resoldre eficientment amb l'algorisme d'Euclides. Resulta que podem ampliar l'algorisme d'Euclides de manera que trobi els nombres x i y que compleixin l'equació següent:

$$ax + by = \gcd(a, b)$$

Una equació diofàntica es pot resoldre si c és divisible per $\gcd(a, b)$, i en cas contrari no es pot resoldre.

Per exemple, busquem els nombres x i y que compleixin l'equació següent:

$$39x + 15y = 12$$

L'equació es pot resoldre, perquè $\gcd(39, 15) = 3$ i $3 \mid 12$. Quan l'algorisme d'Euclides calcula el màxim comú divisor de 39 i 15, produeix la següent seqüència de crides:

$$\gcd(39, 15) = \gcd(15, 9) = \gcd(9, 6) = \gcd(6, 3) = \gcd(3, 0) = 3$$

This corresponds to the following equations:

$$\begin{aligned} 39 - 2 \cdot 15 &= 9 \\ 15 - 1 \cdot 9 &= 6 \\ 9 - 1 \cdot 6 &= 3 \end{aligned}$$

Amb aquestes equacions, trobem

$$39 \cdot 2 + 15 \cdot (-5) = 3$$

i multiplicant això per 4, el resultat és

$$39 \cdot 8 + 15 \cdot (-20) = 12,$$

per tant, una solució de l'equació és $x = 8$ i $y = -20$.

Les solucions d'equacions diofàniques mai no són úniques, perquè podem formar un nombre infinit de solucions a partir d'una de donada. Si un parell (x, y) és una solució, aleshores tots els parells

$$\left(x + \frac{kb}{\gcd(a, b)}, y - \frac{ka}{\gcd(a, b)}\right)$$

són també solucions, on k és qualsevol nombre enter.

Teorema xinès del residu

El **teorema xinès del residu** resol un grup d'equacions de la forma

$$\begin{aligned}x &= a_1 \bmod m_1 \\x &= a_2 \bmod m_2 \\&\dots \\x &= a_n \bmod m_n\end{aligned}$$

on els nombres m_1, m_2, \dots, m_n són coprimers dos a dos.

Sigui x_m^{-1} la inversa de x mòdul m , i

$$X_k = \frac{m_1 m_2 \cdots m_n}{m_k}.$$

Amb aquesta notació, una solució a les equacions és

$$x = a_1 X_1 X_{1m_1}^{-1} + a_2 X_2 X_{2m_2}^{-1} + \cdots + a_n X_n X_{nm_n}^{-1}.$$

En aquesta solució, per a cada $k = 1, 2, \dots, n$,

$$a_k X_k X_{km_k}^{-1} \bmod m_k = a_k,$$

perquè

$$X_k X_{km_k}^{-1} \bmod m_k = 1.$$

Com que tots els altres termes de la suma són divisibles per m_k , no tenen cap efecte sobre el residu, i $x \bmod m_k = a_k$.

Per exemple, una solució per

$$\begin{aligned}x &= 3 \bmod 5 \\x &= 4 \bmod 7 \\x &= 2 \bmod 3\end{aligned}$$

és

$$3 \cdot 21 \cdot 1 + 4 \cdot 15 \cdot 1 + 2 \cdot 35 \cdot 2 = 263.$$

Un cop hem trobat una solució x , podem crear-ne un nombre infinit, ja que tots els nombres de la forma

$$x + m_1 m_2 \cdots m_n$$

són solucions.

21.4 Altres resultats

Teorema de Lagrange

El **teorema de Lagrange** afirma que cada nombre enter positiu es pot representar com una suma de quatre quadrats, és a dir, $a^2 + b^2 + c^2 + d^2$. Per exemple, el número 123 es pot representar com la suma $8^2 + 5^2 + 5^2 + 3^2$.

Teorema de Zeckendorf

El **teorema de Zeckendorf** afirma que cada enter positiu té una representació única com a suma de nombres de Fibonacci de manera que no hi ha dos nombres iguals o consecutius de Fibonacci. Per exemple, el número 74 es pot representar com la suma $55 + 13 + 5 + 1$.

Terna pitagòrica

Una **terna pitagòrica** és un terna (a, b, c) que compleix el teorema de Pitàgores $a^2 + b^2 = c^2$, és a dir, que hi ha un triangle rectangle amb longituds de costat a , b i c . Per exemple, $(3, 4, 5)$ és un terna pitagòrica.

Si (a, b, c) és una terna pitagòrica, totes les ternes de la forma (ka, kb, kc) amb $k > 1$ també són ternes pitagòriques. Una terna pitagòrica és *primitiva* si a , b i c són coprims, i totes les ternes pitagòriques es poden construir a partir de ternes primitives amb multiplicador k .

La fórmula d'Euclides es fa servir per a produir totes les ternes pitagòriques primitives. Cada terna és de la forma

$$(n^2 - m^2, 2nm, n^2 + m^2),$$

on $0 < m < n$, n i m són coprims, i almenys un de n i m és parell. Per exemple, quan $m = 1$ i $n = 2$, la fórmula produeix la terna pitagòrica més petita

$$(2^2 - 1^2, 2 \cdot 2 \cdot 1, 2^2 + 1^2) = (3, 4, 5).$$

Teorema de Wilson

El **teorema de Wilson** afirma que un nombre n és primer exactament quan

$$(n - 1)! \bmod n = n - 1.$$

Per exemple, el nombre 11 és primer, perquè

$$10! \bmod 11 = 10,$$

però el nombre 12 no és primer, perquè

$$11! \bmod 12 = 0 \neq 11.$$

Per tant, el teorema de Wilson es pot fer servir per esbrinar si un nombre és primer. Tanmateix, a la pràctica, el teorema no es pot aplicar per valors grans de n , perquè és difícil calcular valors de $(n - 1)!$ quan n és gran.

Capítol 22

Combinatòria

La **combinatòria** estudia mètodes per comptar combinacions d'objectes. Normalment, l'objectiu és trobar una manera de comptar les combinacions de manera eficient sense generar cada combinació per separat.

Per exemple, considereu el problema de comptar el nombre de maneres de representar un nombre enter n com a suma de nombres enters positius. Per exemple, hi ha 8 representacions per 4:

- $1 + 1 + 1 + 1$
- $1 + 1 + 2$
- $1 + 2 + 1$
- $2 + 1 + 1$
- $2 + 2$
- $3 + 1$
- $1 + 3$
- 4

Sovint, un problema combinatori es pot resoldre mitjançant una funció recursiva. En aquest problema, podem definir una funció $f(n)$ que dona el nombre de representacions de n . Per exemple, $f(4) = 8$ segons l'exemple anterior. Els valors de la funció es poden calcular recursivament de la següent manera:

$$f(n) = \begin{cases} 1 & n = 0 \\ f(0) + f(1) + \dots + f(n-1) & n > 0 \end{cases}$$

El cas base és $f(0) = 1$, perquè la suma buida representa el nombre 0. Aleshores, si $n > 0$, considerem totes les maneres d'escollir el primer nombre de la suma. Si el primer nombre és k , hi ha $f(n - k)$ representacions per a la part restant de la suma. Així, calculem la suma de tots els valors de la forma $f(n - k)$ on $k < n$.

Els primers valors de la funció són:

$$\begin{aligned} f(0) &= 1 \\ f(1) &= 1 \\ f(2) &= 2 \\ f(3) &= 4 \\ f(4) &= 8 \end{aligned}$$

De vegades, una fórmula recursiva es pot substituir per una fórmula de forma tancada. En aquest problema,

$$f(n) = 2^{n-1},$$

que es basa en el fet que hi ha $n - 1$ posicions possibles per als signes + a la suma i podem triar-ne qualsevol subconjunt.

22.1 Coeficients binomials

El **coeficient binomial** $\binom{n}{k}$ és igual al nombre de maneres de triar un subconjunt de k elements d'un conjunt de n elements. Per exemple, $\binom{5}{3} = 10$, perquè el conjunt $\{1, 2, 3, 4, 5\}$ té 10 subconjunts de 3 elements:

$$\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 4\}, \{2, 3, 5\}, \{2, 4, 5\}, \{3, 4, 5\}$$

Fórmula 1

Els coeficients binomials es poden calcular recursivament de la següent manera:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

La idea és fixar un element x al conjunt. Si x s'inclou al subconjunt, hem de triar $k - 1$ elements d'entre $n - 1$ elements, i si x no s'inclou al subconjunt, hem de triar k elements de $n - 1$ elements.

Els casos base de la recursivitat són

$$\binom{n}{0} = \binom{n}{n} = 1,$$

perquè sempre hi ha exactament una manera de construir un subconjunt buit i un subconjunt que conté tots els elements.

Fórmula 2

Una altra manera de calcular els coeficients binomials és:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Hi ha $n!$ permutacions de n elements. Passem per totes les permutacions i sempre incloem els primers k elements de la permutació al subconjunt. Com que l'ordre dels elements dintre i fora del subconjunt no importa, el resultat es divideix per $k!$ i $(n - k)!$

Propietats

Per als coeficients binomials,

$$\binom{n}{k} = \binom{n}{n-k},$$

perquè en realitat dividim un conjunt de n elements en dos subconjunts: el primer conté k elements i el segon conté $n - k$ elements.

La suma dels coeficients binomials és

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n} = 2^n.$$

La raó del nom "coeficient binomial" es pot veure quan elevem el binomi $(a + b)$ a la potència n -èsima:

$$(a + b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \dots + \binom{n}{n-1}a^1b^{n-1} + \binom{n}{n}a^0b^n.$$

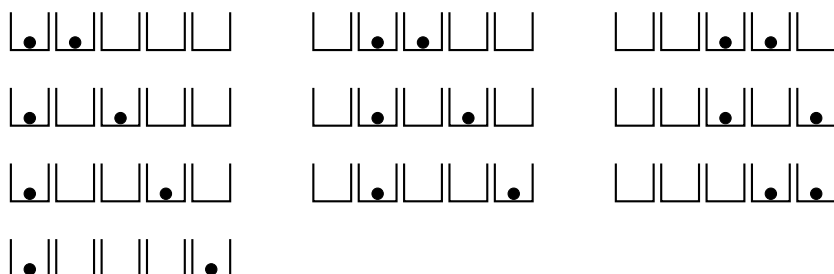
Els coeficients binomials també apareixen al **triangle de Pascal**, on cada valor és igual a la suma dels dos valors anteriors:

$$\begin{array}{ccccccc} & & & & 1 & & & & \\ & & & & 1 & & 1 & & \\ & & & 1 & & 2 & & 1 & \\ & & 1 & & 3 & & 3 & & 1 \\ 1 & & 4 & & 6 & & 4 & & 1 \\ \dots & & \dots & & \dots & & \dots & & \dots \end{array}$$

Caixes i boles

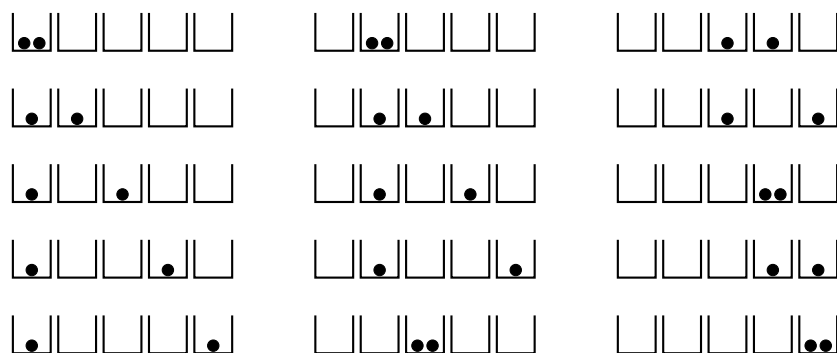
"Caixes i boles" és un model útil, on comptem les maneres de col·locar k boles en n caixes. Considerem tres escenaris:

Escenari 1: cada caixa pot contenir com a màxim una bola. Per exemple, quan $n = 5$ i $k = 2$, hi ha 10 solucions:



En aquest escenari, la resposta és directament el coeficient binomial $\binom{n}{k}$.

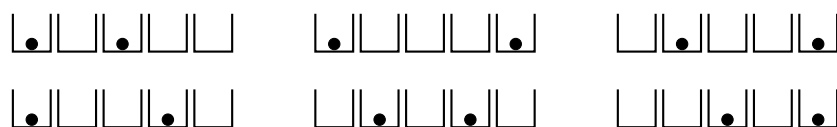
Escenari 2: una caixa pot contenir diverses boles. Per exemple, quan $n = 5$ i $k = 2$, hi ha 15 solucions:



El procés de col·locació de les boles a les caixes es pot representar com una cadena amb els símbols "o" i "→". Inicialment, suposem que estem a la casella més a l'esquerra. El símbol "o" significa que col·loquem una bola a la casella actual, i el símbol "→" significa que ens movem a la casella següent a la dreta.

Amb aquesta notació, cada solució és una cadena que conté k vegades el símbol "o" i $n - 1$ vegades el símbol "→". Per exemple, la solució superior dreta de la imatge de dalt correspon a la cadena "→ → o → o →". Així, el nombre de solucions és $\binom{k+n-1}{k}$.

Escenari 3: Cada caixa pot contenir com a màxim una bola i, a més, no permetem que dues caixes adjacents continguin ambdues una bola. Per exemple, quan $n = 5$ i $k = 2$, hi ha 6 solucions:



En aquest escenari, podem suposar que k boles es col·loquen inicialment en caixes i que hi ha una caixa buida entre dues caixes adjacents. El que queda és triar les posicions de les caixes buides restants. Hi ha $n - 2k + 1$ caixes d'aquest tipus i $k + 1$ posicions per a elles. Així, utilitzant la fórmula de l'escenari 2, el nombre de solucions és $\binom{n-k+1}{n-2k+1}$.

Coeficients multinomials

El coeficient multinomial

$$\binom{n}{k_1, k_2, \dots, k_m} = \frac{n!}{k_1! k_2! \dots k_m!},$$

és el nombre de maneres en què podem dividir n elements en subconjunts de mides k_1, k_2, \dots, k_m , on $k_1 + k_2 + \dots + k_m = n$. Els coeficients multinomials es poden veure com una generalització dels coeficients binomials; si $m = 2$, la fórmula anterior correspon a la fórmula del coeficient binomial.

22.2 Nombres de Catalan

El **nombre de Catalan**¹ C_n és igual al nombre d'expressions de parèntesis vàlides amb n parèntesis esquerres i n drets.

Per exemple, $C_3 = 5$, perquè podem construir les següents expressions de parèntesis vàlides utilitzant tres parèntesis esquerres i drets:

- $()()()$
- $((()))$
- $()(())$
- $((()))$
- $((()()))$

Expressions de parèntesis

Què és exactament una *expressió de parèntesis vàlida*? Les regles següents defineixen amb precisió quines són les expressions vàlides:

- L'expressió buida és vàlida.
- Si l'expressió A és vàlida, l'expressió (A) també és vàlida.
- Si les expressions A i B són vàlides, l'expressió AB també ho és.

Una altra manera de caracteritzar les expressions de parèntesis vàlides és que si triem qualsevol prefix d'aquesta expressió, ha de contenir almenys tants parèntesis esquerre com parèntesis dret. A més, l'expressió completa ha de contenir un nombre igual de parèntesis esquerre i dret.

Fórmula 1

Els nombres de Catalan es poden calcular mitjançant la fórmula

$$C_n = \sum_{i=0}^{n-1} C_i C_{n-i-1}.$$

La suma recorre les maneres de dividir l'expressió en dues parts de manera que ambdues parts siguin expressions vàlides i la primera part sigui tan curta com sigui possible, però no buida. Per a qualsevol i , la primera part conté $i + 1$ parells de parèntesis i el nombre d'expressions és el producte dels valors següents:

- C_i : maneres de construir una expressió de parèntesis per la primera part, sense comptar els parèntesis que envolten l'expressió.
- C_{n-i-1} : maneres de construir una expressió de parèntesis per la segona part.

El cas base és $C_0 = 1$, perquè podem construir una expressió de parèntesi buida fent servir zero parells de parèntesis.

¹E. C. Catalan (1814–1894) va ser un matemàtic belga.

Fórmula 2

Els nombres de Catalan també es poden calcular amb coeficients binomials:

$$C_n = \frac{1}{n+1} \binom{2n}{n}$$

La fórmula es pot explicar de la següent manera:

Hi ha $\binom{2n}{n}$ maneres de construir una expressió de parèntesi (no necessàriament vàlida) que conté n parèntesis esquerre i n parèntesis dret. Calculem el nombre d'expressions d'aquest tipus que *no* són vàlides.

Si una expressió de parèntesis no és vàlida, ha de contenir un prefix on el nombre de parèntesis dret superi el nombre de parèntesis esquerre. La idea és invertir cada parèntesi que pertany a aquest prefix. Per exemple, l'expressió $()()()$ conté un prefix $()()$, i després d'invertir el prefix, l'expressió es converteix en $)((())$.

L'expressió resultant consta de $n+1$ parèntesis esquerre i $n-1$ parèntesis dret. El nombre d'aquestes expressions és $\binom{2n}{n+1}$, que és igual al nombre d'expressions de parèntesis no vàlides. Així, el nombre d'expressions de parèntesis vàlides es pot calcular mitjançant la fórmula

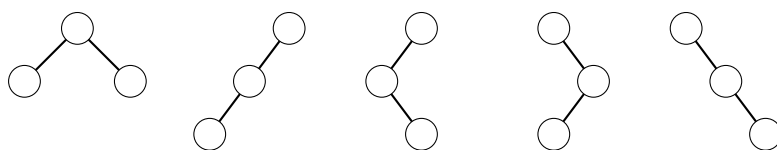
$$\binom{2n}{n} - \binom{2n}{n+1} = \binom{2n}{n} - \frac{n}{n+1} \binom{2n}{n} = \frac{1}{n+1} \binom{2n}{n}.$$

Comptar arbres

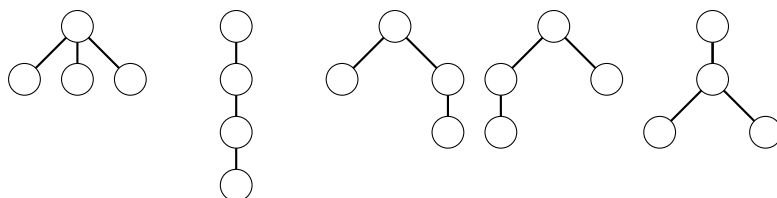
Els nombres catalans també estan relacionats amb els arbres:

- hi ha C_n arbres binaris de n nodes
- hi ha C_{n-1} arbres arrelats de n nodes

Per exemple, per a $C_3 = 5$, els arbres binaris són



i els arbres arrelats són



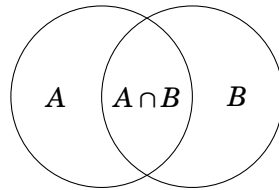
(N. del T.) Pels arbres binaris, un arbre amb subarbres fills A i B es correspon amb l'expressió $(S_A)S_B$, on S_X és l'expressió corresponent al subarbre X . Per exemple, els arbres binaris anteriors es corresponen amb $()()()$, $((()))$, $(())$, $()()$, $()()$. Pels arbres arrelats, un arbre amb subarbres fill A_1, \dots, A_k es correspon amb l'expressió $(S_{A_1}) \dots (S_{A_k})$. Per exemple, els arbres arrelats anteriors es corresponen amb $()()()$, $((()))$, $()()$, $()()$, $()()$.

22.3 Inclusió-exclusió

La **inclusió-exclusió** és una tècnica que es fa servir per comptar la mida d'una unió de conjunts quan es coneixen les mides de les interseccions, i viceversa. Un exemple senzill de la tècnica és la fórmula

$$|A \cup B| = |A| + |B| - |A \cap B|,$$

on A i B són conjunts i $|X|$ indica la mida de X . La fórmula es pot il·lustrar de la següent manera:

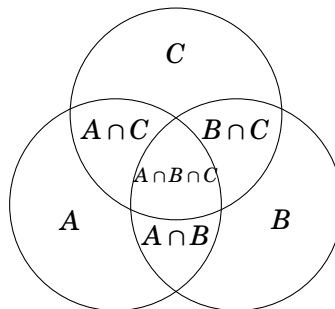


El nostre objectiu és calcular la mida de la unió $A \cup B$ que correspon a l'àrea de la regió que pertany a almenys un cercle. La imatge mostra que podem calcular l'àrea de $A \cup B$ sumant primer les àrees de A i B i després restant l'àrea de $A \cap B$.

La mateixa idea es pot aplicar quan el nombre de conjunts és més gran. Quan hi ha tres conjunts, la fórmula d'inclusió-exclusió és

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

i la imatge corresponent és



En el cas general, la mida de la unió $X_1 \cup X_2 \cup \dots \cup X_n$ es pot calcular passant per totes les interseccions possibles que contenen alguns dels conjunts X_1, X_2, \dots, X_n . Si la intersecció conté un nombre senar de conjunts, la seva mida s'afegeix a la resposta i, en cas contrari, la seva mida es resta de la resposta.

Tingueu en compte que hi ha fórmules similars per calcular la mida d'una intersecció a partir de les mides de les unions. Per exemple,

$$|A \cap B| = |A| + |B| - |A \cup B|$$

i

$$|A \cap B \cap C| = |A| + |B| + |C| - |A \cup B| - |A \cup C| - |B \cup C| + |A \cup B \cup C|.$$

Desarranjament

Com a exemple, comptem el nombre de **desarranjaments** (*derangements*) dels elements $\{1, 2, \dots, n\}$, és a dir, permutacions on cap element roman al seu lloc original. Per exemple, quan $n = 3$, hi ha dos desarranjaments: $(2, 3, 1)$ i $(3, 1, 2)$.

Una manera de resoldre el problema és utilitzar la inclusió-exclusió. Sigui X_k el conjunt de permutacions que contenen l'element k a la posició k . Per exemple, quan $n = 3$, els conjunts són els següents:

$$\begin{aligned}X_1 &= \{(1, 2, 3), (1, 3, 2)\} \\X_2 &= \{(1, 2, 3), (3, 2, 1)\} \\X_3 &= \{(1, 2, 3), (2, 1, 3)\}\end{aligned}$$

Amb aquests conjunts, el nombre de desarranjaments és igual

$$n! - |X_1 \cup X_2 \cup \dots \cup X_n|,$$

i per tant n'hi ha prou amb calcular la mida de la unió. Amb l'inclusió-exclusió això es redueix a calcular les mides de les interseccions, i això es pot fer de manera eficient. Per exemple, quan $n = 3$, la mida de $|X_1 \cup X_2 \cup X_3|$ és

$$\begin{aligned}&|X_1| + |X_2| + |X_3| - |X_1 \cap X_2| - |X_1 \cap X_3| - |X_2 \cap X_3| + |X_1 \cap X_2 \cap X_3| \\&= 2 + 2 + 2 - 1 - 1 - 1 + 1 \\&= 4,\end{aligned}$$

per tant, el nombre de solucions és $3! - 4 = 2$.

Aquest problema també es pot resoldre sense fer servir inclusió-exclusió. Sigui $f(n)$ el nombre de desarranjaments per a $\{1, 2, \dots, n\}$. Podem utilitzar la següent fórmula recursiva:

$$f(n) = \begin{cases} 0 & n = 1 \\ 1 & n = 2 \\ (n-1)(f(n-2) + f(n-1)) & n > 2 \end{cases}$$

La fórmula es deriva considerant com canvia l'element 1 en un desarranjament. Hi ha $n - 1$ maneres de triar un element x que substitueixi l'element 1. En cadascuna d'aquestes eleccions, n'hi ha dues possibilitats:

Opció 1: Substituïm l'element x per l'element 1. Per tant, la tasca restant és construir un desarranjament de $n - 2$ elements.

Opció 2: Substituïm l'element x per algun altre element que no sigui 1. Per tant, la tasca restant és construir un desarranjament de $n - 1$ elements, ja que no podem substituir l'element x per l'element 1, i tots els altres elements s'han de canviar.

22.4 Lema de Burnside

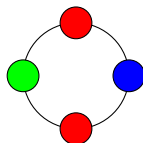
El **lema de Burnside** es pot utilitzar per comptar el nombre de combinacions de manera que només es compti un representant per a cada grup de combinacions

simètriques. El lema de Burnside afirma que el nombre de combinacions és

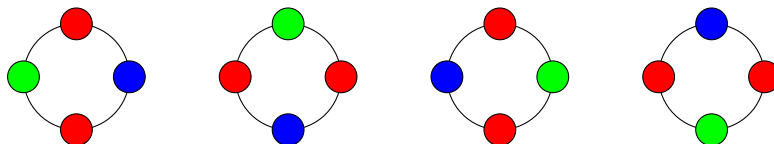
$$\sum_{k=1}^n \frac{c(k)}{n},$$

on hi ha n maneres de canviar la posició d'una combinació, i hi ha $c(k)$ combinacions que no canvien quan s'aplica la k -èssima manera de canviar la posició d'una combinació.

Com a exemple, calculem el nombre de collarets de n perles, on cada perla té m colors possibles. Dos collarets són simètrics si són semblants després de girar-los. Per exemple, el collaret



té els següents collarets simètrics:



Hi ha n maneres de canviar la posició d'un collaret, perquè el podem girar $0, 1, \dots, n-1$ passos en sentit horari. Si el nombre de passos és 0, tots els m^n collarets romanen iguals, i si el nombre de passos és 1, només els m collarets on cada perla té el mateix color romandran iguals.

De manera més general, quan el nombre de passos és k , un total de

$$m^{\gcd(k,n)}$$

collarets segueixen sent els mateixos, on $\gcd(k,n)$ és el màxim comú divisor de k i n . La raó d'això és que els blocs de perles de mida $\gcd(k,n)$ es substituiran mútuament. Així, segons el lema de Burnside, el nombre de collarets és

$$\sum_{i=0}^{n-1} \frac{m^{\gcd(i,n)}}{n}.$$

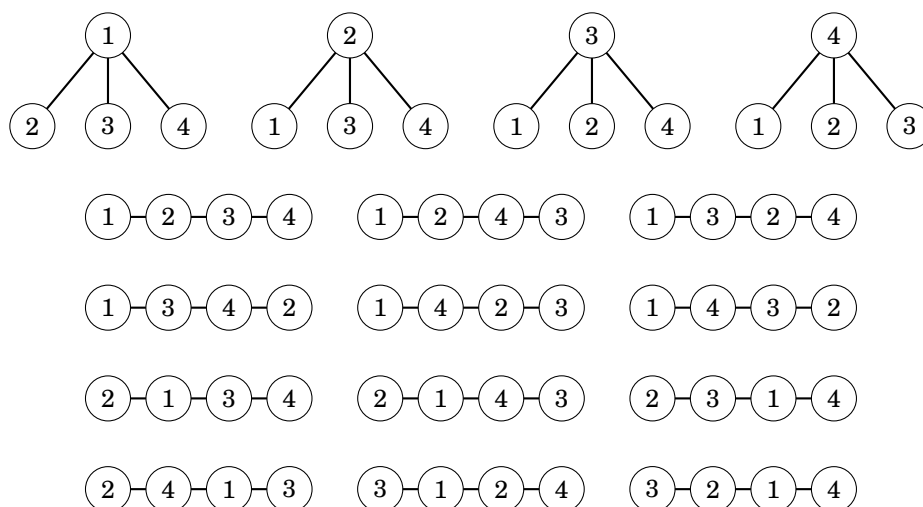
Per exemple, el nombre de collarets de longitud 4 amb 3 colors és

$$\frac{3^4 + 3 + 3^2 + 3}{4} = 24.$$

22.5 Fórmula de Cayley

Fórmula de Cayley afirma que hi ha n^{n-2} arbres etiquetats amb n nodes. Els nodes s'etiqueten $1, 2, \dots, n$, i dos arbres són diferents si la seva estructura o l'etiquetatge és diferent.

Per exemple, quan $n = 4$, el nombre d'arbres etiquetats és $4^{4-2} = 16$:

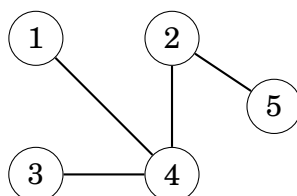


A continuació veurem com es pot derivar la fórmula de Cayley mitjançant els codis de Prüfer.

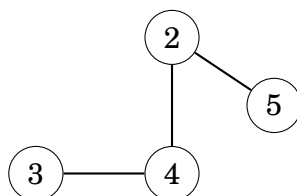
Codi Prüfer

Un **codi Prüfer** és una seqüència de $n-2$ nombres que descriu un arbre etiquetat. La codificació es construeix seguint un procés que elimina $n-2$ fulles de l'arbre. A cada pas, s'elimina la fulla amb l'etiqueta més petita i s'afegeix a la codificació l'etiqueta del seu únic veí.

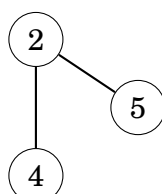
Per exemple, calculem el codi Prüfer del graf següent:



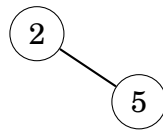
Primer eliminem el node 1 i afegim el node 4:



A continuació, eliminem el node 3 i afegim el node 4:



Finalment eliminem el node 4 i afegim el node 2:



Així, el codi Prüfer del graf és $[4, 4, 2]$.

Podem construir un codi Prüfer per a qualsevol arbre, i el que és més important, l'arbre original es pot reconstruir a partir d'un codi Prüfer. Per tant, el nombre d'arbres etiquetats de n nodes és igual a n^{n-2} , que és el nombre de codis Prüfer de mida n .

Capítol 23

Matrius

Una **matriu** és un concepte matemàtic que es correspon amb una taula bidimensional en programació. Per exemple,

$$A = \begin{bmatrix} 6 & 13 & 7 & 4 \\ 7 & 0 & 8 & 2 \\ 9 & 5 & 4 & 18 \end{bmatrix}$$

és una matriu de mida 3×4 , és a dir, té 3 files i 4 columnes. La notació $[i, j]$ fa referència a l'element de la fila i i la columna j d'una matriu. Per exemple, a la matriu anterior, $A[2, 3] = 8$ i $A[3, 1] = 9$.

Un cas especial d'una matriu és un **vector** que és una matriu unidimensional de mida $n \times 1$. Per exemple,

$$V = \begin{bmatrix} 4 \\ 7 \\ 5 \end{bmatrix}$$

és un vector que conté tres elements.

La **matriu transposta** A^T d'una matriu A s'obté quan s'intercanvien les files i columnes de A , és a dir, $A^T[i, j] = A[j, i]$:

$$A^T = \begin{bmatrix} 6 & 7 & 9 \\ 13 & 0 & 5 \\ 7 & 8 & 4 \\ 4 & 2 & 18 \end{bmatrix}$$

Una matriu és una **matriu quadrada** si té el mateix nombre de files i columnes. Per exemple, la matriu següent és una matriu quadrada:

$$S = \begin{bmatrix} 3 & 12 & 4 \\ 5 & 9 & 15 \\ 0 & 2 & 4 \end{bmatrix}$$

23.1 Operacions

La suma $A + B$ de les matrius A i B només es defineix quan les matrius són de la mateixa mida. El resultat és una matriu on cada element és la suma dels elements corresponents en A i B .

Per exemple,

$$\begin{bmatrix} 6 & 1 & 4 \\ 3 & 9 & 2 \end{bmatrix} + \begin{bmatrix} 4 & 9 & 3 \\ 8 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 6+4 & 1+9 & 4+3 \\ 3+8 & 9+1 & 2+3 \end{bmatrix} = \begin{bmatrix} 10 & 10 & 7 \\ 11 & 10 & 5 \end{bmatrix}.$$

Multiplicar una matriu A per un valor x dóna una matriu xA on cada element de A es multiplica per x . Per exemple,

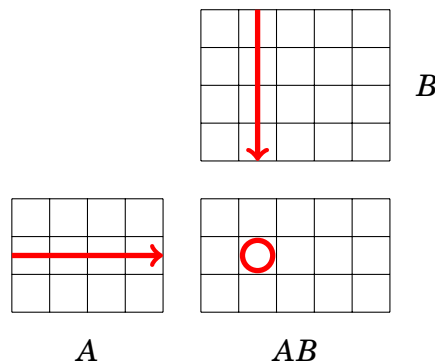
$$2 \cdot \begin{bmatrix} 6 & 1 & 4 \\ 3 & 9 & 2 \end{bmatrix} = \begin{bmatrix} 2 \cdot 6 & 2 \cdot 1 & 2 \cdot 4 \\ 2 \cdot 3 & 2 \cdot 9 & 2 \cdot 2 \end{bmatrix} = \begin{bmatrix} 12 & 2 & 8 \\ 6 & 18 & 4 \end{bmatrix}.$$

Multiplicació de matrius

El producte AB de les matrius A i B es defineix si A és de mida $a \times n$ i B és de mida $n \times b$, és a dir, l'amplada (mida d'una fila) de A és igual a l'alçada (mida d'una columna) de B . El resultat és una matriu de mida $a \times b$ els elements de la qual es calculen mitjançant la fórmula

$$AB[i,j] = \sum_{k=1}^n A[i,k] \cdot B[k,j].$$

La idea és que cada element de AB és la suma de productes dels elements de A i B segons la imatge següent:



Per exemple,

$$\begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 8 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 & 6 \\ 2 & 9 \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 + 4 \cdot 2 & 1 \cdot 6 + 4 \cdot 9 \\ 3 \cdot 1 + 9 \cdot 2 & 3 \cdot 6 + 9 \cdot 9 \\ 8 \cdot 1 + 6 \cdot 2 & 8 \cdot 6 + 6 \cdot 9 \end{bmatrix} = \begin{bmatrix} 9 & 42 \\ 21 & 99 \\ 20 & 102 \end{bmatrix}.$$

La multiplicació de matrius és associativa, i per tant es compleix $A(BC) = (AB)C$, però no és commutativa, de manera que, en general, no es compleix que $AB = BA$.

La **matriu identitat** de mida n és una matriu quadrada on cada element de la diagonal és 1 i tots els altres elements són 0. Per exemple, la matriu següent és la matriu identitat de mida 3:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Multiplicar una matriu per la matriu identitat és el mateix que no fer res. Per exemple,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 8 & 6 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 8 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 8 & 6 \end{bmatrix}.$$

Fent servir un algorisme senzill, podem calcular el producte de dues matrius $n \times n$ en temps $O(n^3)$. També hi ha algorismes més eficients per a la multiplicació de matrius¹, però són sobretot d'interès teòric i aquests algorismes no són necessaris en la programació competitiva.

Potència d'una matriu

La potència A^k d'una matriu A es defineix si A és una matriu quadrada. La definició es basa en la multiplicació de matrius:

$$A^k = \underbrace{A \cdot A \cdot A \cdots A}_{k \text{ times}}$$

Per exemple,

$$\begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix}^3 = \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 48 & 165 \\ 33 & 114 \end{bmatrix}.$$

A més, A^0 és la matriu identitat. Per exemple,

$$\begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix}^0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

La matriu A^k es pot calcular eficientment en temps $O(n^3 \log k)$ usant l'algorisme del capítol 21.2. Per exemple,

$$\begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix}^8 = \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix}^4 \cdot \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix}^4.$$

Determinant

El **determinant** $\det(A)$ d'una matriu A es defineix si A és una matriu quadrada. Si A és de mida 1×1 , llavors $\det(A) = A[1, 1]$. El determinant d'una matriu més gran es calcula recursivament mitjançant la fórmula

$$\det(A) = \sum_{j=1}^n A[1, j] C[1, j],$$

¹El primer algorisme d'aquest tipus va ser l'algorisme de Strassen, publicat el 1969 [63], la complexitat temporal del qual és $O(n^{2.80735})$; el millor algorisme actual [27] funciona en temps $O(n^{2.37286})$.

on $C[i, j]$ és el **cofactor** de A a $[i, j]$. El cofactor es calcula mitjançant la fórmula

$$C[i, j] = (-1)^{i+j} \det(M[i, j]),$$

on $M[i, j]$ s'obté eliminant la fila i i la columna j de A . A causa del coeficient $(-1)^{i+j}$ del cofactor, tots els altres determinants són positius i negatius. Per exemple,

$$\det\begin{pmatrix} 3 & 4 \\ 1 & 6 \end{pmatrix} = 3 \cdot 6 - 4 \cdot 1 = 14$$

i

$$\det\begin{pmatrix} 2 & 4 & 3 \\ 5 & 1 & 6 \\ 7 & 2 & 4 \end{pmatrix} = 2 \cdot \det\begin{pmatrix} 1 & 6 \\ 2 & 4 \end{pmatrix} - 4 \cdot \det\begin{pmatrix} 5 & 6 \\ 7 & 4 \end{pmatrix} + 3 \cdot \det\begin{pmatrix} 5 & 1 \\ 7 & 2 \end{pmatrix} = 81.$$

El determinant de A ens indica si hi ha una **matriu inversa** A^{-1} tal que $A \cdot A^{-1} = I$, on I és la matriu identitat. Resulta que A^{-1} existeix exactament quan $\det(A) \neq 0$, i es pot calcular mitjançant la fórmula

$$A^{-1}[i, j] = \frac{C[j, i]}{\det(A)}.$$

Per exemple,

$$\underbrace{\begin{pmatrix} 2 & 4 & 3 \\ 5 & 1 & 6 \\ 7 & 2 & 4 \end{pmatrix}}_A \cdot \underbrace{\frac{1}{81} \begin{pmatrix} -8 & -10 & 21 \\ 22 & -13 & 3 \\ 3 & 24 & -18 \end{pmatrix}}_{A^{-1}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_I.$$

23.2 Recurrències lineals

Una **recurrència lineal** és una funció $f(n)$ els valors inicials de la qual són $f(0), f(1), \dots, f(k-1)$ i els valors més grans es calculen recursivament mitjançant la fórmula

$$f(n) = c_1 f(n-1) + c_2 f(n-2) + \dots + c_k f(n-k),$$

on c_1, c_2, \dots, c_k són coeficients constants.

La programació dinàmica es pot fer servir per calcular qualsevol valor de $f(n)$ en temps $O(kn)$ calculant tots els valors de $f(0), f(1), \dots, f(n)$ un després un altre. Tanmateix, si k és petit, és possible calcular $f(n)$ de manera molt més eficient en temps $O(k^3 \log n)$ fent servir operacions amb matrius.

Nombres de Fibonacci

Un exemple senzill de recurrència lineal és la funció següent que defineix els nombres de Fibonacci:

$$\begin{aligned} f(0) &= 0 \\ f(1) &= 1 \\ f(n) &= f(n-1) + f(n-2) \end{aligned}$$

En aquest cas, $k = 2$ i $c_1 = c_2 = 1$.

Per calcular de manera eficient els nombres de Fibonacci, representem el F ormula de Fibonacci com un matriu quadrada X de mida 2×2 , on es compleix el seg ent:

$$X \cdot \begin{bmatrix} f(i) \\ f(i+1) \end{bmatrix} = \begin{bmatrix} f(i+1) \\ f(i+2) \end{bmatrix}$$

Aix , els valors $f(i)$ i $f(i+1)$ es donen com a "entrada" per X , i X calcula els valors $f(i+1)$ i $f(i+2)$ de sortida. Resulta que aquesta matriu  s

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Per exemple,

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} f(5) \\ f(6) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 8 \\ 13 \end{bmatrix} = \begin{bmatrix} f(6) \\ f(7) \end{bmatrix}.$$

Aix , podem calcular $f(n)$ mitjan ant la f ormula

$$\begin{bmatrix} f(n) \\ f(n+1) \end{bmatrix} = X^n \cdot \begin{bmatrix} f(0) \\ f(1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^n \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

El valor de X^n es pot calcular en temps $O(\log n)$, de manera que el valor de $f(n)$ tamb  es pot calcular en temps $O(\log n)$.

Cas general

Considerem ara el cas general on $f(n)$  s qualsevol recurr ncia lineal. De nou, el nostre objectiu  s construir una matriu X per a la qual

$$X \cdot \begin{bmatrix} f(i) \\ f(i+1) \\ \vdots \\ f(i+k-1) \end{bmatrix} = \begin{bmatrix} f(i+1) \\ f(i+2) \\ \vdots \\ f(i+k) \end{bmatrix}.$$

Aquesta matriu  s

$$X = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ c_k & c_{k-1} & c_{k-2} & c_{k-3} & \cdots & c_1 \end{bmatrix}.$$

A les primeres $k-1$ files, cada element  s 0, excepte per a un element que  s 1. Aquestes files substitueixen $f(i)$ per $f(i+1)$, $f(i+1)$ per $f(i+2)$, i aix  successivament. L  ltima fila cont  els coeficients de la recurr ncia per calcular el nou valor $f(i+k)$.

Ara, $f(n)$ es pot calcular en temps $O(k^3 \log n)$ fent servir la fórmula

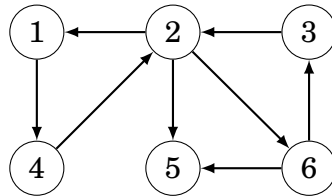
$$\begin{bmatrix} f(n) \\ f(n+1) \\ \vdots \\ f(n+k-1) \end{bmatrix} = X^n \cdot \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(k-1) \end{bmatrix}.$$

23.3 Grafs i matrius

Comptar camins

Les potències de la matriu d'adjacència d'un graf tenen una propietat interessant. Quan V és una matriu d'adjacència en un graf sense pesos ponderat, la matriu V^n conté el nombre de camins de n arestes entre els nodes del graf.

Per exemple, per al graf



la matriu d'adjacència és

$$V = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

Ara, per exemple, la matriu

$$V^4 = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 & 2 & 2 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

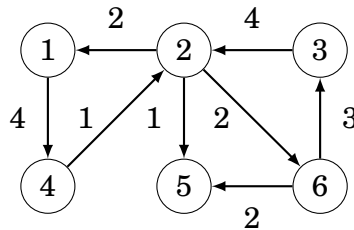
conté el nombre de camins de 4 arestes entre els nodes. Per exemple, $V^4[2,5] = 2$, perquè hi ha dos camins de 4 arestes des del node 2 fins al node 5: $2 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 5$ i $2 \rightarrow 6 \rightarrow 3 \rightarrow 2 \rightarrow 5$.

Camins més curts

Fent servir una idea semblant per als grafs amb pesos, podem calcular per a cada parell de nodes la longitud mínima d'un camí entre ells que conté exactament n

arestes. Per calcular-ho, hem de definir la multiplicació de matrius d'una manera nova, de manera que no calculem el nombre de camins sinó que minimitzem les longituds dels camins.

Com a exemple, considereu el graf següent:



Construïm una matriu d'adjacència on ∞ significa que no existeix cap aresta i els altres valors són els pesos de les arestes. La matriu és

$$V = \begin{bmatrix} \infty & \infty & \infty & 4 & \infty & \infty \\ 2 & \infty & \infty & \infty & 1 & 2 \\ \infty & 4 & \infty & \infty & \infty & \infty \\ \infty & 1 & \infty & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty & \infty & \infty \\ \infty & \infty & 3 & \infty & 2 & \infty \end{bmatrix}.$$

En lloc de la fórmula

$$AB[i,j] = \sum_{k=1}^n A[i,k] \cdot B[k,j]$$

ara fem servir la fórmula

$$AB[i,j] = \min_{k=1}^n A[i,k] + B[k,j]$$

per a la multiplicació de matrius, per tant calculem un mínim en lloc d'una suma i una suma d'elements en lloc d'un producte. Després d'aquesta modificació, els elements de les potències de la matriu es corresponen amb els camins més curts del graf.

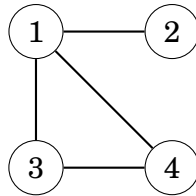
Per exemple, com

$$V^4 = \begin{bmatrix} \infty & \infty & 10 & 11 & 9 & \infty \\ 9 & \infty & \infty & \infty & 8 & 9 \\ \infty & 11 & \infty & \infty & \infty & \infty \\ \infty & 8 & \infty & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty & \infty & \infty \\ \infty & \infty & 12 & 13 & 11 & \infty \end{bmatrix},$$

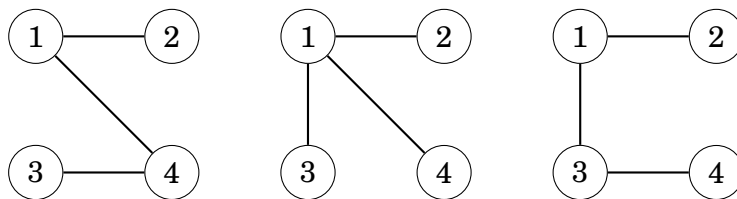
podem concloure que la longitud mínima d'un camí de 4 arestes des del node 2 fins al node 5 és 8. Aquest camí és $2 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 5$.

Teorema de Kirchhoff

El **teorema de Kirchhoff** proporciona una manera de calcular el nombre d'arbres d'expansió d'un graf com el determinant d'una matriu especial. Per exemple, el graf



té tres arbres d'expansió:



Per calcular el nombre d'arbres d'expansió, construïm una **matriu Laplaciana** L , on $L[i, i]$ és el grau del node i i $L[i, j]$ és -1 si hi ha una aresta entre els nodes i i j , i 0 . La matriu Laplaciana del graf anterior és la següent:

$$L = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

Es pot demostrar que el nombre d'arbres d'expansió és igual al determinant de la matriu que s'obté quan eliminem qualsevol fila i qualsevol columna de L . Per exemple, si eliminem la primera fila i columna, el resultat és

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 3.$$

El determinant és sempre el mateix, sense importar quina fila i columna eliminem de L .

La fórmula de Cayley del capítol 22.5 és un cas especial del teorema de Kirchhoff, ja que en un graf complet de n nodes es compleix

$$\det \begin{pmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & n-1 \end{pmatrix} = n^{n-2}.$$

Capítol 24

Probabilitat

Una **probabilitat** és un nombre real entre 0 i 1 que indica quan probable és un esdeveniment. Si un esdeveniment passarà segur, la seva probabilitat és 1, i si un esdeveniment és impossible, la seva probabilitat és 0. La probabilitat d'un esdeveniment s'indica $P(\dots)$ on els tres punts descriuen l'esdeveniment.

Per exemple, quan es llança un dau, el resultat és un nombre enter entre 1 i 6, i la probabilitat de cada resultat és $1/6$. Per exemple, podem calcular les probabilitats següents:

- $P(\text{"el resultat és 4"}) = 1/6$
- $P(\text{"el resultat no és 6"}) = 5/6$
- $P(\text{"el resultat és parell"}) = 1/2$

24.1 Càlcul

Per a calcular la probabilitat d'un esdeveniment fem servir la combinatòria o simulem el procés que genera l'esdeveniment. Per exemple, calculem la probabilitat de treure tres cartes amb el mateix valor d'una baralla de cartes barrejades (per exemple, $\spadesuit 8$, $\clubsuit 8$ i $\diamondsuit 8$).

Mètode 1

Podem calcular la probabilitat mitjançant la fórmula

$$\frac{\text{nombre de resultats desitjats}}{\text{nombre total de resultats}}.$$

En aquest problema, els resultats desitjats són aquells en què el valor de cada carta és el mateix. Hi ha $13 \binom{4}{3}$ resultats d'aquesta mena, perquè hi ha 13 possibilitats pel valor de les cartes i $\binom{4}{3}$ maneres de triar 3 colls (pals) entre els 4 colls possibles.

Hi ha un total de $\binom{52}{3}$ resultats, perquè triem 3 cartes d'entre 52 cartes. Per tant, la probabilitat de l'esdeveniment és

$$\frac{13 \binom{4}{3}}{\binom{52}{3}} = \frac{1}{425}.$$

Mètode 2

Una altra manera de calcular la probabilitat és simular el procés que genera l'esdeveniment. En aquest exemple, treiem tres cartes, de manera que el procés consta de tres passos. Exigim que cada pas del procés tingui èxit.

El coll de la primera carta sempre té èxit, perquè no hi ha restriccions. El segon pas té èxit amb una probabilitat de $3/51$, perquè queden 51 cartes i 3 d'elles tenen el mateix valor que la primera. De la mateixa manera, el tercer pas té èxit amb una probabilitat de $2/50$.

La probabilitat que tot el procés tingui èxit és

$$1 \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{425}.$$

24.2 Esdeveniments

Un esdeveniment en teoria de probabilitats es pot representar com un conjunt

$$A \subset X,$$

on X conté tots els resultats possibles i A és un subconjunt de resultats. Per exemple, quan es llença un dau, els resultats són

$$X = \{1, 2, 3, 4, 5, 6\}.$$

Per exemple, l'esdeveniment "el resultat és parell" es correspon al conjunt

$$A = \{2, 4, 6\}.$$

A cada resultat x li assignem una probabilitat $p(x)$. La probabilitat $P(A)$ d'un esdeveniment A es pot calcular com una suma de probabilitats dels resultats mitjançant la fórmula

$$P(A) = \sum_{x \in A} p(x).$$

Per exemple, quan es llança un dau, $p(x) = 1/6$ per a cada resultat x , de manera que la probabilitat de l'esdeveniment "el resultat és parell" és

$$p(2) + p(4) + p(6) = 1/2.$$

La probabilitat total dels resultats en X ha de ser 1, és a dir, $P(X) = 1$.

Com que els esdeveniments de la teoria de la probabilitat són conjunts, podem manipular-los mitjançant operacions de conjunts estàndard:

- El **complement** \bar{A} significa "A no passa". Per exemple, quan es llança un dau, el complement de $A = \{2, 4, 6\}$ és $\bar{A} = \{1, 3, 5\}$.
- La **unión** $A \cup B$ significa "A o B passa". Per exemple, la unió de $A = \{2, 5\}$ i $B = \{4, 5, 6\}$ és $A \cup B = \{2, 4, 5, 6\}$.
- La **intersecció** $A \cap B$ significa "A i B passen". Per exemple, la intersecció de $A = \{2, 5\}$ i $B = \{4, 5, 6\}$ és $A \cap B = \{5\}$.

Complement

La probabilitat del complement \bar{A} es calcula mitjançant la fórmula

$$P(\bar{A}) = 1 - P(A).$$

De vegades, podem resoldre un problema fàcilment amb complements resolent el problema contrari. Per exemple, la probabilitat d'obtenir almenys un sis en llançar un dau deu vegades és

$$1 - (5/6)^{10}.$$

Aquí $5/6$ és la probabilitat que el resultat d'un sol llançament no sigui sis, i $(5/6)^{10}$ és la probabilitat que cap dels deu llançaments sigui un sis. El complement d'això és la resposta al problema.

Unió

La probabilitat de la unió $A \cup B$ es calcula mitjançant la fórmula

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Per exemple, quan es llança un dau, la unió dels esdeveniments

$$A = \text{"el resultat és parell"}$$

i

$$B = \text{"el resultat és inferior a 4"}$$

és

$$A \cup B = \text{"el resultat és parell o inferior a 4",}$$

i la seva probabilitat és

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/2 - 1/6 = 5/6.$$

Si els esdeveniments A i B són **disjunts**, és a dir, $A \cap B$ està buit, la probabilitat de l'esdeveniment $A \cup B$ és simplement

$$P(A \cup B) = P(A) + P(B).$$

Probabilitat condicionada

La **probabilitat condicionada**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

és la probabilitat que passi A assumint que passa B . Per tant, quan es calcula la probabilitat de A , només considerem els resultats que també pertanyen a B .

Utilitzant els conjunts anteriors,

$$P(A|B) = 1/3,$$

perquè els resultats de B són $\{1, 2, 3\}$, i un d'ells és parell. Aquesta és la probabilitat d'un resultat parell si sabem que el resultat està entre $1 \dots 3$.

Intersecció

Utilitzant la probabilitat condicionada, la probabilitat de la intersecció $A \cap B$ es pot calcular mitjançant la fórmula

$$P(A \cap B) = P(A)P(B|A).$$

Els esdeveniments A i B són **independents** si

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B),$$

el que significa que el fet que passi B no modifica la probabilitat de A , i viceversa. En aquest cas, la probabilitat de la intersecció és

$$P(A \cap B) = P(A)P(B).$$

Per exemple, en treure una carta d'una baralla, els esdeveniments

$$A = \text{"el coll és piques"}$$

i

$$B = \text{"el valor és quatre"}$$

són independents. Per tant, l'esdeveniment

$$A \cap B = \text{"la carta és el quatre de piques"}$$

passa amb probabilitat

$$P(A \cap B) = P(A)P(B) = 1/4 \cdot 1/13 = 1/52.$$

24.3 Variables aleatòries

Una **variable aleatòria** és un valor generat per un procés aleatori. Per exemple, quan es llancen dos daus, una possible variable aleatòria és

$$X = \text{"the sum of the outcomes"}.$$

Per exemple, si els resultats són $[4, 6]$ (és a dir, primer tirem un quatre i després un sis), aleshores el valor de X és 10.

Denotem $P(X = x)$ la probabilitat que el valor d'una variable aleatòria X sigui x ¹. Per exemple, en llançar dos daus, $P(X = 10) = 3/36$, perquè el nombre total de resultats és 36 i hi ha tres maneres possibles d'obtenir la suma 10: $[4, 6]$, $[5, 5]$ i $[6, 4]$.

¹(N. del T.) Les variables aleatòries ens permeten representar esdeveniments (conjunts de resultats) de manera compacta. En aquest cas, $P(X = x)$ és una manera d'abreujar $P(\{r | X(r) = x\})$, és a dir, la probabilitat de l'esdeveniment format per tots aquells resultats r que tenen valor assignat $X(r)$ igual a una constant x .

Valor esperat

La **valor esperat** $E[X]$ indica el valor mitjà d'una variable aleatòria X . El valor esperat es pot calcular com la suma

$$\sum_x P(X = x)x,$$

on x itera per tots els valors possibles de X .

Per exemple, quan es llança un dau, el resultat esperat és

$$1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot 3 + 1/6 \cdot 4 + 1/6 \cdot 5 + 1/6 \cdot 6 = 7/2.$$

Una propietat útil dels valors esperats és que són **lineals**. Això vol dir que la suma $E[X_1 + X_2 + \dots + X_n]$ sempre és igual a la suma $E[X_1] + E[X_2] + \dots + E[X_n]$. Aquesta fórmula és certa encara que les variables aleatòries depenguin les unes de les altres.

Per exemple, quan es llança dos daus, la suma esperada és

$$E[X_1 + X_2] = E[X_1] + E[X_2] = 7/2 + 7/2 = 7.$$

Considerem ara un problema en què n boles es col·loquen aleatòriament en n caixes, i la nostra tasca és calcular el nombre esperat de caixes buides. Cada bola té la mateixa probabilitat de ser col·locada en qualsevol de les caixes. Per exemple, si $n = 2$, les probabilitats són les següents:



En aquest cas, el nombre esperat de caixes buides és

$$\frac{0 + 0 + 1 + 1}{4} = \frac{1}{2}.$$

En el cas general, la probabilitat que una sola caixa estigui buida és

$$\left(\frac{n-1}{n}\right)^n,$$

perquè no hem de posar cap bola. Per tant, utilitzant la linealitat, el nombre esperat de caixes buides és

$$n \cdot \left(\frac{n-1}{n}\right)^n.$$

Distribucions

La **distribució** d'una variable aleatòria X indica la probabilitat de tots els valors que X pot tenir. La distribució consta de valors $P(X = x)$. Per exemple, quan es llança dos daus, la distribució de la seva suma és:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

En una **distribució uniforme**, la variable aleatòria X té n valors possibles $a, a+1, \dots, b$ i la probabilitat de cada valor és $1/n$. Per exemple, quan es llança un dau, $a = 1$, $b = 6$ i $P(X = x) = 1/6$ per a cada valor x .

El valor esperat de X en una distribució uniforme és

$$E[X] = \frac{a+b}{2}.$$

En una **distribució binomial**, es fan n intents i la probabilitat que un sol intent tingui èxit és p . La variable aleatòria X compta el nombre d'intents amb èxit i la probabilitat d'un valor x és

$$P(X = x) = p^x(1-p)^{n-x} \binom{n}{x},$$

on p^x i $(1-p)^{n-x}$ es corresponen amb els intents exitosos i infructuosos, i $\binom{n}{x}$ és el nombre de maneres de triar l'ordre dels intents.

Per exemple, quan es llança un dau deu vegades, la probabilitat de treure un sis exactament tres vegades és $(1/6)^3(5/6)^7 \binom{10}{3}$.

El valor esperat de X en una distribució binomial és

$$E[X] = pn.$$

En una **distribució geomètrica**, la probabilitat que un intent tingui èxit és p , i continuem fins que es produeix el primer èxit. La variable aleatòria X compta el nombre d'intents necessaris i la probabilitat d'un valor x és $(1-p)^{x-1}p$ on $(1-p)^{x-1}$ es correpon amb els intents infructuosos i p es correspon al primer intent amb èxit.

Per exemple, si llancem un dau fins que treiem un sis, la probabilitat que el nombre de llançaments sigui exactament 4 és $(5/6)^3 1/6$.

El valor esperat de X en una distribució geomètrica és

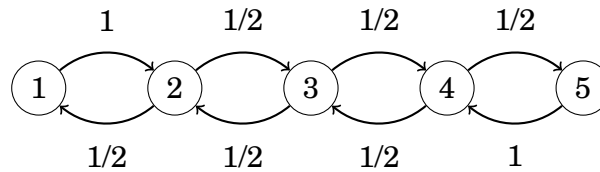
$$E[X] = \frac{1}{p}.$$

24.4 Cadenes de Markov

Una **cadena de Markov** és un procés aleatori que consta d'estats i transicions entre ells. Per a cada estat coneixem les probabilitats de passar als altres estats. Una cadena de Markov es pot representar com un graf els nodes del qual són estats i les arestes són transicions.

Per exemple, considereu el problema on estem a la planta 1 d'un edifici de n plantes. A cada pas, caminem aleatòriament un pis cap amunt o un pis cap avall, excepte si estem al primer o últim pis, on sempre caminem cap amunt o cap avall respectivament. Quina és la probabilitat d'estar al pis m després de k passos?

En aquest problema, cada pis de l'edifici es correspon a un estat en una cadena de Markov. Per exemple, si $n = 5$, el graf és el següent:



La distribució de probabilitat d'una cadena de Markov és un vector $[p_1, p_2, \dots, p_n]$, on p_k és la probabilitat que l'estat actual sigui k . La fórmula $p_1 + p_2 + \dots + p_n = 1$ sempre és vàlida.

En l'escenari anterior, la distribució inicial és $[1, 0, 0, 0, 0]$, perquè sempre comencem al pis 1. La següent distribució és $[0, 1, 0, 0, 0]$, perquè només podem passar del pis 1 al pis 2. Després d'això, podem moure un pis cap amunt o un pis cap avall, de manera que la següent distribució és $[1/2, 0, 1/2, 0, 0]$ i així successivament.

Una manera eficient de simular recorreguts en cadenes de Markov és amb la programació dinàmica. La idea és guardar-nos la distribució de probabilitat i, a cada pas, iterar per totes les possibles maneres de moure's des de cada estat. Amb aquest mètode podem simular un recorregut de m passos en temps $O(n^2m)$.

Les transicions d'una cadena de Markov també es poden representar com una matriu que actualitza la distribució de probabilitat. En l'escenari anterior, la matriu és

$$\begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix}.$$

Quan multipliquem una distribució de probabilitat per aquesta matriu, obtenim la nova distribució després de moure'ns un pas. Per exemple, passem de la distribució $[1, 0, 0, 0, 0]$ a la distribució $[0, 1, 0, 0, 0]$ de la manera següent:

$$\begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Si calculem les potències de la matriu de manera eficient, podem calcular la distribució després de m passos en temps $O(n^3 \log m)$.

24.5 Algorismes aleatoris

De vegades podem fer servir l'atzar per resoldre un problema, encara que el problema no estigui relacionat amb probabilitats. Un **algorisme aleatori** (*randomized algorithm*) és un algorisme que pren decisions aleatòries durant la seva execució.

Un **algorisme de Monte Carlo** és un algorisme aleatori que de vegades pot donar una resposta incorrecta. Aquests algorismes són útils quan la probabilitat de donar una resposta incorrecta és petita.

Un **algorisme de Las Vegas** és un algorisme aleatori que sempre dona la resposta correcta, però que el seu temps d'execució varia aleatòriament. L'objectiu és dissenyar un algorisme que sigui eficient amb alta probabilitat.

A continuació, mostrem tres exemples de problemes que es poden resoldre fent servir l'atzar.

Estadístiques d'ordre

La k -èssima **estadística d'ordre** d'un vector és l'element que resulta a la posició k després d'ordenar el vector en ordre creixent. És fàcil calcular les estadístiques d'ordre en temps $O(n \log n)$ ordenant primer el vector, però és realment necessari ordenar tot el vector per a trobar un sol element?

Resulta que podem trobar l'estadística d'ordre amb un algorisme aleatori sense ordenar el vector. L'algorisme, anomenat **quickselect**², és un algorisme de Las Vegas: el seu temps d'execució sol ser $O(n)$ però és $O(n^2)$ en el pitjor dels casos.

L'algorisme tria un element aleatori x del vector i mou els elements més petits que x a la part esquerra del vector i tots els altres elements a la part dreta. Això triga temps $O(n)$ quan hi ha n elements. Suposem que la part esquerra conté a elements i la part dreta conté b elements. Si $a = k$, l'element x és la k -èssima l'estadística d'ordre. En cas contrari, si $a > k$, trobem recursivament l'estadística d'ordre k -èssima a la part esquerra, i si $a < k$, trobem recursivament l'estadística d'ordre r -èssima per a la part dreta, on $r = k - a$. La recerca continua de manera similar, fins que s'ha trobat l'element.

Quan cada element x es tria aleatòriament, la mida del vector es redueix aproximadament a la meitat a cada pas, de manera que la complexitat temporal de trobar la k -èssima estadística d'ordre és d'aproximadament

$$n + n/2 + n/4 + n/8 + \dots < 2n = O(n).$$

El pitjor cas es quan l'algorisme requereix $O(n^2)$ temps, perquè és possible que x sempre es triï de manera que sigui un dels elements més petits o més grans del vector, i siguin necessaris $O(n)$ passos. Tanmateix, la probabilitat que passi això és tan petita que mai passa a la pràctica.

Verificar multiplicacions de matrius

El nostre següent problema és *verificar* si passa $AB = C$ on A , B i C són matrius de mida $n \times n$. Per descomptat, podem resoldre el problema calculant de nou el producte AB (en temps $O(n^3)$ amb l'algorisme bàsic), però es podria esperar que verificar la resposta hauria de ser més fàcil que calcular-la des de zero.

²L'any 1961, CAR Hoare va publicar dos algorismes eficients en mitjana: **quicksort** [36] per ordenar vectors i **quickselect** [37] per trobar estadístiques d'ordre.

Resulta que podem resoldre el problema amb un algorisme de Monte Carlo³, la complexitat temporal del qual només és $O(n^2)$. La idea és senzilla: triem un vector aleatori X de n elements, i calculem els vectors ABX i CX . Si $ABX = CX$, informem que $AB = C$, i en cas contrari informem que $AB \neq C$.

La complexitat temporal de l'algorisme és $O(n^2)$, perquè podem calcular els matrius ABX i CX en el temps $O(n^2)$. Podem calcular el vector ABX de manera eficient fent servir la representació $A(BX)$, de manera que només calen dues multiplicacions de matrius de mida $n \times n$ i $n \times 1$.

L'inconvenient de l'algorisme és que hi ha una petita possibilitat que l'algorisme cometi un error quan troba que $AB = C$. Per exemple,

$$\begin{bmatrix} 6 & 8 \\ 1 & 3 \end{bmatrix} \neq \begin{bmatrix} 8 & 7 \\ 3 & 2 \end{bmatrix},$$

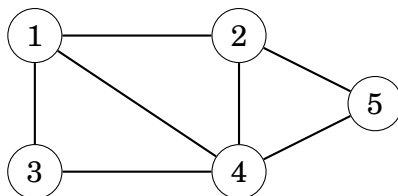
però

$$\begin{bmatrix} 6 & 8 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 8 & 7 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix}.$$

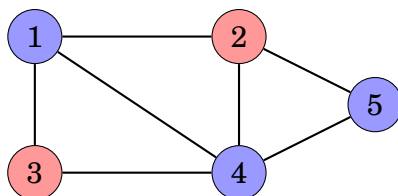
A la pràctica, però, la probabilitat que l'algorisme cometi un error és molt petita, i podem fer que encara sigui més petita verificant el resultat per a diversos vectors aleatoris X abans d'informar que $AB = C$.

Acolorir un graf

Donat un graf amb n nodes i m arestes, la nostra tasca és acolorir els nodes del graf amb dos colors de manera que com a mínim $m/2$ arestes tinguin extrems amb colors diferents. Per exemple, al graf



una coloració vàlida és la següent:



El graf anterior conté 7 arestes i 5 d'elles tenen extrems amb colors diferents, de manera que l'acoloració és vàlida.

El problema es pot resoldre amb un algorisme de Las Vegas que genera coloracions aleatòries fins que troba una coloració vàlida. En una coloració aleatòria,

³R. M. Freivalds va publicar aquest algorisme l'any 1977 [26], i de vegades s'anomena **Algorisme de Freivalds**.

el color de cada node s'escull independentment amb probabilitat $1/2$ per a cada color.

En una coloració aleatòria, la probabilitat que els dos extrems d'una aresta tinguin colors diferents és $1/2$. Per tant, el nombre esperat d'arestes amb colors diferents als extrems és $m/2$. Com que s'espera que una coloració aleatòria sigui vàlida, ràpidament trobarem una coloració vàlida a la pràctica.

Capítol 25

Teoria de jocs

En aquest capítol, ens centrarem en els jocs de dos jugadors sense elements aleatoris. El nostre objectiu és trobar una estratègia que per guanyar el joc independentment del que faci l'oponent, si aquesta estratègia existeix.

Resulta que hi ha una estratègia general per a aquests jocs, i podem analitzar els jocs mitjançant la **teoria dels jocs de nim**. Primer, analitzarem jocs senzills on els jugadors treuen palets d'una pila, i després generalitzarem l'estratègia utilitzada en aquests jocs a altres jocs.

25.1 Estats del joc

Considerem un joc on inicialment hi ha una pila de n palets. Els jugadors A i B es mouen alternativament i el jugador A comença. En cada moviment, el jugador ha de treure 1, 2 o 3 palets de la pila, i el jugador que treu l'últim palet guanya la partida.

Per exemple, si $n = 10$, el joc pot procedir de la següent manera:

- El jugador A treu 2 palets (en queden 8).
- El jugador B treu 3 palets (en queden 5).
- El jugador A treu 1 palet (queden 4).
- El jugador B treu 2 palets (queden 2 palets).
- El jugador A treu 2 palets i guanya.

Aquest joc consta d'estats $0, 1, 2, \dots, n$, on el nombre de l'estat és el nombre de palets que queden.

Estats guanyadors i perdedors

Un **estat guanyador** és un estat on el jugador guanyarà la partida si juga de manera òptima, i un **estat perdedor** és un estat on el jugador perdrà la partida si l'oponent juga de manera òptima. Resulta que podem classificar tots els estats d'un joc de manera que cada estat sigui un estat guanyador o un estat perdedor.

En el joc anterior, l'estat 0 és clarament un estat perdedor, perquè el jugador no pot fer cap moviment. Els estats 1, 2 i 3 són estats guanyadors, perquè podem treure 1, 2 o 3 palets i guanyar la partida. L'estat 4, al seu torn, és un

estat perdedor, perquè qualsevol moviment condueix a un estat que és un estat guanyador per a l'oponent.

De manera més general, si hi ha un moviment que condueix de l'estat actual a un estat perdedor, l'estat actual és un estat guanyador i, en cas contrari, l'estat actual és un estat perdedor. Amb aquesta observació, podem classificar tots els estats d'un joc començant pels estats perdedors on no hi ha moviments possibles.

Els estats 0...15 del joc anterior es poden classificar de la següent manera (*W* denota un estat guanyador i *L* denota un estat perdedor):

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>L</i>	<i>W</i>	<i>W</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>W</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>W</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>W</i>	<i>W</i>

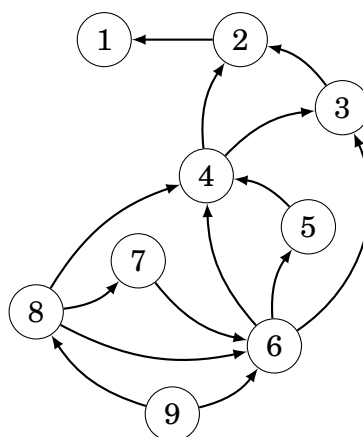
És fàcil analitzar aquest joc: un estat k és un estat perdedor si k és divisible per 4, i en cas contrari és un estat guanyador. Una manera òptima de jugar és triar sempre un moviment després del qual el nombre de palets de la pila és divisible per 4. Finalment, ja no queden palets i l'oponent ha perdut.

Per descomptat, aquesta estratègia requereix que el nombre de palets sigui *no* divisible per 4 quan és el nostre moviment. Si és així, no podem fer res, i el rival guanyarà la partida si juga de manera òptima.

Graf d'estats

Considerem ara un altre joc de palets, on en cada estat k , es permet eliminar qualsevol nombre x de palets de manera que x sigui més petit que k i divideixi k . Per exemple, a l'estat 8 podem treure 1, 2 o 4 palets, però a l'estat 7 l'únic moviment permès és treure 1 palet.

La imatge següent mostra els estats 1...9 del joc com a **graf d'estats**, els nodes del qual són els estats i les arestes són els moviments entre ells:



L'estat final d'aquest joc és sempre l'estat 1, que és un estat perdedor, perquè no hi ha moviments vàlids. La classificació dels estats 1...9 és la següent:

1	2	3	4	5	6	7	8	9
<i>L</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>L</i>	<i>W</i>	<i>L</i>

Sorprenentment, en aquest joc, tots els estats parells són estats guanyadors i tots els estats senars són estats perdedors.

25.2 Joc de nim

El **joc de nim** és un joc senzill que té un paper important en la teoria de jocs, perquè molts altres jocs es poden jugar amb la mateixa estratègia. Primer ens centrem en nim, i després generalitzem l'estratègia a altres jocs.

En el nim hi ha n piles, i cada pila conté un cert nombre de palets. Els jugadors prenen torns per a triar una pila no buida i treure qualsevol nombre de palets de la pila. El guanyador és el jugador que treu l'últim palet.

Els estats a nim tenen la forma $[x_1, x_2, \dots, x_n]$, on x_k denota el nombre de palets a la pila k . Per exemple, $[10, 12, 5]$ és un joc on hi ha tres piles amb 10, 12 i 5 palets. L'estat $[0, 0, \dots, 0]$ és un estat perdedor, perquè no és possible eliminar cap palet, i és l'únic estat final.

Anàlisi

Resulta que podem classificar fàcilment qualsevol estat del joc de nim calculant la seva **suma de nim** $s = x_1 \oplus x_2 \oplus \dots \oplus x_n$, on \oplus és l'operació xor¹. Els estats la suma de nim quals és 0 són estats perdedors i tots els altres estats són estats guanyadors. Per exemple, la suma de nim de $[10, 12, 5]$ és $10 \oplus 12 \oplus 5 = 3$, de manera que l'estat és un estat guanyador.

Però, com es relaciona la suma de nim amb el joc nim? Podem explicar-ho mirant com canvia la suma de nim quan canviem l'estat del joc.

Estats perdedors: L'estat final $[0, 0, \dots, 0]$ és un estat perdedor, i la seva suma de nim és 0, com s'esperava. En els altres estats perdedors, qualsevol moviment condueix a un estat guanyador, perquè quan un sol valor x_k canvia, la suma de nim també canvia, de manera que la suma de nim és diferent de 0 després del moviment.

Estats guanyadors: Ens podem moure a un estat perdedor si existeix una pila k per a la qual $x_k \oplus s < x_k$. En aquest cas, podem treure palets de la pila k fins que contingui $x_k \oplus s$ palets, i per tant sigui un estat perdedor. Sempre hi ha pila d'aquesta mena, on x_k té un bit 1 a la posició del bit més esquerre de s .

Com a exemple, considereu l'estat $[10, 12, 5]$. Aquest estat és un estat guanyador, perquè la seva suma de nim és 3. Per tant, hi ha d'haver un moviment que condueixi a un estat perdedor. A continuació descobrirem aquest moviment.

La suma de nim de l'estat és la següent:

10		1010
12		1100
5		0101
3		0011

En aquest cas, la pila amb 10 palets és l'única pila que té un bit 1 a la posició del bit més esquerre de la suma de nim:

¹L'estratègia òptima per al joc de nim va ser publicada el 1901 per CL Bouton [10].

10	10 <u>1</u> 0
12	1100
5	0101
3	00 <u>1</u> 1

La nova mida de la pila ha de ser de $10 \oplus 3 = 9$, de manera que només treurem un pal. Després d'això, l'estat resultant és $[9, 12, 5]$, que és un estat perdedor:

9	1001
12	1100
5	0101
0	0000

Joc de Misère

En un **joc de misère**, l'objectiu del joc és oposat, de manera que el jugador que treu l'últim pal perd la partida. Resulta que el joc de misère es pot jugar de manera òptima gairebé com el joc de nim estàndard.

La idea és jugar primer al joc de misère com si fos el joc estàndard, però canviar l'estratègia al final del joc. La nova estratègia s'introdueix en la situació on cada pila conté com a màxim un pal després del següent moviment.

Al joc estàndard, hauríem de triar un moviment que ens deixes amb un nombre parell de piles. Però en el joc de misère triem un moviment que ens deixi un nombre senar de piles amb un palet.

Aquesta estratègia funciona perquè en el joc sempre apareix un estat on es possible canviar l'estratègia, i aquest estat és un estat guanyador, perquè conté exactament una sola pila amb més d'un palet, de manera que la suma de nim no és 0.

25.3 Teorema de Sprague–Grundy

El **Teorema de Sprague–Grundy**² generalitza l'estratègia utilitzada al joc de nim a tots els jocs que compleixen els requisits següents:

- Dos jugadors prenen torns.
- El joc consta d'estats. Els moviments que es poden fer des d'un estat no depenen de quin jugador té el torn.
- El joc acaba quan un jugador no pot fer un moviment.
- Es garanteix que el joc acaba tard o d'hora.
- Els jugadors tenen informació completa sobre els estats i els moviments, i no hi ha aleatorietat.

La idea és calcular per a cada estat del joc un nombre de Grundy que correspongui al nombre de palets en una pila de nim. Quan coneixem els nombres de Grundy de tots els estats, podem jugar el joc com si fos el joc de nim.

²El teorema va ser descobert independentment per R. Sprague [61] i PM Grundy [31].

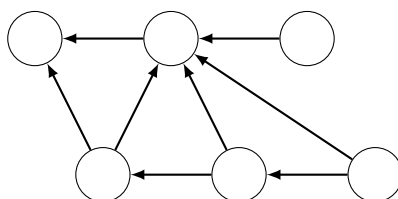
Nombres de Grundy

El **nombre de Grundy** d'un estat del joc és

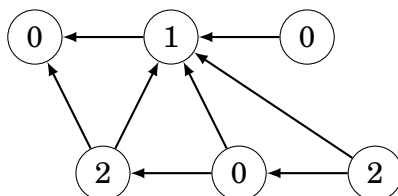
$$\text{mex}(\{g_1, g_2, \dots, g_n\}),$$

on g_1, g_2, \dots, g_n són els nombres de Grundy dels estats als quals ens podem moure, i la funció mex dona el nombre no negatiu més petit que no pertany al conjunt. Per exemple, $\text{mex}(\{0, 1, 3\}) = 2$. Si no hi ha moviments possibles en un estat, el seu nombre de Grundy és $\text{mex}(\emptyset) = 0$.

Per exemple, en el graf d'estats



els nombres de Grundy són els següents:

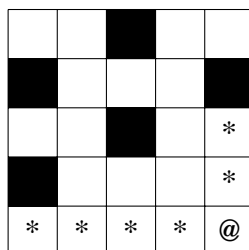


El nombre de Grundy d'un estat perdedor és 0 i el nombre Grundy d'un estat guanyador és un nombre positiu.

El nombre de Grundy d'un estat s'assembla al nombre de palets en una pila del joc de nim³. Si el nombre de Grundy és 0, només ens podem moure a estats els nombres de Grundy dels quals són positius (és a dir, afegir palets), i si el nombre de Grundy és $x > 0$, ens podem moure a estats els nombres de Grundy dels quals inclouen tots els nombres $0, 1, \dots, x - 1$ (és a dir, podem treure qualsevol nombre de palets).

Com a exemple, considereu un joc on els jugadors mouen una figura en un laberint. Cada casella del laberint és buida o paret. En cada torn, el jugador ha de moure la figura uns quants passos cap a l'esquerra o cap amunt. El guanyador del joc és el jugador que fa l'últim moviment.

La imatge següent mostra un possible estat inicial del joc, on @ indica la figura i * indica una casella on es pot moure.



³(N. del T.) Excepte que, a diferència del joc de nim, ara serà possible afegir palets a la pila.

Els estats del joc són totes les caselles buides. Al laberint anterior, els nombres de Grundy són els següents:

0	1		0	1
	0	1	2	
0	2		1	0
	3	0	4	1
0	4	1	3	2

Per exemple, el nombre de Grundy del quadrat inferior dret és 2, de manera que és un estat guanyador. Podem arribar a un estat perdedor i guanyar el joc movent-nos quatre passos cap a l'esquerra o dos passos cap amunt.

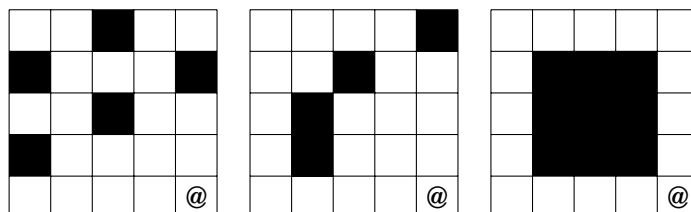
Tingueu en compte que és possible moure's a un estat el nombre de Grundy del qual sigui més gran que el nombre de Grundy de l'estat actual. Tanmateix, aquesta estratègia mai és útil, ja que l'oponent sempre pot triar un moviment que cancel·li aquest moviment.

Subjocs

A continuació, assumirem que el nostre joc consta de subjocs i, a cada torn, el jugador escull primer un subjoc i després un moviment al subjoc. El joc acaba quan no és possible fer cap moviment en cap subjoc.

En aquest cas, el nombre de Grundy d'un joc és la suma de nim dels nombres de Grundy dels subjocs. El joc es pot jugar com un joc de nim calculant tots els nombres de Grundy dels subjocs i després la seva suma de nim.

Com a exemple, considereu un joc que consta de tres laberints. En aquest joc, a cada torn, el jugador tria un dels laberints i després mou la figura al laberint. Suposem que l'estat inicial del joc és el següent:



Els nombres de Grundy dels laberints són els següents:

0	1		0	1
	0	1	2	
0	2		1	0
	3	0	4	1
0	4	1	3	2

0	1	2	3	
1	0		0	1
2		0	1	2
3		1	2	0
4	0	2	5	3

0	1	2	3	4
1				0
2				1
3				2
4	0	1	2	3

A l'estat inicial, la suma de nim dels nombres de Grundy és $2 \oplus 3 \oplus 3 = 2$, de manera que el primer jugador pot guanyar la partida. Un moviment òptim és moure dos passos cap amunt en el primer laberint, que produeix la suma de nim $0 \oplus 3 \oplus 3 = 0$.

El joc de Grundy

De vegades, un moviment en un joc divideix el joc en subjocs que són independents els uns dels altres. En aquest cas, el nombre de Grundy del joc és

$$\text{mex}(\{g_1, g_2, \dots, g_n\}),$$

on n és el nombre de moviments possibles i

$$g_k = a_{k,1} \oplus a_{k,2} \oplus \dots \oplus a_{k,m},$$

on el moviment k genera subjocs amb nombres de Grundy $a_{k,1}, a_{k,2}, \dots, a_{k,m}$.

Un exemple d'aquest joc és **el joc de Grundy**. Inicialment, hi ha un sola pila que conté n sticks. En cada torn, el jugador tria una pila i el divideix en dos piles no buides de manera que les piles siguin de diferent mida. El jugador que fa l'última jugada guanya la partida.

Sigui $f(n)$ el nombre Grundy d'una pila que conté n palets. El nombre de Grundy es pot calcular passant per totes les maneres de dividir la pila en dos piles. Per exemple, quan $n = 8$, les possibilitats són $1 + 7$, $2 + 6$ i $3 + 5$, de manera que

$$f(8) = \text{mex}(\{f(1) \oplus f(7), f(2) \oplus f(6), f(3) \oplus f(5)\}).$$

En aquest joc, el valor de $f(n)$ es basa en els valors de $f(1), \dots, f(n-1)$. Els casos base són $f(1) = f(2) = 0$, perquè no és possible dividir les piles d'1 i 2 palets. Els primers nombres de Grundy són:

$$\begin{aligned} f(1) &= 0 \\ f(2) &= 0 \\ f(3) &= 1 \\ f(4) &= 0 \\ f(5) &= 2 \\ f(6) &= 1 \\ f(7) &= 0 \\ f(8) &= 2 \end{aligned}$$

El nombre de Grundy de $n = 8$ és 2, així que és possible guanyar el joc. El moviment guanyador és crear piles $1 + 7$, perquè $f(1) \oplus f(7) = 0$.

Capítol 26

Algorismes de cadenes

Aquest capítol tracta sobre algorismes eficients per al processament de cadenes. Molts problemes de cadenes es poden resoldre fàcilment en temps $O(n^2)$, però el repte és trobar algorismes que funcionin en temps $O(n)$ o $O(n \log n)$.

Per exemple, un problema fonamental de processament de cadenes és el problema de cercar de patró (**pattern matching**): donada una cadena de longitud n i un patró de longitud m , la nostra tasca és trobar les ocurrencies del patró a la cadena. Per exemple, el patró ABC apareix dues vegades a la cadena ABABCBABC.

El problema de cerca de patrons es pot resoldre fàcilment en temps $O(nm)$ mitjançant un algorisme de força bruta que prova totes les posicions on es pot produir el patró a la cadena. En aquest capítol veurem que hi ha algorismes més eficients que només requereixen temps $O(n + m)$.

26.1 Terminologia de cadenes

Al llarg del capítol, assumim que la indexació en base zero s'utilitza a les cadenes. Així, una cadena s de longitud n consta de caràcters $s[0], s[1], \dots, s[n-1]$. El conjunt de caràcters que poden aparèixer a les cadenes s'anomena **alfabet**. Per exemple, l'alfabet $\{A, B, \dots, Z\}$ consta de les majúscules de l'anglès.

Una **subcadena** és una seqüència de caràcters consecutius en una cadena. Utilitzem la notació $s[a \dots b]$ per referir-nos a una subcadena de s que comença a la posició a i acaba a la posició b . Una cadena de longitud n té $n(n+1)/2$ subcadenaes. Per exemple, les subcadenaes de ABCD són A, B, C, D, AB, BC, CD, ABC, BCD i ABCD.

Una **subseqüència** és una seqüència de caràcters (no necessàriament consecutius) en una cadena en el seu ordre original. Una cadena de longitud n té $2^n - 1$ subseqüències. Per exemple, les subseqüències de ABCD són A, B, C, D, AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD i ABCD.

Un **prefix** és una subcadena que comença al principi d'una cadena, i un **sufix** és una subcadena que acaba al final d'una cadena. Per exemple, els prefixos de ABCD són A, AB, ABC i ABCD, i els sufixos de ABCD són D, CD, BCD i ABCD.

Es pot generar una **rotació** movent els caràcters d'una cadena un per un del principi al final (o viceversa). Per exemple, les rotacions de ABCD són ABCD, BCDA, CDAB i DABC.

Un **període** és un prefix d'una cadena de manera que la cadena es pot construir repetint el període. L'última repetició pot ser parcial i contenir només un prefix del període. Per exemple, el període més curt de ABCABCA és ABC.

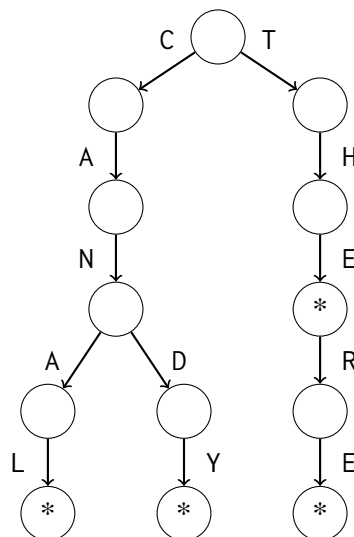
Una **vora** (*border*) és una subcadena que és alhora un prefix i un sufix d'una cadena. Per exemple, les vores de ABACABA són A, ABA i ABACABA.

Les cadenes es comparen utilitzant l'**ordre lexicogràfic** (que correspon a l'ordre alfabètic). Diem que $x < y$ si $x \neq y$ i x és un prefix de y , o si hi ha una posició k tal que $x[i] = y[i]$ quan $i < k$ i $x[k] < y[k]$.

26.2 Tipus Trie

Un **trie**, o arbre de prefixos, és un arbre arrelat que representa un conjunt de cadenes. Cada cadena del conjunt s'emmagatzema com una cadena de caràcters que comença a l'arrel. Si dues cadenes tenen un prefix comú, també tenen una cadena comuna a l'arbre.

Per exemple, considereu el trie següent:



Aquest trie correspon al conjunt {CANAL, CANDY, THE, THERE}. El caràcter * en un node significa que una cadena del conjunt acaba al node. Aquest caràcter és necessari, perquè una cadena pot ser un prefix d'una altra cadena. Per exemple, a la prova anterior, THE és un prefix de THERE.

Podem comprovar en temps $O(n)$ si un trie conté una cadena de longitud n , perquè podem seguir la cadena començant al node arrel. També podem afegir una cadena de longitud n al trie en temps $O(n)$, seguint primer la cadena i després afegint nous nodes al trie si cal.

Si representem un conjunt com a trie, podem trobar, donada una cadena, quin és el prefix més llarg que pertanyi al conjunt. Si afegim informació addicional als nodes del trie també podem calcular el nombre de cadenes del conjunt que tenen un prefix donat.

Un trie es pot emmagatzemar en una matriu

```
int trie[N][A];
```

on N és el nombre màxim de nodes (la longitud total màxima de les cadenes del conjunt) i A és la mida de l'alfabet. Els nodes d'un trie estan numerats $0, 1, 2, \dots$ de manera que el número de l'arrel és 0 , i $\text{trie}[s][c]$ és el següent node de la cadena quan ens movem del node s utilitzant el caràcter c .

26.3 Hashing de cadenes

El **hashing de cadenes** és una tècnica que ens permet comprovar de manera eficient si dues cadenes són iguals¹. La idea del hash de cadenes és comparar els valors hash de les cadenes en lloc dels seus caràcters individuals.

Càlcul dels valors hash

Un **valor hash** d'una cadena és un nombre que es calcula a partir dels caràcters de la cadena. Si dues cadenes són iguals, els seus valors hash també són els mateixos, cosa que permet comparar cadenes en funció dels seus valors hash.

Una manera habitual d'implementar un hash de cadena és amb el **hashing polinomial**, que significa que el valor hash d'una cadena s de longitud n és

$$(s[0]A^{n-1} + s[1]A^{n-2} + \dots + s[n-1]A^0) \bmod B,$$

on $s[0], s[1], \dots, s[n-1]$ són els codis dels caràcters de s , i A i B són constants pre-determinades.

Per exemple, els codis dels caràcters de ALLEY són:

A	L	L	E	Y
65	76	76	69	89

Així, si $A = 3$ i $B = 97$, el valor hash de ALLEY és

$$(65 \cdot 3^4 + 76 \cdot 3^3 + 76 \cdot 3^2 + 69 \cdot 3^1 + 89 \cdot 3^0) \bmod 97 = 52.$$

Preprocessament

Utilitzant el hash polinomial, podem calcular el valor hash de qualsevol subcadena d'una cadena s en temps $O(1)$ després d'un preprocessament de temps $O(n)$. La idea és construir un vector h tal que $h[k]$ contingui el valor hash del prefix $s[0 \dots k]$. Els valors del vector es poden calcular recursivament de la següent manera:

$$\begin{aligned} h[0] &= s[0] \\ h[k] &= (h[k-1]A + s[k]) \bmod B \end{aligned}$$

¹La tècnica va ser popularitzada per l'algorisme de cerca de patrons Karp-Rabin [42].

A més, construïm un vector p on $p[k] = A^k \bmod B$:

$$\begin{aligned} p[0] &= 1 \\ p[k] &= (p[k-1]A) \bmod B. \end{aligned}$$

La construcció d'aquestes matrius requereix temps $O(n)$. Després d'això, el valor hash de qualsevol subcadena $s[a \dots b]$ es pot calcular en temps $O(1)$ mitjançant la fórmula

$$(h[b] - h[a-1]p[b-a+1]) \bmod B$$

suposant que $a > 0$. Si $a = 0$, el valor hash és simplement $h[b]$.

Ús dels valors hash

Podem comparar cadenes de manera eficient mitjançant valors hash. En lloc de comparar els caràcters individuals de les cadenes, la idea és comparar els seus valors hash. Si els valors hash són iguals, les cadenes són *probablement* iguals, i si els valors hash són diferents, les cadenes són *certament* diferents.

Fent servir hashing, sovint podem fer eficient un algorisme de força bruta. Per exemple, considerem el problema de cerca de patrons: donada una cadena s i un patró p , trobem les posicions on es produeix p a s . Un algorisme de força bruta passa per totes les posicions on es pot produir p i compara les cadenes caràcter per caràcter. La complexitat temporal d'aquest algorisme és $O(n^2)$.

Podem fer que l'algorisme de força bruta sigui més eficient utilitzant hashing, perquè l'algorisme compara subcadena de cadenes. Utilitzant el hash, cada comparació només triga $O(1)$ temps, perquè només es comparen els valors hash de les subcadena. Això resulta en un algorisme amb complexitat temporal $O(n)$, que és la millor complexitat temporal possible per a aquest problema.

Combinant hashing i *cerca binària*, també és possible esbrinar l'ordre lexicogràfic de dues cadenes en temps logarítmic. Això es pot fer calculant la longitud del prefix comú de les cadenes mitjançant la cerca binària. Un cop sabem la longitud del prefix comú, només podem comprovar el caràcter següent després del prefix, perquè això determina l'ordre de les cadenes.

Col·lisions i paràmetres

Un risc evident quan es comparen els valors hash és una **col·lisió**, que vol dir que dues cadenes tenen continguts diferents però valors hash iguals. En aquest cas, un algorisme que es basa en els valors hash conclou que les cadenes són iguals, però en realitat no ho són, i l'algorisme pot donar resultats incorrectes.

Les col·lisions sempre són possibles, perquè el nombre de cadenes diferents és més gran que el nombre de valors hash diferents. Tanmateix, la probabilitat d'una col·lisió és petita si es trien amb cura les constants A i B . Una manera habitual és triar constants aleatòries properes a 10^9 , per exemple de la següent manera:

$$\begin{aligned} A &= 911382323 \\ B &= 972663749 \end{aligned}$$

Utilitzant aquestes constants, el tipus `long` es pot fer servir quan es calculen els valors hash, perquè els productes AB i BB s'ajustaran a `long`. Però n'hi ha prou amb tenir 10^9 valors hash diferents?

Considerem tres escenaris en què es pot utilitzar el hash:

Escenari 1: Les cadenes x i y es comparen entre si. La probabilitat d'una col·lisió és $1/B$ assumint que tots els valors hash són igualment probables.

Escenari 2: Es compara una cadena x amb les cadenes y_1, y_2, \dots, y_n . La probabilitat d'una o més col·lisions és

$$1 - \left(1 - \frac{1}{B}\right)^n.$$

Escenari 3: Tots els parells de cadenes x_1, x_2, \dots, x_n es comparen entre si. La probabilitat d'una o més col·lisions és

$$1 - \frac{B \cdot (B-1) \cdot (B-2) \cdots (B-n+1)}{B^n}.$$

La taula següent mostra les probabilitats de col·lisió quan $n = 10^6$ i el valor de B varia:

constant B	scenario 1	scenario 2	scenario 3
10^3	0.001000	1.000000	1.000000
10^6	0.000001	0.632121	1.000000
10^9	0.000000	0.001000	1.000000
10^{12}	0.000000	0.000000	0.393469
10^{15}	0.000000	0.000000	0.000500
10^{18}	0.000000	0.000000	0.000001

La taula mostra que a l'escenari 1, la probabilitat d'una col·lisió és insignificant quan $B \approx 10^9$. En l'escenari 2, és possible una col·lisió, però la probabilitat és encara bastant petita. Tanmateix, a l'escenari 3 la situació és molt diferent: gairebé sempre es produirà una col·lisió quan $B \approx 10^9$.

El fenomen de l'escenari 3 es coneix com la **paradoxa de l'aniversari**: si hi ha n persones a una habitació, la probabilitat que *algun* parell de persones tingui el mateix aniversari és gran encara que n sigui bastant petit. En hash, en conseqüència, quan es comparen tots els valors hash entre si, la probabilitat que uns dos valors hash siguin iguals és gran.

Podem reduir la probabilitat d'una col·lisió calculant *múltiples* valors hash amb paràmetres distints, ja que és poc probable que es produeixi una col·lisió en tots els valors hash al mateix temps. Per exemple, dos valors hash amb el paràmetre $B \approx 10^9$ corresponen a un valor hash amb el paràmetre $B \approx 10^{18}$, la qual cosa fa que la probabilitat d'una col·lisió sigui molt petita.

Algunes persones utilitzen constants $B = 2^{32}$ i $B = 2^{64}$, cosa que és convenient, perquè les operacions amb enters de 32 i 64 bits es calculen mòdul 2^{32} i 2^{64} . Tanmateix, aquesta *no* és una bona opció, perquè és possible construir entrades que sempre generin col·lisions quan s'utilitzen constants de la forma 2^x [51].

26.4 Algorisme Z

El **Z-vector** z d'una cadena s de longitud n conté per a cada $k = 0, 1, \dots, n-1$ la longitud de la subcadena més llarga de s que comença a la posició k i és un prefix de s . Així, $z[k] = p$ ens diu que $s[0 \dots p-1]$ és igual a $s[k \dots k+p-1]$. Molts problemes de processament de cadenes es poden resoldre de manera eficient mitjançant el Z-vector.

Per exemple, el Z-vector de ACBACDACBACBACDA és el següent:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	7	0	0	2	0	0	1

En aquest cas, per exemple, $z[6] = 5$, perquè la subcadena ACBAC de longitud 5 és un prefix de s , però la subcadena ACBACB de La longitud 6 no és un prefix de s .

Descripció de l'algorisme

A continuació es descriu un algorisme, anomenat **algorisme Z**², que construeix de manera eficient el Z-vector en temps $O(n)$ temps. L'algorisme calcula els valors del Z-vector d'esquerra a dreta fent servir la informació ja emmagatzemada al Z-vector i comparant subcadena caràcter per caràcter.

Per calcular de manera eficient els valors del Z-vector, l'algorisme manté un rang $[x, y]$ tal que $s[x \dots y]$ és un prefix de s i y és tan gran com sigui possible. Com que sabem que $s[0 \dots y-x]$ i $s[x \dots y]$ són iguals, podem fer servir aquesta informació per calcular els valors Z per a les posicions $x+1, x+2, \dots, y$.

A cada posició k , primer comprovem el valor de $z[k-x]$. Si $k+z[k-x] < y$, sabem que $z[k] = z[k-x]$. Tanmateix, si $k+z[k-x] \geq y$, aleshores $s[0 \dots y-k]$ és igual a $s[k \dots y]$, i per determinar la valor de $z[k]$ hem de comparar les subcadena caràcter per caràcter. Tot i així, l'algorisme funciona en temps $O(n)$, perquè comencem a comparar a les posicions $y-k+1$ i $y+1$.

Per exemple, construïm el Z-vector següent:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Després de calcular el valor $z[6] = 5$, l'interval actual de $[x, y]$ és $[6, 10]$:

						$\begin{array}{c} x \qquad \qquad \qquad y \\ \hline \end{array}$									
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	?	?	?	?	?	?	?	?	?

²L'algorisme Z es va presentar a [32] com el mètode conegut més senzill per a la concordança de patrons en temps lineal, i la idea original es va atribuir a citeimai84.

Ara podem calcular els valors posteriors del Z-vector de manera eficient, perquè sabem que $s[0 \dots 4]$ i $s[6 \dots 10]$ són iguals. En primer lloc, com que $z[1] = z[2] = 0$, de seguida sabem que també $z[7] = z[8] = 0$:

						x		y							
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	?	?	?	?	?	?	?

Aleshores, com que $z[3] = 2$, sabem que $z[9] \geq 2$:

						x		y							
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	?	?	?	?	?	?	?

Tanmateix, no tenim informació sobre la cadena després de la posició 10, de manera que hem de comparar les subcadena caràcter per caràcter:

						x		y							
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	?	?	?	?	?	?	?

Resulta que $z[9] = 7$, de manera que el nou rang $[x, y]$ és $[9, 15]$:

									x		y				
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	7	?	?	?	?	?	?

Després d'això, es poden determinar tots els valors restants del Z-vector fent servir la informació ja emmagatzemada al Z-vector:

									x		y				
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	C	B	A	C	D	A	C	B	A	C	B	A	C	D	A
-	0	0	2	0	0	5	0	0	7	0	0	2	0	0	1

Ús del Z-vector

Sovint és qüestió de gustos si s'ha d'utilitzar string hashing o l'algorisme Z. A diferència del hashing, l'algorisme Z sempre funciona i no hi ha cap risc de col·lisions. D'altra banda, l'algorisme Z és més difícil d'implementar i alguns problemes només es poden resoldre mitjançant hashing.

Com a exemple, considerem de nou el problema de cerca de patrons, on la nostra tasca és trobar les ocurrencies d'un patró p en una cadena s . Ja hem resolt aquest problema de manera eficient utilitzant el hashing de cadena, però l'algorisme Z proporciona una altra manera de resoldre el problema.

Una idea habitual en el processament de cadenes és construir una cadena que consta de múltiples cadenes separades per caràcters especials. En aquest problema, podem construir una cadena $p\#s$, on p i s estan separats per un caràcter especial $\#$ que no apareix a les cadenes. El Z-vector de $p\#s$ ens indica les posicions on p apareix a s , perquè aquestes posicions contenen la longitud de p .

Per exemple, si $s = \text{HATTIVATTI}$ i $p = \text{ATT}$, el Z-vector és els següent:

0	1	2	3	4	5	6	7	8	9	10	11	12	13
A	T	T	#	H	A	T	T	I	V	A	T	T	I
–	0	0	0	0	3	0	0	0	0	3	0	0	0

Les posicions 5 i 10 contenen el valor 3, la qual cosa significa que el patró ATT apareix a les posicions corresponents de HATTIVATTI.

La complexitat temporal de l'algorisme resultant és lineal, perquè n'hi ha prou per construir el Z-vector i repassar els seus valors.

Implementació

Aquí hi ha una breu implementació de l'algorisme Z que retorna el Z-vector.

```
vector<int> z(string s) {
    int n = s.size();
    vector<int> z(n);
    int x = 0, y = 0;
    for (int i = 1; i < n; i++) {
        z[i] = max(0, min(z[i-x], y-i+1));
        while (i+z[i] < n && s[z[i]] == s[i+z[i]]) {
            x = i; y = i+z[i]; z[i]++;
        }
    }
    return z;
}
```

Capítol 27

Algorismes d'arrel quadrada

Un **algorisme d'arrel quadrada** és un algorisme que té una arrel quadrada en la seva complexitat temporal. Una arrel quadrada es pot veure com un "logaritme de pobretons": la complexitat $O(\sqrt{n})$ és millor que $O(n)$ però pitjor que $O(\log n)$. En qualsevol cas, molts algorismes d'arrel quadrada són ràpids i utilitzables a la pràctica.

Per exemple, considereu el problema de crear una estructura de dades que admeti dues operacions en vector: modificar un element en una posició determinada i calcular la suma d'elements en un interval donat. Prèviament hem resolt el problema fent servir arbres binaris indexats i segmentats, que admeten ambdues operacions en temps $O(\log n)$. Tanmateix, ara resoldrem el problema d'una altra manera fent servir una estructura d'arrel quadrada que ens permeti modificar elements en temps $O(1)$ i calcular sumes en temps $O(\sqrt{n})$.

La idea és dividir el vector en *blocs* de mida \sqrt{n} de manera que cada bloc contingui la suma d'elements dins del bloc. Per exemple, un vector de 16 elements es dividirà en blocs de 4 elements de la següent manera:

21				17				20				13			
5	8	6	3	2	7	2	6	7	1	7	5	6	2	3	2

En aquesta estructura, és fàcil modificar els elements del vector, perquè només cal actualitzar la suma d'un sol bloc després de cada modificació, i això es pot fer en temps $O(1)$. Per exemple, la imatge següent mostra com canvia el valor d'un element i la suma del bloc corresponent:

21				15				20				13			
5	8	6	3	2	5	2	6	7	1	7	5	6	2	3	2

Aleshores, per calcular la suma dels elements d'un rang, dividim el rang en tres parts de manera que la suma consta de valors d'elements individuals i sumes de blocs entre ells:

21				15				20				13			
5	8	6	3	2	5	2	6	7	1	7	5	6	2	3	2

Com que el nombre d'elements individuals és $O(\sqrt{n})$ i el nombre de blocs també és $O(\sqrt{n})$, calcular la suma triga temps $O(\sqrt{n})$. Per què triem \sqrt{n} com a mida de bloc? Perquè és la mida que *equilibra* ambdues operacions: el vector es divideix en \sqrt{n} blocs, cadascun dels quals conté \sqrt{n} elements.

A la pràctica, no és necessari triar el valor exacte de \sqrt{n} , i en canvi podem fer servir k i n/k on k és diferent de \sqrt{n} . El paràmetre òptim depèn del problema i de l'entrada. Per exemple, si un algorisme passa sovint pels blocs però rarament inspecciona elements únics dins dels blocs, seria preferible triar $k < \sqrt{n}$ blocs, cadascun dels quals conté $n/k > \sqrt{n}$ elements.

27.1 Combinació d'algorismes

En aquesta secció discutim dos algorismes d'arrel quadrada que es basen en combinar dos algorismes en un de sol. En ambdós casos, podríem fer servir qualsevol dels algorismes sense l'altre i resoldre el problema en temps $O(n^2)$. Tanmateix, combinant els algorismes, el temps d'execució és només $O(n\sqrt{n})$.

Processament per casos

Suposem que se'ns dona una taula bidimensional que conté n cel·les. Cada cel·la té una lletra assignada, i la nostra tasca és trobar dues cel·les amb la mateixa lletra la distància de les quals sigui mínima, on la distància entre les cel·les (x_1, y_1) i (x_2, y_2) és $|x_1 - x_2| + |y_1 - y_2|$. Per exemple, considereu la taula següent:

A	F	B	A
C	E	G	E
B	D	A	F
A	C	B	D

En aquest cas, la distància mínima entre les dues lletres "E" és 2.

Podem resoldre el problema considerant cada lletra per separat. Amb aquest enfocament, el nou problema és calcular la distància mínima entre dues cel·les amb una lletra *fixa* c . Ens centrem en dos algorismes que resolen aquest problema:

Algorisme 1: Recorre tots els parells de cel·les amb la lletra c i calcula la distància mínima entre aquestes cel·les. Això triga $O(k^2)$, on k és el nombre de cel·les amb la lletra c .

Algorisme 2: Realitza una cerca d'amplada que comenci simultàniament a cada cel·la amb la lletra c . Trobar la distància mínima entre dues cel·les amb la lletra c triga temps $O(n)$.

Una manera de resoldre el problema és escollir qualsevol dels algorismes i utilitzar-lo per a totes les lletres. Si fem servir l'algorisme 1, el temps d'execució és $O(n^2)$, perquè totes les cel·les poden contenir la mateixa lletra, i en aquest cas $k = n$. A més, si fem servir l'algorisme 2, el temps d'execució és $O(n^2)$, perquè totes les cel·les poden tenir lletres diferents, i en aquest cas necessitem n cerques.

Tanmateix, podem *combinar* els dos algorismes i utilitzar algorismes diferents per a lletres diferents en funció de quantes vegades apareix cada lletra a la taula. Suposem que una lletra c apareix k vegades. Si $k \leq \sqrt{n}$, fem servir l'algorisme 1, i si $k > \sqrt{n}$, fem servir l'algorisme 2. Veurem ara que si fem això, el temps total d'execució de l'algorisme és només $O(n\sqrt{n})$.

Primer, suposem que fem servir l'algorisme 1 per a una lletra c . Com que c apareix com a màxim \sqrt{n} vegades a la taula, comparem cada cel·la amb la lletra c com a molt $O(\sqrt{n})$ vegades amb altres cel·les. Així, el temps utilitzat per processar les com a molt n caselles amb lletres on es fa servir l'algorisme 1 és $O(n\sqrt{n})$. Ara, suposem que fem servir l'algorisme 2 per a una lletra c . Hi ha com a màxim \sqrt{n} d'aquestes lletres, de manera que processar aquestes lletres també triga $O(n\sqrt{n})$ temps.

Processament per lots

El nostre següent problema també tracta d'una taula bidimensional que conté n cel·les. Inicialment, cada cel·la excepte una és blanca. Realitzem $n - 1$ operacions, cadascuna de les quals calcula primer la distància mínima des d'una cel·la blanca donada a una cel·la negra, i després pinta la cel·la blanca de negra.

Per exemple, considereu l'operació següent:

		*	

Primer, calculem la distància mínima des de la cel·la blanca marcada amb * fins a una cel·la negra. La distància mínima és 2, perquè podem moure dos passos a l'esquerra fins a una cel·la negra. Aleshores, pintem la cel·lula blanca de negra:

Considereu els dos algorismes següents:

Algorisme 1: Fem cerca en amplada per calcular per a cada cel·la blanca la distància a la cel·la negra més propera. Això triga temps $O(n)$, i després de la cerca, podem trobar la distància mínima des de qualsevol cel·la blanca a una cel·la negra en temps $O(1)$.

Algorisme 2: Mantenim una llista de cel·les pintades de negra. Per a cada operació, recorrem la llista per trobar la cel·la negra més propera i, a continuació, afegim una nova cel·la a la llista. Això triga $O(k)$ on k és la longitud de la llista.

Podem combinar els algorismes anteriors dividint les operacions en $O(\sqrt{n})$ lots (batches), cadascun dels quals consta de $O(\sqrt{n})$ operacions. Al començament de cada lot, fem servir l'algorisme 1. A continuació, utilitzem l'algorisme 2 per processar les operacions del lot. Netegem la llista de l'algorisme 2 entre els lots. En cada operació, la distància mínima a una cel·la negra és la distància calculada per l'algorisme 1 o la distància calculada per l'algorisme 2.

L'algorisme resultant triga temps $O(n\sqrt{n})$. Primer, l'algorisme 1 s'executa $O(\sqrt{n})$ vegades, i cada cerca triga temps $O(n)$. En segon lloc, quan s'utilitza l'algorisme 2 en un lot, la llista conté $O(\sqrt{n})$ cel·les (ja que netegem la llista entre els lots) i cada operació triga temps $O(\sqrt{n})$.

27.2 Particions senceres

Alguns algorismes d'arrel quadrada es basen en la següent observació: si un nombre enter positiu n es representa com a suma d'enters positius, aquesta suma sempre conté com a màxim $O(\sqrt{n})$ nombres *diferents*. La raó d'això és que per a construir una suma que contingui un nombre màxim de nombres diferents, hauríem de triar nombres *petits*. Si triem els nombres $1, 2, \dots, k$, la suma resultant és

$$\frac{k(k+1)}{2}.$$

Així, la quantitat màxima de nombres diferents és $k = O(\sqrt{n})$. A continuació mostrem dos problemes que es poden resoldre de manera eficient amb aquesta observació.

Motxilla

Suposem que se'ns dona una llista de pesos enters la suma dels quals és n . La nostra tasca és esbrinar totes les sumes que es poden formar mitjançant un subconjunt de pesos. Per exemple, si els pesos són $\{1, 3, 3\}$, les sumes possibles són les següents:

- 0 (empty set)
- 1
- 3
- $1 + 3 = 4$
- $3 + 3 = 6$
- $1 + 3 + 3 = 7$

Utilitzant l'enfocament estàndard de la motxilla (vegeu el capítol 7.4), el problema es pot resoldre de la següent manera: definim una funció $\text{possible}(x, k)$ el valor de la qual és 1 si la suma x es pot formar mitjançant els primers k pesos, i 0 en cas contrari. Com que la suma dels pesos és n , hi ha com a màxim n pesos i tots els valors de la funció es poden calcular en temps $O(n^2)$ mitjançant la programació dinàmica.

Tanmateix, podem fer que l'algorisme sigui més eficient utilitzant el fet que com a màxim hi ha $O(\sqrt{n})$ pesos *diferents*. Així, podem processar els pesos en

grups que consisteixen en pesos similars. Podem processar cada grup en temps $O(n)$, i això produeix un algorisme de temps $O(n\sqrt{n})$.

La idea és utilitzar un vector que enregistri les sumes de pesos que es poden formar utilitzant els grups processats fins ara. El vector conté n elements: l'element k és 1 si es pot formar la suma k i 0 en cas contrari. Per processar un grup de pesos, escanegem el vector d'esquerra a dreta i registrem les noves sumes de pesos que es poden formar utilitzant aquest grup i els grups anteriors¹

Construcció de cadenes

Donada una cadena s de longitud n i un conjunt de cadenes D la longitud total de les quals és m , considereu el problema de comptar el nombre de maneres en què s es pot formar com una concatenació de cadenes en D . Per exemple, si $s = \text{ABAB}$ i $D = \{\text{A}, \text{B}, \text{AB}\}$, hi ha 4 maneres:

- $\text{A} + \text{B} + \text{A} + \text{B}$
- $\text{AB} + \text{A} + \text{B}$
- $\text{A} + \text{B} + \text{AB}$
- $\text{AB} + \text{AB}$

Podem resoldre el problema mitjançant la programació dinàmica: Sigui $\text{count}(k)$ el nombre de maneres de construir el prefix $s[0 \dots k]$ utilitzant les cadenes de D . Ara $\text{count}(n-1)$ dóna la resposta al problema, i podem resoldre el problema en $O(n^2)$ temps utilitzant una estructura trie.

No obstant això, podem resoldre el problema de manera més eficient mitjançant l'ús d'un hash de cadenes i el fet que com a màxim hi ha $O(\sqrt{m})$ longituds de cadenes diferents a D . Primer, construïm un conjunt H que conté tots els valors hash de les cadenes de D . Aleshores, quan calculem un valor de $\text{count}(k)$, passem per tots els valors de p de manera que hi hagi una cadena de longitud p a D , calculem el valor hash de $s[k-p+1 \dots k]$ i comprovem si pertany a H . Com que hi ha com a màxim $O(\sqrt{m})$ longituds de cadena diferents, això resulta en un algorisme de temps $O(n\sqrt{m})$.

27.3 Algorisme de Mo

Algorisme de Mo² es pot utilitzar en molts problemes que requereixen processar consultes d'interval en un vector *estàtic*, és a dir, un vector amb valors que no canvien entre consultes. En cada consulta, se'ns dóna un rang $[a, b]$, i hauríem de calcular un valor basant-nos en els elements del vector entre les posicions a i b .

¹(N. del T.) Suposem que tenim un nou grup de t pesos de pes x . Creem un vector v de n elements, l'element $v[k]$ del qual és m si es pot formar la suma k amb els pesos anteriors, m pesos del nou grup, i no és possible fer-ho amb $m-1$ pesos, i $v[k] = \infty$ si no es pot formar la suma k de cap manera. En particular, $v[k] = 0$ si k es podia formar abans, i $v[k] = v[k-x] + 1$ altrament. Amb això podem omplir aquest vector d'esquerra a dreta en temps $O(n)$. Les sumes k que ens interessin són aquelles que compleixen $v[k] \leq t$.

²Segons [12], aquest algorisme rep el nom de Mo Tao, un programador competitiu xinès, però la tècnica ha aparegut anteriorment a la literatura [44].

Com que el vector és estàtic, les consultes es poden processar en qualsevol ordre i l'algorisme de Mo processa les consultes en un ordre especial que garanteix que l'algorisme funciona de manera eficient.

L'algorisme de Mo manté un *rang actiu* de la matriu, i la resposta a una consulta sobre l'interval actiu es coneix en cada moment. L'algorisme processa les consultes una per una i sempre mou els punts finals de l'interval actiu inserint i eliminant elements. La complexitat temporal de l'algorisme és $O(n\sqrt{n}f(n))$ on la matriu conté n elements, hi ha n consultes i cada inserció i eliminació d'un element triga temps $O(f(n))$.

El truc de l'algorisme de Mo és l'ordre en què es processen les consultes: el matriu es divideix en blocs de $k = O(\sqrt{n})$ elements i una consulta $[a_1, b_1]$ es processa abans d'una consulta $[a_2, b_2]$ si es compleix

- $\lfloor a_1/k \rfloor < \lfloor a_2/k \rfloor$ o
- $\lfloor a_1/k \rfloor = \lfloor a_2/k \rfloor$ i $b_1 < b_2$.

Així, totes les consultes els extrems esquerres de les quals es troben en un bloc determinat es processen una darrere l'altra ordenades segons els seus extrems dret. Utilitzant aquest ordre, l'algorisme només realitza $O(n\sqrt{n})$ operacions, perquè el punt final esquerre es mou $O(n)$ vegades $O(\sqrt{n})$ passos, i el punt final dret es mou $O(\sqrt{n})$ vegades $O(n)$ passos. Així, ambdós extrems mouen un total de $O(n\sqrt{n})$ passos durant l'algorisme.

Exemple

Com a exemple, considerem un problema on se'ns dona un conjunt de consultes, cadascuna d'elles corresponent a un interval del vector, i la nostra tasca és calcular per a cada consulta el nombre d'elements *diferents* de l'interval.

En l'algorisme de Mo, les consultes sempre s'ordenen de la mateixa manera, però segons el problema mantenim la resposta d'una manera o d'altra. En aquest problema, mantenim un vector count on $\text{count}[x]$ indica el nombre de vegades que apareix un element x a l'interval actiu.

Quan passem d'una consulta a una altra, l'interval actiu canvia. Per exemple, si l'interval actual és

4	2	5	4	2	4	3	3	4
---	---	---	---	---	---	---	---	---

i el següent rang és

4	2	5	4	2	4	3	3	4
---	---	---	---	---	---	---	---	---

hi haurà tres passos: el punt final esquerre es mou un pas cap a la dreta i el punt final dret es mou dos passos cap a la dreta.

Després de cada pas, el vector count s'ha d'actualitzar. Després d'afegir un element x , augmentem el valor de $\text{count}[x]$ en 1, i si $\text{count}[x] = 1$ després d'això, també augmentem la resposta a la consulta en 1. De la mateixa manera, després

d'eliminar un element x , disminuïm el valor de $\text{count}[x]$ en 1, i si $\text{count}[x] = 0$ després d'això, també disminueix la resposta a la consulta en 1.

En aquest problema, el temps necessari per realitzar cada pas és $O(1)$, de manera que l'algorisme triga temps $O(n\sqrt{n})$.

Capítol 28

Arbres de segments revisats

Un arbre de segment és una estructura de dades versàtil que es pot fer servir per a resoldre un gran nombre de problemes algorísmics. Tanmateix, hi ha molts temes relacionats amb els arbres de segments que encara no hem tocat. Ara és el moment de parlar d'algunes variants d'arbres de segments més avançades.

Fins ara, hem implementat les operacions d'un arbre de segments caminant *bottom-up* (de baix cap amunt) de l'arbre. Per exemple, hem calculat les sumes d'interval de la manera següent (Capítol 9.3):

```
int sum(int a, int b) {
    a += n; b += n;
    int s = 0;
    while (a <= b) {
        if (a%2 == 1) s += tree[a++];
        if (b%2 == 0) s += tree[b--];
        a /= 2; b /= 2;
    }
    return s;
}
```

Tanmateix, en els arbres de segments més avançats, sovint és necessari implementar les operacions de manera *top-down* (de dalt a baix). Amb aquest enfocament, la funció es converteix en:

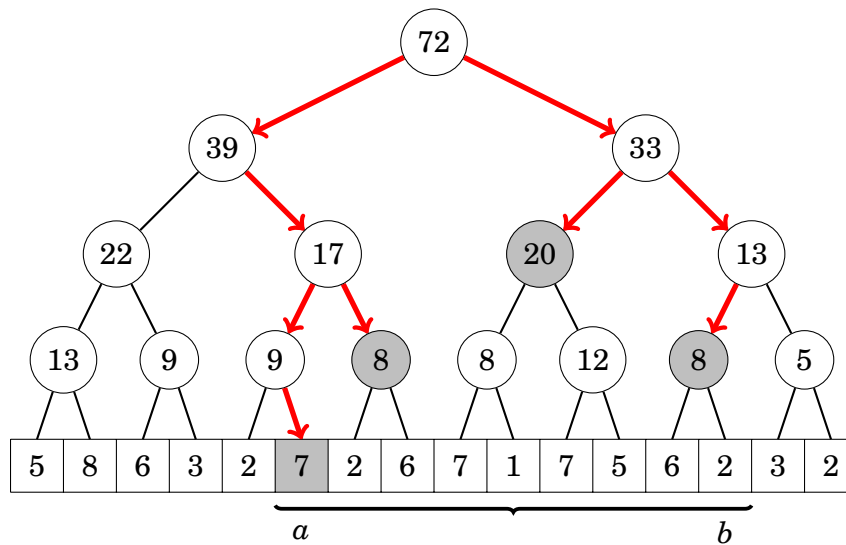
```
int sum(int a, int b, int k, int x, int y) {
    if (b < x || a > y) return 0;
    if (a <= x && y <= b) return tree[k];
    int d = (x+y)/2;
    return sum(a,b,2*k,x,d) + sum(a,b,2*k+1,d+1,y);
}
```

Ara podem calcular qualsevol valor de $\text{sum}_q(a, b)$ (la suma dels valors del vector en el rang $[a, b]$) de la següent manera:

```
int s = sum(a, b, 1, 0, n-1);
```

El paràmetre k indica la posició actual a arbre. Inicialment k és igual a 1, perquè comencem per l'arrel de l'arbre. L'interval $[x, y]$ es correspon amb k i inicialment és $[0, n - 1]$. Quan es calcula la suma, si $[x, y]$ està fora de $[a, b]$, la suma és 0, i si $[x, y]$ està completament dins de $[a, b]$, la suma es troba en tree. Si $[x, y]$ es troba parcialment dins de $[a, b]$, la cerca continua recursivament a la meitat esquerra i dreta de $[x, y]$. La meitat esquerra és $[x, d]$ i la meitat dreta $[d + 1, y]$ on $d = \lfloor \frac{x+y}{2} \rfloor$.

La imatge següent mostra com es desenvolupa la cerca quan es calcula el valor de $\text{sum}_q(a, b)$. Els nodes grisos indiquen nodes on s'atura la recursió i la suma es troba en tree.



En aquesta implementació les operacions també triguen temps $O(\log n)$ temps, perquè el nombre total de nodes visitats és $O(\log n)$.

28.1 Propagació mandrosa

Utilitzant **propagació mandrosa** (lazy), podem construir un arbre de segments que admeti actualitzacions d'interval i consultes d'interval en temps $O(\log n)$. La idea és fer les actualitzacions i consultes de dalt a baix i fer les actualitzacions de manera *mandrosa*, de manera que es propaguin per l'arbre només quan sigui necessari.

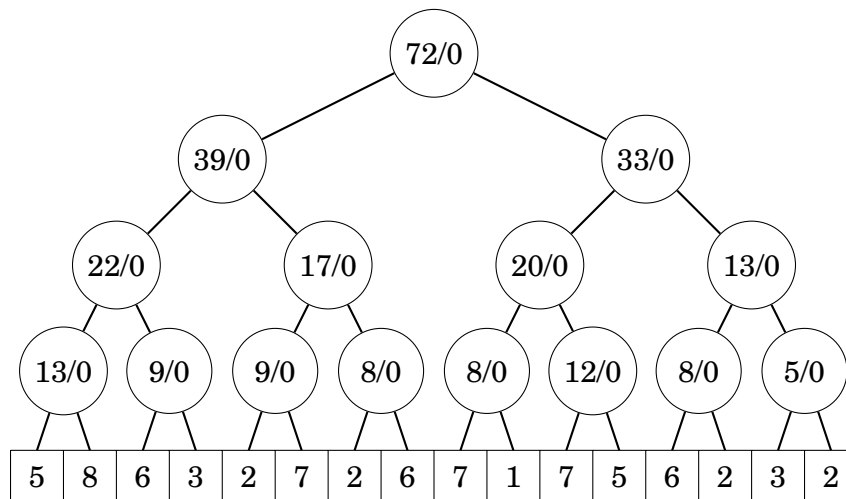
Els nodes d'un arbre de segments mandrós contenen dos tipus d'informació. Com en un arbre de segments ordinari, cada node conté la suma o algun altre valor relacionat amb el subvector corresponent. A més, el node pot contenir informació relacionada amb les actualitzacions mandroses, que no s'ha propagat encara als seus fills.

Hi ha dos tipus d'actualitzacions d'interval: cada valor de l'interval s'*augmenta* en una certa quantitat o se li *assigna* algun valor. Ambdues operacions es poden implementar amb idees similars, i fins i tot és possible construir un arbre que suporti ambdues operacions alhora.

Arbre de segments mandrós

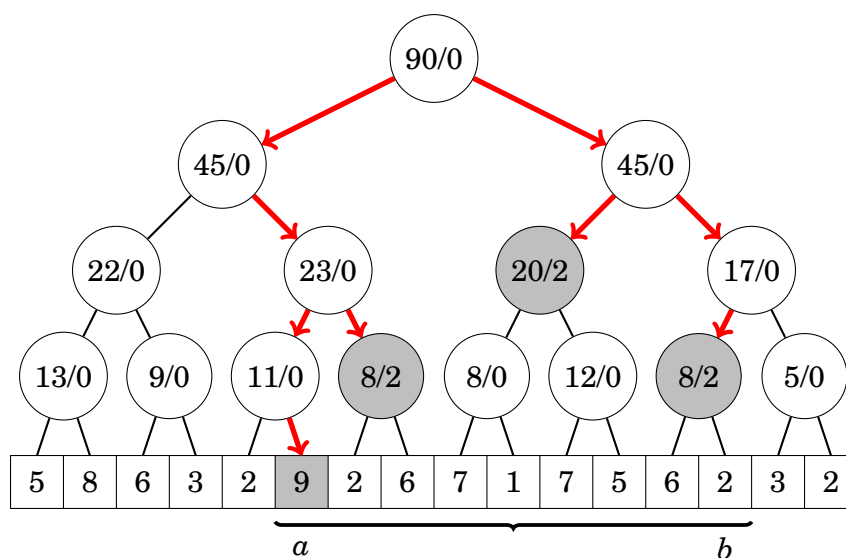
Considerem un exemple on el nostre objectiu és construir un arbre de segments que admeti dues operacions: augmentar cada valor de $[a, b]$ per una constant i calcular la suma de valors en $[a, b]$.

Construïm un arbre on cada node té dos valors s/z : s és la suma de valors de l'interval i z és el valor d'una actualització mandrosa, el que significa que tots els valors de l'interval haurien d'augmentar en z . En l'arbre següent es té $z = 0$ a tots els nodes, de manera que no hi ha actualitzacions mandroses en curs.

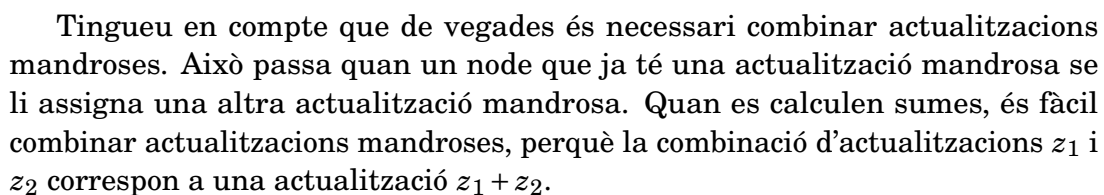


Quan els elements de $[a, b]$ s'incrementen en u , anem des de l'arrel cap a les fulles i modifiquem els nodes de l'arbre de la següent manera. Si el rang $[x, y]$ d'un node és completament dins de $[a, b]$, augmentem el valor z del node en u i ens aturem. Si $[x, y]$ només pertany parcialment a $[a, b]$, augmentem el valor s del node en hu , on h és la mida de la intersecció de $[a, b]$ i $[x, y]$, i continuem la nostra caminada recursivament per l'arbre.

Per exemple, la imatge següent mostra l'arbre després d'augmentar els elements de $[a, b]$ en 2:



La imatge següent mostra com canvia l'arbre quan calculem el valor de $\text{sum}_a(a, b)$. El rectangle mostra els nodes els valors dels quals canvien, perquè una actualització mandrosa es propaga cap avall.



Les actualitzacions mandroses es poden generalitzar de manera que sigui possible actualitzar els intervals amb polinomis de la forma

En aquest cas, l'actualització d'un valor a la posició i a $[a, b]$ és $p(i - a)$. Per exemple, afegir el polinomi $p(u) = u + 1$ a $[a, b]$ vol dir que el valor a la posició a augmenta en 1, el valor a la posició $a + 1$ augmenta en 2, etcètera.

276

Ara, la suma de valors d'un interval $[x, y]$ és igual

$$s + \sum_{u=0}^{y-x} z_k u^k + z_{k-1} u^{k-1} + \dots + z_0.$$

El valor d'aquesta suma es pot calcular de manera eficient amb fórmules. Per exemple, el terme z_0 correspon a la suma $(y-x+1)z_0$ i el terme $z_1 u$ correspon a la suma

$$z_1(0 + 1 + \dots + y - x) = z_1 \frac{(y-x)(y-x+1)}{2}.$$

Quan es propaga una actualització a l'arbre, els índexs de $p(u)$ canvien, perquè en cada rang $[x, y]$, els valors es calculen per a $u = 0, 1, \dots, y - x$. Tanmateix, això no és un problema, perquè $p'(u) = p(u + h)$ és un polinomi d'igual grau que $p(u)$. Per exemple, si $p(u) = t_2 u^2 + t_1 u - t_0$, aleshores

$$p'(u) = t_2(u + h)^2 + t_1(u + h) - t_0 = t_2 u^2 + (2ht_2 + t_1)u + t_2 h^2 + t_1 h - t_0.$$

28.2 Arbres dinàmics

Un arbre de segments normal és estàtic, la qual cosa vol dir que cada node té una posició fixa al vector i l'arbre requereix una quantitat fixa de memòria. En un **arbre de segments dinàmic**, la memòria s'assigna només per als nodes als quals s'accedeix realment durant l'algorisme, cosa que pot estalviar una gran quantitat de memòria.

Els nodes d'un arbre dinàmic es poden representar com a estructures:

```
struct node {
    int value;
    int x, y;
    node *left, *right;
    node(int v, int x, int y) : value(v), x(x), y(y) {}
};
```

Aquí `value` és el valor del node, $[x, y]$ és l'interval corresponent i `left` i `right` apunten als subarbres esquerre i dret.

Després d'això, es poden crear nodes de la següent manera:

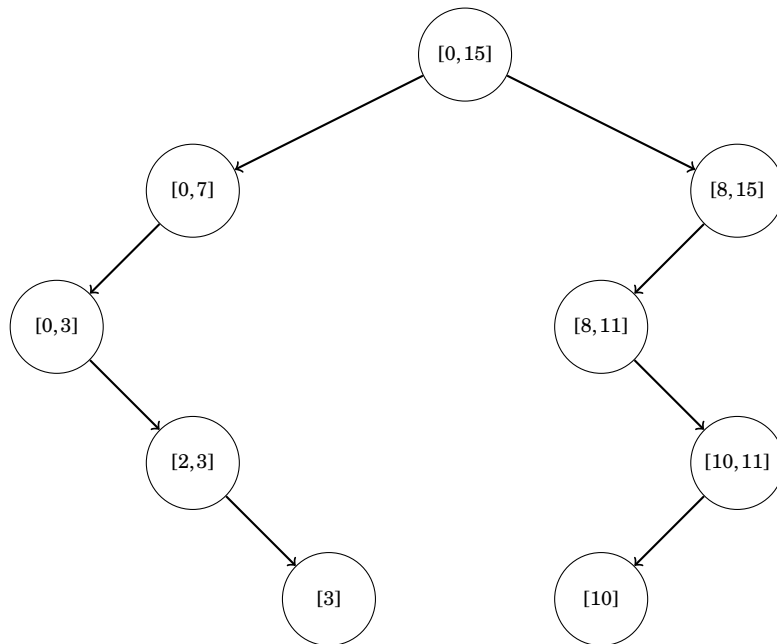
```
// create new node
node *x = new node(0, 0, 15);
// change value
x->value = 5;
```

Arbres de segments dispersos

Un arbre de segments dinàmic és útil quan el vector subjacent és *espars*, és a dir, el rang $[0, n-1]$ dels índexs permesos és gran, però la majoria dels valors del vector són zeros. Mentre que un arbre de segments normal necessita espai

$O(n)$, un arbre de segments dinàmics només necessita espai $O(k \log n)$, on k és el nombre d'operacions realitzades.

Un **arbre de segments dispersos** té inicialment només un node $[0, n - 1]$ el valor del qual és zero, la qual cosa significa que cada valor del vector és zero. Després de les actualitzacions, s'afegeixen nous nodes de manera dinàmica a l'arbre. Per exemple, si $n = 16$ i s'han modificat els elements de les posicions 3 i 10, l'arbre conté els nodes següents:



Qualsevol camí des del node arrel fins a una fulla conté $O(\log n)$ nodes, de manera que cada operació afegeix com a màxim $O(\log n)$ nous nodes a l'arbre. Així, després de k operacions, l'arbre conté com a màxim $O(k \log n)$ nodes.

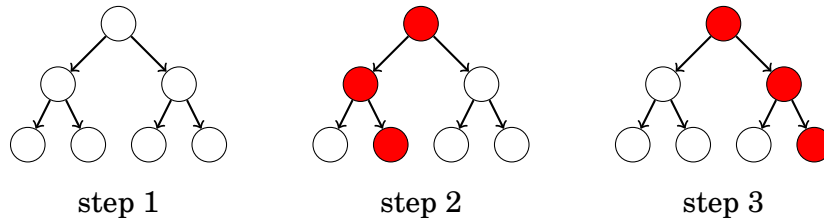
Tingueu en compte que si sabem tots els elements que s'han d'actualitzar al principi de l'algorisme, no és necessari un arbre de segment dinàmic, ja que podem utilitzar un arbre de segment normal amb compressió d'índexs (Capítol 9.4). Això no és possible quan els índexs es generen durant l'algorisme.

Arbres de segments persistent

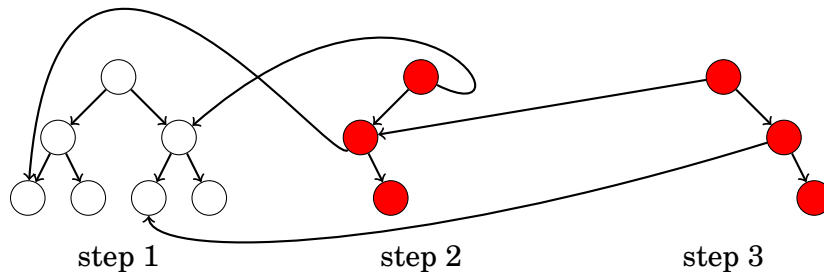
Fent servir una implementació dinàmica, també és possible crear un **arbre de segments persistent** que emmagatzema l'*historial de modificacions* de l'arbre. En aquesta implementació, podem accedir de manera eficient a totes les versions de l'arbre que han existit durant l'algorisme.

Quan l'historial de modificacions està disponible, podem fer consultes en qualsevol arbre anterior com en un arbre de segments normal, perquè s'emmagatzema l'estructura completa de cada arbre. També podem crear nous arbres basats en arbres anteriors i modificar-los de manera independent.

Considereu la següent seqüència d'actualitzacions, on els nodes vermells canvien i els altres nodes es mantenen iguals:



Després de cada actualització, la majoria dels nodes de l'arbre segueixen sent els mateixos, de manera que una manera eficient d'emmagatzemar l'historial de modificacions és representar cada arbre de l'historial com una combinació de nous nodes i subarbres d'arbres anteriors. En aquest exemple, l'historial de modificacions es pot emmagatzemar de la següent manera:



L'estructura de cada arbre anterior es pot reconstruir seguint els punters que comencen al node arrel corresponent. Com que cada operació només afegeix $O(\log n)$ nous nodes a l'arbre, és possible emmagatzemar l'historial complet de modificacions de l'arbre.

28.3 Estructures de dades

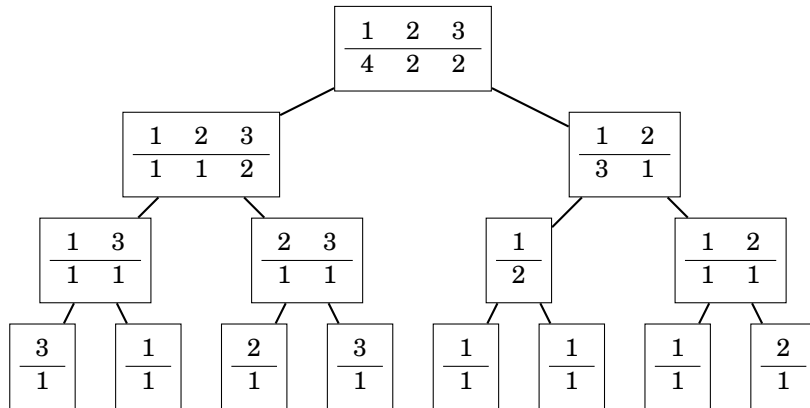
En lloc de valors únics, els nodes d'un arbre de segments també poden contenir *estructures de dades* que mantenen informació sobre els intervals corresponents. En aquest arbre, les operacions prenen temps $O(f(n)\log n)$, on $f(n)$ és el temps necessari per processar un sol node durant una operació.

Com a exemple, considereu un arbre de segments que admet consultes de la forma "quantes vegades apareix un element x a l'interval $[a, b]$?" Per exemple, l'element 1 apareix tres vegades en el rang següent:

3	1	2	3	1	1	1	2
---	---	---	---	---	---	---	---

Per a suportar aquestes consultes, construïm un arbre de segments on a cada node li assignem una estructura de dades que porta el compte de quantes vegades apareix un element x a l'interval corresponent. La resposta a una consulta es pot calcular combinant els resultats dels nodes que pertanyen a l'interval.

Per exemple, l'arbre de segments següent correspon al vector anterior:



Podem construir l'arbre de manera que cada node contingui una estructura map. En aquest cas, el temps necessari per processar cada node és $O(\log n)$, de manera que la complexitat de temps total d'una consulta és $O(\log^2 n)$. L'arbre utilitza espai $O(n \log n)$, perquè hi ha $O(\log n)$ nivells i cada nivell conté $O(n)$ elements.

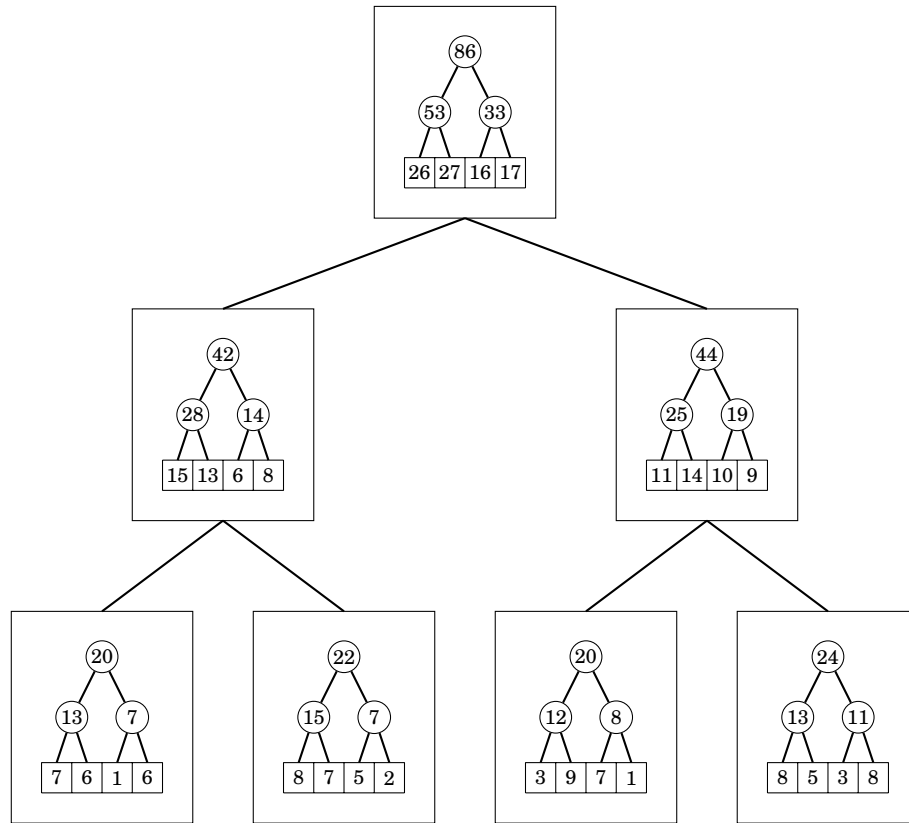
28.4 Bidimensionalitat

Un **arbre de segments bidimensionals** admet consultes relacionades amb submatrius rectangulars d'una matriu bidimensional. Aquest arbre es pot implementar com a arbres de segments niats: un arbre gran correspon a les files de la matriu i cada node conté un arbre petit que correspon a una columna.

Per exemple, en la matriu

7	6	1	6
8	7	5	2
3	9	7	1
8	5	3	8

la suma de qualsevol submatriu es pot calcular a partir del següent arbre de segments:



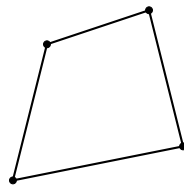
Les operations d'un arbre de segment bidimensional triguen temps $O(\log^2 n)$, perquè l'arbre gran i cada arbre petit tenen $O(\log n)$. L'arbre requereix espai $O(n^2)$, perquè cada arbre petit conté $O(n)$ valors.

Capítol 29

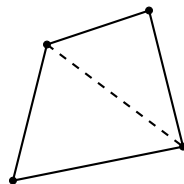
Geometria

En problemes geomètrics és sovint difícil trobar una manera d'abordar el problema de manera que la seva solució sigui senzilla d'implementar i el nombre de casos especials sigui petit.

Per exemple, considerem un problema on se'ns donen els vèrtexs d'un quadrilàter (un polígon amb quatre vèrtexs) i hem de calcular la seva àrea. Una possible entrada per al problema és la següent:



Una manera d'abordar el problema és dividir el quadrilàter en dos triangles mitjançant una línia recta entre dos vèrtexs oposats:

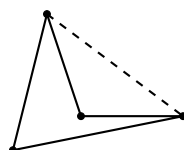


Després d'això, n'hi ha prou amb sumar les àrees dels triangles. L'àrea d'un triangle es pot calcular, per exemple, utilitzant **la fórmula d'Heron**,

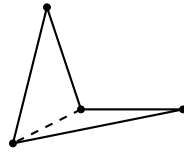
$$\sqrt{s(s-a)(s-b)(s-c)},$$

on a , b i c són les longituds dels costats del triangle i $s = (a + b + c)/2$.

Aquesta és una possible manera de resoldre el problema, però hi ha una trampa: com dividir el quadrilàter en triangles? Resulta que de vegades no podem escollir només dos vèrtexs contraris arbitraris. Per exemple, en la situació següent, la línia de divisió és *fora* del quadrilàter:



Tanmateix, una altra manera de dibuixar la línia funciona:



Per a un humà és clar quina de les línies és l'opció correcta, però la situació és difícil per a un ordinador. Tanmateix, resulta que podem resoldre el problema utilitzant un altre mètode més convenient per al programador. Hi ha una fórmula general

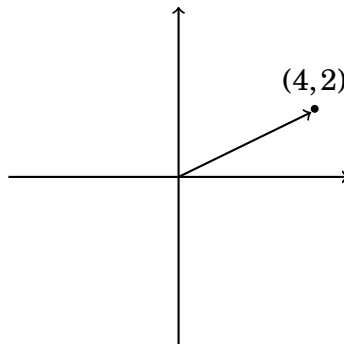
$$x_1y_2 - x_2y_1 + x_2y_3 - x_3y_2 + x_3y_4 - x_4y_3 + x_4y_1 - x_1y_4,$$

que calcula l'àrea d'un quadrilàter els vèrtexs del qual són (x_1, y_1) , (x_2, y_2) , (x_3, y_3) i (x_4, y_4) . Aquesta fórmula és fàcil d'implementar, no hi ha casos especials, i fins i tot podem generalitzar la fórmula a *tots* els polígons.

29.1 Nombres complexos

Un **nombre complex** és un nombre de la forma $x + yi$, on $i = \sqrt{-1}$ és la **unitat imaginària**. Una interpretació geomètrica d'un nombre complex és que representa un punt bidimensional (x, y) o un vector des de l'origen fins a un punt (x, y) .

Per exemple, $4 + 2i$ es correspon amb el punt i el vector següents:



En C++ la classe de nombre complexos `complex` és útil per a resoldre problemes geomètrics. Utilitzant la classe podem representar punts i vectors com a nombres complexos, i la classe conté eines que són útils en geometria.

En el codi següent, `C` és el tipus d'una coordenada i `P` és el tipus d'un punt o vector. A més, el codi defineix macros `X` i `Y` que es poden fer servir per a fer referència a les coordenades x i y .

```
typedef long long C;  
typedef complex<C> P;  
#define X real()  
#define Y imag()
```

Per exemple, el codi següent defineix un punt $p = (4, 2)$ i imprimeix les seves coordenades x i y :

```
P p = {4, 2};  
cout << p.X << " " << p.Y << "\n"; // 4 2
```

El codi següent defineix els vectors $v = (3, 1)$ i $u = (2, 2)$, i després calcula la seva suma $s = v + u$.

```
P v = {3, 1};  
P u = {2, 2};  
P s = v+u;  
cout << s.X << " " << s.Y << "\n"; // 5 3
```

A la pràctica, un tipus de coordenada adequat sol ser `long` (`enter`) o `long double` (`nombre real`). És bona idea fer servir nombres enters sempre que sigui possible, perquè els càlculs amb nombres enters són exactes. Si necessitem nombres reals, s'ha de tenir en compte els errors de precisió a l'hora de comparar nombres. Una manera segura de comprovar si els nombres reals a i b són iguals és comparar-los mitjançant $|a - b| < \epsilon$, on ϵ és un nombre petit (per exemple, $\epsilon = 10^{-9}$).

Funcions

En els exemples següents, el tipus de coordenades és `long double`.

La funció $\text{abs}(v)$ calcula la longitud $|v|$ d'un vector $v = (x, y)$ mitjançant la fórmula $\sqrt{x^2 + y^2}$. La funció també es pot utilitzar per calcular la distància entre els punts (x_1, y_1) i (x_2, y_2) , perquè aquesta distància és igual a la longitud del vector $(x_2 - x_1, y_2 - y_1)$.

El codi següent calcula la distància entre els punts $(4, 2)$ i $(3, -1)$:

```
P a = {4, 2};  
P b = {3, -1};  
cout << abs(b-a) << "\n"; // 3.16228
```

La funció $\text{arg}(v)$ calcula l'angle d'un vector $v = (x, y)$ respecte a l'eix x . La funció dona l'angle en radians, on r radians són $180r/\pi$ graus. L'angle d'un vector que apunta cap a la dreta és 0, i els angles disminueixen en sentit horari i augmenten en sentit contrari.

La funció $\text{polar}(s, a)$ construeix un vector la longitud del qual és s i que apunta a un angle a . Un vector es pot girar per un angle a multiplicant-lo per un vector de longitud 1 i d'angle a .

El codi següent calcula l'angle del vector $(4, 2)$, el gira $1/2$ radians en sentit contrari a les agulles del rellotge i després torna a calcular l'angle:

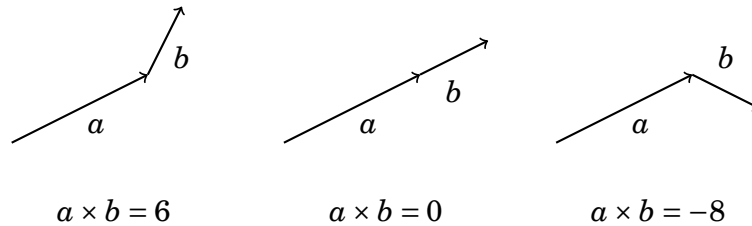
```
P v = {4, 2};  
cout << arg(v) << "\n"; // 0.463648  
v *= polar(1.0, 0.5);
```

```
cout << arg(v) << "\n"; // 0.963648
```

29.2 Punts i línies

El **producte creuat** $a \times b$ dels vectors $a = (x_1, y_1)$ i $b = (x_2, y_2)$ es calcula mitjançant la fórmula $x_1y_2 - x_2y_1$. El producte creuat ens indica si b gira a l'esquerra (valor positiu), no gira (zero) o gira a la dreta (valor negatiu) quan es col·loca directament després de a .

La següent imatge il·lustra els casos anteriors:



Per exemple, en el primer cas $a = (4, 2)$ i $b = (1, 2)$. El codi següent calcula el producte creuat amb la classe complex:

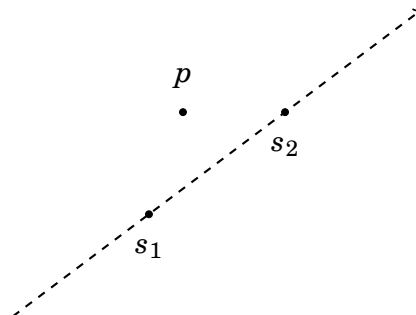
```
P a = {4, 2};  
P b = {1, 2};  
C p = (conj(a)*b).Y; // 6
```

El codi anterior funciona, perquè la funció `conj` nega la coordenada y d'un vector, i quan els vectors $(x_1, -y_1)$ i (x_2, y_2) es multipliquen junts, la coordenada y del vector resultant és $x_1y_2 - x_2y_1$.

Ubicació d'un punt

Els productes creuats es poden fer servir per comprovar si un punt es troba al costat esquerre o dret d'una línia. Suposem que la línia passa pels punts s_1 i s_2 , mirant des de s_1 cap a s_2 , i el punt és p .

Per exemple, a la imatge següent, p es troba al costat esquerre de la línia:

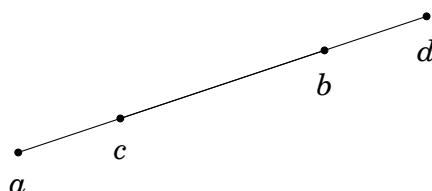


El producte creuat $(p - s_1) \times (p - s_2)$ ens indica la ubicació del punt p . Si el producte creuat és positiu, p es troba al costat esquerre, i si el producte creuat és negatiu, p es troba al costat dret. Finalment, si el producte creuat és zero, els punts s_1 , s_2 i p estan a la mateixa línia.

Intersecció d'un segment amb una línia

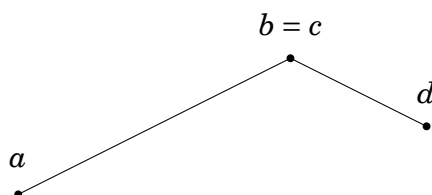
A continuació considerem el problema de comprovar si dos segments de línia ab i cd es tallen. Els casos possibles són:

Cas 1: Els segments de línia es troben a la mateixa línia i se superposen. En aquest cas, hi ha un nombre infinit de punts d'intersecció. Per exemple, a la imatge següent, tots els punts entre c i b són punts d'intersecció:



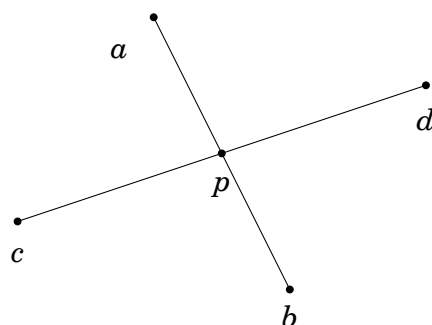
Podem fer servir productes creuats per comprovar si tots els punts estan a la mateixa línia. Després d'això, podem ordenar els punts i comprovar si els segments de línia se superposen.

Cas 2: Els segments de línia tenen un vèrtex comú que és l'únic punt d'intersecció. Per exemple, a la imatge següent el punt d'intersecció és $b = c$:



Aquest cas és fàcil de comprovar, perquè només hi ha quatre possibilitats per al punt d'intersecció: $a = c$, $a = d$, $b = c$ i $b = d$.

Cas 3: Hi ha exactament un punt d'intersecció que no és un vèrtex de cap segment de línia. A la imatge següent, el punt p és el punt d'intersecció:



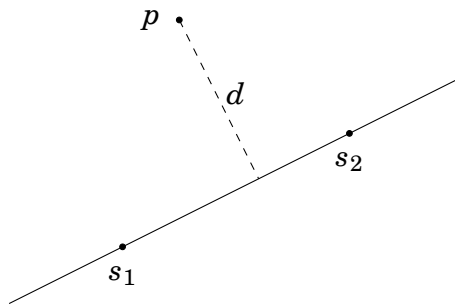
En aquest cas, els segments de línia es tallen exactament quan els dos punts c i d es troben a diferents costats de la línia que passa per a i b , i els punts a i b es troben a diferents costats de la línia que passa per c i d . Podem fer servir productes creuats per comprovar-ho.

Distància d'un punt a una línia

Una altra característica dels productes creuats és que l'àrea d'un triangle es pot calcular mitjançant la fórmula

$$\frac{|(a - c) \times (b - c)|}{2},$$

on a , b i c són els vèrtexs del triangle. Amb això podem derivar una fórmula per calcular la distància més curta entre un punt i una recta. Per exemple, a la imatge següent, d és la distància més curta entre el punt p i la línia que està definida pels punts s_1 i s_2 :

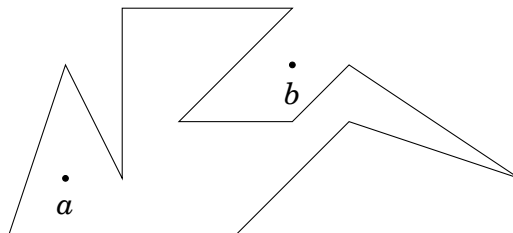


L'àrea del triangle els vèrtexs del qual són s_1 , s_2 i p es pot calcular de dues maneres: és $\frac{1}{2}|s_2 - s_1|d$ i $\frac{1}{2}((s_1 - p) \times (s_2 - p))$. Per tant, la distància més curta és

$$d = \frac{(s_1 - p) \times (s_2 - p)}{|s_2 - s_1|}.$$

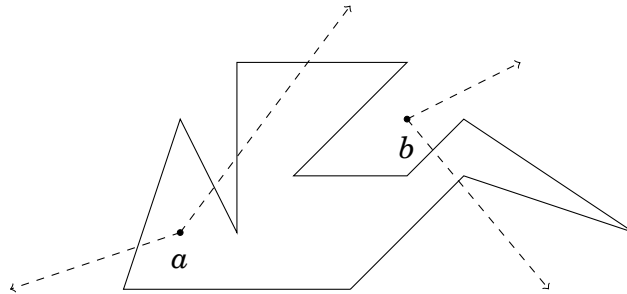
Punt dins d'un polígon

Considerem ara el problema de comprovar si un punt està situat dins o fora d'un polígon. Per exemple, a la imatge següent, el punt a es troba dins del polígon i el punt b està fora del polígon.



Una manera convenient de resoldre el problema és enviar un *raig* des del punt en una direcció arbitrària i calcular el nombre de vegades que toca la frontera del polígon. Si el nombre és senar, el punt està dins del polígon, i si el nombre és parell, el punt està fora del polígon.

Per exemple, podem enviar els rajos següents:



Els raigs de a toquen 1 i 3 vegades la frontera del polígon, de manera que a es troba dintre. Els raigs de b toquen 0 i 2 vegades la frontera del polígon, de manera que b es troba a fora.

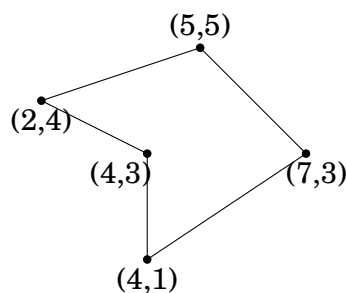
29.3 Àrea del polígon

Una fórmula general per calcular l'àrea d'un polígon, de vegades anomenada **fórmula de Gauss** o **fórmula de la llaçada** (de cordons de sabates), és la següent:

$$\frac{1}{2} \left| \sum_{i=1}^{n-1} (p_i \times p_{i+1}) \right| = \frac{1}{2} \left| \sum_{i=1}^{n-1} (x_i y_{i+1} - x_{i+1} y_i) \right|,$$

on $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, ..., $p_n = (x_n, y_n)$ és la seqüència de vèrtexs adjacents a la frontera del polígon i el primer i l'últim vèrtex són els mateixos, és a dir, $p_1 = p_n$.

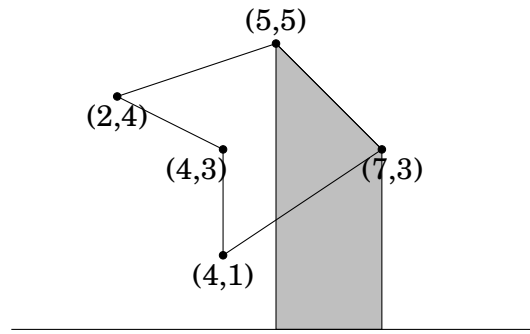
Per exemple, l'àrea del polígon



és

$$\frac{|(2 \cdot 5 - 5 \cdot 4) + (5 \cdot 3 - 7 \cdot 5) + (7 \cdot 1 - 4 \cdot 3) + (4 \cdot 3 - 4 \cdot 1) + (4 \cdot 4 - 2 \cdot 3)|}{2} = 17/2.$$

La idea de la fórmula és passar pels trapezis un costat dels quals és un costat del polígon i un altre costat es troba a la línia horitzontal $y = 0$. Per exemple:



L'àrea d'aquest trapezi és

$$(x_{i+1} - x_i) \frac{y_i + y_{i+1}}{2},$$

on els vèrtexs del polígon són p_i i p_{i+1} . Si $x_{i+1} > x_i$, l'àrea és positiva, i si $x_{i+1} < x_i$, l'àrea és negativa.

L'àrea del polígon és la suma de les àrees de tots aquests trapezis, la qual cosa dóna la fórmula

$$\left| \sum_{i=1}^{n-1} (x_{i+1} - x_i) \frac{y_i + y_{i+1}}{2} \right| = \frac{1}{2} \left| \sum_{i=1}^{n-1} (x_i y_{i+1} - x_{i+1} y_i) \right|.$$

Es pren el valor absolut de la suma perquè el seu valor pot ser positiu o negatiu, depenent de si caminem en sentit horari o en sentit contrari per la frontera del polígon.

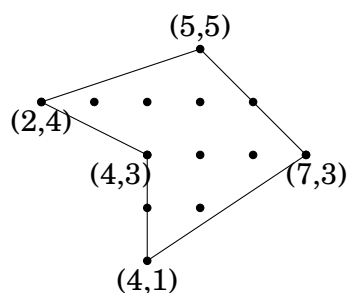
Teorema de Pick

El teorema de Pick proporciona una altra manera de calcular l'àrea d'un polígon, sempre que tots els vèrtexs del polígon tinguin coordenades enteres. Segons el teorema de Pick, l'àrea del polígon és

$$a + b/2 - 1,$$

on a és el nombre de punts amb coordenades enteres dins del polígon i b és el nombre de punts amb coordenades enteres a la frontera del polígon.

Per exemple, l'àrea del polígon



és $6 + 7/2 - 1 = 17/2$.

29.4 Funcions distància

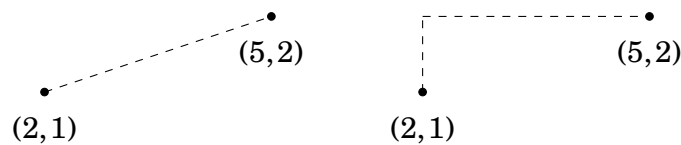
Una **funció distància** defineix la distància entre dos punts. La funció distància habitual és la **distància euclidiana** on la distància entre els punts (x_1, y_1) i (x_2, y_2) és

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Una funció distància alternativa és la **Distància de Manhattan** on la distància entre els punts (x_1, y_1) i (x_2, y_2) és

$$|x_1 - x_2| + |y_1 - y_2|.$$

Per exemple, considerem la imatge següent:



Euclidean distance

Manhattan distance

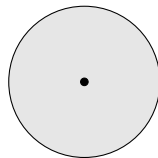
La distància euclidiana entre els punts és

$$\sqrt{(5-2)^2 + (2-1)^2} = \sqrt{10}$$

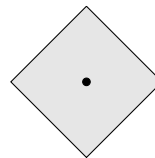
i la distància de Manhattan és

$$|5-2| + |2-1| = 4.$$

La imatge següent mostra regions que es troben a distància 1 del punt central, fent servir les distàncies euclidianes i de Manhattan:



distància Euclidiana

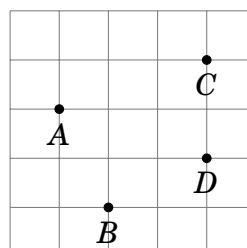


distància de Manhattan

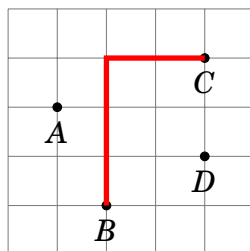
Coordenades giratòries

Alguns problemes són més fàcils de resoldre si es fa servir distàncies de Manhattan en lloc de distàncies euclidianes. Per exemple, considerem uel problema on se'ns donen n punts en el pla bidimensional i la nostra tasca és calcular la distància de Manhattan màxima entre dos punts qualsevol.

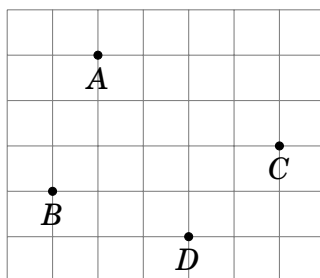
Per exemple, considereu el següent conjunt de punts:



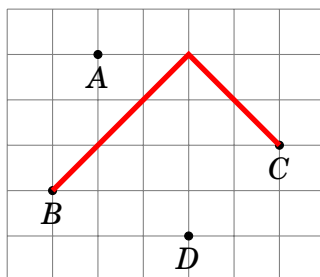
La distància de Manhattan màxima és 5, entre els punts B i C :



Una tècnica útil relacionada amb les distàncies de Manhattan és girar totes les coordenades 45 graus de manera que un punt (x, y) esdevingui $(x + y, y - x)$. Per exemple, després de girar els punts anteriors, el resultat és:



I la distància màxima és la següent:



Considereu dos punts $p_1 = (x_1, y_1)$ i $p_2 = (x_2, y_2)$ les coordenades girades dels quals són $p'_1 = (x'_1, y'_1)$ i $p'_2 = (x'_2, y'_2)$. Ara hi ha dues maneres d'expressar la distància de Manhattan entre p_1 i p_2 :

$$|x_1 - x_2| + |y_1 - y_2| = \max(|x'_1 - x'_2|, |y'_1 - y'_2|)$$

Per exemple, si $p_1 = (1, 0)$ i $p_2 = (3, 3)$, les coordenades girades són $p'_1 = (1, -1)$ i $p'_2 = (6, 0)$ i la distància de Manhattan és

$$|1 - 3| + |0 - 3| = \max(|1 - 6|, |-1 - 0|) = 5.$$

Les coordenades girades proporcionen una manera senzilla d'operar amb distàncies de Manhattan, perquè podem considerar les coordenades x i y per separat. Per maximitzar la distància de Manhattan entre dos punts, hauríem de trobar dos punts les coordenades girades dels quals maximitzin el valor de

$$\max(|x'_1 - x'_2|, |y'_1 - y'_2|).$$

Això és fàcil, perquè la diferència horitzontal o vertical de les coordenades girades ha de ser màxima.

Capítol 30

Algorismes d'escombrat de línies

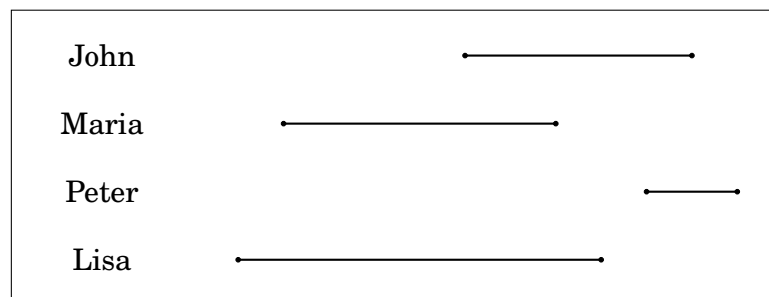
Molts problemes geomètrics es poden resoldre fent servir algorismes de **línia d'escombrat** (*sweep line*). La idea en aquests algorismes és representar una instància del problema com un conjunt d'esdeveniments que es corresponen amb punts del pla. Els esdeveniments es processen en ordre creixent segons les seves coordenades x o y .

A tall d'exemple, considereu el problema següent: una empresa té n empleats, i sabem per a cada empleat a quina hora ha arribat i quina hora ha sortit un dia donat. La nostra tasca és calcular el nombre màxim d'empleats que hi havia a l'oficina alhora.

El problema es pot resoldre modelant la situació de manera que a cada empleat se li assignin dos esdeveniments que corresponen als seus horaris d'arribada i sortida. Després d'ordenar els esdeveniments, els repassem i fem un seguiment del nombre de persones a l'oficina. Per exemple, la taula

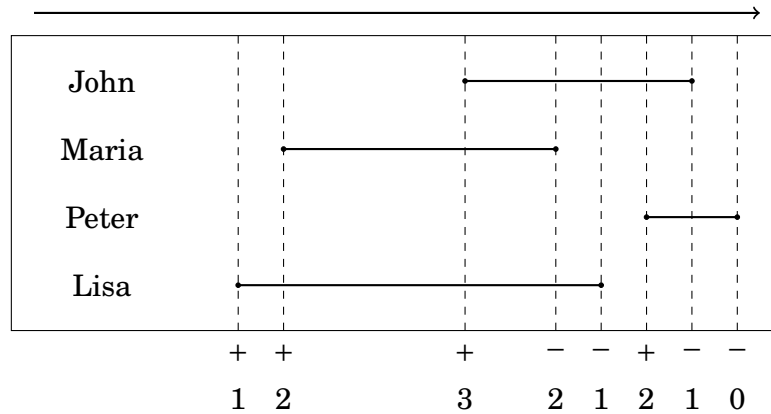
persona	temps d'arribada	temps de sortida
John	10	15
Maria	6	12
Peter	14	16
Lisa	5	13

correspon als següents esdeveniments:



Repassem els esdeveniments d'esquerra a dreta i mantenim un comptador. Sempre quan arriba una persona, augmentem el valor del comptador en un, i quan una persona marxa, reduïm el valor del comptador en un. La resposta al problema és el valor màxim del comptador durant l'algorisme.

En aquest exemple els esdeveniments es processen de la següent manera:

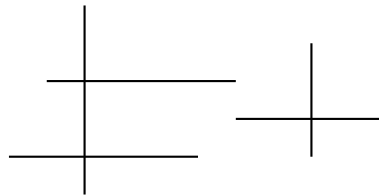


Els símbols + i – indiquen si el valor del comptador augmenta o disminueix, i el valor del comptador es mostra a sota. El valor màxim del comptador és 3 entre que arriba el John i es marxa la Maria.

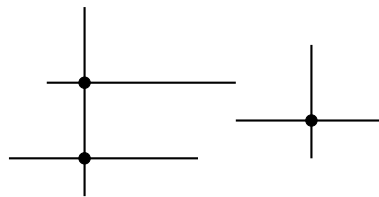
El temps d'execució de l'algorisme és $O(n \log n)$, perquè l'ordenació dels esdeveniments triga temps $O(n \log n)$ i la resta de l'algorisme triga temps $O(n)$.

30.1 Punts d'intersecció

Donat un conjunt de n segments de línia, cadascun d'ells horitzontal o vertical, considereu el problema de comptar el nombre total de punts d'intersecció. Per exemple, quan els segments de línia són



hi ha tres punts d'intersecció:

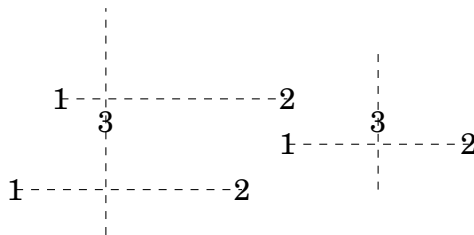


És fàcil resoldre el problema en temps $O(n^2)$, perquè podem recórrer tots els parells possibles de segments i comprovar si es tallen. Tanmateix, podem resoldre el problema de manera més eficient en temps $O(n \log n)$ fent servir un algorisme de línia d'escombrat i una estructura de dades de consulta d'interval.

La idea és processar els punts finals dels segments de línia d'esquerra a dreta i centrar-se en tres tipus d'esdeveniments:

- (1) comença un segment horitzontal
- (2) un segment horitzontal acaba
- (3) apareix un segment vertical

Els esdeveniments següents corresponen a l'exemple:



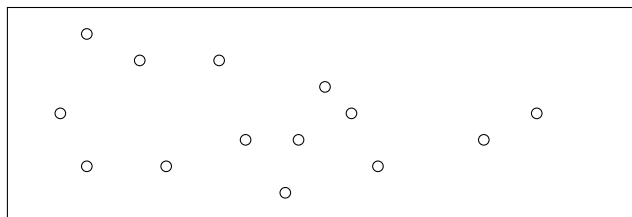
Repassem els esdeveniments d'esquerra a dreta i fem servir una estructura de dades que manté un conjunt de coordenades y on hi ha un segment horitzontal actiu. A l'esdeveniment 1, afegim la coordenada y del segment al conjunt, i a l'esdeveniment 2, eliminem la coordenada y del conjunt.

Els punts d'intersecció es calculen a l'esdeveniment 3. Quan hi ha un segment vertical entre els punts y_1 i y_2 , comptem el nombre de segments horitzontals actius la coordenada y dels quals està entre y_1 i y_2 , i afegim aquest nombre al nombre total de punts d'intersecció.

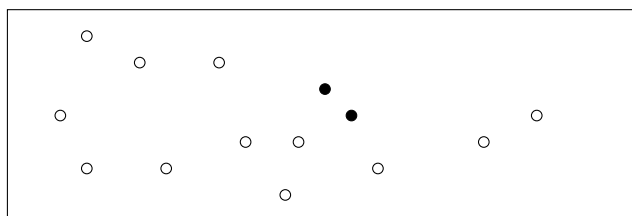
Per emmagatzemar les coordenades y dels segments horitzontals, podem utilitzar un arbre binari indexat o arbre de segments, possiblement amb compressió d'índex. Quan s'utilitzen aquestes estructures, el processament de cada esdeveniment requereix temps $O(\log n)$, de manera que el temps total d'execució de l'algorisme és $O(n \log n)$.

30.2 Problema de la parella més propera

Donat un conjunt de n punts, el nostre problema és trobar dos punts la distància euclidiana dels quals sigui mínima. Per exemple, si els punts són



hauríem de trobar els punts següents:



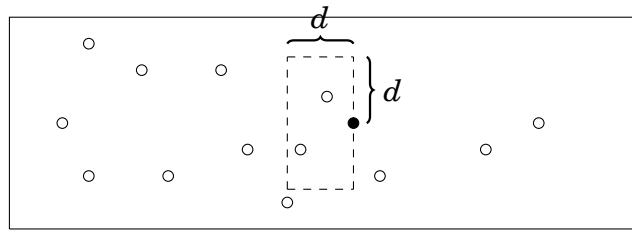
Aquest és un altre exemple d'un problema que es pot resoldre en temps $O(n \log n)$ fent servir un algorisme de línia d'escombrat¹. Recorrem els punts

¹A més d'aquest enfocament, també hi ha un algorisme en temps $O(n \log n)$ de dividir i vèncer [56] que divideix els punts en dos conjunts i resol recursivament el problema dels dos conjunts.

d'esquerra a dreta i mantenim un valor d : la distància mínima entre dos punts vist fins ara. Per a cada punt, trobem el punt més proper a l'esquerra. Si la distància és inferior a d , és la nova distància mínima i actualitzem el valor de d .

Si el punt actual és (x, y) i hi ha un punt a l'esquerra a una distància inferior a d , la coordenada x d'aquest punt ha d'estar entre $[x - d, x]$ i la coordenada y ha d'estar entre $[y - d, y + d]$. Per tant, n'hi ha prou amb considerar només els punts que es troben en aquests intervals, la qual cosa fa que l'algorisme sigui eficient.

Per exemple, a la imatge següent, la regió marcada amb línies discontinúes conté els punts que estan a distància d o inferior del punt actiu:



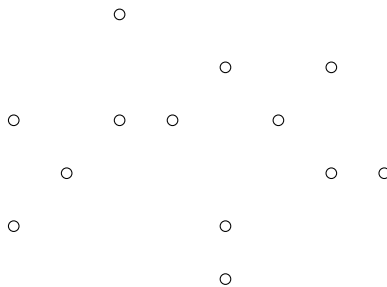
L'eficiència de l'algorisme es basa en el fet que la regió sempre conté només $O(1)$ punts². Podem recórrer aquests punts en temps $O(\log n)$ mantenint un conjunt de punts la coordenada x dels quals està entre $[x - d, x]$, en ordre creixent segons les seves coordenades y .

La complexitat temporal de l'algorisme és $O(n \log n)$, perquè passem per n punts i trobem per a cada punt el punt més proper a l'esquerra en temps $O(\log n)$ ³.

30.3 Problema de l'envolupant convexa

L'**envolupant convexa** (*convex hull*) és el polígon convex més petit que conté tots els punts d'un conjunt donat. Un polígon és convex si un segment de línia entre dos vèrtexs qualsevols del polígon està completament dintre del polígon.

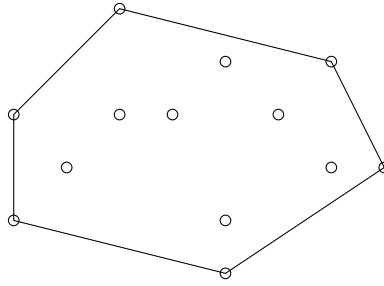
Per exemple, els punts



tenen la següent envolupant convexa:

²(N. del T.) Això és degut que la regió marcada amb línies discontinúes té alçada $2d$, i tots els punts vistos fins ara estan a distància d o superior.

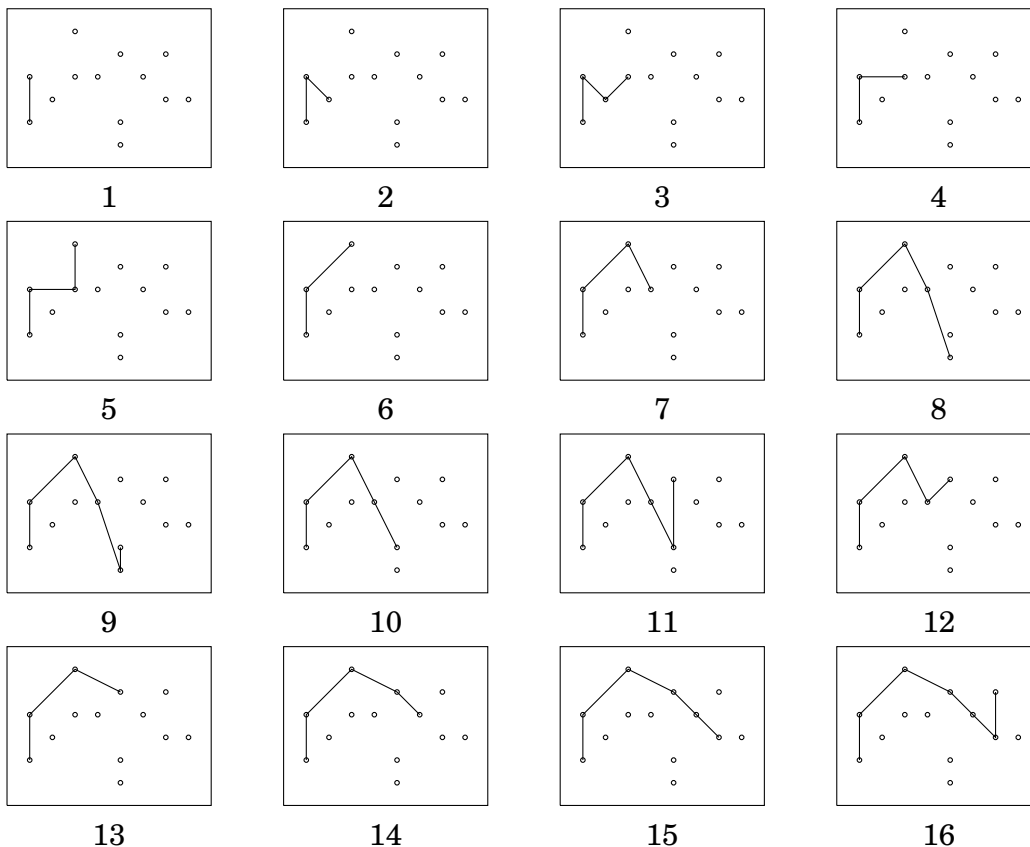
³(N. del T.) També hem de treure els punts d'aquest conjunt ordenat quan d es fa més petita o ens movem cap a la dreta de (x, y) , però això també triga temps $O(\log n)$.

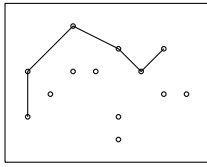


L'algorisme d'Andrew [3] proporciona una manera senzilla de construir l'envolupant convexa d'un conjunt de punts en temps $O(n \log n)$. L'algoritme localitza primer els punts més a l'esquerra i més a la dreta, i després construeix l'envolupant convexa en dues parts: primer la part superior i després la part inferior. Les dues parts són semblants, de manera que ens podem centrar a construir la part superior.

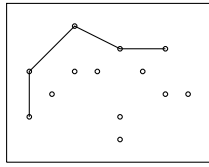
En primer lloc, ordenem els punts segons la coordenada x i, en cas d'empat, segons la coordenada y . Després d'això, recorrem els punts i els afegim un a un a l'envolupant convexa, però sempre que n'afegim un, ens assegurem que l'últim segment de línia de l'envolupant no giri a l'esquerra. Sempre que giri a l'esquerra treiem repetidament el penúltim punt de l'envolupant.

Les imatges següents mostren com funciona l'algorisme d'Andrew:

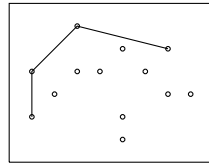




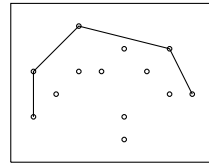
17



18



19



20

Bibliografia

- [1] A. V. Aho, J. E. Hopcroft and J. Ullman. *Data Structures and Algorithms*, Addison-Wesley, 1983.
- [2] R. K. Ahuja and J. B. Orlin. Distance directed augmenting path algorithms for maximum flow and parametric maximum flow problems. *Naval Research Logistics*, 38(3):413–430, 1991.
- [3] A. M. Andrew. Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, 9(5):216–219, 1979.
- [4] B. Aspvall, M. F. Plass and R. E. Tarjan. A linear-time algorithm for testing the truth of certain quantified boolean formulas. *Information Processing Letters*, 8(3):121–123, 1979.
- [5] R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.
- [6] M. Beck, E. Pine, W. Tarrat and K. Y. Jensen. New integer representations as the sum of three cubes. *Mathematics of Computation*, 76(259):1683–1690, 2007.
- [7] M. A. Bender and M. Farach-Colton. The LCA problem revisited. In *Latin American Symposium on Theoretical Informatics*, 88–94, 2000.
- [8] J. Bentley. *Programming Pearls*. Addison-Wesley, 1999 (2nd edition).
- [9] J. Bentley and D. Wood. An optimal worst case algorithm for reporting intersections of rectangles. *IEEE Transactions on Computers*, C-29(7):571–577, 1980.
- [10] C. L. Bouton. Nim, a game with a complete mathematical theory. *Annals of Mathematics*, 3(1/4):35–39, 1901.
- [11] Croatian Open Competition in Informatics, <http://hsin.hr/coci/>
- [12] Codeforces: On "Mo's algorithm", <http://codeforces.com/blog/entry/20032>
- [13] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein. *Introduction to Algorithms*, MIT Press, 2009 (3rd edition).

- [14] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.
- [15] K. Diks et al. *Looking for a Challenge? The Ultimate Problem Set from the University of Warsaw Programming Competitions*, University of Warsaw, 2012.
- [16] M. Dima and R. Ceterchi. Efficient range minimum queries using binary indexed trees. *Olympiad in Informatics*, 9(1):39–44, 2015.
- [17] J. Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17(3):449–467, 1965.
- [18] J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2):248–264, 1972.
- [19] S. Even, A. Itai and A. Shamir. On the complexity of time table and multi-commodity flow problems. *16th Annual Symposium on Foundations of Computer Science*, 184–193, 1975.
- [20] D. Fanding. A faster algorithm for shortest-path – SPFA. *Journal of Southwest Jiaotong University*, 2, 1994.
- [21] P. M. Fenwick. A new data structure for cumulative frequency tables. *Software: Practice and Experience*, 24(3):327–336, 1994.
- [22] J. Fischer and V. Heun. Theoretical and practical improvements on the RMQ-problem, with applications to LCA and LCE. In *Annual Symposium on Combinatorial Pattern Matching*, 36–48, 2006.
- [23] R. W. Floyd Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.
- [24] L. R. Ford. Network flow theory. RAND Corporation, Santa Monica, California, 1956.
- [25] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8(3):399–404, 1956.
- [26] R. Freivalds. Probabilistic machines can use less running time. In *IFIP congress*, 839–842, 1977.
- [27] F. Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, 296–303, 2014.
- [28] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, 1979.
- [29] Google Code Jam Statistics (2017), <https://www.go-hero.net/jam/17>

- [30] A. Grønlund and S. Pettie. Threesomes, degenerates, and love triangles. In *Proceedings of the 55th Annual Symposium on Foundations of Computer Science*, 621–630, 2014.
- [31] P. M. Grundy. Mathematics and games. *Eureka*, 2(5):6–8, 1939.
- [32] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, 1997.
- [33] S. Halim and F. Halim. *Competitive Programming 3: The New Lower Bound of Programming Contests*, 2013.
- [34] M. Held and R. M. Karp. A dynamic programming approach to sequencing problems. *Journal of the Society for Industrial and Applied Mathematics*, 10(1):196–210, 1962.
- [35] C. Hierholzer and C. Wiener. Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Mathematische Annalen*, 6(1), 30–32, 1873.
- [36] C. A. R. Hoare. Algorithm 64: Quicksort. *Communications of the ACM*, 4(7):321, 1961.
- [37] C. A. R. Hoare. Algorithm 65: Find. *Communications of the ACM*, 4(7):321–322, 1961.
- [38] J. E. Hopcroft and J. D. Ullman. A linear list merging algorithm. Technical report, Cornell University, 1971.
- [39] E. Horowitz and S. Sahni. Computing partitions with applications to the knapsack problem. *Journal of the ACM*, 21(2):277–292, 1974.
- [40] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.
- [41] The International Olympiad in Informatics Syllabus, <https://people.ksp.sk/~misof/ioi-syllabus/>
- [42] R. M. Karp and M. O. Rabin. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2):249–260, 1987.
- [43] P. W. Kasteleyn. The statistics of dimers on a lattice: I. The number of dimer arrangements on a quadratic lattice. *Physica*, 27(12):1209–1225, 1961.
- [44] C. Kent, G. M. Landau and M. Ziv-Ukelson. On the complexity of sparse exon assembly. *Journal of Computational Biology*, 13(5):1013–1027, 2006.
- [45] J. Kleinberg and É. Tardos. *Algorithm Design*, Pearson, 2005.
- [46] D. E. Knuth. *The Art of Computer Programming. Volume 2: Seminumerical Algorithms*, Addison–Wesley, 1998 (3rd edition).

- [47] D. E. Knuth. *The Art of Computer Programming. Volume 3: Sorting and Searching*, Addison–Wesley, 1998 (2nd edition).
- [48] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.
- [49] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
- [50] M. G. Main and R. J. Lorentz. An $O(n \log n)$ algorithm for finding all repetitions in a string. *Journal of Algorithms*, 5(3):422–432, 1984.
- [51] J. Pachocki and J. Radoszewski. Where to use and how not to use polynomial string hashing. *Olympiads in Informatics*, 7(1):90–100, 2013.
- [52] I. Parberry. An efficient algorithm for the Knight’s tour problem. *Discrete Applied Mathematics*, 73(3):251–260, 1997.
- [53] D. Pearson. A polynomial-time algorithm for the change-making problem. *Operations Research Letters*, 33(3):231–234, 2005.
- [54] R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.
- [55] 27-Queens Puzzle: Massively Parallel Enumeration and Solution Counting. <https://github.com/preusser/q27>
- [56] M. I. Shamos and D. Hoey. Closest-point problems. In *Proceedings of the 16th Annual Symposium on Foundations of Computer Science*, 151–162, 1975.
- [57] M. Sharir. A strong-connectivity algorithm and its applications in data flow analysis. *Computers & Mathematics with Applications*, 7(1):67–72, 1981.
- [58] S. S. Skiena. *The Algorithm Design Manual*, Springer, 2008 (2nd edition).
- [59] S. S. Skiena and M. A. Revilla. *Programming Challenges: The Programming Contest Training Manual*, Springer, 2003.
- [60] SZKOpuł, <https://szkopul.edu.pl/>
- [61] R. Sprague. Über mathematische Kampfspiele. *Tohoku Mathematical Journal*, 41:438–444, 1935.
- [62] P. Stańczyk. *Algorytmika praktyczna w konkursach Informatycznych*, MSc thesis, University of Warsaw, 2006.
- [63] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13(4):354–356, 1969.
- [64] R. E. Tarjan. Efficiency of a good but not linear set union algorithm. *Journal of the ACM*, 22(2):215–225, 1975.

- [65] R. E. Tarjan. Applications of path compression on balanced trees. *Journal of the ACM*, 26(4):690–715, 1979.
- [66] R. E. Tarjan and U. Vishkin. Finding biconnected components and computing tree functions in logarithmic parallel time. In *Proceedings of the 25th Annual Symposium on Foundations of Computer Science*, 12–20, 1984.
- [67] H. N. V. Temperley and M. E. Fisher. Dimer problem in statistical mechanics – an exact result. *Philosophical Magazine*, 6(68):1061–1063, 1961.
- [68] USA Computing Olympiad, <http://www.usaco.org/>
- [69] H. C. von Warnsdorf. *Des Rösselsprunges einfachste und allgemeinste Lösung*. Schmalkalden, 1823.
- [70] S. Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, 1962.

