Hašování (*Vyhledávání metodou transformace klíče*)

Velmi účinnou vyhledávací metodou je hašování. Pro její anglický název *hashing* se poměrně obtížně hledá vhodný překlad do českého jazyka, proto zůstaneme u používání mírně češtině přizpůsobeného anglického názvu.

Datová struktura použitá v hašování pro uložení prvků je tabulka. Tabulka se skládá z řádků. U hašování pro řádky tabulky používáme označení *přihrádky* (v angličtině jsou označovány pojmem *buckets*). V každé přihrádce je místo pro uložení jednoho datového prvku. Počet přihrádek v tabulce, tedy kapacitu tabulky označme *m*. Na jednotlivé přihrádky v tabulce se odkazujeme (adresujeme je) čísly 0 až *m*-1. Při implementaci hašování se tabulka snadno realizuje pomocí pole. Datový typ prvků pole se zvolí takový, aby se do něho daly uložit údaje, které ukládáme do přihrádek tabulky. Tím každý prvek pole reprezentuje jednu přihrádku tabulky a indexování prvků pole odpovídá adresování jednotlivých přihrádek v tabulce.

Základem hašování je hašovací funkce. Je to zobrazení, které hodnotě prvku (nebo vyhledávacímu klíči prvku, pokud prvek je strukturovaný typ) přiřadí číslo některé z přihrádek v tabulce, tedy číslo v rozmezí 0 až *m*-1. Hašovací funkce se typicky sestaví ze dvou funkcí. Ta první hodnotu prvku zobrazí na celé (nezáporné) číslo. Druhá celé číslo zobrazí na číslo přihrádky v tabulce, tedy na celé číslo z intervalu <0,*m*-1>. Je zřejmé, že první funkce závisí na datovém typu prvku a sestavujeme si ji sami. Druhá už je víceméně standardní a jen ji použijeme.

Cílem hašovací funkce je rovnoměrné rozmístění prvků v tabulce. Z toho plyne, že první funkce, která převádí hodnotu prvku na celé číslo, by měla mít vlastnosti:

- Zobrazovat hodnoty prvků na co největší počet různých celých čísel.
- Zobrazení na celá čísla by mělo být rovnoměrné (na jednotlivá čísla by se měl zobrazovat přibližně stejný počet prvků, které chceme do hašovací tabulky uložit).

Dalším přirozeným požadavkem na hašovací funkci je, aby její výpočet nebyl příliš časově náročný.

Hašovací funkce pro řetězce

Řetězce jsou častým vyhledávacím klíčem. Řetězec je posloupnost znaků. Jsou různé možnosti, jak tuto posloupnost zobrazit na celá čísla. Standardně se používá ASCII tabulka nebo kódování Unicode. ASCII tabulka znaky zobrazuje na čísla z intervalu <0,255>. Unicode je nejčastěji zobrazuje na čísla z intervalu <0,65535> (kódování UTF-16).

Řetězec si označme

$$z_1 z_2 \ldots z_k \quad ,$$

kde z_i jsou jednotlivé znaky v řetězci a k je počet těchto znaků v řetězci (délka řetězce).

Jedna z jednodušších funkcí zobrazující řetězec na celé číslo je

$$c_1(z_1 z_2 ... z_k) = p * asc(z_1) + q * asc(z_2) + asc(z_k) + k \quad ,$$

kde p a q jsou zvolené konstanty, nejlépe prvočísla (např. p=127, q=31), protože ty mají nejlepší předpoklady pro rovnoměrné zobrazení do množiny celých čísel. Funkce *asc* převádí znak na jeho ASCII hodnotu nebo Unicode hodnotu.

Dokonalejší, ale na druhé straně náročnější na výpočet, je funkce

$$c_2(z_1 z_2 ... z_k) = p^{k-1} * asc(z_1) + p^{k-2} * asc(z_2) + p^{k-3} * asc(z_3) + ... + p * asc(z_{k-1}) + asc(z_k) \ ,$$

kde p je konstanta, opět nejlépe prvočíslo (např. p=31). Pro její výpočet je účelné přepsat ji do tvaru

$$c_2(z_1 z_2 ... z_k) =$$
$$p * (... p * (p * (p * asc(z_1) + asc(z_2)) + asc(z_3)) ... + asc(z_{k-1})) + asc(z_k) \ .$$

Druhá část hašovací funkce, která převádí celé číslo na číslo přihrádky v hašovací tabulce, je velmi jednoduchá. Používá se pro ni operace *mod*, což je zbytek po celočíselném dělení. Obecný zápis hašovací funkce je

$$h(x) = c(x) \bmod m \quad ,$$

kde *c(x)* je první část hašovací funkce, která nám převádí hodnotu prvku na celé číslo, m je rozsah (počet přihrádek) hašovací tabulky. Opět je nejlepší zvolit m prvočíslo, protože to nemá žádného netriviálního vlastního dělitele, čímž poskytuje nejlepší předpoklady pro rovnoměrné rozmístění prvků v tabulce.

Vedle prvočíselného počtu přihrádek v tabulce se v praxi používá i počet přihrádek, který je mocninou čísla 2. Tento počet nemá tak dobré předpoklady pro rovnoměrné rozmístění prvků v tabulce, ale výpočet hašovací funkce je snadnější, protože místo operace modulo lze použít jednodušší operaci bitového součinu:

$$h(x) = c(x) \,\&\, (m-1), \quad m = 2^s \quad .$$

Jednoduchá hašovací funkce bez náročných operací (násobení, dělení):

$$h(z_1 z_2 ... z_k) = \left(asc(z_1) << 7 - asc(z_1) + asc(z_2) << 5 - asc(z_2) + asc(z_k) + k\right) \& (m-1) \ ,$$

kde 127=128-1, 31=32-1 a $m = 2^s$.

Hašovací funkce pro datum

$$c(den, mesic, rok) = p * den + q * mesic + rok \quad ,$$

kde p a q jsou nejlépe prvočísla.

Jiná hašovací funkce pro datum

$$c(den, mesic, rok) = den << 16 + mesic << 12 + rok \quad .$$

Příklad. Máme navrhnout hašování pro uložení řetězců. Velikost tabulky požadujeme přibližně 500 přihrádek.
Nebližší prvočísla jsou 499 a 503, velikost tabulky zvolíme třeba 503 přihrádek. Pro sestavení hašovací funkce použijeme již uvedenou funkci c_1. Hašovací funkce bude mít tvar:

$$h(z_1 z_2 ... z_k) = (127 * asc(z_1) + 31 * asc(z_2) + asc(z_k) + k) \bmod 503$$

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

Příklad. Do tabulky velikosti 11 budeme ukládat řetězce. Hašovací funkci zvolíme

$$h(z_1 z_2 ... z_k) = c(z_1 z_2 ... z_{k1}) \bmod 11 \quad ,$$

kde

$$c(z_1 z_2 ... z_k) = 7 * asc(z_1) + 3 * asc(z_2) + asc(z_k) + k \quad .$$

Do tabulky uložíme jména

h(*Eva*) = (7*69+3*118+97+3) mod 11 = 2

h(*Irena*) = (7*73+3*114+97+5) mod 11 = 9

h(*Pavel*) = (7*80+3*97+108+5) mod 11 = 7

h(*Marta*) = (7*77+3*97+97+5 mod 11 = 8

h(*Ivan*) = (7*73+3*118+110+4) mod 11 = 0

h(*Nina*) = (7*78+3*105+97+4) mod 11 = 5

Číslo přihrádky	Uložený prvek
0	Ivan
1	
2	Eva
3	
4	
5	Nina
6	
7	Pavel
8	Marta
9	Irena
10	

Kdybychom nyní chtěli do tabulky uložit jméno *Helena*, jehož hodnota hašovací funkce je

h(*Helena*) = (7*72+3*101+97+6) mod 11 = 8 ,

zjistíme, že tato pozice už je obsazená jménem *Marta*. Takové kolize, kdy více prvků se zobrazuje na stejnou přihrádku hašovací tabulky, se v hašování vyskytují zcela běžně. Uvedeme si nyní nejpoužívanější způsoby jejich řešení.

Otevřené adresování

Metoda otevřeného adresování počítá v případě, kdy pozice v tabulce vypočítaná hašovací funkcí je obsazena, další pozice tak dlouho, dokud se nenajde volná pozice anebo se nezjistí, že tabulka už je zaplněna.

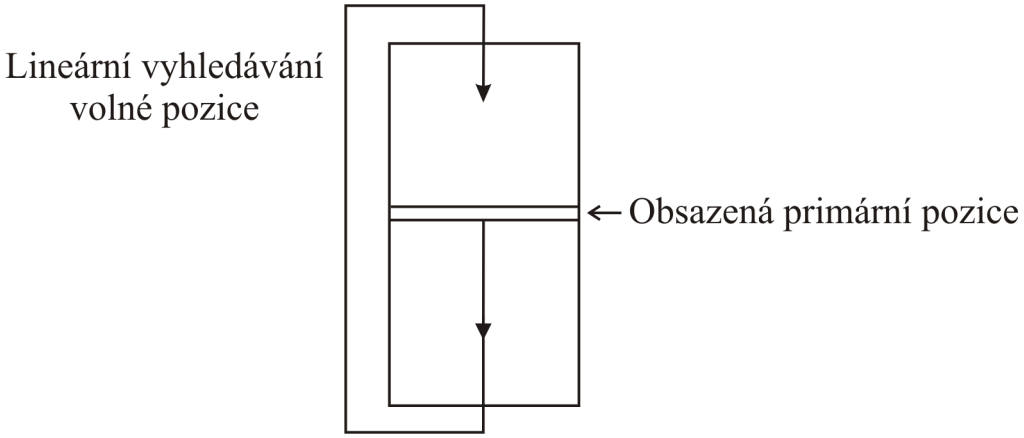
Lineární hledání

Nejjednodušší je lineární hledání, kdy nové pozice počítáme funkcí

$$H(x,i) = (h(x) + i) \bmod m \quad ,$$

kde *h(x)* je výchozí hašovací funkce, *i* je celočíselný parametr a *m* je rozsah tabulky. Je-li tedy primární pozice (*h(x) = H(x,0)*) obsazena, prohledávají se postupně další pozice

$$H(x,1), H(x,2), \ldots, H(x,m-1)$$



Lineární umísťování za sebou vede k vytváření nežádoucích shluků. Shlukem nazýváme větší počty za sebou následujících obsazených přihrádek tabulky. Pokud je vypočítaná primární pozice obsazená a je přitom uvnitř takového shluku, znamená to při lineárním hledání, že musíme projít všechny přihrádky v tomto shluku za primární pozicí, než se dostaneme k nějaké volné sekundární pozici, abychom prvek do ní mohli uložit. Přitom shluky prodlužují nejen operaci přidání prvku do tabulky, ale také vyhledávání prvku v tabulce. Při vyhledávání začínáme na primární pozici a pokud na ní prvek není a tato pozice je přitom obsazena (je na ní jiný prvek), procházíme sekundární pozice tak dlouho, dokud prvek nenajdeme nebo se nedostaneme k volné pozici, což je příznakem toho, že hledaný prvek v tabulce není.

Kvadratické hledání

Proto místo lineárního hledání se často používá kvadratické hledání. U něho sice také vznikají shluky, ale už v menší míře. Hašovací funkce používaná pro kvadratické hledání má většinou jednoduchý tvar

$$H(x,i) = (h(x) + i^2) \bmod m \ .$$

U ní je už určitý problém, že během hledání se můžeme touto funkcí dostat znovu na stejnou pozici, kterou jsme již prošli, aniž jsme přitom vyčerpali celou tabulku. Nechť například hodnota hašovací funkce *h* pro nějaký prvek *v* je *h*(*v*)=3 a mějme tabulku s 5 přihrádkami (*m*=5). Pak u kvadratického hledání dostáváme pozice:

$H(v,0) = (3+0^2) \bmod 5 = 3$
$H(v,1) = (3+1^2) \bmod 5 = 4$
$H(v,2) = (3+2^2) \bmod 5 = 2$
$H(v,3) = (3+3^2) \bmod 5 = 2$

Je zřejmé, že při čtvrtém pokusu (pro i=3) vypočítat novou pozici jsme se dostali na pozici, která už předtím byla. Ukažme, že to nastane (za předpokladu, že m je prvočíslo), až když prohledáme nejméně polovinu tabulky. V našem příkladu to nastalo, až když jsme prohledali 3 pozice, což je více než polovina z 5. V běžném použití, pokud už tabulka není téměř zaplněna, najdeme nějakou volnou pozici většinou dříve, než projdeme polovinu tabulky.

Nechť v kvadratickém hledání dva různé pokusy nalezení nové pozice dají stejný výsledek, tj.

$$H(x, j) = H(x, i), \quad \text{kde} \quad j > i \quad .$$

Dosadíme hašovací funkci

$$(h(x) + j^2) \bmod m = (h(x) + i^2) \bmod m$$

Odstraníme operaci *mod*

$$h(x) + j^2 = h(x) + i^2 + K * m \quad , \quad \text{kde } K \text{ je nějaké celé číslo.}$$

Odtud

$$j^2 - i^2 = K * m$$

$$(j + i) * (j - i) = K * m \quad .$$

Protože podle předpokladu je $j>i$, je i $K\neq 0$. Dále protože m je prvočíslo, je buďto $j+i$ dělitelné m nebo je $j-i$ je dělitelné m.

Uvažujme, že $j+i$ je dělitelné m, pak

$$j + i = L * m, \text{ kde } L\neq 0 \text{ je celé číslo} \Rightarrow j + i \geq m$$

$$j + i \geq m \ \wedge \ j > i \ \Rightarrow \ 2 * j + i > m + i \ \Rightarrow \ 2 * j > m \ \Rightarrow \ j > \left\lfloor \frac{m}{2} \right\rfloor \quad .$$

Pokud $j-i$ je dělitelné m, pak

$$j - i = L * m \ \Rightarrow \ j - i \geq m \ \Rightarrow \ j \geq m + i$$

$$j \geq m + i \ \wedge \ i \geq 0 \ \Rightarrow \ j \geq m \quad .$$

Závěr: $j > \left\lfloor \frac{m}{2} \right\rfloor$.

Dvojí hašování

Ještě propracovanější je metoda dvojího hašování. U ní funkce pro hledání má obecný tvar

$$H(x, i) = (h(x) + i * h_2(x)) \bmod m \quad ,$$

kde $h_2(x)$ je další (sekundární) hašovací funkce, která ale nabývá jen hodnoty v rozmezí 1 až m-1 (tedy ne hodnotu 0).
V praxi se sekundární hašovací funkce $h_2(x)$ nejčastěji vytváří přímo z primární hašovací $h(x)$, která sama je sestavena z nějaké výchozí funkce $c(x)$ a má tvar

$$h(x) = c(x) \bmod m \quad .$$

Z ní použijeme její základní část, funkci $c(x)$, a hašovací funkci $h_2(x)$ vytvoříme ve tvaru

$$h_2(x) = 1 + (c(x) \bmod (m-1)) \quad .$$

Funkci

$$H(x,i) = (h(x) + i * h_2(x)) \bmod m$$

upravíme na tvar

$$H(x,i) = (h(x) + (i-1) * h_2(x) + h_2(x)) \bmod m =$$
$$= ((h(x) + (i-1) * h_2(x)) \bmod m + h_2(x)) \bmod m = (H(x,i-1) + h_2(x)) \bmod m \quad .$$

Pozice vložení nyní můžeme počítat rekurzívně

$$H(x,0) = h(x)$$
$$H(x,i) = (H(x,i-1) + h_2(x)) \bmod m \quad \text{pro } i = 1,2,3,\ldots.$$

Příklad. Do tabulky vytvořené v předchozím příkladu chceme uložit další jména

h(*Helena*) = c(*Helena*) mod 11 = 8,
kde c(*Helena*) = 7*72+3*101+97+6 = 910

h(*Bohumil*) = c(*Bohumil*) mod 11 = 8,
kde c(*Bohumil*) = 7*66+3*111+108+7 = 910

h(*Jana*) = c(*Jana*) mod 11 = 8,
kde c(*Jana*) = 7*74+3*97+97+4 = 910

U všech je hodnota hašovací funkce rovna 8, přičemž tato pozice je již obsazena. Zvolíme-li kvadratické hledání, pak další možné pozice jsou

$(8+1^2)$ mod 11 = 9
$(8+2^2)$ mod 11 = 1
$(8+3^2)$ mod 11 = 6
$(8+4^2)$ mod 11 = 2
$(8+5^2)$ mod 11 = 0
$(8+6^2)$ mod 11 = 0
$(8+7^2)$ mod 11 = 2
$(8+8^2)$ mod 11 = 6
$(8+9^2)$ mod 11 = 1

$(8+10^2) \bmod 11 = 9$
$(8+11^2) \bmod 11 = 8$
$(8+12^2) \bmod 11 = 9$
$(8+13^2) \bmod 11 = 1$
$(8+14^2) \bmod 11 = 6$
$(8+15^2) \bmod 11 = 2$
$(8+16^2) \bmod 11 = 0$
$(8+17^2) \bmod 11 = 0$
$(8+18^2) \bmod 11 = 2$
$(8+19^2) \bmod 11 = 6$

Z vypočtených pozic jsou volné pozice 1, 6. Do nich umístíme 2 nová jména:

Číslo přihrádky	Uložený prvek
0	Ivan
1	Helena
2	Eva
3	
4	
5	Nina
6	Bohumil
7	Pavel
8	Marta
9	Irena
10	

V dalším výpočtu pozic se už pozice opakují a pro třetí jméno se volné místo už nenašlo.

Příklad. Do tabulky ukládáme stejná jména jako v předchozím příkladě, pro výpočet pozic použijeme dvojí hašování. Sekundární hašovací funkce bude

$h_2(x) = 1 + (c(x) \bmod 10)$.

Její hodnota pro ukládaná jména je

$h_2(Helena) = h_2(Bohumil) = h_2(Jana) = 1 + 910 \bmod 10 = 1$

Sekundární pozice budou

$(8+1) \bmod 11 = 9$
$(9+1) \bmod 11 = 10$

(10+1) mod 11 = 0
(0+1) mod 11 = 1
(1+1) mod 11 = 2
(2+1) mod 11 = 3

Z vypočtených pozic jsou volné pozice 10, 1, 3. Do nich umístíme 3 nová jména:

Číslo přihrádky	Uložený prvek
0	Ivan
1	Bohumil
2	Eva
3	Jana
4	
5	Nina
6	
7	Pavel
8	Marta
9	Irena
10	Helena

Vyhledávání v tabulce

Při vyhledávání v hašovací tabulce nejprve vypočítáme hodnotu hašovací funkce pro hledaný prvek *x*. Podíváme se do tabulky na přihrádku, na kterou ukazuje hodnota hašovací funkce. Mohou nastat případy:

- Přihrádka je prázdná – hledaný prvek není v tabulce.
- V přihrádce je hledaný prvek *x* – vyhledávání tím úspěšně končí.
- V přihrádce je jiný prvek než *x*. Začneme postupně počítat další možné pozice a srovnávat prvky na nich s hledaným prvkem *x*, dokud buďto hledaný prvek nenalezneme anebo se nedostaneme na prázdnou přihrádku anebo nevyčerpáme všechny možné pozice.

Příklad. Máme vyhledat jméno *Robert* v tabulce z předminulého příkladu.

h(*Robert*)=(7*82+3*111+116+6) mod 11 = 6

Na této pozici je ale jiné jméno - *Jana*. Začneme počítat a prohledávat další možné pozice

$(6+1^2) \bmod 11 = 7$ - *Pavel*

$(6+2^2) \bmod 11 = 10$

Vyhledávání skončí na pozici 10, která je prázdná. Jméno *Robert* tedy v tabulce není.

Pseudokódy:

```
SearchLin(T, m, x)
  h ← h0 ← hash(x)
  do
    if T[h] = NIL
      return -1
    if T[h] = x
      return h
    h ← (h+1) mod m
  while h ≠ h0
  return -1

SearchQuadr(T, m, x)
  h0 ← hash(x)
  for i ← 0 to m-1
    h ← (h0 + i*i) mod m
    if T[h] = NIL
      return -1
    if T[h] = x
      return h
  return -1

SearchDHash(T, m, x)
  h ← hash(x)
  h2 ← hash2(x)
  for i ← 0 to m-1
    if T[h] = NIL
      return -1
    if T[h] = x
      return h
    h ← (h + h2) mod m
  return -1

InsertQuadr(T, m, x)
```

```
h0 ← hash(x)
for i ← 0 to m-1
  h ← (h0 + i*i) mod m
  if T[h] = NIL
    T[h] ← x
    return
error
```

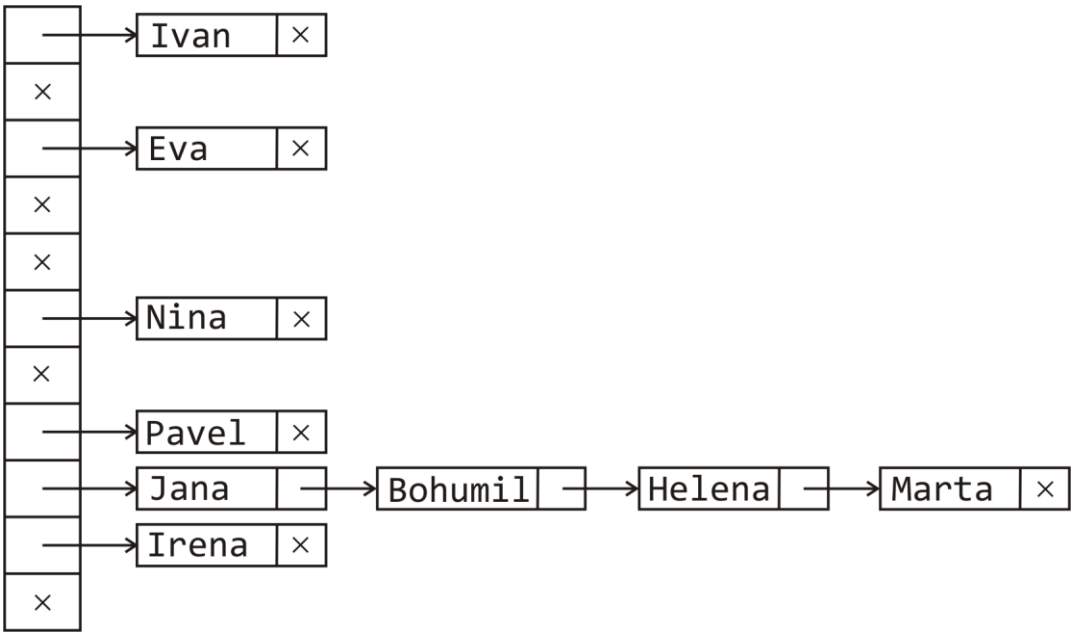
Zřetězení

Předchozí metoda otevřeného adresování má dvě nevýhody:

- Počet prvků, jež lze do tabulky uložit, je omezen její velikostí. Pokud dopředu neznáme, kolik prvků bude do tabulky ukládáno, může se stát, že ji stanovíme malou a dojde k jejímu přeplnění. Následné zvětšení velikosti tabulky je většinou časově náročně.
- Při vyhledávání, zejména v dost zaplněné tabulce, procházíme v důsledku otevřeného adresování i prvky, které mají jinou hodnotu hašovací funkce, čímž se doba vyhledávání zvětšuje.

Tyto nevýhody odstraňuje metoda zřetězení, která k ukládání dalších prvků se stejnou hodnotou hašovací funkce využívá seznamy. Hašovací tabulka v tomto případě obsahuje ukazatele na začátek (první uzel) jednotlivých seznamů.

Příklad. Tabulka z předchozího příkladu se zřetězením. Všechny prvky jsou nyní uloženy v seznamech.



Pseudokódy:

```
Search(T, x)
```

```
  u ← T[hash(x)]
  while u ≠ NIL
    if u.item = x
      return u
    u ← u.next
  return NIL

Insert(T, x)
  u ← new Node
  u.item ← x
  h ← hash(x)
  u.next ← T[h]
  T[h] ← u
```