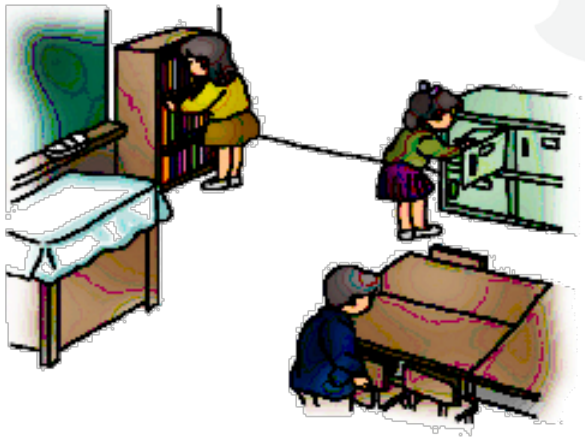
CSE2010 자료구조론

Week 14: Hashing

ICT융합학부 조용우

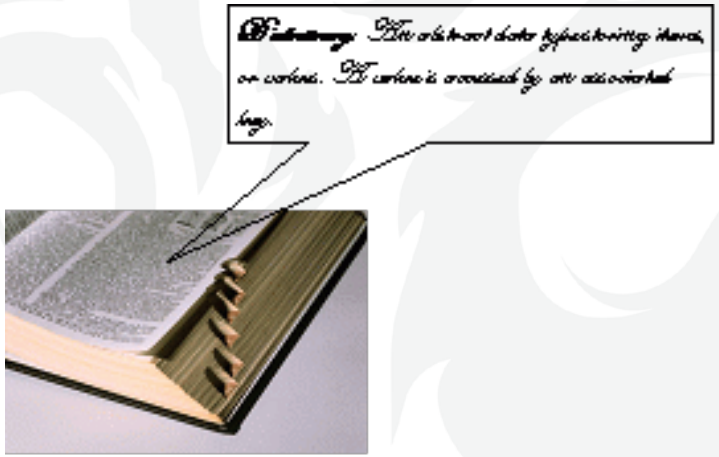
해싱이란?

- 대부분의 탐색 방법들은 키 값 비교로 탐색하고자 하는 항목에 접근
- 해싱(hashing)
  - 키 값에 대한 산술적 연산에 의해 테이블의 주소를 계산하여 항목에 접근
- 해시 테이블(hash table)
  - 키 값의 연산에 의해 직접 접근이 가능한 구조
- 해싱은 물건을 정리하는 것과 유사함



사전(Dictionary)

- 사전(dictionary)
  - 맵(map) 또는 테이블(table)로 불리움
  - 탐색 키와 관련된 값의 2가지 필드로 구성
    - 영어 단어나 사람의 이름 같은 탐색 키(search key)
    - 단어의 정의나 주소 또는 전화 번호 같은 탐색 키와 관련된 값(value)

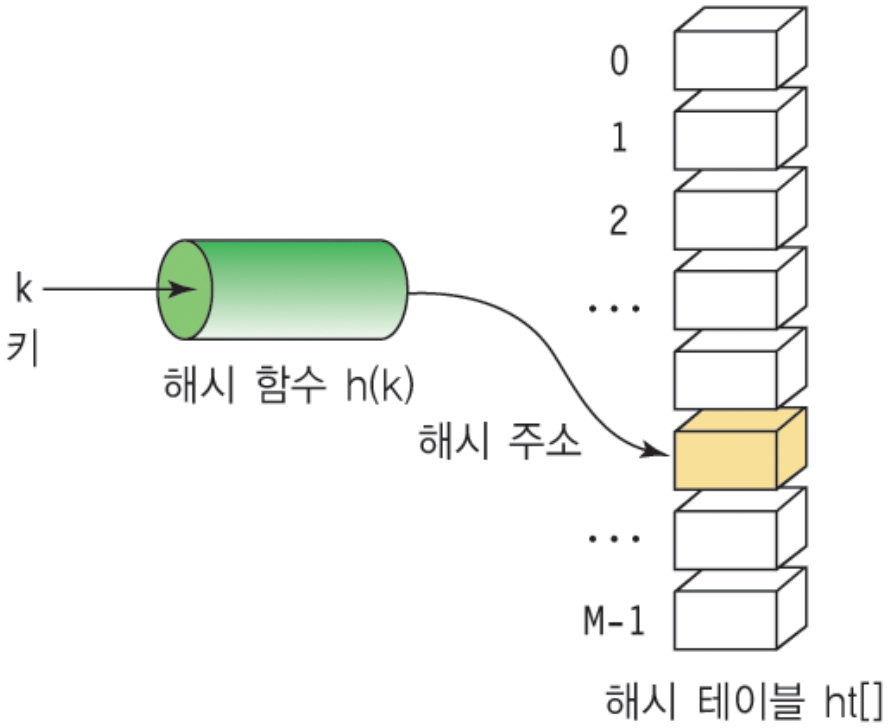


사전 ADT

·객체: 일련의 (key, value) 쌍의 집합
·연산:
- add(key, value) ::= (key, value)를 사전에 추가한다.
- delete(key) ::= key에 해당되는 (key, value)를 찾아서 삭제한다.
  관련된 value를 반환한다.만약 탐색이 실패하면 NULL를 반환한다.
- search(key) ::= key에 해당되는 value를 찾아서 반환한다.
  만약 탐색이 실패하면 NULL를 반환한다.

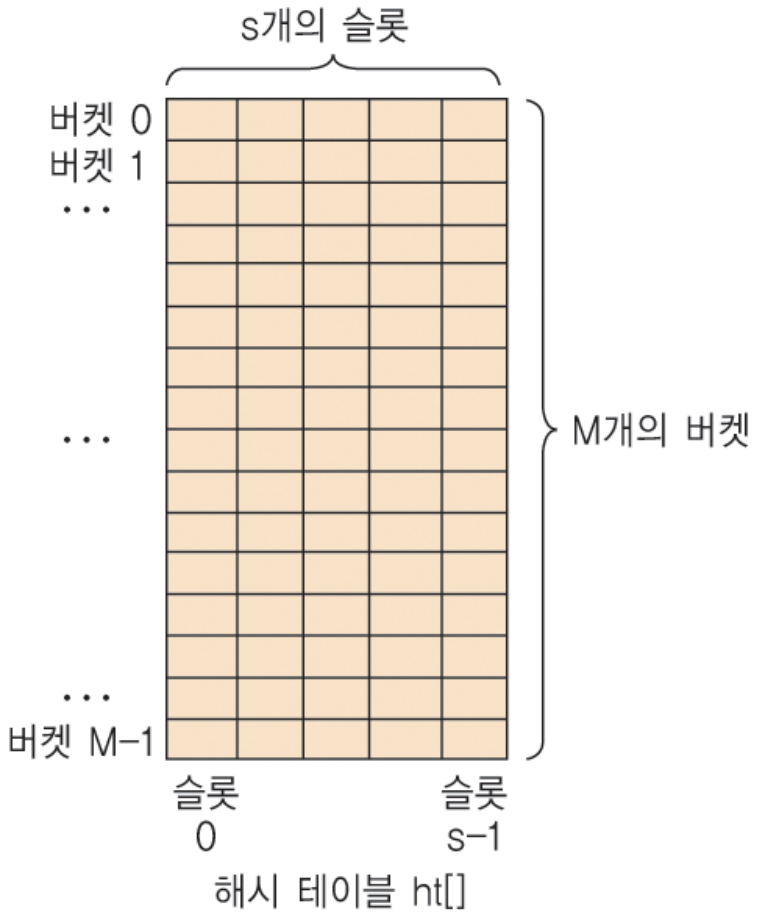
해싱의 구조

- 해시 함수(hash function)
  - 탐색키를 입력 받아 해시 주소(hash address) 생성
  - 해시 주소: 배열로 구현된 해시 테이블(hash table)의 인덱스



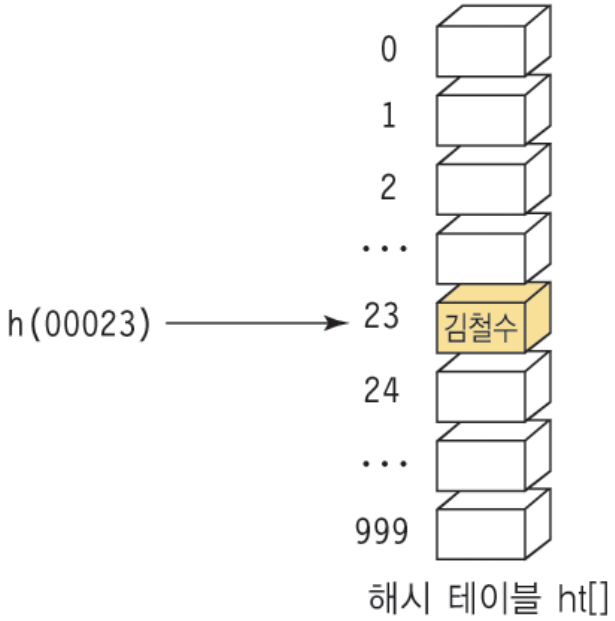
해시 테이블의 구조

- 해시테이블 ht
  - M개의 버켓(bucket)으로 구성된 테이블
  - ht[0], ht[1], ...,ht[M-1]의 원소를 가짐
  - 하나의 버켓에 s개의 슬롯(slot) 가능
- 충돌(collision)
  - 서로 다른 두 개의 탐색키 k1과 k2에 대하여 h(k1) = h(k2)인 경우
- 오버플로우(overflow)
  - 충돌이 버켓에 할당된 슬롯 수보다 많이 발생하는 것
  - 오버플로우 해결 방법 반드시 필요



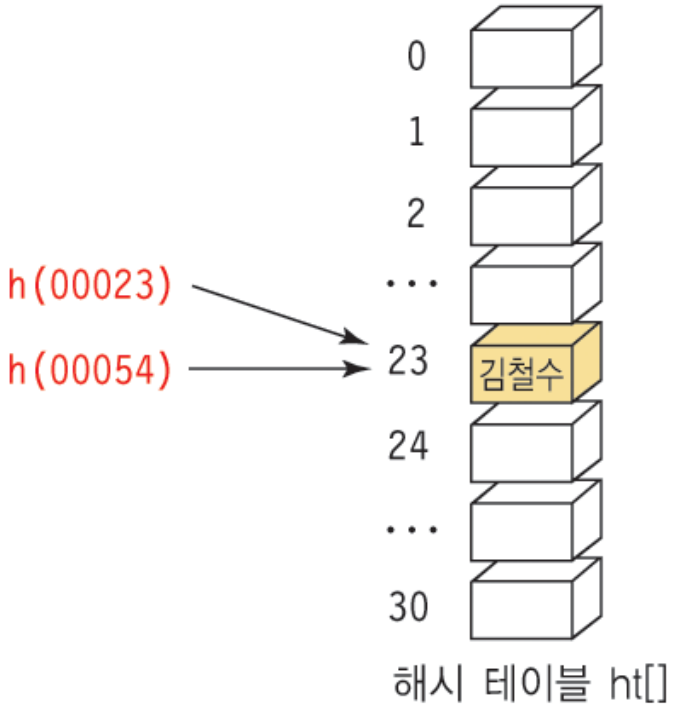
이상적인 해싱

- 학생 정보를 해싱으로 저장, 탐색해보자
  - 5자리 학번 중에 앞 2자리가 학과 번호, 뒤 3자리가 각 학과의 학생 번호
  - 같은 학과 학생들만 가정하면 뒤의 3자리만 사용해서 탐색 가능
  - 학번이 00023이라면 이 학생의 인적사항은 해시테이블 ht[23]에 저장
  - 만약 해시테이블이 1000개의 공간을 가지고 있다면 탐색 시간이 O(1) 되므로 이상적임



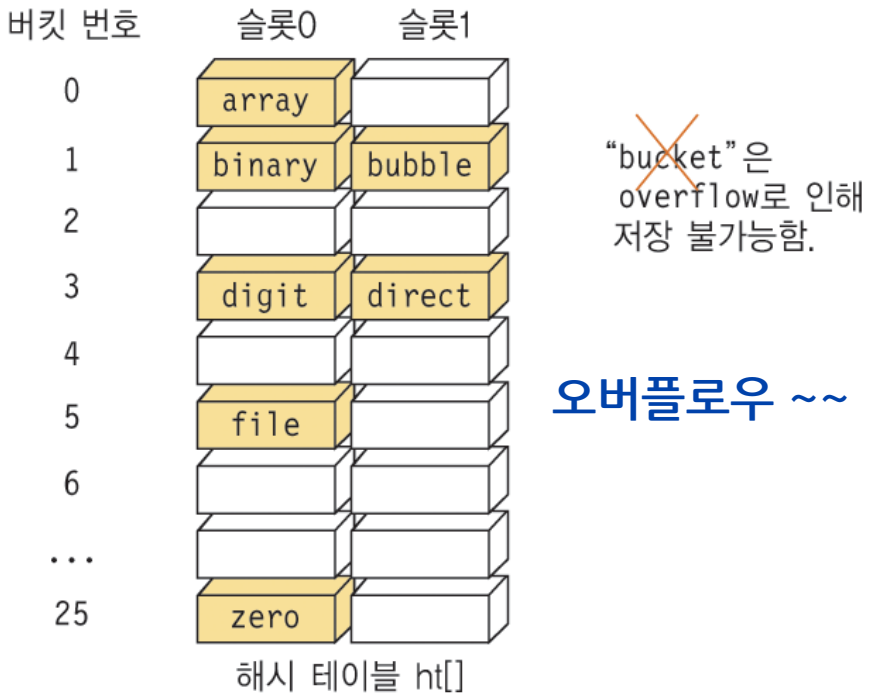
실제 해싱(1)

- 실제로는 해시테이블의 크기가 제한되므로, 존재 가능한 모든 키에 대해 저장 공간을 할당할 수 없음
- h(k)= k mod M 의 예에서 보듯이 필연적으로 충돌과 오버플로우 발생함



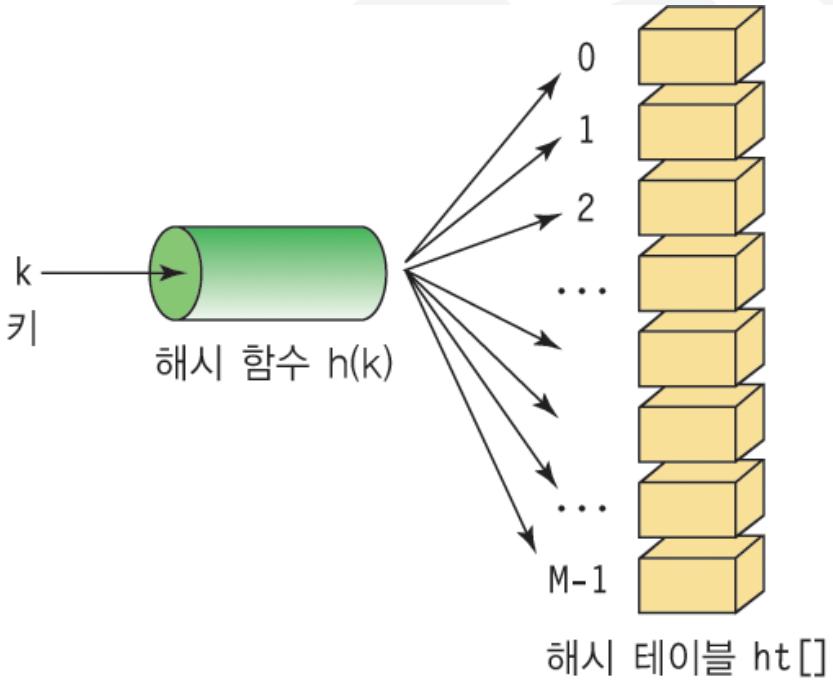
실제 해싱(2)

- 알파벳 문자열 키의 해시함수가 키의 첫 번째 문자의 순서
  - h("array")=1
  - h("binary")=2
- 입력데이터: array, binary, bubble, file, digit, direct, zero, bucket



해시 함수(1)

- 좋은 해시 함수의 조건
  - 충돌이 적어야 함
  - 해시함수 값이 해시테이블의 주소 영역 내에서 고르게 분포되어야 함
  - 계산이 빨라야 함

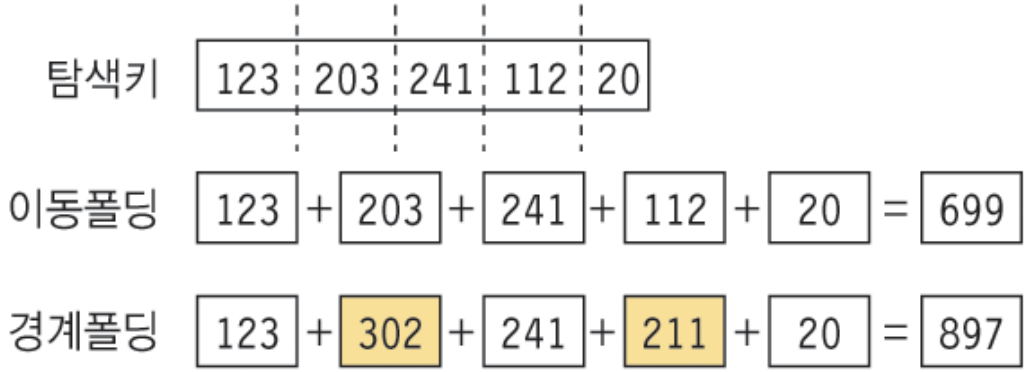


해시함수(2)

- 제산 함수
  - h(k)=k mod M
  - 해시 테이블의 크기 M은 소수(prime number) 선택 <- 흩뿌리기 위해서 짝수로 하면 너무 몰린다~

나머지로 연산함

- 폴딩 함수
  - 이동 폴딩(shift folding)과 경계 폴딩(boundary folding)



해시함수(3)

- 중간제곱 함수
  - 탐색키를 제곱한 다음, 중간의 몇 비트를 취해서 해시 주소 생성
- 비트추출 함수
  - 탐색키를 이진수로 간주하여 임의의 위치의 k개의 비트를 해시 주소로 사용
- 숫자 분석 방법
  - 키 중에서 편중되지 않는 수들을 해시테이블의 크기에 적합하게 조합하여 사용

충돌(Collision)

- 충돌(collision)
  - 서로 다른 탐색 키를 갖는 항목들이 같은 해시 주소를 가지는 현상
  - 충돌이 발생하면 해시 테이블에 항목 저장 불가능
  - 충돌을 효과적으로 해결하는 방법 반드시 필요

해시테이블

- 충돌해결책
  - 선형조사법: 충돌이 일어난 항목을 해시 테이블의 다른 위치에 저장
  - 체이닝: 각 버켓에 삽입과 삭제가 용이한 연결 리스트 할당

선형조사법(Linear Probing)

- 충돌이 ht[k]에서 발생했다면,
  - ht[k+1]이 비어 있는지 조사
  - 만약 비어있지 않다면 ht[k+2] 조사
  - 비어있는 공간이 나올 때까지 계속 조사
  - 테이블의 끝에 도달하게 되면 다시 테이블의 처음부터 조사
  - 조사를 시작했던 곳으로 다시 되돌아오게 되면 테이블이 가득 찬 것임
  - 조사되는 위치: h(k), h(k)+1, h(k)+2,...
- 군집화(clustering)과 결합(Coalescing) 문제 발생

선형조사법 예(1)

- 예: h(k)=k mod 7

1단계 (8) : h(8) = 8 mod 7 = 1(저장)
2단계 (1) : h(1) = 1 mod 7 = 1(충돌발생)
(h(1)+1) mod 7 = 2(저장)
3단계 (9) : h(9) = 9 mod 7 = 2(충돌발생)
(h(9)+1) mod 7 = 3(저장)
4단계 (6) : h(6) = 6 mod 7 = 6(저장)
5단계 (13) : h(13) = 13 mod 7 = 6(충돌 발생)
(h(13)+1) mod 7 = 0(저장)

	1단계	2단계	3단계	4단계	5단계
[0]					13
[1]	8	8	8	8	8
[2]		1	1	1	1
[3]			9	9	9
[4]					
[5]					
[6]				6	6

선형조사법 예(2)

- 예: “do”, “for”, “if”, “case”, “else”, “return”, “function”

탐색 키	덧셈식 변환 과정	덧셈 합계	해시 주소
do	100+111	211	3
for	102+111+114	327	2
if	105+102	207	12
case	99+97+115+101	412	9
else	101+108+115+101	425	9
return	114+101+116+117+114+110	672	9
function	102+117+110+99+116+105+111+110	870	12

버켓	1단계	2단계	3단계	4단계	5단계	6단계	7단계
[0]							function
[1]							
[2]		for	for	for	for	for	for
[3]	do	do	do	do	do	do	do
[4]							
[5]							
[6]							
[7]							
[8]							
[9]				case	case	case	case
[10]					else	else	else
[11]						return	return
[12]			if	if	if	if	if

이차 조사법(Quadratic Probing)

- 선형 조사법과 유사하지만, 다음 조사할 위치를 아래 식 사용
  - (h(k)+inc*inc) mod M
- 조사되는 위치는 다음과 같음
  - h(k), h(k)+1, h(k)+4,…
- 선형 조사법에서의 문제점인 군집과 결합 크게 완화 가능

이중해싱법(Double Hashing)

- 재해싱(rehashing)이라고도 함
  - 오버플로우가 발생하면 원 해시함수와 다른 별개의 해시 함수 사용
  - step=C-(k mod C)
  - h(k), h(k)+step, h(k)+2*step, ...

- 예: 크기가 7인 해시테이블에서,
  - 첫 번째 해시 함수가 k mod 7
  - 오버플로우 발생시의 해시 함수는 step=5-(5 mod 5)
  - 입력 (8, 1, 9, 6, 13) 적용

이중해싱법 과정

1단계 (8) : h(8) = 8 mod 7 = 1(저장)
2단계 (1) : h(1) = 1 mod 7 = 1(충돌발생)
(h(1)+h'(1)) mod 7 = (1+5-(1 mod 5)) mod 7 = 5(저장)
3단계 (9) : h(9) = 9 mod 7 = 2(저장)
4단계 (6) : h(6) = 6 mod 7 = 6(저장)
5단계 (13) : h(13) = 13 mod 7 = 6(충돌 발생)
(h(13)+h'(13)) mod 7 = (6+5-(13 mod 5)) mod 7= 1(충돌발생)
(h(13)+2*h'(13)) mod 7 = (6+2*2) mod 7= 3(저장)

	1단계	2단계	3단계	4단계	5단계
[0]					
[1]	8	8	8	8	8
[2]			9	9	9
[3]					13
[4]					
[5]		1	1	1	1
[6]				6	6

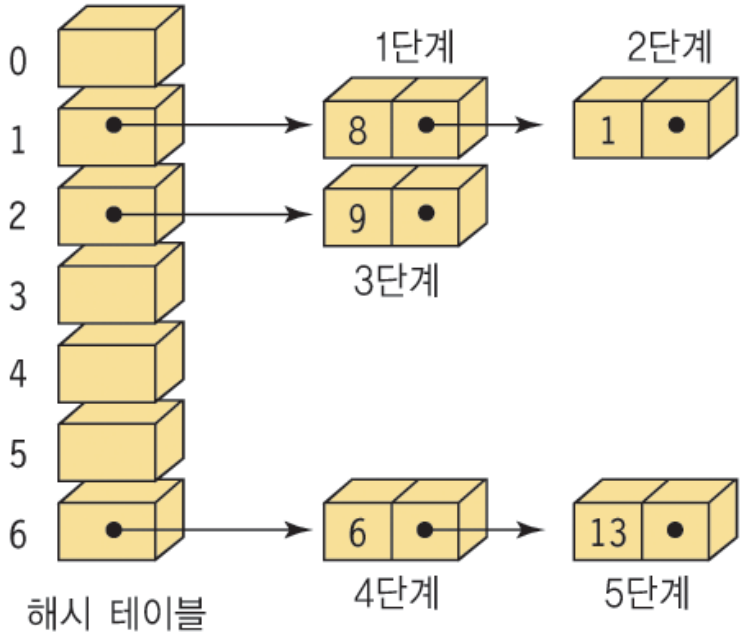
체이닝(Chaining)

- 오버플로우 문제를 연결 리스트로 해결
  - 각 버켓에 고정된 슬롯이 할당되어 있지 않음
  - 각 버켓에, 삽입과 삭제가 용이한 연결 리스트 할당
  - 버켓 내에서는 연결 리스트 순차 탐색

- 예: 크기가 7인 해시테이블에서
  - h(k)=k mod 7의 해시 함수 사용
  - 입력 (8, 1, 9, 6, 13) 적용

체이닝 과정

1단계 (8) : h(8) = 8 mod 7 = 1(저장)
2단계 (1) : h(1) = 1 mod 7 = 1(충돌발생->새로운 노드 생성 저장)
3단계 (9) : h(9) = 9 mod 7 = 2(저장)
4단계 (6) : h(6) = 6 mod 7 = 6(저장)
5단계 (13) : h(13) = 13 mod 7 = 6(충돌 발생->새로운 노드 생성 저장)



해싱의 성능분석

- 적재 밀도(loading density) 또는 적재 비율(loading factor)
  - 저장되는 항목의 개수 n과 해시 테이블의 크기 M의 비율

$$\alpha = \frac{\text{저장된 항목의 개수}}{\text{해싱 테이블의 버켓의 개수}} = \frac{n}{M}$$

- 선형 조사법에서의 비교 연산
  - 실패한 탐색: $\frac{1}{2}\left\{1+\frac{1}{(1-\alpha)^2}\right\}$
  - 성공한 탐색: $\frac{1}{2}\left\{1+\frac{1}{(1-\alpha)}\right\}$
- 체이닝에서의 비교 연산
  - 실패한 탐색: α
  - 성공한 탐색: $1+\alpha/2$

해싱의 성능분석: 선형조사법에서 비교연산 횟수

α	실패한 탐색	성공한 탐색
0.1	1.1	1.1
0.3	1.5	1.2
0.5	2.5	1.5
0.7	6.1	2.2
0.9	50.5	5.5

데이터 개수가 커지면 대부분 실패하게 됨

앞에서의 충돌이 뒤에서의 충돌의 확률을 높임

해싱의 성능분석: 체이닝에서 비교연산 횟수

α	실패한 탐색	성공한 탐색
0.1	0.1	1.1
0.3	0.3	1.2
0.5	0.5	1.3
0.7	0.7	1.4
0.9	0.9	1.5
1.3	1.3	1.7
1.5	1.5	1.8
2.0	2.0	2.0

앞에서의 충돌이 뒤에서의 충돌의 확률을 높이지 않는다는 점!

Week 14: Hashing