

# Cichelli's Minimal Perfect Hash Function algorithm example

## Cichelli의 최소 완전 해시 함수 알고리즘 예제

박혜웅

키 집합이 {DO, DOWNTON, ELSE, END, IF, IN, TYPE, VAR, WITH} 일 때,

$$h(s) = ( s.length() + g( first(s) ) + g( last(s) ) ) \% m$$

$s.length()$  = 문자열의 길이

$first(s) = s.charAt(0)$  = 문자열의 첫 문자

$last(s) = s.charAt(s.length()-1)$  = 문자열의 끝 문자

$m$  = 문자열(키)의 개수

를 이용하여 최소 완전 해시 함수를 작성하라.

<단계 1> 키 집합의 모든 키에 대하여 첫 문자와 끝 문자를 뽑아 중복없이 나열하고, 각 문자에 대한 빈도수를 계산하여 문자 빈도수 테이블을 만든다.

키 테이블		
키	첫문자	끝문자
DO	D	O
DOWNTON	D	O
ELSE	E	E
END	E	D
IF	I	F
IN	I	N
TYPE	T	E
VAR	V	V
WITH	W	H

문자 빈도수 테이블	
문자	빈도수
D	3
O	2
E	4
I	2
F	1
N	1
T	1
V	1
R	1
W	1
H	1

<단계 2> 각 키에 대하여 첫 문자와, 마지막 문자에 대한 빈도수의 합을 계산한다.

키 빈도수합 테이블					
키	첫문자	끝문자	첫문자 빈도수	끝문자 빈도수	빈도수합
DO	D	O	2	3	5
DOWNT0	D	O	2	3	5
ELSE	E	E	4	4	8
END	E	D	4	3	7
IF	I	F	2	1	3
IN	I	N	2	1	3
TYPE	T	E	1	4	5
VAR	V	V	1	1	2
WITH	W	H	1	1	2

<단계 3> 키 빈도수합 테이블을 빈도수합 내림차순으로 정렬한다.

정렬된 키 빈도수합 테이블			
키	첫문자	끝문자	빈도수합
ELSE	D	O	8
END	D	O	7
DOWNT0	E	E	5
DO	E	D	5
TYPE	I	F	5
IN	I	N	3
IF	T	E	3
VAR	V	V	2
WITH	W	H	2

<단계 4> 정렬된 키 빈도수합 테이블에서 해쉬값 집합의 원소가 중복되지 않도록, g(문자) 함수를 정의한다. 중복될 경우에는 백트래킹(backtracking)을 이용하여, g(문자) 함수를 재정의 한다. 해쉬값은 아래 공식을 이용하여 구한다.

$$h(s) = ( s.length() + g( first(s) ) + g( last(s) ) ) \% m$$

s.length() = 문자열의 길이

first(s) = s.charAt(0) = 문자열의 첫 문자

last(s) = s.charAt(s.length()-1) = 문자열의 끝 문자

m = 문자열(키)의 개수

계산 과정															
키	길이	g(문자)										해쉬값=h(s)	해쉬값 집합		
		E	D	O	T	I	N	F	V	R	W			H	
ELSE	4	0												4+0+0=4%9=4	{4}
END	3		0											3+0+0=3%9=3	{3,4}
DOWNT0	6			0										6+0+0=6%9=6	{3,4,6}
DO	2													2+0+0=2%9=2	{2,3,4,6}
TYPE	4				0									4+0+0=4%9=4	
TYPE	4				1									4+1+0=5%9=5	{2,3,4,5,6}
IN	2					0	0							2+0+0=2%9=2	
IN	2					1	0							2+1+0=3%9=3	
IN	2					2	0							2+2+0=4%9=4	
IN	2					3	0							2+3+0=5%9=5	
IN	2						1							2+3+1=6%9=6	
IN	2						2							2+3+2=7%9=7	{2,3,4,5,6,7}
IF	2							0						2+3+0=5%9=5	
IF	2							1						2+3+1=6%9=6	
IF	2							2						2+3+2=7%9=7	
IF	2							3						2+3+3=8%9=8	{2,3,4,5,6,7,8}
...	...													...	....
VAR	3								3	3				3+3+3=9%9=0	{0,2,3,4,5,6,7,8}
WITH	4										3	3		4+3+3=10%9=1	{0,1,2,3,4,5,6,7,8}

- 공식에서 m 은 문자열의 개수 9 이다.
- **붉은색** 숫자는 기존 해쉬값 집합의 원소와 해쉬값이 중복되었음을 의미한다.
- **파란색** 숫자는 각 문자에 대하여 결정된 g(문자) 값이다.

< 최종 > 완성된 해쉬 테이블과 g() 함수

문자->g(문자)	
문자	g(문자)
E	0
D	0
O	0
T	1
I	3
N	2
F	3
V	3
R	3
W	3
H	3

해쉬테이블	
키	h(키)
VAR	0
WITH	1
DO	2
END	3
ELSE	4
TYPE	5
DOWNTO	6
IN	7
IF	8