14장 Heaps and Priority Queues

14.5 허프만코드

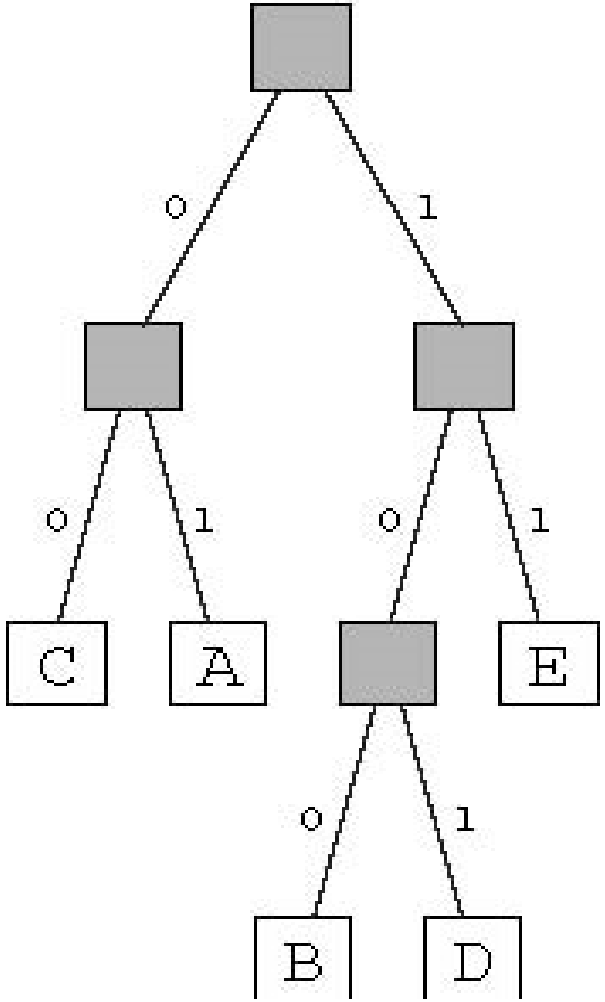
허프만 코드(1)

- 허프만 코드
  - 문서를 인코딩하는 최적의 알고리즘
  - 이 알고리즘은 가장 자주 나타나는 문자가 가장 짧은 코드를 갖도록 문자에 이진 코드를 부여하는데, 이것은 텍스트 문서를 최소 길이로 인코딩함
  - 허프만 코드(Huffman code)는 팩스, 모뎀, 컴퓨터 네트워크, 고해상도 텔레비전(high-definition television) 등의 실제적인 응용에 널리 사용되고 있음

허프만 코드(2)

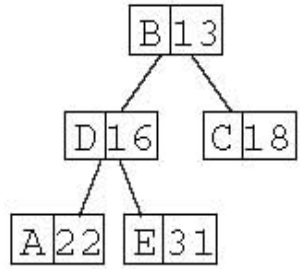
- 허프만 코드 생성 과정
  - 문서를 위한 허프만 코드는 문서에 나타나는 서로 다른 문자에 대해 각각 하나의 리프를 갖는 이진트리로부터 생성됨
  - 각각의 문자에 대한 코드는 그 문자에 대한 루트-리프 경로에 의해 결정되는데, 왼쪽 가지는 “0”으로 표시되고, 오른쪽 가지는 “1”로 표시됨
  - 오른쪽-왼쪽-오른쪽으로 가는 루트-리프 경로는 코드 101로 정해짐

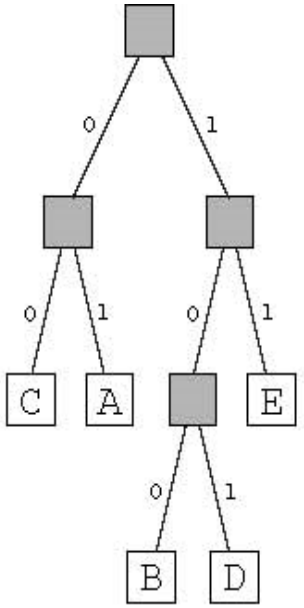
허프만 코드의 예



허프만 알고리즘

Letter	Freq.
A	22%
B	13%
C	18%
D	16%
E	31%





Letter	Code
A	01
B	100
C	00
D	101
E	11

빈도수 테이블 최소 히프 허프만 트리 허프만 코드

허프만 코드의 특성

- 허프만 트리가 만들어지면, 문서는 유일하게 인코드되고 디코드될 수 있다.
- 하나의 문자 코드가 다른 어떤 문자 코드의 접두부(prefix)와도 겹치지 않게 되는데, 이것을 유일 접두부 특성(unique prefix property)라고 한다

허프만 코드의 생성

- 허프만 코드 생성 알고리즘
  - 입력 : 문자의 시이퀀스.
  - 출력 : 입력 문자에 대한 비트 코드.
  - 후조건 : 비트 코드가 유일 접두부 특성을 가지고 최적임.

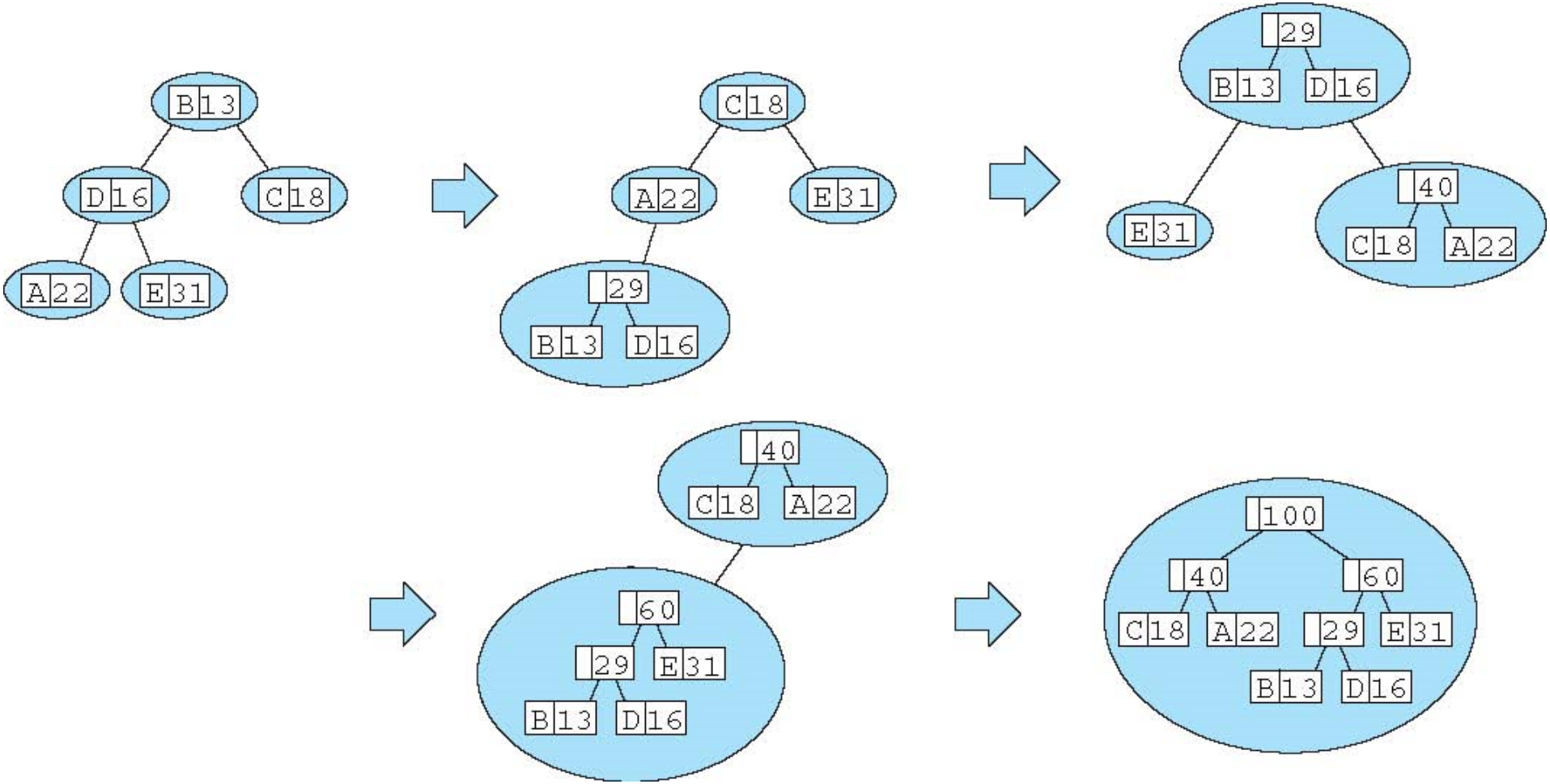
  1. 입력 문자에 대한 빈도수 표를 작성.
  2. 최소 우선순위 큐에 문자-빈도수 쌍을 적재.
  3. 이 쌍들을 허프만 트리로 합병.
  4. 루트-리프 경로 상의 비트 시이퀀스로 각각의 리프에 있는 문자를 인코드.

허프만 트리의 합병

- 허프만 트리 합병 알고리즘
  - 입력 : 정수로 이루어진 최소 히프 Q.
  - 출력 : 정수로 이루어진 허프만 트리 H.
  - 후조건 : Q의 원소가 H의 리프가 됨.

1. 각각의 원소를 자체가 단독 트리인 것으로 해석하여 *Q*를 재구성.
2. *Q*가 하나 이상의 원소를 가지고 있는 동안 단계 3-5를 반복.
3. *Q*로부터 최고 우선순위를 가지는 트리 *x*와 *y*를 삭제.
4. 자식 *x*와 *y*를 가지는 허프만 트리 *z*를 생성.
5. *Q*에 *z*를 추가.
6. *Q*의 나머지 원소를 리턴.

허프만 트리의 우선순위 포리스트 합병



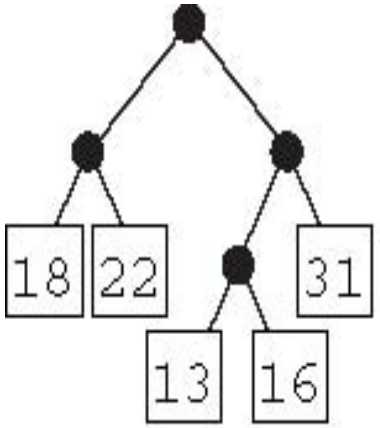
가중치 외부경로길이

- n개의 양의 가중치 $q_1,\ldots,q_n$ 이 n개의 외부노드에 1:1로 대응될 때, 가중치 외부 경로 길이(weighted external path length):

$$\sum_{1\le i\le n} q_i\, k_i$$

(k_i는 루트노드에서 가중치 q_i를 갖는 외부노드까지의 거리)

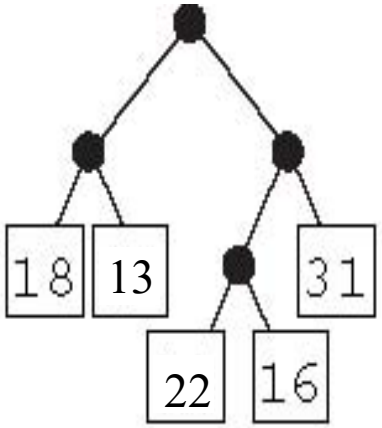
가중치 외부 경로 길이



18+22+31=71

13+16=29

WEPL = 2(71) + 3(29) = 229



18+13+31=62

22+16=38

WEPL = 2(62) + 3(38) = 238

- 허프만트리는 최소의 가중치 외부 경로 길이를 가지는 이진트리이다