

## 5. 확률적 분석과 랜덤화된 알고리즘

### 5.1 고용 문제

직업 소개소에서 면접자를 추천 받으려면 적은 소개료를 지불해야 한다. 소개받아 면담한 후 일을 잘할 수 있는 사람이면 해고하고 새로운 지원자를 고용하면 된다. 고용하면 많은 소개료를 지불해야 한다. 이 때 비용을 알고자 한다.

Hire-ASSISTANT( $n$ )

```
1 best = 0 // 0번은 가장 낮은 점수를 갖는 가상의 지원자
2 for i = 1 to n
3     지원자 i를 면접한다.
4     if 지원자 i가 지원자 best보다 나은가?
5         best = i
6     지원자 i를 고용한다.
```

위 의사코드가 비용을 의미한다.

여기서 말하는 비용은 그동안 말했던, 특히 2장에서 다룬 것과 다르다. HIRE-ASSISTANT의 수행시간이 아니라 면접과 고용에 드는 비용을 살펴보는 것이다. 하지만 수행된 기본 연산의 횟수를 세는 것이기 때문에 분석은 큰 차이가 없다.

면접 비용을  $c_i$ , 고용 비용을  $c_k$ 라 하면  $c_k$ 가  $c_i$ 보다 당연히 크다. 고용된 사람의 수를  $m$ 이라 한다면 이 알고리즘의 총 비용은  $O(nc_i + mck)$ 이다. 면접 보는 인원  $n$ 은 고용과 상관없이 고정되어있으므로  $mck$ 에 초점을 맞춘다.

### 최악의 경우 분석

최악의 경우는 모든 지원자를 고용해야 하는 경우이다.  $n$ 번 고용하므로 총  $O(nck)$ 의 비용이 든다. 이때는 지원자가 오는 순서대로 점수가 점점 높아질 때를 의미한다.

### 확률적 분석

**확률적 분석**은 문제를 분석할 때 확률을 사용하는 것이다. 주로 알고리즘의 수행시간을 분석하기 위해 사용된다. 확률적 분석을 하기 위해서는 입력의 분포에 대한 정보를 이용하거나 그에 관한 가정을 해야 한다. 평균 수행시간을 계산하게 되는데, 그러한 기댓값은 모든 가능한 입력의 분포에 따라 결정된다. 그러므로 가능한 모든 입력에 대한 수행시간의 평균을 구하게 되는데 이를 **평균 수행시간**이라고 한다. 적절한 입력 분포를 서술할 수 없다면 확률적 분석을 사용할 수 없다.

지원자가 무작위 순서로 온다고 가정해보자. 그렇다면 우리는 두 지원자를 비교해 누가 더 뛰어난지를 알 수 있다고 가정하는 것이다. 다시 말해, 지원자들의 전체 순서가 존재하게 된다. 이 순서를 1부터  $n$ 까지 정할 수 있는데, 순열 리스트를 의미하게 된다. 이 순열리스트는  $n!$ 가지가 존재하는데, 각각 같은 확률을 가진다. 이럴 때 이 순위는 **균등 임의의 순열(uniform random permutation)**을 이룬다고 한다. 가능한 각  $n!$  순열이 모두 같은 확률로 일어날 수 있다.

### 랜덤화된 알고리즘

확률적 분석을 이용하기 위해서는 입력의 분포에 대해 어느정도 알아야 하는데 쉽지 않다. 그래서 화굴과 무작위성을 통해 알고리즘의 일부가 랜덤하게 동작하도록 함으로써 알고리즘 설계와 분석을 위한 도구로 종종 사용한다.

지원자가 소개소에서 정말로 임의의 순서로 오는건지는 우리 입장에서 알 수 없다. 가정을 바꿔서, 직업 소개소에는  $n$ 명의 지원자가 있고 그들의 목록을 미리 보내준다고 해보자. 면접할 지원자는 매일 무작위로 고른다. 이 때 생기는 가장 중요한 변화는, 지원자가 무작위 순서로 오게 된다는 추측에 의존하는 대신 진행 과정을 제어할 수 있게 되어 무작위 순서로 오도록 시행할 수 있다.

일반화하면, 어떤 알고리즘의 동작이 입력뿐만 아니라 난수 **생성기**로 얻은 값들로 결정되면 이 알고리즘을 **랜덤화**되었다고 한다. 난수 생성기 RANDOM이 있을 때,  $\text{RANDOM}(a,b)$ 는  $a$  이상  $b$  이하의 정수를 동일한 확률로 반환한다.

랜덤화된 아목리즘의 수행시간을 분석할 때 난수 생성기가 리턴하는 값의 분포에 대한 수행시간의 평균 값을 구한다. 이 수행시간을 입력이 랜덤한 알고리즘의 평균 수행시간과 구분하기 위해 **기대 수행시간**이라고 부른다.

대체로 입력의 확률적 분포를 알 때는 평균 수행시간, 알고리즘 그 자체가 랜덤화된 경우에는 기대 수행시간을 분석한다.