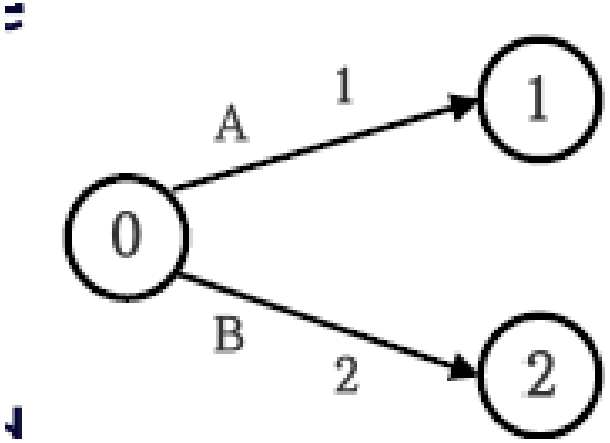
마르코프 과정

Making a single Decision

- 0이라는 state에서 A와 B라는 여러 개의 action이 있다.
  - Reward를 maximizing한다 : B라는 action을 취한다.



하지만 하나의 선택으로만 모델링 되는 경우는 거의 없다.

Markov Decision Processes

- (강화학습 기반모델)
- S : state들의 집합
- P : state transition function(state를 받아 다른 state로 매핑하는 함수), n개의 state가 있을 때 nxn 행렬로 표현 가능
- Memoryless random process : 현재 state만 알면 그 이전의 history는 몰라도 된다.
- Markov property : 현재 state가 주어지면 이 현재 state의 미래 일과 과거의 일이 서로 독립이다.

Example of Student MP

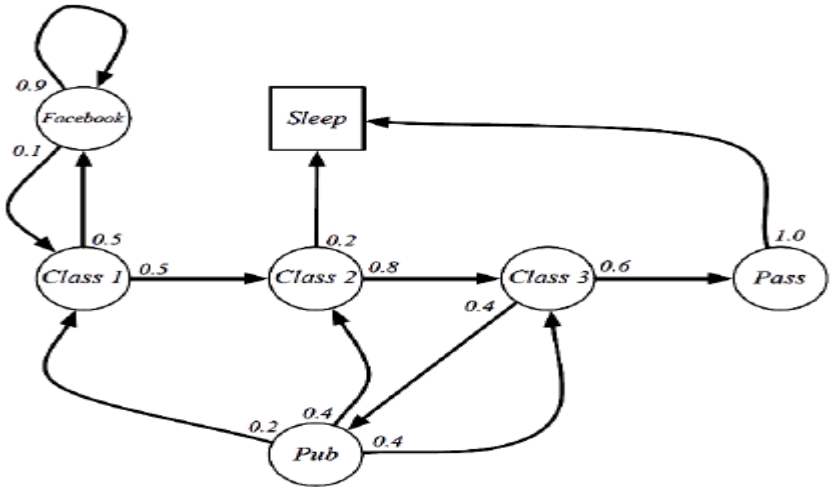
- **Sample episodes starting from $S_1 = C_1$**
  - C1 C2 C3 Pass Sleep
  - C1 FB FB C1 C2 Sleep
  - C1 C2 C3 Pub C2 C3 Pass Sleep
  - C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

- Episode : 현재에서 다음 state로 가는데 동전 던지기 같은 거로 결정해서 감.(random하게 끝나는 state까지 다음 state로 감)

ㅁ : 종료 state

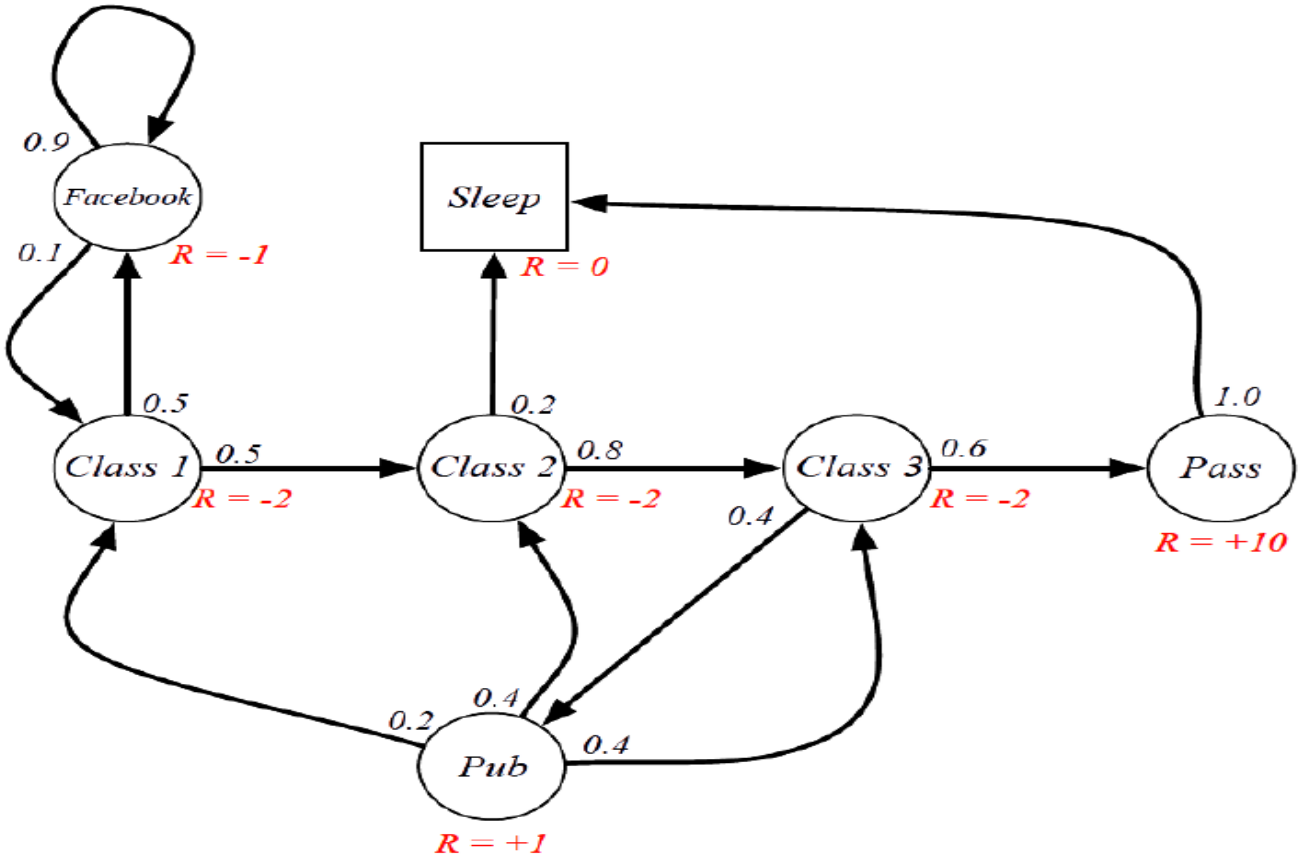
ㅇ : state

P : state transition function(Episode생성가능)



$$\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array} \begin{array}{c} \begin{array}{cccccccc} C1 & C2 & C3 & Pass & Pub & FB & Sleep \end{array} \\ \begin{bmatrix} & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{bmatrix} \end{array}$$

Markov Reward Processes

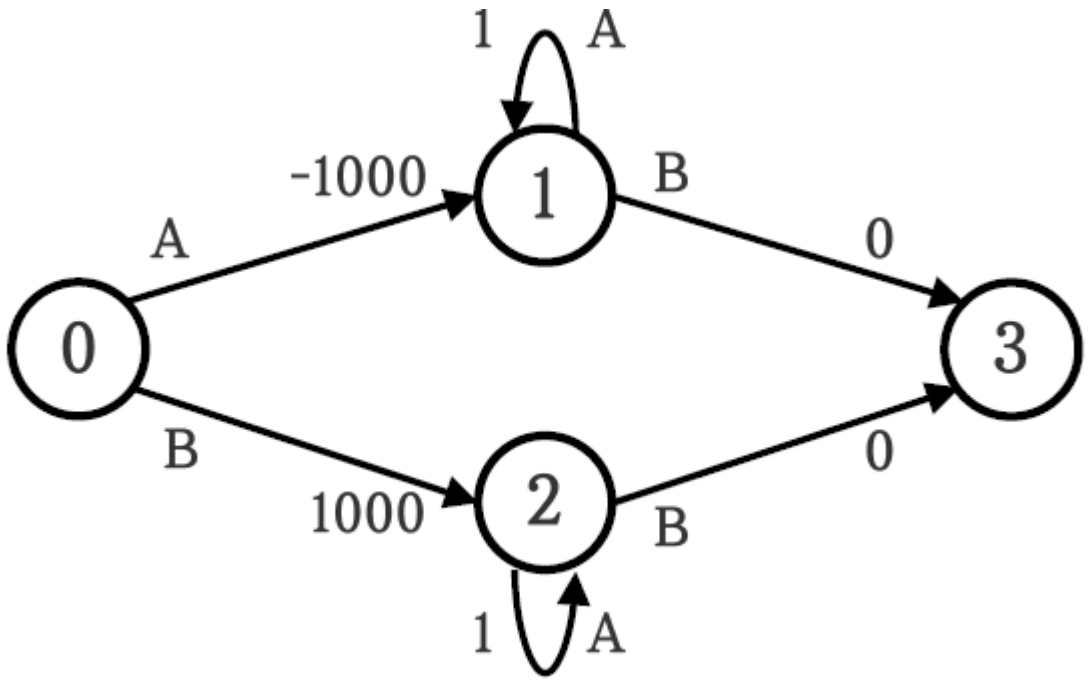


- Reward function : 어떤 state가 주어졌을 때 그 state에서 받을 reward 기댓값
- Discount factor : 스케일러 값, 0~1 값을 가짐

Return and Value function

- Return
  - Total discount reward from time step t
  - 현재 time step t로부터 해서 미래에 얻을 수 있는 reward를 다 합함
- Value function
  - 현재 state의 value는 내가 만약에 t라는 step에 s라는 state에 있을 때 내가 return을 Gt라고 했을 때 얻을 수 있는 return의 expectation 값
  - Expected return starting from state s

Why Discount Factor?

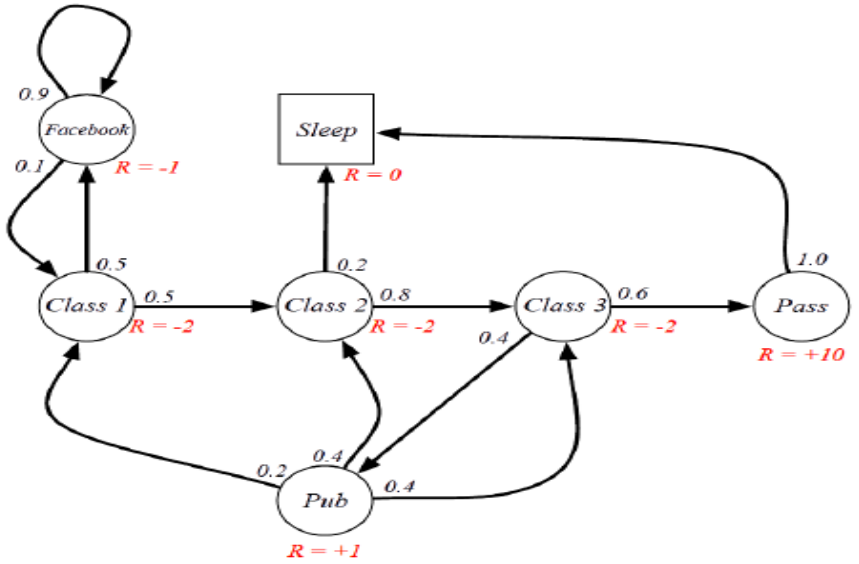


A만 계속 infinity 하게 취함

- 에이전트가 시간이 지날수록 살아남을 확률이 줄어듦
- 미래에 대한 불확실성 존재
- 당장의 reward는 아주 먼 미래의 reward보다는 더 가치 있게 평가받음
- Discount factor가 있다면 특정 값으로 수렴함

Example of Student MRP

- **Sample returns starting from $S_1 = C_1$ with $\gamma = 0.5$**
  - C1 C2 C3 Pass Sleep $\quad | V_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8} = -2.25$
  - C1 FB FB C1 C2 Sleep $\quad | V_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} = -3.125$
  - C1 C2 C3 Pub C2 C3 Pass Sleep $\quad | V_1 = \cdots$
  - C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep $\quad | V_1 = \cdots$



$$G_1 = R_2 + \gamma R_3 + \cdots + \gamma^{T-2} R_T$$

$$V(s) = \mathbb{E}[G_t | s_t = s]$$

G1 : state1에서의 return은 다음 step의 reward로부터 해서 계속 람다만큼 discount 하면서 끝까지 갔을 때의 reward의 총 합
V : 그것에 대한 expectation
100개를 샘플링 했을 때 V1값을 다 구하고 평균을 내면 V of C1, 즉 C1일 때 value function을 구할 수 있다.