PERSONAL PROJECT

reported by

20112096

Joonyoung Choi

**Design of data-sturcture**

# 0 프로젝트 개요

실생활에서 찾을 수 있는 dynamic graph로 표현될 수 있는 자료를 선택하고 graph 구조로 표현하는 방법과 실질적 데이터를 취합한다. 그렇게 구축한 주제의 graph가 나타내는 바를 파악하고 평가할 수 있도록 한다.

* 진행에 대한 내용은 아래에 링크된 Github를 통해 업데이트한다.
* [20112096 GitHub](https://github.com/mydream757/Design-of-DataStructure)
* 프로젝트 진행 내역은 지속적으로 첨가/수정될 수 있다.
* 다음에 진행 계획을 첨부한다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 9월 | | | | | | |
| S | M | T | W | T | F | S |
|  |  |  |  |  |  | 1 |
| 2 | 3 | 4 | 5 | 6 | 7  Problem definition & modeling | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20  Problem solving | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| 30 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 10월 | | | | | | |
| S | M | T | W | T | F | S |
| 30 | 1 | 2 | 3 | 4  Implementation & Data collection | 5 | 6 |
| 7 | 8 | 9 | 10 | 11  Testing & Evaluation | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 |  |  |  |

1. Problem definition & modeling (9/5~9/12)
   * 주제 조건 확인
   * 주제 선정 및 목표
   * 모델링
2. Problem Solving (9/13~9/30)
   * 주제 조건 확인
   * 주제 선정 및 목표
   * 모델링

# Problem definition & modeling

**주제 조건 확인**

주제(문제) 선정을 위해 다음과 같은 조건을 확인하였다.

* 데이터 간의 관계를 일정한 기준으로 비교할 수 있는가
* 실질적인 데이터를 수집할 수 있는가
* 그래프로 구성하였을 때 유용한가

**주제 선정 및 목표**

\*정재은 교수님의 피드백에 영향을 받아 주제의 방향성을 수정.

초기에 국내 방송 연예인들을 대상으로, 공통된 출연작이 있을 경우 엣지를 연결하는 그래프를 구성하려 하였다. 그것을 통해 공통된 출연작으로 연결되어있는 연예인들을 탐색하는 목적을 설정하였으나, 강의 시간 발표 때 받은 피드백을 통해 주제의 목적을 수정하여 프로젝트를 재구성하였다.

**케빈 베이컨 지수**

미국의 남자 배우인 Kevin Bacon은 굉장히 많은 작품에 출연한 것으로 유명한데, 이러한 특징에 기반해, 배우들을 출연작이라는 비교 데이터를 가지고서 케빈 베이컨과 관계적으로 얼마나 가까이에 있는지를 나타내는 지표이다.

이를 계산하는 규칙은 다음과 같다.

1.모든 배우들을 각각 그래프의 꼭짓점(vertex)으로 삼는다.  
2. 임의의 두 배우가 같이 출연한 작품이 있다면, 두 배우를 변(edge)로 잇는다.  
3. 임의의 배우의 베이컨 지수는, 해당 배우의 꼭짓점과 케빈 베이컨의 꼭짓점 사이의 거리로 정의한다.

4. 케빈베이컨과 연결되지 않는 배우는 무한대의 거리를 갖는다.(즉, 연결되지 않는다.)

<http://oracleofbacon.org/> 에서 케빈베이컨과 특정 배우 간 베이컨 지수를 확인할 수 있으며, <https://oracleofbacon.org/onecenter.php>에서 특정 배우의 평균 베이컨 지수를 확인할 수 있다.

이에 따르면, 케빈 베이컨은 할리우드 영화배우들에 대해 3.027의 평균 지수를 갖는다. 그렇다면 다작을 많이 한 배우가 평균 지수가 높은가? 케빈 베이컨보다 작품 수가 적은 Brad Pitt는 평균 지수가2.955였다.

**Problem Definition**

케빈 베이컨 지수를 확인하던 중 다음과 같은 사실을 발견했다.

* 작품에 많이 출연했다고 해서 실제로 평균 지수가 가장 작지는 않다.

이에 다음과 같은 의문이 발생한다.

* 같은 데이터 그룹군(영화배우)에 있는 구성원들의 평균 지수는 미세한 범위에서 일정한가? 그렇지 않다면 그 차이는 얼마나 벌어질까?
* 평균 지수가 작을수록, 유명한 배우인가? 혹은 많은 수입을 거두었는가? 평균 지수가 낮을수록 나타나는 특징에는 무엇이 있을까?

이러한 의문을 해결하기 위해, 이 문제를 본인의 상황에 맞게 모델링할 것이다. 그 내용은 다음과 같다.

**Modelling**

주제: 한국 영화 배우들의 관계를 출연작을 통해 표현하여 데이터를 분석한다.

데이터 모집군: 한국 영화 배우

데이터 수집: 네이버 인물 검색, 나무위키,한국 영화 진흥회

자료 구조: 무방향 그래프

* Node: 배우
* Edge: 배우 간 공통 출연작이 있으면 연결
* Weight: 없음
* 국내 모든 배우의 데이터를 수집하기에는 현실적 어려움이 따르므로, ‘케빈 베이컨 지수’의 사례와 유사하게 국내에서 다작을 가장 많이 했다고 평가되는 배우 ‘ 이경영’을 기준으로 자료를 수집할 것.
* 모인 데이터를 그래프로 구성한다.
* 다음은 구현 예정 목록이다.

0 그래프에 Node와 Edge를 추가/삭제할 수 있다.

1임의의 배우A와 다른 배우B의 최단 경로를 탐색/출력한다. 이를 통해 얻은 수치를 ‘A 지수’라 부른다.

2 A지수 계산을 A를 제외한 모든 다른 배우들에 대해 실행하여 평균을 구한다. 이를 ‘평균 A지수’라 한다.

3 모든 배우들에 대해 ‘평균X 지수’를 구해 ‘오름차순’으로 정리한다.

# Problem Solving

**개발 언어: C++**

1. 데이터 파싱

Ex)

|  |  |
| --- | --- |
| 배우 | 영화 |
| Person.name | Movie.name |
| Person.name | Movie.name |
| Person.name | Movie.name |
| …… | …… |

1. 자료 구조
   * 노드

|  |  |
| --- | --- |
| Person(배우) | Movie(영화) |
| Name | Name |
| Adj[] - 출연 영화 | Adj[] - 출연 배우 |
| 신상 정보 | 영화 정보 |

* + 결과값

|  |
| --- |
| Result |
| Distance |
| Route |

1. Methods
   * 노드 탐색

Search(node);

* + 데이터 입력

If(Search(node)=NULL) newNode(node);

Node.SetAdj(edge=연결 노드의 주소);

* + 데이터 삭제

Delete(Node);

(edge를 따로 삭제해야하는가?)

* + 기능구현

1 A노드와 B노드의 최단거리 및 루트 찾기

Result = FindbyBFS(A,B);

1. 결과 저장

Result.setResult(결과값);

1. 결과 출력

경로: Result.showRoute();

거리: Result.showDistance();