Kingo Manager

Kim junwoo,Kang Shingyu, Park Chunghyeon, Park Juhyeon

Sungkyunkwan University SWE3028_41
https://github.com/rlawnsdn/SWE3028_41_team_H

Abstract. 대학 생활을 하면서 본인의 진로를 위해 학점관리뿐만 아니라 다른 다양한 활동들을 하는 학생들이 많은데, 학기 중에는 과제나 프로젝트 등을 진행하기 바빠 이런 활동들에 대한 정보를 찾아보기 쉽지 않습니다. 또한 이런 활동들이 여러 홈페이지에 널리 퍼져 있기 때문에 자기 자신에게 맞는 활동이 무엇인지 찾는데 많은 시간을 소비하게 됩니다. 우리는 이러한 활동들을 찾는 시간과 노력을 줄여주기 위해 개개인에게 맞는 활동들을 추천해주는 어플리케이션을 개발하고자 합니다. 저희 프로젝트 kingo manager는 GNN을 기반으로 한 추천시스템을 이용하여 본인에게 맞는 추천과목, 동아리 모집공고, 학술대회, 채용공고 등의 정보를 제공해줍니다. 각 학생들의 학년, 관심분야를 고려하여 추천 시스템이 개개인별로 알맞은 정보를 추천해줍니다. 추천 받은 정보들은 메신저 어플처럼 stack형식으로 쌓이게 되고 언제든지 다시 찾아볼 수 있습니다. 어플리케이션은 Android studio를 이용하여 안드로이드 어플만 만들 예정이고, Figma를 이용하여 디자인하고, Aws를 이용하여 서버를 구성할 예정입니다. 현재는 성균관대학교 소프트웨어학과 학생들만을 대상으로 하는 어플을 만들 계획이지만 추후 프로젝트가 성공적으로 마무리되면 다른 과, 다른 학교를 대상으로 확장을 할 계획도 가지고 있습니다.

Keywords: GNN(Graph Neural Network)

1 Introduction

1.1 Suggested Problem

소프트웨어학과는 졸업 후 여러가지 분야로 취업할 수 있습니다. 예를 들어 프론트엔드, 백엔드, ios, 안드로이드, 임베디드 시스템, AI 등 다양한 종류의 커리큘럼이 존재하고 그에 따라 취업 시 요구되는 조건도 달라집니다. 하지만 저학년 학생들은 이에 대한 정보를 얻기 어렵고, 특히나 복수전공생인 경우에는 더 어렵습니다. 이런 학생들이 조금 더 일찍, 조금 더 쉽게 커리큘럼에 대한 정보를 얻을 수 있다면, 자신의 진로에 대해 좀 더 구체적인 계획을 세우고 대비해나갈 수 있을 것이라고 생각됩니다.

1.2 Previous Approaches

물론 각종 기업 공고가 올라오는 잡코리아, 워크넷 등의 사이트가 존재하지만 그런 사이트들은 성균관대 학생들이 주로 원하는 직장에 대한 공고보다는 알바, 일일직 등의 공고도 많이 섞여있기 때문에 아무리 키워드로 검색을 한다고 해도 본인에게 알맞은 공고들을 직접 찾아보아야 한다는 불편함이 존재합니다. 또한 교내 정보를

알림으로 제공해주는 kingo-M 어플이나 kingo-M안에 있는 기능인 kingo bot도 있지만, kingo-M의 경우 커리큘럼 관련 정보뿐만 아니라 다른 각종 학교 행사 정보 등의 원하지 않는 정보들도 같이 제공되고 알림의 내용을 보려면 게시판처럼 한번 클릭해서 들어가야 한다는 점에서 불편함이 있습니다. 카카오톡 플러스 친구처럼 원하는 정보를 메세지로 보내면 그에 알맞은 답변을 해주는 kingo bot의 경우 학교 졸업요건, 식단, 셔틀버스, 전화번호 조회, 캠퍼스 안내 등의 정보만 제공해줄 뿐 학교 내에서 일시적으로 발생하는 동아리 활동, 외부 대회, 기업 공고 등의 정보는 전혀 제공해주지 못합니다. 그 뿐만 아니라 kingo bot이 때때로 사용자가 무엇을 원하는지 정확히 파악하지 못하는 경우가 많아 학생들에게 큰 도움을 주지 못한다고 판단됩니다.

1.3 Proposed Solution

그래서 저희 프로젝트 kingo manager는 각 학년별, 커리큘럼별로 추천하는 과목, 동아리, 학술대회, 취업정보 등의 정보를 어플 하나에서 알림으로 받아볼 수 있게 하려고 합니다. kingo manager는 저학년 학생들에게는 커리큘럼별로 어떤 과목을 수강하면 좋은지, 기업에서는 어떤 능력을 요구하는지 등의 정보를 제공받아 좀 더 일찍 본인의 커리어를 준비해나갈 수 있고, 고학년 학생들에게는 본인이 혹시라도 알지 못했던 기업의 취업공고나 학술대회등의 정보를 제공하고 기업 홈페이지에서 일일히 취업 공고를 찾을 수고를 덜어줄 수 있습니다.

2 Motivation

친구와 대화를 하다가 친구가 창의품을 따기 위해서 Etest를 본다는 것을 들었습니다. 창의품을 따는 방법에는 여러가지가 있지만, 그중에서 Etest가 가장 쉽다는 사실을 듣고는 이 친구는 이런 정보를 어떻게 알고 있을까? 라는 생각이 들었습니다. 생각을 해보면 대학생활을 하면서 대부분의 정보를 스스로 찾아야 하는 경우가 많습니다. 간단하게는 수강신청 날짜부터, 졸업요건, 인턴십이나 동아리활동 같은 것들은 누군가가 알려주는 것이 아닌, 스스로 찾아야 하거나, 선후배간의 커넥션을 통해서 듣는 경우가 많다는 것을 알았습니다. 특히 코로나 시기로 선후배간 연계가 끊어진 점, 복수전공생에게는 선후배를 만나기 쉽지 않은 점 등을 고려해보면 구전으로 전해지는 정보들에 대해서 명문화의 필요성을 느꼈습니다.

3 Objective

1. 저희는 대학생활간 필요한 정보를 알람의 형태로 대학생에게 전달해주며 대학생에게 정보의 존재를 보여주는 것을 목표로 합니다. 예를 들어 톨게이트에 기업 인턴십이 올라왔다거나, 원하는 상담을 받을 수 있다는 정보를 알려주기만 할 뿐, 정보의 취사선택은 사용자에게 맡깁니다.

2. Kingo-M, 에브리타임 등의 다른 어플과 달리, 일반적인 대학생이 아닌, 성균관대학교 소프트웨어학과의 학생에게 맞는 정보만을 제공할 것이며 최소한의 정보를 제공하는 것을 목표로 합니다.

3. 정보를 취합하는 과정에서는 크롤링 기술을, 정보를 추천하는 과정에서 인공지능을 활용하는 것을 목표로합니다.

4. 학교 홈페이지, 챌린지스퀘어, 소프트웨어학과 홈페이지, 학생성공 게이트웨이 등 나누어져있는 정보를 취합해서 사용자에게 연결해주는 것을 목표로 합니다.

5. 구전으로 전해오는 정보들, 일부 학생만 활용하던 서비스 및 정보를 널리 알리는 것을 목표로 합니다.

6. 실제로 작동하는 어플리케이션을 만들어서 배포하는 것까지를 목표로합니다.

4 Background

4.1 Android Studio

안드로이드 스튜디오는 안드로이드 앱개발에 사용되는 IDE입니다. 이번 프로젝트에서는 안드로이드 언어로 Java를 사용할 것이며, 안드로이드 스튜디오는 Native 개발 패키지이기 때문에 크로스 플랫폼인 React 혹은 Flutter보다 구체적인 디자인의 어플을 구현할 수 있습니다. 깔끔한 디자인의 사용자 UI를 제공할 예정이며, 서버와 통신을 통하여 어플리케이션이 빠른 속도로 반응할 수 있습니다.

4.2 Aws

Aws(Amazon web services)는 웹 서버, 데이터베이스 및 클라우드 서비스를 제공합니다. 이번 프로젝트에서는 안드로이드 스튜디오에서 요청하는 데이터 정보를 송수신 하는 데에 aws 서버를 사용할 계획이며, aws 서버에는 GNN 기반의 추천 알고리즘이 함수 형태로 deploy 되어있습니다. 서버에 함수를 deploy하는 것은 aws lambda라는 기능을 이용할 것이며, python 기반의 Flask 서버를 먼저 제작한 후, aws 서버에 올리는 과정을 통해 AI 추천 알고리즘이 서버상에 등록되게 됩니다. 최종적으로, 서버에서 받은 어플리케이션의 요청에 맞추어 aws lambda 함수가 연산을 진행한 후, 다시 어플리케이션에 데이터를 전송합니다. 이를 통해 어플리케이션과 인공지능 알고리즘의 연동이 구현됩니다.

4.3 GNN

Graph는 vertices(node)와 edge로 구성된 구조를 말합니다.

$$G = (V, E) \tag{1}$$

(1)과 같은 식으로 나타낼 수 있는데, 각 node는 feature vector를 가지고 있습니다. 기존의 모델과 다르게 그래프를 사용해서 얻을 수 있는 차별점으로는 노드들 사이의 패턴을 찾는 데 집중하는 것이 아니라 노드들 사이의 관계에 집중한다는 것입니다.

GNN의 task로는 Node classification, Link prediction, Feature prediction이 존재합니다. Adjacency matrix(A)와 각 feature vector로 구성된 feature matrix(B)를 통해 GNN layer를 통과할 때마다 각 matrix의 정보를 업데이트 해줍니다.

$$h^{(l+1)} = \sigma(AH^{(l)}W^{(l)} + b^{(l)}) \tag{2}$$

$$H = AH \tag{3}$$

추가로, Adjacency matrix에서 각 노드에 연결된 가중치는 서로 다를 것이므로, Pagerank를 사용해 각 노드 사이에 어느 정도 연관성이 있느지를 가중치를 통해 나타냅니다. 이 Pagerank를 사용해서 GNN layer를 지날 때마다 가중치를 최신화 시켜 각 노드의 가중치가 수렴을 하게 됩니다. 가중치를 적용하기 위해서는 A에 가중치 matrix인 R을 곱해주면 됩니다.[1]

5 Related Work

5.1 Application

대학생활에 도움이 되는 어플리케이션으로 가장 잘 알려진 것들은 에브리타임, 캠퍼스픽, 링커리어, KINGO-M의 5개를 꼽을 수 있습니다. 가장 유명한 어플리케이션인 에브리타임은 처음에는 시간표 어플로 시작해서 점차 그 기능을 확대하여 지금에서는 게시판 베이스의 커뮤니티로서의 역할도 하며, 학교 중고나라, 취업과 진로에 대한 정보, 장학금이나 동아리, 학회에 대한 정보를 제공하고 있습니다. 다만 어플이 수동적인 면이 있어서 게시판을 활용할 생각을 하지 못하는 사용자들에게는 평범한 시간표 어플이 되며, 전국의 대학생들에게 일괄적으로 추천되는 정보들이 노출되어서 정보의 질이 정확하지 않은 편입니다. 같은 회사에서 출시한 캠퍼스픽 어플의 경우 에브리타임에서 동아리, 대외활동, 공모전, 스터디, 취업정보와 이벤트의 6가지 정보에 좀 더 집중한 어플로 보이지만 에브리타임과 마찬가지로 동아리의 경우에는 등록한 동아리만 노출되기 때문에 대학교 내 동아리 정보를 제대로 전달하지 못할 뿐더러, 같은 학교가 아닌 인근지역 동아리도 같이 노출되는 문제가 있었습니다. 또 대외활동이나 공모전 역시 전국 대학생을 대상으로 일반적이고 유명한 대회들에 초점을 맞추며 교내 대회 정보를 제공하지 못합니다. 링커리어는 대학생활 전반에 사용하기 보다는 취업시즌에 도움을 주는 링크드인 같은 모습을 보입니다. 마지막으로 KINGO-M은 우리학교에서 공식적으로 지원하는 어플리케이션으로 중요한 정보에대해서 알람의 형태로 알려주고, GLS 에 접근이 가능하지만 대부분의 사람들이 GLS 와 같은 기능을 제대로 활용하지 못하고 있는 점, 산학협력 톨게이트, 학생성공 게이트웨이, 챌린지스퀘어와 같은 곳에 올라오는 정보를 일부만 제공하는 점 등을 문제로 꼽을 수 있습니다. 또한 많은 알람은 사용자를 피곤하게 한다는 문제가 있습니다.

5.2 GNN(Graph Neural Network)

인터넷과 광고, E-commerce가 발달하면서 다양한 추천 알고리즘이 소개되고 발달하고 있습니다. 그중에 인터넷에서 대표적으로 사용되는 알고리즘은 Session에서 주어지는 정보를 바탕으로 하며(Session-based Recommendation with Graph

[1] http://web.stanford.edu/class/cs224w/를 통해 공부한 내용 요약

Neural Networks Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, Tieniu Tan) 실시간으로 웹 페이지를 사용하고 있는 사용자의 Session 정보를 이용해서 general interests 와 current interests를 구하는 것에 초점을 두고 있습니다. 하지만 저희는 보다 정적이며, 사전에 입력된 정보를 바탕으로 추천하는 알고리즘을 기획하는 것으로 보다 전통적이고 정적인 그래프를 바탕으로 GNN을 적용할 생각입니다. Directed Mapped Graph가 적용된 GNN([Scarselli et al. 2009] The graph neural network model.)과 거기에 더해서 back-propagation through time(BPTT)이 적용된 GNN([Gori, Monfardini and Scarselli 2005. A new model for learning in graph domains.)을 참고할 것이며, script event prediction (Li, Ding, and Liu 2018), situation recognition (Li et al. 2017b), 그리고 image classification (Marino, Salakhutdinov, and Gupta 2017)에 사용된 예시가 있습니다.

6 Problem Statement and Proposed Solution

Problem Statement (Kingo-M) 학생들은 대학생활을 하면서, 학년이 올라갈 수록 졸업요건과 대외활동에 대해서 더욱 신경을 쓰려고 합니다. 하지만, Kingo-M 같은 경우 대외 활동이나 취업 도움이 되는 정보들을 알림으로 알려주기는 하지만, 정작 학생들이 가장 원하는 정보를 알려주는 것이 아닌, 모든 사람에게 어떤 프로그램이 열리는 것을 전체적으로 공지하는 방식으로 알려줍니다. 그래서, Kingo-M 알람을 보지 않는 사람들이 많이 존재합니다. 저희는 이러한 문제점에서 시작하여 이러한 단점을 해결할 수 있는 방법을 찾고자 했습니다. 학생들이 자신이 가장 필요로 하고 있는 정보를 알 수 있다면, 학생들이 대학생활을 하면서, 신경을 쓸 부분을 덜어줄 수 있기 때문입니다. 이러한 추천 시스템을 만들기 위해서 저희는 GNN을 활용하기로 결정을 했습니다.

Proposed Solution (4 steps) 먼저, GNN을 적용하기에 앞서서, 학생들의 학업 정보를 데이터로 받아와 라벨링을 통해 컴퓨터가 학생들의 정보를 저희가 원하는대로 해주는 작업이 필요합니다. 이 작업을 해줘야 각 학생들을 그래프에서 node라 했을 때, feature vector를 나타낼 수 있기 때문입니다. 라벨링을 통해 node와 feature vector를 정하는 것이 완료가 되었다면, node embedding 작업까지는 완료가 된 것으로 볼 수 있습니다.

두 번째로, 추천 알고리즘을 만들어야 합니다. NGNN[2]을 사용하여 proper outfit을 추천하는 것처럼 GCN을 활용하여 입력받은 새로운 node의 feature마다 딥러닝을 기반으로 알맞은 활동을 추천해 주는 것을 목표로 합니다. 이것을 학습시키기 위해서 많은 수의 학생들의 정보가 필요합니다.

세 번째로, 앞에 정리한 데이터를 가지고 모델을 학습시킵니다. 이것을 잘 학습시키기 위해서 '학년', '들은 학점 수', '3품 충족 여부' 등과 관련된 정보들로 데이터를 잘 나눠야합니다. 이 정보들은 feature vector에 존재합니다. 그리고 가능하면 데이터를 test와 valid로 나눠서 만든 GNN 모델의 성능을 계속 테스트해 볼 것입니다.

마지막으로, F1, accuracy, 또는 AUC 같은 다양한 평가 방식이 있지만, 모델의

[2] https://github.com/CRIPAC-DIG/NGNN

성능평가는 Cross Entropy로 진행을 할 예정입니다. 처음에는 데이터를 모두 라벨링을 할 수 있을 것이라고 가정을 하고 진행을 할 예정이어서 supervised learning이라고 가정을 한 상태로 진행을 하겠지만, 라벨링 과정 중에 데이터를 원하는 대로 작업이 불가능할 경우에 다른 방식을 생각해 나아갈 것입니다.

7 Plan Considerations

7.1 Plan

저희 팀에서는 전체 개발 과제를 어플리케이션 개발 파트와 AI 알고리즘 개발 파트로 나누었습니다. 또한 어플리케이션 개발 파트는 프론트엔드와 벡엔드로 나눠지며, AI 알고리즘 개발은 데이터셋 제작 및 크롤링 파트와 AI 모델 개발 파트로 나누어집니다. 프론트엔드는 진로활동 정보를 사용자가 쉽게 보고, 파악할 수 있도록 깔끔한 UI로 보여주는 역할을 하고, 백엔드는 사용자 정보 및 추천 알고리즘 정보를 서로 연동할 수 있도록 서버와 데이터베이스를 생성 및 관리합니다. 또한 데이터 크롤링은 GNN(Graph Neural Network)에서 각각의 Graph node가 가리키는 진로활동 정보를 수집하는 것이며, 데이터셋 제작은 사용자 정보로부터 Graph node 간의 연관성 벡터(adjacent vector)를 생성하는 것을 의미합니다. 마지막으로 AI 모델 개발은 연관성 벡터를 이용해 Graph node들 간의 연관성 parameter를 최신화하고, 추천 진로활동을 선정해 주는 AI model을 개발하는 것을 의미합니다. 최종적으로, 전체 개발과제를 앞서 설명한 4개의 subtask로 나누었고, 저희 팀의 4명의 팀원이 각각 한개의 subtask에 분담되었습니다. 구체적인 계발 계획은 아래와 같습니다.

No.	수행내용	3-4주	5-6주	7-8주	9-10주	11-12주	13-14주	15-16주
1	안드로이드 앱 프론트엔드 개발							
2	안드로이드 앱 백엔드 개발							
3	알고리즘 모델 선정							
4	데이터셋 크롤링 범위 확정							
5	데이터셋 수집 및 제작							
6	추천 알고리즘 개발							
7	앱 프론트엔드 & 백엔드 연동							
8	추천 알고리즘과 앱 연동							
9	실사용 테스트 및 프로그램 디버깅							

7.2 Considerations

Supervised or Unspervised learning? AI 추천 알고리즘으로 GNN(Graph Neural Network) 기반의 모델을 사용할 것이며, 진로 활동 정보를 각 Graph node로 맵핑하고, 생성된 node 간의 연관성을 사용자로부터 얻어서 주기적으로 학습을 진행할 것입니다. 이는 사용자로부터 얻는 연관성 정보를 GNN의 연관성 백터로 라벨링하여 데이터셋을 제작한 후 Graph node들 간의 연관성 parameter를 학습시키는 작업이므로 Supervised learning모델로 분류할 수 있습니다.

Dataset 앞서 설명드린 것과 같이, 이번 프로젝트에서는 진로활동 정보를 Graph node에 단순히 맵핑시키고, 사용자 정보로부터 연관성 백터를 라벨링하여 데이터셋을 얻을 것입니다. 이것은 진로활동 추천 알고리즘을 자체적으로 개발하는 과정이므로, 오픈소스 데이터셋을 얻는 것은 어려우며, 주기적으로 측정되는 연관성 백터로부터 계속하여 모델을 학습하는 방식을 사용할 것입니다.

Goal 이번 프로젝트의 AI 모델은 사용자의 학사 현황과 관심사 등을 고려하여 맞춤 진로활동을 추천해 주는 것입니다. 처음 데이터 크롤링의 범위를 고려할 때, 각종 취업활동을 포함한 대내 및 대외활동으로 설정하였지만, 광범위한 데이터가 존재하며 서로 다른 정보의 주체에서 생산하는 취업활동 정보나 각종 대외활동 정보를 자동화된 크롤링 기법으로 얻는 것은 어렵다고 여겨져, 서비스 범위를 성균관대학교 대내활동으로 축소하였습니다. 특히, 서비스 대상은 소프트웨어학과 학생들로 하여 1차적으로 서비스를 개발한 후, 서비스의 문제점을 계속해서 보완하면서 점차 서비스 범위를 확대할 계획입니다. 다음으로, 모델 선정은 추천 알고리즘을 구현하기 위해서 GNN 기반의 모델을 사용할 계획이며, 인접한 Graph node들 간의 연관성을 측정하기 위해서 Objective Function으로는 cross entropy loss 를 사용할 계획입니다. 인접한 노드들 관의 연관성은 기본적으로는 연관되어 있는지 여부를 0과 1로 맵핑하고, 여러 겹의 레이어와 parameter를 사용하여 연관성을 학습하는 방법을 구상하고 있습니다.

GNN model GNN(Graph Neural Network)는 유튜브(Youtube), 핀터레스트(Pinterest) 등의 다양한 추천 알고리즘 서비스에서 사용되는 모델입니다. GNN은 Graph의 각 노드들이 인접한 노드와 연결되어 있는지 여부를 0 과 1로 치환하여 각각의 노드들과 상호 연결 상태를 벡터로 표현합니다. 이후 각 노드들 간의 연관도 정도를 parameter로 설정한 후, 사용자 정보에 맞추어 연관도를 학습합니다. 결과적으로 GNN model은 사용자가 관심있어 할 만한 진로활동 정보를 추천해 주기에 적합하다고 여겨져, 이번 프로젝트에서 사용할 모델로 선정하였습니다. 또한, GNN model은 다양한 버전의 알고리즘으로 사용되는데, 커널 또는 필터를 사용하여 학습하는 GCN(Graph Convolution Network) 모델도 존재하며, 이웃한 Graph node 사이에 서로 다른 attention을 부여하여 자기 자신의 값을 업데이트 하는 GAT 방식도 존재합니다. 프로젝트를 진행하면서 이처럼 다양한 버전의 GNN model을 사용해 보면서 추천 알고리즘에 가장 적합한 모델을 찾아볼 것입니다.

Input and Output Form AI model의 데이터셋은 사용자 정보로부터 얻어진 연관성 벡터를 사용합니다. 따라서 AI model의 입력은 이러한 연관성 벡터입니다. 데이터셋을 만들기 위해서 Labeling되지 않은 사용자 정보로부터 연관성 벡터를 제작하는 전처리(pre-processing)과정을 거칠 것입니다. 이후 데이터셋으로부터 모델을 학습하고, 가장 높은 연관도를 보인 node의 진로활동 정보를 순서대로 출력할 것입니다.

Limitations 학사 진로활동 정보 및 학생의 학사 현황 등을 수집하기 위해서 여러 제약 요건들이 존재합니다. 먼저 학사 진로활동 정보를 얻기 위해서는 학부 공지사항, 동아리 게시판, 톨게이트 게시판 등 다양한 정보 주체로부터 데이터를 크롤링 해와야 합니다. 이러한 과정을 모두 수작업으로 하는 것은 매우 번거롭기 때문에,

반자동화된 크롤링 기법의 개발이 필요하다는 제한사항이 존재합니다. 또한 학생의 졸업요건과 3품 충족 등의 학사현황을 학교 서버와 연동하여 받아오는 것을 계획중인데, 학교와의 기술적, 행정적인 협조를 구해야 하는 어려움이 존재합니다.

References