creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

협업적 필터링을 활용한 추천 채용 시스템의 설계와 구현

Design and Implementation of Employment Recommender System using Collaborative Filtering

2016년 8월

서울대학교 대학원

컴퓨터공학부

박 수 상

협업적 필터링을 활용한
추천 채용 시스템의 설계와 구현

Design and Implementation of
Employment Recommender System
using Collaborative Filtering

지도 교수 문 병 로

이 논문을 공학석사 학위논문으로 제출함
2016년 4월

서울대학교 대학원
컴퓨터공학부
박 수 상

박수상의 공학석사 학위논문을 인준함
2016년 6월

위 원 장 박 근 수 (인)

부위원장 문 병 로 (인)

위 원 Srinivasa Rao Satti (인)

초 록

1990년대 중반 이후 데이터의 양이 폭발적으로 증가함과 동시에 빅데이터 정보 처리 기술이 진화하면서 데이터를 효과적으로 수집, 처리, 분석 가능해졌다. 이러한 데이터 분석 기술 중 하나가 추천 시스템(Recommender System)이다. 책, 상품, 영화 등 다양한 분야에서 사용자의 데이터를 바탕으로 개인화된 항목을 추천해준다. 본 논문에서는 이러한 추천 시스템을 채용 분야에 적용하여, 구직자가 선호할만한 채용 공고를 추천해서 채용 공고를 찾기 위한 불편함을 줄이고자 한다. 나아가 적절한 구직자에게 적절한 구인 기업을 추천하여 취업의 미스매치를 완화하는데 도움이 되고자 한다.

본 연구에서는 여러 추천 알고리즘 중 메모리 기반의 협업 필터링(Memory-based Collaborative Filtering)을 사용해서 각기 다른 데이터를 사용하여 두 가지 추천 시스템을 구축한다. 첫 번째 채용 공고 클릭 로그 데이터를 활용하여 사용자의 기업에 대한 선호도를 바탕으로 기업을 추천하는 시스템이며, 두 번째는 지원한 기업/직무의 자기소개서 작성 이력 데이터를 활용하여 사용자의 기업/직무에 대한 선호도를 바탕으로 기업/직무를 추천하는 시스템이다.

또한, 하이브리드 형식의 추천 시스템도 구현을 하였다. 기업 추천 시스템과 기업/직무 추천 시스템의 선호도 예측 값을 가중합(Weighted Sum)하여 하이브리드 방식의 추천 시스템을 구현하였고, 이런 하이브리드 방식이 추천에 어떤 영향을 주는지 알아보았다.

주요어 : 추천 시스템, 협업적 필터링, 추천 채용 시스템, 취업 미스매치
학 번 : 2014-21768

목 차

제 1 장 서 론 1
제 1 절 연구의 배경 1
제 2 절 연구 목적 2
제 3 절 논문의 구성 4

제 2 장 관련 연구 5
제 1 절 추천 시스템 5
1.1.협업적 필터링 6
1.2.내용 기반 필터링 8
1.3.하이브리드 추천 시스템 8

제 3 장 문제 정의 10
제 1 절 문제 정의 10
1.1 양질의 데이터 수집 10
1.2 추천 시스템 구성 11

제 4 장 추천 알고리즘의 설계와 구현 12
제 1 절 전체 시스템 구성 12
제 2 절 데이터 수집부 12
제 3 절 데이터 전처리부 13
제 4 절 추천 채용 시스템 15

제 5 장 실험 결과 18

제 6 장 결론 및 향후 연구 20
제 1 절 결론 20
제 2 절 향후 연구 20

참고문헌 22

Abstract 24

표 목차

[표 1] 세무직무 카테고리 표..... 15
[표 2] 추천 채용 시스템에 사용된 데이터 종류와 수..... 18
[표 3] 각각 추천 시스템의 RMSE 결과..... 18
[표 4] 하이브리드 추천 시스템의 RMSE 결과..... 19

그림 목차

[그림 1] 아마존닷컴의 추천 시스템을 통한 책 추천..... 2
[그림 2] Netflix의 추천 시스템을 통한 영화 추천 2
[그림 3] 사람인의 광고 중심의 채용 공고..... 3
[그림 4] 전체 시스템 구성 12
[그림 5] 추천 시스템별 RMSE 결과..... 19

제 1 장 서 론

제 1 절 연구의 배경

1990년대 중반 이후 인터넷의 급격한 보급과 모바일, 사물인터넷의 확산으로 정보 생산이 촉진되면서 데이터의 양이 폭발적으로 증가하였다. 동시에 빅데이터 정보 처리 기술이 진화하면서 폭발적으로 증가한 데이터를 효과적으로 수집, 처리, 분석이 가능해졌으며, 많은 데이터의 축적과 빅데이터 처리 기술의 발전을 토대로 다양한 분야에서 데이터를 기반으로 한 부가 가치를 창출하고 있다.[1]

데이터를 분석하여 부가가치를 창출하고 있는 기술 중 하나가 추천 시스템이다. 추천 시스템은 사용자의 과거 행동 데이터를 바탕으로 사용자의 기호와 행동을 예측하고, 더 나아가 새로운 기호나 행동을 추천할 수 있다. 유통, 온라인 커머스, 뉴스 등 다양한 분야에서 추천 시스템을 활용하고 있으며, 개인에게 최적화된 추천을 통해 기존 판매 상품에 차별화된 가치를 제공하고 있다.

온라인 커머스 '아마존닷컴[①]' 의 경우, 추천 시스템을 활용하여 고객이 구입하거나 열람한 상품 정보를 분석하여 구매 예상 상품을 추천하고 개인화된 쿠폰을 제공하고 있다. 이를 통해 연 매출의 35%가 추천 시스템을 통해 추천된 상품에서 발생하며, 매년 이익의 10%를 추천 시스템의 성능 향상에 투자하고 있다.[2]

① Amazon.com

그림 1. 아마존닷컴의 추천 시스템을 통한 책 추천

온라인 영화 서비스인 Netflix는 10만 여 개의 영화 콘텐츠와 3,000만 명 사용자의 대여, 시청 이력, 감상평 등을 분석하는 '시네매치(cinematch)' 라는 자체 영화 추천 시스템을 활용하여 하루 평균 50억 건의 영화 콘텐츠를 추천하고 있다. Netflix는 '시네매치'의 성능을 향상 시키기 위해 추천 시스템의 성능을 개선하는 경연대회인 Netflix Prize를 진행하고, 다양한 추천 알고리즘을 접목하는 등 기술 향상에 집중하고 있다.[3] 그 결과 Netflix 사용자 중 75%가 추천 받은 영화 콘텐츠를 이용하고 있다.

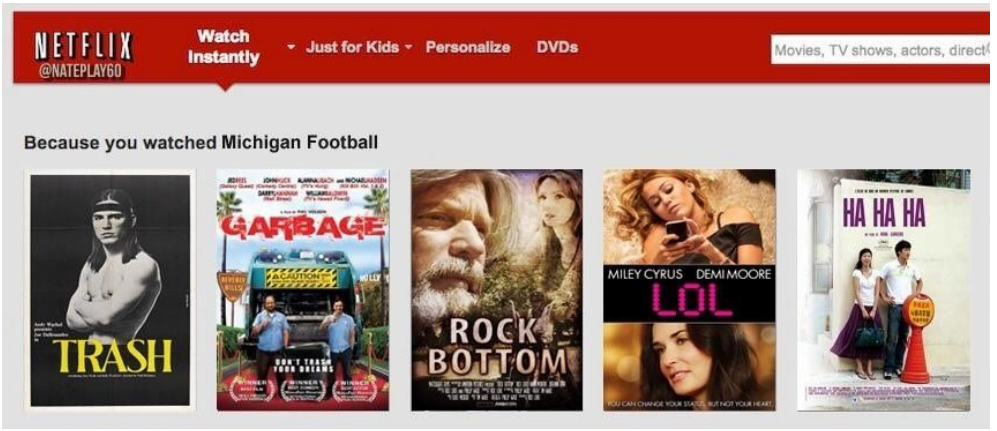

그림 2. Netflix의 추천 시스템을 통한 영화 추천

제 2 절 연구 목적

이러한 추천 시스템을 채용에 활용하고자 한다. 구직자가 선호하는 채용 정보를 제공하거나, 구직자에게 알맞은 채용 정보를 제공해주는

방식으로 구직자가 원하는 채용 정보를 찾기 위해 들이는 노력을 줄여주고, 취업 시장의 미스매치 문제를 완화하는데 도움이 되고자 한다.

현재 청년 실업이 지속적으로 심각해지고 있으며, 취업이 어려워지는 만큼 구직에 대한 수요는 높아졌다. 이러한 수요를 바탕으로 취업 정보를 제공하는 취업포털, 커뮤니티 등은 지속적으로 성장해왔다.

그러나 이러한 산업의 성장에도 불구하고 여전히 구직자는 자신에게 필요한 채용공고와 기업 정보를 찾는데 어려움을 겪고 있다. 이는 기존의 취업사업에서 제공하는 공고와 정보들이 구인/구직의 수요에 따라 제공되는 것이 아니라 광고비를 부담하는 광고주의 수요에 따라 제공되기 때문이다.

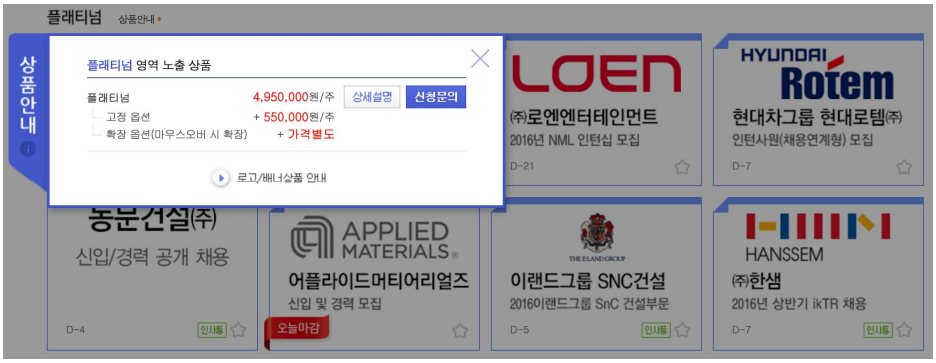

그림 3. 사람인의 광고 중심의 채용 공고

채용공고의 배너광고가 주 수입원인 취업 포털과 커뮤니티의 경우 광고 매출을 높이기 위해 사이트 내에서 유료 광고의 노출을 중심으로 사이트를 개선해왔다. 그러나 이는 구직자 입장에서는 화면에 광고가 더 많아지면서 구직자가 원하는 채용 정보는 더욱 찾기가 어려워졌다. 광고를 기반으로 한 채용 정보 제공이 아닌, 추천 시스템을 활용하여 구직자가 선호하는 채용 정보를 제공하여 구직자가 기업을 찾는데 겪는 어려움을 해소할 수 있다.

또한, 청년 실업이 심화되고 있지만, 아이러니하게도 중소기업에서는 심각한 인력난을 겪고 있다. 이처럼 구직자의 수요와 일자리 공급이 충분한대도 서로 연결이 되지 않아 취업이 되지 않고 있는 현상을 '일자리 미스매치'라고 한다. 인력난을 겪고 있는 중소 기업

중에서는 열악한 근무 환경과 강도 높은 노동으로 인해 구직자가 기피하는 기업도 있지만, 구직자가 선호할만한 좋은 기업임에도 불구하고 구직자에게 잘 알려지지 않아 좋은 인재를 구하지 못하고 있는 기업들도 다수 존재한다. 추천 시스템을 활용하여 구직자에게 이러한 좋은 기업을 추천해준다면 취업의 미스매치를 완화할 수 있을 것이다.

제 3 절 논문의 구성

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문이 다루고자 하는 추천 시스템에 관련된 선행 연구를 살펴본다. 제 3장에서는 본 논문에서 다루고자 하는 문제를 정의하고, 이러한 문제를 해결하기 위한 방법들을 제시한다. 4장에서는 본 논문을 통해서 구현하고자 하는 추천 채용 시스템의 구현 방법에 대해서 자세히 살펴본다.. 5장에서는 구현한 시스템을 사용한 실험 결과를 보이고, 6장에서는 결론 및 앞으로의 연구 방향을 제시한다.

제 2 장 관련 연구

제 1 절 추천 시스템

추천 시스템은 사용자의 과거 행동 패턴을 바탕으로 사용자의 선호도가 높은 서비스나 상품을 제안하는 기술이다. 이러한 제안들은 여러 분야의 결정을 내리는 과정에서 사용되는데, 상품을 구입하거나, 음악을 선택하거나, 읽을만한 뉴스를 선별하는데 사용될 수 있다. 특히 정보나 경험의 부족으로 인해 서비스나 상품을 쉽게 선택하기 어려운 사용자에게 다양한 대체재를 제공하여, 결정에 도움을 줄 수 있다. [4]

즉 사용자가 경험하지 못한 서비스나 상품에 대해서 사용자의 선호도를 예측하여, 높은 선호도를 가진 항목들을 사용자에게 제안하는 것이다. 이는 다음의 식으로 표현이 가능하다.

$$\forall c \in C, \quad s'_c = \underset{s \in S}{argmax}\, u(c, s). \qquad \text{(식 1)}$$

C는 모든 유저의 집합으로 정의하고, S는 추천 가능한 모든 가능한 항목의 집합으로 정의한다. C는 모든 유저의 집합이므로 그 크기는 수백만 또는 그 이상에 달할 수 있고, S 역시도 추천 가능한 모든 항목이므로 수백만 또는 그 이상에 달할 수 있다. u는 항목 s가 c에 얼마나 유용한지를 나타내는 유용성 함수(Utility Function)으로 정의한다. 양의 정수 혹은 특정 범위 안의 실수 등에 해당하는 전순서집합(Totally Ordered Set) R에 대해 u: C × S → R의 관계를 가지게 된다[5]. 유용성 함수는 선호도에 대한 평점(Rating)으로 나타내는 것이 일반적이지만, 경우에 따라서는 평점 없이 선호에 대한 유무에 따라 불(Boolean) 방식으로 나타내기도 한다.

이러한 추천 시스템은 다음의 세가지 방식으로 구분이 가능하다.

- 협업적 필터링(Collaborative Filtering)
- 내용 기반 추천(Content-based Recommendation)
- 하이브리드 추천 시스템(Hybrid Recommender Systems)

1.1. 협업적 필터링

사용자와 유사한 취향을 가진 사용자들이 선호하는 항목을 추천해주는 방식이다. 이는 동일한 항목에 유사한 평가를 내린 사용자들은 새로운 항목에 대해서도 유사한 평가를 내릴 것이라는 전제를 두고 있다. 즉, 추천을 받을 사용자가 항목들에 대해 평가한 선호도와 다른 사용자들이 항목들에 대해 평가한 선호도가 유사하다면, 다른 사용자가 선호한 항목들도 추천을 받을 사용자는 선호할 것으로 가정하고 다른 사용자가 선호한 항목들을 추천하는 것이다.

수식적으로 표현하자면, 식 1에서 항목 s의 유용성 함수 u(c,s)는 사용자 c와 유사한 사용자인 c_j의 유용성 함수 u(c_j, s)에 따라서 항목 s의 선호도가 결정된다.[5] 사용자 c와 사용자 c_j 간의 유사도는 사용자 c와 사용자 c_j가 내린 과거의 평가들의 유사도에 따라 결정이 된다.

협업적 필터링은 그 방식에 따라서 메모리 기반 협업적 필터링(Memory-based Collaborative Filtering)과 모델 기반 협업적 필터링(Model-based Collaborative Filtering)로 나눌 수 있다.

메모리 기반 협업적 필터링.

메모리 기반 협업적 필터링은 이웃 기반 협업적 필터링(Neighborhood-based Collaborative Filtering)으로 불리기도 하는데, 그 이유는 사용자에게 추천되는 항목은 이웃의 선호도를 토대로 결정되기 때문이다. 이웃은 두 사용자가 여러 항목에 내린 평가가 유사할 경우, 사용자 간에 유사도가 크다고 판단한다. 사용자 간 유사도는 코사인 유사도(Cosine Similiarity)나 피어슨 상관 계수(Pearson Correlation), 타니모토 상관 계수(Tanimoto

Coefficient) 등을 사용하며, 계산된 유사도를 바탕으로 K-최근접 이웃(K-Nearest Neighbors), K-평균 (K-means), K-d Tree, Locality Sensitive Hashing 등의 알고리즘을 통해 이웃을 결정하게 된다.[6] 사용자가 평가하지 않은 항목에 대해서 이웃들의 평가와 유사할 것으로 가정하여, 선호도를 예측한다.

그러나 메모리 기반의 협업적 필터링은 몇 가지 한계를 가지고 있다. 첫 번째로 희박성(Sparsity) 문제이다. 유사한 성향을 보이는 이웃을 찾고, 항목에 대한 선호도를 예측하기 위해서는 데이터가 충분히 모여야 정확한 추천이 가능하다. 이를 해결 하기 위해 내용 기반 필터링을 결합한 하이브리드 기법으로 희박성을 완화하는 방식이 제안되었다.[1]

두 번째는 확장성(Scalability) 문제이다. 메모리 기반 협업적 필터링은 잠재적인 이웃들을 찾기 위해서 수만 번의 검색을 거쳐야 하는데, 데이터가 많아질 경우 검색 횟수는 수백만 번에 달하기도 한다. 이러한 검색을 실시간으로 처리하는 것은 한계가 있으므로 실시간 처리 시스템에서는 이러한 검색을 하는 것은 불가능해진다. 이를 해결하기 위해 사용자 간의 유사도를 구하는 것이 아닌, 항목 간의 유사도를 구하는 방법과[7], 사용자를 군집한 후 군집간의 유사도를 계산하는 방법 등이 제안되었다[8].

모델 기반 협업적 필터링

두 번째 방법은 모델 기반 협업적 필터링이다. 모델 기반 협업적 필터링은 메모리 기반의 협업적 필터링의 한계를 극복하기 위해 고안이 되었다. 희박성 문제를 해결하기 위해 높은 차원의 행렬을 낮은 차원의 행렬로 분해하는 차원 압축(Dimensionality Reduction) 알고리즘을 활용한다. 대표적으로 특이값분해(Singular Value Decomposition;SVD)를 활용해서 높은 차원의 사용자-항목의 행렬을 낮은 차원의 사용자-특성 벡터 행렬과 항목-특성 벡터 행렬로 분해하는 방식이 있다[9].

1.2. 내용 기반 필터링

컨텐츠 기반의 추천은 사용자가 과거에 선호했던 항목과 가장 유사한 항목을 사용자에게 추천한다. 식 1에서 항목 s의 유용성 함수 u(c, s)는 유저 c가 과거에 선호한 si에 대한 유용성 함수 u(c, si)와의 유사도에 따라 추천이 되는 방식이다.[1] 항목 s와 si 간의 유사도는 각 항목의 특성을 분석하여 추천을 하도록 한다.

항목 간의 유사도를 항목의 각 항목의 특성을 분석하여 구하기 때문에 각 도메인에 따라 항목의 특성을 프로파일링 해야하는 단점이 있다. 예를 들어 유사한 뉴스를 추천한다면 뉴스 분야, 주제, 사용된 단어 등의 특성을 프로파일링 해야 하고, 영화를 추천한다면, 영화의 장르, 감독, 배우 등의 특성을 프로파일링 해야 한다. 그러므로 사전에 도메인에 대한 지식이 필요하다.

문제는 이러한 도메인 지식이 다른 도메인에서는 사용될 수 없기 때문에 협업적 필터링과 달리 도메인이 다른 경우 추천에 사용할 수가 없다. 그러나 협업적 필터링과 달리 다른 사용자가 항목에 대한 평가를 하지 않았더라도 항목의 특성을 통해 유사도를 계산해 바로 추천에 사용될 수 있는 장점이 있다.

1.3. 하이브리드 추천 시스템

하이브리드 추천 시스템은 내용 기반 필터링과 협업적 필터링 방식의 한계를 극복하기 위해서 두 가지 이상의 방식을 혼합적으로 사용하여 추천을 하는 방식이다. 이러한 하이브리드 추천 시스템은 다음의 세 가지 방식으로 분류가 가능하다. [5]

- 협업적 필터링 모델과 내용 기반 필터링 모델을 각각 구현하여 혼합하는 방식
- 협업적 필터링 모델에 내용 기반 필터링의 특성을 추가하는 방식
- 내용 기반 필터링 모델에 협업적 필터링의 특성을 추가하는 방식

협업적 필터링 모델과 내용 기반 필터링 모델을 각각 구현하여 혼합하는 방식

협업적 필터링 모델과 내용 기반 필터링 모델을 각각 구현하여 혼합하는 방식으로는 크게 두 가지로 나눌 수 있다. 첫 번째는 각각 구현된 모델의 결과 값을 선형 결합(Linear Combination) 하거나, 가중합(Weighted Sum)하여 최종 값을 도출하는 방식이다. 각각 모델에 대한 가중치를 설정하는 것이 중요한 요소가 된다. 두 번째는 각각의 모델 중 더 나은 결과 값을 선택하는 방식이다. 역치(Threshold)를 설정하는 등의 조건에 따른 모델 선택 정책을 설정하여 결정을 하게 된다.

협업적 필터링 모델에 내용 기반 필터링의 특성을 추가하는 방식

많은 하이브리드 추천 시스템이 사용하는 방식으로 협업적 필터링 모델에 사용자의 특성 데이터를 추가하여 재평가(rescore)하거나, 항목의 특성에 따라서 재평가 하는 방식이다. 이러한 방식의 장점은 처음 추천 시스템을 사용하는 유저의 경우 내용 기반 필터링을 통해 얻어진 유사 집단의 평균 값으로 선호도를 예측하여 Cold Start Problem을 완화할 수 있으며, 희박성이 높은 협업적 필터링의 경우 내용 기반 필터링을 통해 얻어진 항목 간의 유사도를 통해 정확도 높은 추천을 할 수 있다[11].

내용 기반 필터링 모델에 협업적 필터링의 특성을 추가하는 방식

내용 기반 필터링 모델에 협업적 필터링의 특성을 추가하는 방식 중 가장 많이 사용되는 방식은 내용 기반 필터링 모델을 통해 얻어진 사용자 그룹에 차원 압축을 하여 사용자의 프로필을 벡터 행렬로 표현하는 방식이다.

제 3 장 문제 정의

제 1 절 문제 정의

추천 채용 시스템을 구현하기 위해 다음의 두 가지 독립적인 문제로 구성이 된다.

- 양질의 데이터 수집
- 추천 시스템 구성

1.1 양질의 데이터 수집

국내 대형 취업 포털에서 추천 채용 시스템을 개발하고자 하였으나, 현재까지 눈에 띄는 성과를 보이지 못하고 있다. 그 이유는 취업 포털에서 얻을 수 있는 데이터의 한계가 있기 때문이다.

기존의 취업 포털에서 추천에 사용하고 있는 데이터는 채용 공고 클릭 로그, 채용 공고 즐겨찾기 로그, 검색어 등이다. 그 중 채용 공고 클릭 로그가 가장 많은 데이터를 차지하는데, 취업 포털의 특성상 과도한 광고로 인한 클릭이 많아 인기 편향(Popularity Bias) 문제가 발생한다. 뿐만 아니라, 채용 공고 클릭 로그와 채용 공고 즐겨찾기의 경우 기업에 대한 선호도는 예측할 수 있으나, 직무에 대한 선호도는 예측을 할 수가 없다. 사용자가 기업을 선호하더라도 자신이 지원할 수 있는 직무가 없으면 지원을 할 수 없기 때문에 추천 신뢰도는 급감하게 된다. 이러한 이유로 인해 취업 포털에서 수집한 기존의 데이터로는 추천 채용 알고리즘을 개발하는 것에 어려움이 있다.

이에 반해 본 논문에서 활용하고자 하는 데이터는 다음의 두 가지 이다.

- 광고로 인한 인기 편향 문제가 없는 채용 공고 클릭 로그
- 회사와 직무를 포함한 자기소개서 작성 이력

광고로 인한 인기 편향 문제가 없는 채용 공고 클릭 로그

첫 번째 데이터는 광고로 인한 인기 편향 문제가 없는 채용 공고 클릭 로그이다. 사용자 ID와 채용 공고 실제 사용자가 원하는 기업에 대한 클릭만을 수집하였으므로 더욱 정확도 높은 기업 선호도를 예측할 수 있다.

회사와 직무를 포함한 자기소개서 작성 이력

두 번째 데이터는 회사와 직무를 포함한 자기소개서 작성 이력이다. 자기소개서를 작성한 경우 대부분 실제로 기업에 지원을 하기 때문에 사용자가 실제 기업에 지원한 이력과 상관 관계가 높다.

1.2 추천 시스템 구성

추천 시스템은 내용 기반 필터링, 협업적 필터링, 하이브리드 추천 시스템 방식이 있다. 본 연구의 데이터는 사용자의 특성이나 기업의 특성에 대한 정보가 없으므로 내용 기반 필터링은 불가능하다. 따라서 협업적 필터링 방식을 사용하도록 한다.

각기 다른 두 데이터에 대해서 각각 협업적 필터링 모델을 구현한 뒤, 식 2와 같이 각 모델의 선호도의 가중합(Weighted Sum) 연산을 통해 하이브리드 추천 모델을 도출하도록 한다.

$$u(c,s) = \left[\alpha\,\mathrm{Norm}\left(u_a(c,s)\right) + \beta\,\mathrm{Norm}\left(u_b(c,s)\right)\right] * \left[\max u_b(c,s) - \min u_b(c,s)\right] \quad (\alpha + \beta - 1) \quad \text{(식 2)}$$

이 때, u_a는 기업 클릭 데이터를 바탕으로 한 사용자의 기업 선호도이며, u_b는 자기소개서 작성 이력을 바탕으로 한 사용자의 기업/직무 선호도이다. 이에 따라서 선호하는 기업의 경우 $u_a(c,s)$가 높아지고, 사용자가 선호하지 않는 기업의 경우 $u_b(c,s)$가 낮아지게 된다. α와 β 값은 매개 변수 최적화 과정을 통해 결정이 된다.

제 4 장 추천 알고리즘의 설계와 구현

제 1 절 전체 시스템 구성

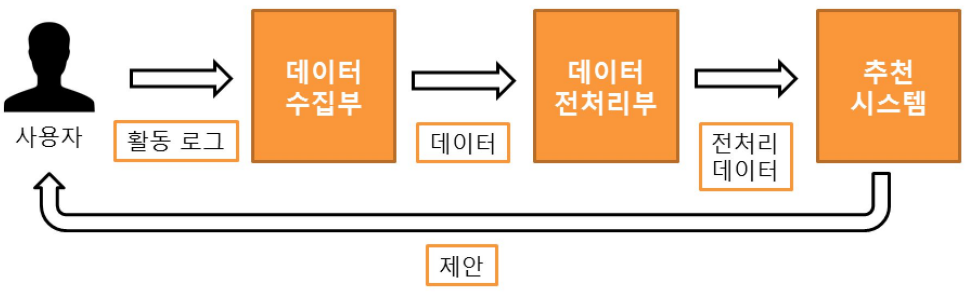

그림 4. 전체 시스템 구성

전체 시스템은 채용 공고 클릭 로그와 기업과 직무를 포함한 자기소개서 작성 이력 데이터를 수집하는 데이터 수집부, 수집된 데이터를 추천 시스템에 알맞게 처리/분류하는 데이터 전처리부, 처리된 데이터를 통해 실제로 기업과 직무의 선호도를 예측하는 추천 채용 시스템으로 이루어져 있다.

제 2 절 데이터 수집부

데이터 수집부에서는 추천 시스템에서 사용할 데이터를 수집하는 역할을 한다. 취업 사이트의 로그를 통해 데이터를 수집하도록 한다.

본 취업 사이트는 취업 자기소개서와 이력서를 온라인에서 작성할 수 있는 사이트이며, 클라우드 관리, 맞춤법 검사, 동일 기업 지원자 간 채팅 등 자기소개서와 이력서 작성을 위한 다양한 편의 기능을 지원한다. 또한, 채용 공고를 확인할 수 있는 채용 달력 기능, 사용자 별로 자신의 채용 일정을 관리할 수 있는 채용 달력 등의 채용 정보 관련 기능을 제공한다. 가입자는 9만 명, 한 달 방문자는 약 60만 명이다. 구직자가 활발하게 사용하고 있는 사이트이므로, 추천 채용 시스템 개발을 위한

데이터를 다량 수집할 수 있다.

본 취업 사이트에서 수집할 데이터는 채용 달력에서 기업을 클릭한 로그 데이터와 기업의 직무에 자기소개서를 작성한 이력 데이터이다.

채용 공고 클릭 로그 데이터

채용 달력에서 기업을 클릭한 데이터이다. 기업의 이름을 클릭하면 채용 공고가 열리고 데이터가 축적이 되게 된다. 해당 기업에 대해서 클릭만으로 데이터가 기록이 되므로 데이터의 수가 많으나, 기업의 선호만을 알 수 있을 뿐, 직무에 대한 정보는 알 수 없다. 데이터의 *{사용자 ID, 채용 공고 ID}* 형식으로 수집이 된다..

2016년 2월 16일부터 2016년 5월 29일까지 수집된 데이터이며, 사용자 수는 89,879명, 채용 공고의 수는 10,254개이다. 채용 공고에 대한 4,735,344개의 클릭 데이터가 수집되었다.

기업의 직무 자기소개서 작성 이력 데이터

구직자가 자기소개서를 작성한 기업의 내역이다. 대부분이 실제로 기업에 지원을 한 내역이며, 기업뿐만 아니라 직무에 대한 선택도 가지고 있다. 데이터는 *{사용자 ID, 직무 ID}* 형식으로 수집된다.

2014년 2월 15일부터 2016년 5월 29일까지 수집된 데이터이며, 마찬가지로 사용자 수는 89,879명, 채용 공고의 수는 10,254개이다. 채용 공고 별로 지원 직무가 있으며, 전체 직무 수는 44,343개이다. 1,314,727개의 데이터 중 기업과 직무 정보가 있는 910,769개의 자기소개서 작성 이력을 수집하였다.

제 3 절 데이터 전처리부

데이터 전처리부에서는 데이터 수집부에서 수집한 데이터를 추천 알고리즘에서 활용할 수 있도록 처리/분류하는 역할을 한다. 다음과

같은 전처리 과정을 거친다.

- 동일한 기업의 채용 공고를 통합하는 과정
- 유사한 직무를 카테고리화 하는 과정
- 직무를 기업과 직무 카테고리로 치환하는 과정

동일한 기업의 채용 공고를 통합하는 과정

한 기업에서 여러 번 채용 공고를 올린 경우가 있다. 예를 들면 A 기업에서 2014년 9월, 2015년 3월, 2016년 3월에 채용을 진행한 경우에는 기업은 하나이지만 채용 공고는 3개가 된다. 수집된 채용 공고 클릭 로그 데이터는 기업 ID를 가지고 있는 것이 아니라, 각 채용 공고마다의 ID를 가지고 있으므로, 각기 다른 기업으로 인식해서 추천이 동작하지 않는다. 동일한 기업의 채용 공고를 하나로 합치고 채용 공고의 ID가 아닌 기업 ID로 치환하도록 한다. 그 결과 10,254개의 채용 공고를 3,899개의 기업으로 합쳤다.

{사용자 ID, 채용 공고 ID} → {사용자 ID, 기업 ID}

유사한 직무를 카테고리화 하는 과정

유사한 직무라고 하더라도 각각의 세부 직무는 다르다. 예를 들면 온라인 마케팅의 경우 온라인 마케팅도 있지만, 커뮤니티 마케팅, 온라인 컨텐츠 마케팅 등 다양하다. 이러한 직무를 모두 따로 처리할 경우 유사한 직무가 각기 다른 직무로 판단이 되므로 추천이 동작하지 않는다. 그러므로 직무를 16개의 카테고리로 분류하였다.

직무 카테고리	세부 직무
경영·사무	사업기획, HR(인사), 전략기획팀, Service Coordinator 등
재무·회계·경리	회계, ERP 전표 입력, 공인회계사, 재무(관세), 재무팀 등
마케팅	마케터, 온라인마케팅, 브랜드관리팀, 서비스운영/마케팅 등
금융	자운용직군, 손해사정, 채권관리, 신입행원, 보험업무 등
영업 ·고객상담	영업관리, 기술 영업, 고객지원팀, 공사영업팀, 건재영업팀 등
유통·무역	자재운영 및 관리, 구매, 수출MD, 온라인 여성MD 등

IT·인터넷	시스템운영/개발, 전산프로그래머, 전산, 정보보안, IT기획 등
디자인	그래픽디자인, 패키지 디자인, 영상 컨텐츠 그래픽 디자이너 등
연구·개발	연구운영팀장, 연구소 운영관리, 연구개발, 이화확 분석 등
생산·제조	제조팀, 생산관리, 품질관리, 생산기술, 전자 설계 등
미디어	언론홍보 및 취재, PD, 방송기자, 광고, 미디어기획/AE 등
서비스	서비스지원, 오퍼레이션, 주방매니저, 홀매니저, 캐빈승무원 등
전문직	변호사, 기록물관리, 통번역, 상품계리, 리스크 관리 등
건설	건축, 설비, 조경, CAD 설계, A/F설계팀, 시공, 안전관리 등
교육	영어강사, 중국어강사, 교육팀, 교육사업 총괄, 제품 강사 등
의료	신약합성, 바이오신약개발, 약효평가, 합성공정, 제제연구 등

표 1. 세무직무 카테고리 표

직무를 기업과 직무 카테고리로 치환하는 과정

채용 공고를 기업으로 통합하고, 각각의 직무를 카테고리화 하였기 때문에 기존의 직무 ID를 기업 ID와 직무 카테고리 ID로 치환을 하여 기업의 직무 관계를 표현하도록 한다. 앞의 4자리를 기업 ID로, 뒤의 2자리를 기업 카테고리 ID로 규정하여, 직무 ID를 표현하도록 한다.

{사용자 ID, 직무 ID}→{사용자 ID, (기업 ID + 기업 카테고리 ID)}

제 4 절 추천 채용 시스템

데이터 전처리부에서 생성한 데이터를 바탕으로 하여 사용자 별 기업, 직무별 선호도를 예측한다. 다음의 세 가지 추천 모델을 구현하도록 한다.

- 기업의 선호도를 예측하는 추천 시스템
- 기업과 직무 선호도를 예측하는 추천 시스템
- 두 추천 시스템을 하이브리드 하여 선호도를 예측하는 추천 시스템

기업의 선호도를 예측하는 추천 시스템

채용 공고 클릭 로그 데이터를 사용하여 기업의 선호도를 예측하는 시스템이다. *{사용자 ID, 기업 ID}* 형태의 4,735,344개의 채용 공고

클릭 로그 데이터가 입력으로 주어진다. 입력 받은 데이터를 메모리 기반 협업적 필터링을 해서 각 사용자 별로 선호도 값이 높은 10개의 기업과 예측 값이 출력되도록 한다. 사용자 별로 {기업 ID, 선호도 값} 형식으로 출력 된다.

사용자 간 유사도는 타니모토 상관 계수를 사용한다. 타니모토 상관 계수는 평점을 무시하고 사용자의 선호 표현 여부만을 가지고 사용자 간 유사도를 계산한다. 채용 공고 클릭 로그 데이터는 평점이 없이 단순히 선호 표현 여부만을 알 수 있으므로, 타니모토 상관 계수가 적절하다.

K-Nearest Neighbors(KNN) 알고리즘을 사용하여 상위 20개의 기업을 추출하도록 학습을 한다.

학습이 끝난 후 실제 적용 시에는 시작 전이거나 이미 마감을 한 채용 공고는 필터링하여 현재 채용 중인 공고만 추천되도록 한다. 기업의 선호도와 직무에 선호도에 따라 사용자에게 기업과 직무를 추천하더라도 이전에 마감이 되었거나, 아직 채용이 시작되지 않은 공고는 지원을 할 수가 없으므로 제외하도록 한다.

기업과 직무 선호도를 예측하는 추천 시스템

자기소개서 작성 이력 데이터를 사용하여 지원하는 기업의 직무 선호도를 예측하는 추천 시스템이다. *{사용자 ID, (기업 ID + 직무 카테고리 ID)}* 형태의 910,769개의 자기소개서 작성 이력 데이터가 입력으로 입력 받은 데이터를 메모리 기반 협업적 필터링을 해서 각 사용자 별로 선호도 값이 높은 20개의 기업과 예측 값이 출력되도록 한다. 사용자 별로 {*(기업 ID + 직무 카테고리 ID)*, 선호도 값} 형식으로 출력 된다.

기업의 선호도를 예측하는 추천 시스템과 기업과 직무의 선호도를 예측하는 추천 시스템은 입력과 출력 값은 다르지만 메모리 기반 협력적 필터링의 구조는 동일하다. 4.1. 추천 시스템과 마찬가지로 타니모토 상관 계수로 유사도를 구하고, K-Nearest Neighbors(KNN) 알고리즘을 사용하여 상위 20개의 기업을 추출하도록 학습을 한다.

마찬가지로 실제 적용 시에는 현재 채용 중인 공고만 추천이 되도록 한다.

두 추천 시스템을 하이브리드하여 선호도를 예측하는 추천 시스템

두 추천 시스템을 하이브리드 하여 선호도를 예측하는 추천시스템이다. 식 2와 같이 기업 추천 시스템과 기업/직무 추천 시스템에서 얻어진 예측 값을 0에서 1 사이의 값으로 정규화를 진행한 뒤, 가중합을 통하여 값을 구하고 다시 기업/직무 추천 시스템의 최대값과 최소값의 차만큼 곱한다. 이를 통해 기업/직무 예측 값의 과도한 수정은 줄이고, 두 추천 시스템에서 동시에 추천되는 기업이 있을 경우 그 예측 값을 강화하는 역할을 하게 된다.

$$u(c,s) = \left[\alpha\,\mathrm{Norm}\left(u_a(c,s)\right) + \beta\,\mathrm{Norm}\left(u_b(c,s)\right)\right] * \left[\max u_b(c,s) - \min u_b(c,s)\right] \quad (\alpha + \beta = 1) \quad \text{(식 2)}$$

제 5 장 실험 결과

학습한 모델은 2016년 5월 29일 기준 입력된 데이터를 기준으로 진행하였다.

항목	개수
사용자 수	89,879
채용 공고 수	10,254
직무 수	44,343
채용 공고 클릭 로그 수	4,735,344
자기소개서 작성 이력 수	910,769

표 2. 추천 채용 시스템에 사용된 데이터 종류와 수

기업 추천 시스템과 기업, 직무 추천시스템 결과

4,735,344개의 기업 클릭 데이터 중 90%는 학습 데이터로 사용하였으며, 10%는 테스트 데이터로 사용하였다. 마찬가지로 910,769개의 자기소개서 작성 이력 데이터 중 90%는 학습 데이터로 사용하였으며, 10%는 테스트 데이터로 사용하였다.

평가 지표는 평균 제곱근의 편차(Root Mean Square Error; RMSE)를 사용하였다. 평균 제곱근의 편차는 예측한 선호도 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용되는 지표이다.. 평균 제곱근의 편차가 낮을수록 실제 환경과 유사하다는 것으로 좋은 품질의 추천 시스템이다.

(n=10)

추천 시스템	데이터	RMSE
기업 추천 시스템	채용 공고 클릭 로고	1.5215
기업/직무 추천 시스템	자기소개서 작성 이력	0.6359

표 3. 각각 추천 시스템의 RMSE 결과

하이브리드 추천 시스템 결과

두 개의 추천 시스템에서 도출된 선호도를 통하여 최종 선호도를 도출하였고, 전처리가 된 자기소개서 작성 이력 데이터 중 10%를 테스트 데이터로 사용하였다. 마찬가지로 평가 지표는 평균 제곱근의 편차(Root Mean Square Error; RMSE)을 사용하였다.

(n=10)

추천 시스템	데이터	RMSE
하이브리드 추천 시스템	기업 추천 시스템, 기업/직무 추천 시스템의 예측 값	0.6254

표 4. 하이브리드 추천 시스템의 RMSE 결과

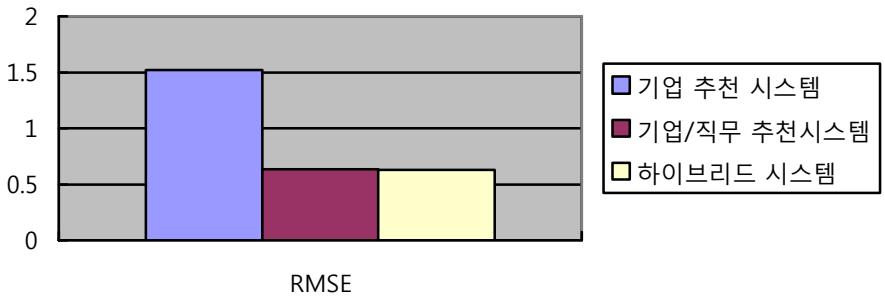

그림 5. 추천 시스템별 RMSE 결과

제 6 장 결론 및 향후 연구

제 1 절 결론

본 논문에서 메모리 기반의 협업적 필터링을 활용하여 채용 공고 클릭 로그 데이터를 입력으로 한 기업 추천 시스템과 자기소개서 작성 이력 데이터를 입력으로 한 기업/직무 추천 시스템을 구현하였고, 두 추천 시스템의 예측 값을 결합하여 더 나은 하이브리드 추천 시스템을 구현하였다.

하이브리드 추천 시스템은 기업/직무 추천 시스템의 성능에 비해 약간의 성능 개선이 있었다. 그 이유는 두 가지로 예상할 수 있다. 첫 번째 이유는 기업 추천 시스템의 성능이 기업/직무 추천 시스템에 비해 좋지 않아서 두 추천 시스템을 하이브리드 하더라도 기업/직무 추천 시스템의 성능이 크게 향상되지 않았기 때문이다. 두 번째 이유는 기업 추천 시스템에서 추천된 기업과 기업/직무 추천 시스템에서 추천된 서비스 사이에 공통된 기업이 적어서 값에 대한 영향이 크지 않았기 때문이다.

제 2 절 향후 연구

본 논문에서 제시한 추천 시스템은 협업적 필터링으로 도출된 결과 값의 하이브리드로 좋은 결과를 내고자 하였다. 추천 시스템에서의 일반적인 하이브리드는 협업적 필터링과 내용 기반 필터링을 하이브리드 하는 경우가 많은데, 그렇게 하지 못한 이유는 사용자의 특성 데이터와 기업의 특성 데이터가 없었기 때문이다. 그러므로 향후 사용자의 학교, 전공, 경력, 나이 등의 사용자 특성 데이터와 기업의 규모, 주력 분야, 관련 뉴스 등 기업의 특성 데이터를 수집하여, 협업적 필터링과 내용 기반 필터링을 하이브리드한 추천 시스템을 구축하고자 한다. 이러한 하이브리드 추천 시스템은 추천 시스템의 성능을 개선할 수 있으며,

사용자의 특성 데이터와 기업의 특성 데이터를 우선적으로 수집하여 Cold Start Problem도 완화시킬 수 있을 것이다[14].

또한, 본 연구에서 도출된 추천 채용 시스템을 실제 서비스에 접목하여 실제 사용자의 반응을 확인하고, 피드백을 통해 지속적으로 모델을 개선하여 정확도를 향상시키고자 한다.

참고 문헌

[1] Ansari, A., S. Essegaier and R. Kohli, “Internet Recommender Systems”, Journal of Marketing Research, Vol.37, No.3(2000): 363-375.

[2] Lamere, Paul, and S. Green. "Project aura: recommendation for the rest of us." Presentation at Sun JavaOne Conference(2008).

[3] Gomez-Uribe, Carlos A., and Neil Hunt. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." ACM Transactions on Management Information Systems (TMIS) 6.4 (2015): 13.

[4] Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Introduction to recommender systems handbook", Springer US(2011).

[5] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17.6 (2005): 734-749..

[6]Dommeti, Ramesh. "Neighborhood based methods for Collaborative Filtering." A Case Study, I (2009): 1-5.

[7] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.

[8] Li, Peng, and Seiji Yamada. "A movie recommender system based on inductive learning." Cybernetics and Intelligent Systems, 2004 IEEE Conference on. Vol. 1. IEEE(2004).

[9] Sarwar, Badrul, et al. "Application of dimensionality reduction in recommender system-a case study". No. TR-00-043. Minnesota

Univ Minneapolis Dept of Computer Science(2000).

[10] Lops, Pasquale, Marco De Gemmis, and Giovanni Semeraro. "Content-based recommender systems: State of the art and trends." Recommender systems handbook. Springer US, (2011): 73-105.

[11] Pazzani, Michael J. "A framework for collaborative, content-based and demographic filtering." Artificial Intelligence Review 13.5-6 (1999): 393-408.

[12] Deshpande, Mukund, and George Karypis. "Item-based top-n recommendation algorithms." ACM Transactions on Information Systems (TOIS) 22.1 (2004): 143-177.

[13] Wang, Jun, Arjen P. De Vries, and Marcel JT Reinders. "Unifying user-based and item-based collaborative filtering approaches by similarity fusion." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM(2006).

[14] Good, Nathaniel, et al. "Combining collaborative filtering with personal agents for better recommendations." AAAI/IAAI. (1999).

Abstract

Design and Implementation of Employment Recommender System using Collaborative Filtering

Susang Park

Department of Computer Science and Engineering

College of Engineering

The Graduate School

Seoul National University

Recommender system is one of big data processing technologies. Recommender system can suggest personalized items to users in many different domains, e.g. books, items, movies.

In this paper, recommender system introduce to employment domain. Many job applicants are difficult to find appropriate job positions. Using recommender system can suggest job positions to job applicants.

Two memory-based collaborative filtering recommender systems are developed. Structure of two models is similar, but only input data are different. Also, hybrid recommender system is developed. The input data of this hybrid system is weighted sum of two collaborative filtering recommender systems

Keywords : Recommender System, Collaborative Filtering, Employment Recommender System

Student Number : 2014-21769