

**추천시스템의 편향 보정 및
공정성 보장 방안 연구 :
측정방법론을 중심으로**

2020. 11. 13.

주관연구기관 건국대학(교)

추천시스템의 편향 보정 및 공정성 보장 방안 연구

2020. 11. 13.

주관연구기관 건국대학(교) 산학협력단
(수행: 디지털커뮤니케이션연구센터 DCRC)



정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

제 출 문

정보통신정책연구원장 귀하

본 연구결과보고서를 “추천시스템의 편향 보정 및 공정성 보장 방안 연구”에 관한 연구의 최종연구보고서로 제출합니다.

2020년 11월 13일

주관연구기관 : 건국대학교 산학협력단 (인)

연구 책임자 : 황 용 석

공동 연구원 : 문 수 복

정 재 선

연구 보조원 : 정 지 완

남 은 수

황 현 정

보 조 원 : 김 승 완

김 형 준

위 희 주

조 경 재

요 약 문

데이터 기반의 자동화된 의사결정을 바탕으로 하는 추천 알고리즘은 현대사회의 사회적 에이전트로서 공식·비공식 영역에서 그 영향력이 커지고 있다. 추천 알고리즘에 대한 기대는 인간의 편향적 판단이나 조작이 배제된 결과를 출력할 수 있을 것이라는 믿음에서 비롯되었다고 할 수 있다. 그러나 머신러닝을 기반으로 하는 추천 알고리즘의 학습과정에서 사회적으로 문제가 발생할 수 있는 행동을 학습을 할 경우 알고리즘은 인간의 기대와 다른 판단을 할 수 있다. 또한 알고리즘은 학습된 데이터를 바탕으로 한다는 점에서 사회적 시스템이 변화할 경우 변화한 시스템을 반영하지 못할 수 있다. 인간의 개입이 최소화된다는 점에서 알고리즘의 추천과정에 대한 검토가 불투명할 수 있다. 이러한 알고리즘의 내재적 문제점에 대한 관심은 공학과 비공학적 접근에서 상이하게 다루어진다. 알고리즘의 내재적 문제점에 대한 이슈의 예를 보면, 정확한 예측의 절차적 방법 도출을 목표로 하는 시스템 중심의 접근, 추천된 아이템이 차별 없이 대우받는지와 관련한 공급자 이슈, 추천시스템이 사용자의 요구를 충족하는지와 관련한 이용자 중심 접근 등이 있을 수 있다. 해당 이슈에 대하여 공학적 관점은 추천한 상품이나 서비스에 대한 사용자의 만족뿐만 아니라 추천 알고리즘에서 나타나는 편향과 불공정의 문제를 제거하는 알고리즘 발견에 관심을 가진다. 비공학적 접근은 공학적 접근이 알고리즘에 내포된 문제를 해결할 수 있게 하는 방향 제시와 정책 마련에 목적이 있다. 그렇다면 공학적 접근은 비공학적 접근에서 제시된 문제점을 바탕으로 효과적인 알고리즘을 개발할 수 있고, 비공학적 접근은 공학적 접근의 한계를 바탕으로 문제점을 제기해야 한다. 이는 추천 알고리즘이 기술적/사회적 측면에서 독립적으로 접근되어야 하는 대상이 아니라 다학제적 접근이 요구된다는 것을 의미한다.

이미 우리나라에 비해 앞서 추천 알고리즘을 도입한 외국의 경우 추천 알고리즘에서 나타나는 학습과정의 문제와 부작용 고찰, 비공학적 관점의 문제점을 반영하는 알고리즘의 보정방법, 알고리즘이 반영해야 할 공정성의 법적·개념적 논의, 알고리즘에 적용되어야 할 조건의 마련 등이 이루어지고 있다. 우리나라에서는 개인정보 활용의 법적 제약으로 인해 개인이 생성하는 데이터를 기반으로 하는 추천 알고리즘이 활성화되지 못했지만 2020년 5월의 데이터 3법 개정으로 인해 추천 알고리즘을 바탕으로 하는 추천시스템 도입의 확대가 예측된다. 2020년이 우리나라 추천 알고리즘의 시작이라는 점, 추천 알고리즘의 문제점 파악과 보완방안 모색은 공학적 접근과 비공학적 접근이 상호보완적으로 적용되어야 한다는 점에서 본 연구는 추천시스템에 요구되는 편향의 보정과 공정성 보장을 개념적, 실증적으로 고찰하고 향후 활발해질 것으로 예상되는 추천시스

템 연구방법의 도출을 목적으로 하였다.

이러한 연구목적 달성을 위해 첫째, 추천시스템의 개념 및 추천시스템에 요구되는 쟁점 사항, 둘째, 공정성을 보장할 수 있는 알고리즘의 유형 관련 연구, 셋째, 머신러닝 기반 추천시스템의 추천 과정 및 편향성 보정 실증 연구를 통해 편향의 불가피성 및 보정의 방법을 실증, 넷째, 추천시스템의 공정성 확보를 위한 제도와 정책을 살펴본 후 다섯째, 제도와 정책 추진방법에 대한 전문가 조사를 바탕으로 추천시스템 편향 보정 및 공정성 보장을 위한 가이드라인을 제시하고자 하였다.

알고리즘 기반의 추천시스템은 어떠한 결과물을 인간이 인위적으로 구조화한 절차에 따라 출력을 하는 것이 아니라, 머신러닝의 과정을 통해 인공지능 혹은 시스템이 분류와 예측 절차를 규정하고 이에 따라 나타나는 추천이라는 결과를 제공하는 것을 의미한다. 추천이 머신러닝의 결과라는 점에서 추천시스템에 요구되는 쟁점은 다양한 사례가 포함되는지와 관련한 훈련 데이터의 문제, 인간이 처리했던 결과를 바탕으로 하는 데이터에는 이미 과거에 이루어졌던 의도적 불공정이 포함되어 있기 때문에 나타날 수 있는 비지각 공정성 문제, 알고리즘이 사회적 상황을 반영해야 한다는 그룹/개별 공정성 문제가 대두된다. 이들 문제점은 추천의 편향이라는 구체적인 결과로 나타날 수 있고, 편향의 제거는 추천시스템의 일차적 목표가 될 수 있다. 이에 해외의 연구는 편향의 사례뿐만 아니라 편향의 법적 고찰, 학습과정에서 편향이 심화되는 과정의 규명, 추천의 내용을 중심으로 각 내용에 편향의 발생배경과 보정 방법, 편향의 탐지방법에 관심을 가진다.

특히 데이터 공학에서는 공정성을 보장할 수 있는 알고리즘의 유형에 관심을 가진다. 알고리즘 기반 추천시스템은 대표적으로 협업필터링, 콘텐츠기반 필터링, 딥러닝 기반 추천시스템, 이용자 선택기반 추천시스템 방식이 있다. 협업필터링은 기존 사용자의 행동정보를 이용해 사용자와 비슷한 성향의 사용자가 선호하는 아이템을 추천하는 방식으로 비지각에서 오는 공정성은 만족할 수 있으나 콜드스타트 문제가 발생한다. 콘텐츠기반 필터링은 추천하는 대상의 내재적 특성을 사용한 프로파일과 이용자 특성을 분석한 프로파일을 추출해서 추천에 활용하기 때문에 콜드스타트 문제를 해결할 수 있지만 프로파일을 기반의 문제점 때문에 비지각에서 오는 공정성을 만족할 수 없다. 딥러닝 기반 추천시스템의 경우 구조화된 절차를 사용하는 것이 아니라 예측의 단계에서 사용된다. 이용자 선택기반은 이용자가 스스로 추천받을 환경을 설정하기 때문에 알고리즘과는 다소 거리가 있다. 따라서 알고리즘 기반 추천시스템의 접근은 주로 협업필터링과 콘텐츠기반 필터링을 사용하게 된다.

두 가지 필터링 중 콘텐츠 기반 필터링의 경우 사용할 수 있는 데이터 필드와 그를 가공하는 방식이 다양하여 표준적인 방식이 없다. 그에 비하여, 협업 필터링의 작동 방식은 대부분 행렬 분해에 기반하고 있어 알고리즘이 정형화된 편이며, 시뮬레이션 실험을 위한 합성 데이터를 구

축하기 용이하다. 이러한 점에서 본 연구는 협업필터링 기반의 추천시스템이 이용자 측면에서 갖는 편향성과 공정성에 초점을 두고 이용자 그룹 규모의 불균형으로 인해서 발생하게 되는 편향성을 보정하는 방법을 제안하기 위해 가상의 사용자와 그들의 평가 정보 데이터를 만들고 이를 이용해 선행연구에서 가장 효과적인 방법으로 언급되는 User-KNN과 SVD 알고리즘을 이용해 추천시스템을 구현하였다. 구현한 추천시스템은 가상으로 생성한 행동기록 데이터(synthetic data generation)를 이용하여 편향성의 차이를 확인했으며 해당 방법론을 MovieLens의 실제데이터에 적용했다. 실험결과, 알고리즘을 이용한 경우 기존의 사용자가 선호했던 정도보다 더 증폭된 정도로 항목을 추천해주는 것이 확인되어 알고리즘으로 인해 편향이 발생할 수 있음을 확인했다. 알고리즘 편향을 보정하기 위한 방안으로 해당 실험에서는 입력데이터를 전처리하여 편향성 차이를 줄일 수 있는 네 가지 방법을 고안했다. 첫 번째는 소수 집단의 크기를 다수 집단과 동일하도록 임의의 유저를 여러 명 선택하고 복제해 늘리는 방법, 두 번째는 소수 집단의 크기가 다수 집단의 크기의 0.25배 보다 작을 때, 소수 집단의 크기가 다수 집단의 크기의 0.25배가 되도록 조정하는 방법, 세 번째는 소수집단에서 임의로 두 유저를 고른 후에 두 유저의 선호하는 항목을 더해서 새로운 유저 값을 만들어 소수 집단에 추가하는 방법으로, 소수 집단과 다수 집단의 크기가 같을 때까지 과정을 반복하여 추가하는 방식, 네 번째는 소수 집단과 다수 집단의 크기를 동일하게 맞추기 위해, 두 명의 사용자를 임의로 뽑고 그 값에서 임의로 항목들을 선택하는 방법으로 두 명의 사용자를 임의로 뽑은 뒤, 두 사용자가 선택한 항목들에 대해서 임의로 10개의 항목들을 뽑고 이를 선택한 항목으로 가지는 가상의 사용자 값을 소수 집단에 추가하며 이 과정을 소수 집단과 다수 집단의 크기가 같아질 때까지 반복하는 방법이다. 이러한 편향성 보정 방법은 데이터 입력과 추천 알고리즘 사이에 데이터 증진 단계를 추가하는 방법으로 편향성을 보정한다. 이는 추천 알고리즘 자체를 수정하지는 않으므로, 추천 알고리즘을 수정하는 것에 비해 매우 적은 비용으로 적용할 수 있다는 장점이 있다. 이에 새로운 공정성 정의를 기계학습 모델의 학습목표(learning objective)로 설정하여 모델을 새롭게 학습시켜 공정성을 향상시키는 방법이 적용되어야 할 것이다. 이러한 실험연구는 데이터 증대 방법이라는 접근의 유용성을 입증한 것일 뿐 매우 다양하게 사용될 수 있는 모든 추천시스템에서 편향성 차이를 감소시킨다고 볼 수는 없다. 하지만 추천시스템이 가지는 편향의 기본적인 발생 과정과 편향의 보정 방법을 보여준다. 즉, 추천시스템에 내재적으로 편향의 발생이 나타나고 보정이 가능한 만큼 실제 추천시스템에서는 시스템이 가지는 내용과 특성을 바탕으로 다양한 접근이 필요하다.

앞서 언급한 바와 같이 우리나라는 2020년에 이르러 알고리즘 기반의 추천시스템이 법적으로 사용 가능한 환경이 된 만큼 추천시스템의 공정성 확보를 위한 제도와 정책이 구체적으로 마련

되지 않았다고 할 수 있다. 기업이나 협회차원에서는 접근이 시작되었다고 할 수 있는데, 카카오는 <알고리즘 윤리 헌장> 발표를 통해 국내에서는 최초로 인공지능 윤리헌장을 제정한 바 있다. 내용을 살펴보면 첫째, ‘알고리즘 기본 원칙’에서 윤리헌장을 도입 목적을 제시하고, 둘째, ‘차별에 대한 경계’를 통해 알고리즘의 결과물로 인한 의도적인 사회적 차별이 일어나지 않도록 함을 이야기하고, 셋째, ‘학습 데이터 운영’을 통해 알고리즘 학습에 사용되는 데이터가 반드시 사회 윤리에 근거하여 수집 및 분석, 활용됨을 강조하고, 넷째, ‘알고리즘의 독립성’을 통해 알고리즘이 타인에 의해 자의적으로 훼손되거나 영향받지 않도록 엄중한 관리가 필요함을 제안하고, 다섯째, ‘알고리즘에 대한 설명’을 통해 이용자와의 신뢰관계 및 기업 경쟁력을 훼손하지 않는 범위에서 알고리즘에 대해 성실히 설명할 것을 약속하고, 여섯째, ‘기술의 포용성’을 통해 알고리즘을 기반으로 제공되는 서비스가 우리 사회 전반을 포용할 수 있도록 명시하며, 마지막으로 ‘아동과 청소년에 대한 보호’를 통해 아동 및 청소년이 알고리즘으로 인해 부적절한 정보와 위험에 노출되지 않도록 함을 강조하고 있다. 한국 인공지능 윤리위원회는 2019년 10월 23일에 인공지능 윤리헌장(The AI Ethics Charter)을 공포하여 인공지능의 원칙, 인공지능의 목적, 인공지능 개발자의 윤리들과 관련된 13가지 원칙, 인공지능의 소비자 관점에서의 원칙, 인공지능의 공동의 책임과 이익공유의 측면에서 사회적 가치 추구를 공포한 바 있다. 하지만 이는 선언적 의미에 지나지 않은 것으로서 아직 공정성 보장과 편향보정 등 추천 알고리즘과 관련한 접근은 아직 개념적 수준에 머무르고 있다. 따라서 공포내용을 업무에 반영하고 오류를 수정하는 구체적 방법과 감시기능의 적용이 구체화 되어야 한다.

이처럼 추천시스템의 공정성 편향의 보정은 그 영역이 방대하고 개별 시스템에서도 다른 접근이 요구되는바 추천시스템 편향 보정 및 공정성 보장을 위한 가이드라인 제시의 기본 단계로서 전문가 조사를 실시하였다. 전문가 조사결과 ‘알고리즘 공정성’ 개념이 보편적으로 정의가 가능한지 여부에 대하여 소수의 전문가들이 ‘추천결과의 편향성’, ‘보편인권에 벗어나지 않음’ 등으로 정의하고, 핵심개념으로 ‘차별과 편향의 배제’, ‘투명성’ 등을 제시하였다. 하지만 다수의 전문가들은 알고리즘이 편향되지 않고 중립적이게 제공하는 것은 매우 어려우며 공정성의 정량적 평가의 어려움, 사회문화적인 배경에 따른 차이 등의 이유에서 공정성의 보편적 정의가 불가능하다고 판단했다. 알고리즘 공정성을 분석하기 위한 가장 적합한 분석방법에 대한 의견도 다양하게 제시됐는데 이용자 기반의 실험, 대규모 온라인 패널 기반 실험, 시스템 기반의 자료를 바탕으로 한 연구, 시뮬레이션 기반 실험 등의 분석방법이 모두 장단점이 있으며 이를 통합적으로 사용하여 각 방법론의 단점을 보완하는 방향으로 나아갈 필요성을 제시했다. 알고리즘 공정성을 보장하기 위한 방안으로 영향성 평가제도에 대한 의견에서는 반드시 필요하다는 의견과 무의미하다

는 의견이 충돌하였고, 다수의 전문가들이 영향성 평가제도에 대한 필요성에 대해서 동의했지만 방향성과 실행방향에 대해서는 다양한 의견을 제시했다. 알고리즘 영향성 평가제도의 주체는 독립성과 전문성을 갖춘 제3의 기관이나 전문가로 구성된 위원회, 시민단체, 정부기관 등을 제시했으며, 평가대상에 대해서도 공공영역 우선적용 후 확대, 일정 규모 이상의 기업 모두 포함 등 다양한 의견이 제시되었다. 그러나 알고리즘 적용되는 분야에서 영향성 평가제도가 무의미하며 평가제도에 대한 연구결과와 사회적 합의가 선행되어야 한다는 의견 등 다양한 평가와 방향성이 제시되었다. 이러한 전문가 조사 결과는 추천 알고리즘에 대한 접근이 개별 사안을 중심으로 접근되어야 하며, 연구의 방법 또한 추천시스템의 내용과 목적을 바탕으로 공학적 접근과 비공학적 접근이 다학제적으로 보완하여 적절한 알고리즘과 보정방법 및 감시기능이 적용되어야 함을 의미한다.

본 연구는 시뮬레이션 기반 실험연구를 통해 데이터 단계의 보정을 통해 추천시스템의 편향을 줄일 수 있다는 것을 실증하고, 전문가 조사를 통해 공정성 평가 및 보정을 위한 방법론적 시도의 필요성과 의미를 확인하였다. 이를 바탕으로 첫째, 추천시스템 공정 연구에 대한 메타분석적 연구와 접근을 통해 개념 차원의 하위 요소들을 분류하고 개념 차원 분류와 기존 연구의 효과 등을 종합적으로 분석할 필요성을 제시한다. 이러한 접근을 바탕으로 정책적 차원에서 필요한 공정성의 평가적 개념을 구성할 수 있을 것이다. 둘째, 확대된 시뮬레이션 연구방법과 실제 데이터 기반의 검증을 통해 추천시스템에 적용되는 다양한 알고리즘의 효과를 확장적으로 비교할 수 있을 것이다. 셋째, 데이터 수집 플랫폼의 구축으로 제3자 영역에서의 데이터를 수집함으로써 대규모 데이터를 확보를 통해 시스템 기반적인 실험 구현이 가능할 것이다. 넷째, 본 연구결과는 추천시스템 알고리즘의 직접 및 간접적 효과를 통계적으로 증명하였는데 이를 기반으로 대규모 통제된 온라인 실험연구가 병행된다면 추천시스템의 편향이 가져오는 지각 및 행동적 영향을 탐지할 수 있을 것이다. 다섯째, 본 연구의 결과는 민간 추천시스템에 의한 이용자 보호 침해 사례가 발생했을 때, 그 영향의 크기나 방향을 가늠할 수 있는 평가적 방법론으로 활용 가능하다. 여섯째, 기업의 자율규제활동을 진작하기 위해 개발 및 운영단계별 투명성 가이드라인을 개발하는데 지원함으로써 추천시스템의 공정성을 확보하고 운영의 투명성을 높일 수 있을 것이다. 일곱째, 협의체 구성 등 공동의 협력 플랫폼을 통해 자율활동의 성과를 공유하고 확산시킬 수 있을 것이다. 이용자보호제도의 하나로, 추천 알고리즘의 투명성을 위한 자율활동을 포함하고 이를 지원하는 제도적 노력도 고려될 수 있다. 다음으로, 다양한 교육 프로그램과 이용자의 인식제고 캠페인 활동을 통해 이용자를 대상으로 추천시스템의 작동원리와 영향을 이해시킬 수 있을 것이다.

본 연구를 통해 추천시스템 알고리즘의 공정성 문제는 추천시스템 제공자가 시스템이 사회적 가치를 훼손하지 않는지를 모니터링할 필요가 있으며, 이를 위한 윤리적 가이드라인 등을 마련할 필요성이 제시되었다. 연구의 결과를 토대로 향후 이용자 보호 정책으로서 첫째, 추천시스템의 영향성 평가를 공공분야에 적용하는 노력이 필요하다. 추천시스템은 민간영역에서 기업 등의 경쟁력을 좌우하는 지적 재산으로서, 제도를 통해 공정성과 같은 사회적 가치를 강제하기는 어렵지만 공적 시스템에 한해서는 그 정당성이 확보됨으로써, 이를 우선적으로 적용할 필요가 있다. 영향성 평가방법을 고안하게 되면, 민간의 영향력 있는 추천시스템이 이용자 및 시장 보호를 침해하는 사안이 발생했을 때, 그 영향의 크기나 방향을 가늠할 수 있는 평가적 개념과 도구로 활용할 수 있다. 둘째, 추천시스템 알고리즘의 투명성 강화를 위한 자율노력 지원의 필요성이다. 기업의 자율규제 활동의 하나로 추천시스템의 투명성을 강화 및 이용자 보호 노력을 지속하기 위한 자율규약 및 자율적 모니터링과 실험결과의 공표 등의 노력을 진작시키는 것이 필요하다. 이를 위해 자율규제를 잘 수행하는 기업 등에 인센티브를 제공하는 방안을 고민해야 하며, 자율노력의 확산을 위한 다양한 협의체 등의 구성이 필요하다. 셋째, 이용자의 경각심 고취와 자율능력을 기르기 위한 정책의 확대이다. 추천시스템 알고리즘이 갖는 편향의 문제를 이용자들이 지각하고 이를 합리적으로 이용할 수 있도록 다양한 캠페인 활동과 교육사업 등이 확대될 필요가 있다. 마지막으로, 추천시스템의 작동방식 등에 대한 이용자 대상 고지 노력이다. 알고리즘이 적용된 검색 시스템 이용자의 권익을 위해 검색 알고리즘이 상세하고 복합적으로 제공되어야 할 필요가 있다. 본 연구의 결과를 통해 추천시스템이 이용자에게 그 작동원리 등 결괏값이 미치는 요인을 고지하고 설명하는 것이 중요하다는 것을 알 수 있다. 따라서 이와 관련한 고지 노력에 관심을 기울일 필요가 있다.

목 차

제1장 서 론	1
제1절 연구의 필요성 및 목적	1
1. 연구의 필요성	1
2. 연구내용	4
제2장 추천시스템 알고리즘의 현황 및 문제점	5
제1절 알고리즘 기반의 추천시스템	5
1. 알고리즘 기반 추천시스템의 개념	5
2. 알고리즘 기반 추천시스템의 사용 현황	6
제2절. 추천시스템의 공정성 개념과 탐지방법	16
1. 추천시스템 공정성의 개념	16
2. 추천시스템 공정성 기존 연구	20
3. 공정성 탐지 노력의 예: 순위의 공정성(fair ranking)	23
4. 추천시스템 공정성 탐지방법	26
5. 추천시스템 공정성 평가에 대한 기존 접근	32
제3장 추천시스템 공정성 확보를 위한 정책 및 제도적 논의	36
제1절 개발단계에서의 투명성 확보 방안	36
1. 개발자를 위한 오픈소스 툴킷	36
제2절 자율규제활동과 공정성	40
1. 기업차원의 자율활동	40
2. 정부차원의 제도적 활동	46
3. 알고리즘 영향성 평가	49

제4장 시뮬레이션 기반 실험방법론의 적용과 공정성 평가	52
제1절 추천시스템 공정성 관련 선행 실험	52
1. (주요연구1) Gajane, P. and Pechenizkiy, M.(2018) 연구	52
2. (주요연구2) Yao, S. and Huang, B(2017) 연구	53
3. (주요연구3) Tsintzou, V 외 (2018) 연구	54
4. (주요연구4) Mansoury, M. 외 (2019) 연구	54
제2절 추천시스템 공정성 실험설계와 절차	55
1. 연구문제 및 연구방법	55
제3절 실험결과와 함의	56
1. 정의 및 모델	56
2. 가상의 데이터를 이용한 편향성 차이 실험 결과	59
3. 데이터 증대 방법을 통한 편향성 보정	63
제5장 추천시스템 공정성을 평가 및 보정 방법론에 대한 전문가의견	72
제1절. 전문가조사 방법 및 연구문제	72
제2절. 전문가조사 결과	73
1. (연구문제1:개념적 접근) 알고리즘 공정성에 관한 보편적 정의가 가능한가?	73
2. (연구문제2:방법론적 접근) 알고리즘 공정성을 분석하기 위한 가장 적합한 분석방법은 무엇인가?	76
3. (연구문제3:제도적 접근) 알고리즘 공정성 보장방안을 위한 영향성 평가 제도가 필요한가?	78
4. 추가적인 전문가의견	81
제6장 결론 및 제언	84
제1절 주요 연구요약	84
1. 추천시스템 알고리즘의 현황 및 문제점	84
2. 추천시스템 공정성 확보를 위한 정책 제도적 방향	86

3. 추천시스템 공정성 보정을 위한 시뮬레이션 기반 실험	87
4. 추천시스템 공정성을 평가 및 보정 방법론에 대한 전문가의견	87
제2절 제언	88
1. 추천시스템 공정성 보정을 위한 방법론적 제언	88
2. 추천시스템의 공정성 보정을 위한 정책적 제언	90
참 고 문 헌	93
부 록	96

표 목 차

<표 II-1> 대표적인 추천시스템에서의 공정성 문제	11
<표 II-2> 편향 개념	19
<표 II-3> 각 성별 구분 알고리즘의 피부색과 성별에 따른 인식률(eg., DF=Darker Female)	30
<표 IV-1> 실제 데이터에 대한 User-KNN(위)과 SVD(아래)의 편향성 차이	67
<표 IV-2> 실제 데이터를 이용한 User-KNN(위)와 SVD(아래)과 우리가 제안한 4가지 방법을 적용한 추천시스템의 편향성 차이 비교	68
<표 V-1> 전문가조사 리스트	73
<표 V-2> 개념적 측면 전문가 의견	75
<표 V-3> 방법론적 측면 전문가 의견	78
<표 V-4> 제도적 측면 영향성 평가제도에 대한 전문가 의견	81

그 림 목 차

[그림 II-1] UserKNN과 GULM	21
[그림 II-2] TREC 2020 Fair Ranking Track	25
[그림 II-3] 닌넨버스와 윌렘센(2015)의 이용자 중심 평가 모델(User-Centric Evaluation Framework: UCEF)	27
[그림 II-4] 대규모 온라인 통제실험방법	27
[그림 II-5] COMPASS에 의한 백인과 흑인의 위험평가 점수분포	28
[그림 II-6] COMPASS의 스코어에 따른 인종별 재범확률	29
[그림 II-7] 인물 데이터세트별 표본의 성별 및 피부색	31
[그림 II-8] 커뮤니티와 미디어 채널에 대한 활동량 증가	33
[그림 II-9] 사용자 교차점 분석	33
[그림 II-10] 사용자의 콘텐츠 이동양상	34
[그림 II-11] 채널과 영상에 대한 임의 행로(random-walks)시뮬레이션 결과	34
[그림 III-1] <AI Fairness 360>의 AI 모델 사이클 전반에 대한 툴킷 작동원리 ..	37
[그림 III-2]<what-if>공정성 툴킷의 시각화 기능	38
[그림 III-3]<LiFT>의 공정성 측정 프레임워크 플로우 차트	39
[그림 III-4] 캐나다의 알고리즘 영향성 평가(Algorithmic Impact Assessment) ..	50
[그림IV-1] 사용자의 선호도에 대한 추천 결과물 선호도	60
[그림IV-2] 비대칭적 선호도를 가지는 경우, 두 집단의 사용자 선호도에 따른 추천 결과물 선호도.	61
[그림IV-3] 두 집단의 크기 비율에 따른 추천 결과물 선호도	62
[그림IV-4] 두 카테고리의 크기 비율에 따른 추천 결과물 선호도	62
[그림IV-5] 데이터 증대 방법을 통한 추천알고리즘(User-KNN알고리즘)의 결과물 선호도 보정결과	65
[그림IV-6] 데이터 증대방법을 통한 추천알고리즘(SVD알고리즘)의 결과물 선호도	

보정 결과	66
[그림Ⅳ-7] 실제 데이터 사용자 집단 비율을 조절한 데이터 SVD 추천알고리즘 이용한 편향성 차이	69
[그림Ⅳ-8] 실제 데이터 사용자 집단 비율을 조절한 데이터 User-KNN 편향성 차이 결과	70

제1장 서론

제1절 연구의 필요성 및 목적

1. 연구의 필요성

가. 추천 알고리즘의 영향력 확대 및 문제점 보완의 필요성

추천 알고리즘은 데이터 기반의 자동화된 의사결정(automated individual decision-making)을 의미한다. 과거 인간의 의사결정에 개입하는 외부의 추천행위는 인간의 분석과 판단에 의해 이루어졌다. 그러나 인간을 둘러싼 환경과 행위들이 데이터화 되고 이를 분석하는 기법이 발전하면서 인간의 의사결정에 개입하는 외부의 추천행위에 추천 알고리즘의 역할은 커지고 있다. 추천 알고리즘은 다양한 영역에서 프로파일링 등 개인의 선호와 성향을 예측하는 기법을 통해서 인간행위를 유도하는 결정을 한다.

예를 들어 내비게이션은 최적의 주행경로를 빠른 시간 안에 알려주며 주행 도중 주행환경을 재분석하여 끊임없는 최적 주행경로 추천을 지속한다. 아마존이나 이베이는 엄청난 종류의 상품 안에서 소비자가 관심을 가질만한 상품을 추천해 주고, 넷플릭스와 유튜브는 관심을 가질만한 영상을 추천한다¹⁾. 금융권에서는 고객의 몇 가지 정보를 바탕으로 금융상품의 적용방법을 추천하기도 한다. 이처럼 현대사회에서 추천 알고리즘은 사회적 에이전트로서 공식 또는 비공식 영역에서 그 영향력이 커지고 있다.

우리나라에서는 개인정보 활용의 법적 제약에 따라 추천 알고리즘 적용에 한계가 있었다. 그러나 2020년 5월의 데이터 3법 개정으로 이전, 온라인상에서 수집된 개인정보를 총괄하는 모법(母法)이 정보통신망법에서 개인정보보호법으로 바뀌게 되었다. 개정된 개인정보보호법의 제28조의 2는 과학적 연구를 위해 정보주체의 동의 없이 가명정보 처리를 가능하게 하였다. 여기에서 ‘과학적 연구’의 목적은 ‘기술의 개발과 실증, 기초연구, 응용연구 및 민간 투자 연구 등 과학적 방법을 적용하는 연구’라고 규정함으로써 데이터를 기반으로 하는 새로운 기술·제품·서비스 개발 등에 개인정보를 사용 가능하도록 하였다. 신용정보법에서는 금융이력 부족자도 통신료 등의 요금 납부 내역이나 온라인 쇼핑 내역, SNS 정보 등 비금융정보를 활용해 개인 신용을 평가받을 수 있도록 하여 금융사업에서 개인정보 활용의 제약을 완화시켰다. 이에 따라 프로파일링과 알고리즘 기반의 의사결정을 바탕으로 하는 추천시스템 도입의 확대가 예측된다.

추천 알고리즘에 대한 기본적인 믿음은 수학적 원리에 따라 작동하며 동일한 입력에 대해서 항상 동일한

1) 넷플릭스(Netflix) 내 영상 시청의 75%는 알고리즘 추천서비스를 통해 제공되고 있고, 넷플릭스 추천서비스를 사용하는 소비자들은 다른 OTT 플랫폼 서비스가 제공하는 추천 콘텐츠보다 넷플릭스가 제공하는 추천 콘텐츠에 대한 만족도에 더 높은 점수를 주었다고 보고된바 있다(Troy, 2019).

결과를 출력하는 알고리즘²⁾에 대한 기대에서 비롯되었을 수 있다. 그러나 개인의 선호와 성향을 예측하는 기법을 통해서 인간행위를 유도하는 추천시스템에 편향과 불공정이 나타나는 문제가 대두되고 있는 실정이다. 머신러닝을 기반으로 하는 학습과정에서 사회적으로 문제가 발생할 수 있는 학습을 할 경우 알고리즘은 인간의 기대와 다른 판단을 할 수 있다. 또한 알고리즘이 변화한 사회적 시스템을 반영하지 못할 수 있으며, 알고리즘의 추천과정에 대한 검토가 불투명 할 수 있다. 알고리즘의 비정상적 결과출력은 정보편식 현상, 에코챔버 등 다양한 사회적 문제를 야기할 수 있다(황용석·김기태, 2019/2020³⁾). 이 때문에 유럽연합 위원회(European Commission)는 ‘디지털 형태변형 모니터 (Digital Transformation Monitor)’ 차원에서 추천시스템의 부작용을 추적하고 있다. 또한 유럽의 입법조사기관(European Parliamentary Research Service: EPRS)에서는 알고리즘 시스템이 대규모의 자료를 사용한 복잡한 추론의 결과라는 점에서 결정과정이 가지는 불투명성을 우려하여 2019년 4월, 알고리즘 시스템에 적용되어야 할 조건인 “A governance framework for algorithmic accountability and transparency”를 발표한 바 있다. 따라서 추천 알고리즘 연구와 적용 확대가 예측되는 우리나라에서도 추천 알고리즘의 문제점을 파악하고 이를 보완할 수 있는 방안 모색이 이루어져야 한다.

나. 추천시스템의 사회적 영향 및 방법론 연구의 필요성

내비게이션이나 자율주행 등의 의사결정에 이용되는 추천 알고리즘은 안전과 시간이라는 명확한 목적을 가지고 추천으로 인한 효익도 비교적 명확하게 측정된다. 하지만 인간의 선호나 기업의 경영 관련 의사결정에 개입하는 추천 알고리즘은 ‘합리적 의사결정’이라는 목적에 비해 추천을 통해 발생하는 상대적 효익의 측정방법이 명확하지 않다. 이 때문에 추천시스템이 적절한 결과를 제공할 수 있는 구조를 갖추고 있는지에 대한 분석은 매우 중요하다.

가장 기본적인 것은 공급자 측면의 추천 공정성이다. 상업적 추천시스템을 가정하면 다수의 공급자와 상품, 사용자가 추천을 통해 연결되는 과정에서 배제되는 비추천의 케이스는 추천과 비교되지 않을 만큼 많다. 비교적 소수의 상품이나 서비스가 추천된다는 점에서 추천시스템에는 기본적으로 불공정한 추천이 가지고 오는 비합리적 결과가 감추어 질 수 있는 가능성이 있다. 따라서 추천의 공정성 문제에서 특정 공급자와 관련한 추천에 편향되는지의 문제는 적절한 추천이 이루어지는지와 관련한 기본적 문제가 될 수 있다. 다음으로는 사용자의 특성에 기반한 추천의 편향 문제이다. 개인화된 추천을 제공한다는 점에서 사용자 특성은 개인의 선호를 판단하는 일반적 기준으로 사용할 수 있다. 그러나 선호가 아닌 평가에서는 사용자 특성을 기준으로 사용하는 것이 문제가 될 수 있다. 은행 거래에 있어서 단순한 사용자 특성을 사용하는 것은

2) Drösser, C. (2016). Total berechenbar? Wenn Algorithmen für uns entscheiden. München: Hanser.

3) 황용석·김기태 (2019. 개인화 서비스 진전에 따른 자동추천시스템 연구 동향과 방법론적 특성 연구. 『사이버커뮤니케이션학보』, 36(2), 221 ~ 253.; 황용석·김기태, 2020 황용석·김기태 (2020. 알고리즘 기반 자동화된 의사결정의 설명 가능성에 대한 연구. 『언론정보연구』, 57(3), 41-80.

불공정한 대우로 이어질 수 있다. 또한 사회적 통합을 위해 별도의 접근을 하고있는 사안은 추천이나 평가 시스템에 부가적으로 적용되어야 한다. 공급자 측면의 추천 공정성과 사용자의 특성에 기반한 추천의 편향 문제는 추천시스템의 확대로 인해 영향력이 확대될 수 있고, 불공정이 고착화될 수 있다는 점에서 지속적으로 수정 보완해야 할 대상이다.

이와 함께 추천시스템의 연구방법도 적극적 고민이 필요하다. 공정한 추천과 관련한 접근은 크게 정확한 예측의 절차적 방법 도출을 목표로 하는 시스템 중심 접근(System-centric approaches), 아이템이 차별 없이 대우받는지와 관련한 공급자 이슈(Content provider-centric approaches), 사용자의 요구 충족에 관해 분석하는 이용자 중심 접근(User-centric approaches)으로 구분될 수 있다. 이러한 접근은 추천시스템의 종류에 따라 재정의되고 재구성되어야 한다.

다. 추천시스템 연구의 융합적 접근

추천시스템 관련 접근은 공학적 접근과 비공학적 접근으로 분류할 수 있다. 공학적 관점의 추천시스템 연구는 추천한 상품이나 서비스에 대한 사용자의 만족뿐만 아니라 추천 알고리즘에서 나타나는 편향과 불공정의 문제를 제거하는 알고리즘 발견에 관심을 가진다. 추천시스템은 과거의 데이터를 바탕으로 패턴을 수집하고 분석하여 알고리즘을 찾아 활용하는 단계를 거치게 된다. 이러한 점에서 다수의 선호를 반영하는 협업 필터링, 새로운 방식의 업무, 플랫폼을 처음 이용하는 이용자에서 나타나는 상태인 콜드스타트는 추천시스템의 기본적인 문제점으로 대두된다. 또한 학습 데이터 자체의 문제도 고려되어야 한다. 알고리즘 성능에 직결된 핵심 요소는 학습데이터이다. 하지만 학습데이터로서 학습의 기반이 되는 개인의 접근 기록 및 관련 자료에는 사회 혹은 이슈에 관심을 가지는 사용자의 사회적 편향(bias)이 반영될 가능성이 높다. 따라서 학습의 단계에서 수집될 데이터에 반영될 편향을 미리 예측하고 이를 보정할 방법이 적용되어야 한다.

비공학적 접근은 사회적 시스템 안에서 편향과 불공정 문제의 유형, 나아가 추천시스템의 사회적 영향을 발견하고 공학적 접근이 알고리즘에 내포된 문제를 해결할 수 있게 하는 방향 제시와 정책 마련에 목적이 있다고 할 것이다. 이와 같은 접근이 후행적 접근을 의미하는 것은 아니다. 예를 들어 추천시스템의 부작용 해소에 있어서 학습과정에서 나타나는 오류, 학습의 결과에서 나타난 결정 내용의 법적/윤리적 문제는 비공학적 접근을 통해 검토되어 지속적으로 알고리즘에 반영되어야 한다. 이 경우 부작용은 학습과정의 오류, 결정 내용 고지의 문제로 분류하여 학습과정의 경우 학습의 중단, 결정내용의 경우 고지의 가이드라인이 현행법 수준에서 제시되어야 한다. 이는 사회적 합의가 반영되기 어려운 알고리즘 중심의 추천시스템에 사회적 합의를 반영시키는 과정으로 이해할 수 있다.

이처럼 추천 알고리즘의 문제점 파악과 보완방안 모색은 공학적 접근과 비공학적 접근이 상호보완적으로 적용되어야 한다. 나아가 추천시스템과 관련한 사회적 합의 수준, 국제수준의 논의를 위한 접근도 이루어져야 한다.

2. 연구내용

가. 연구내용

연구의 필요성에서 밝힌 바와 같이 추천시스템은 단순한 기술의 영역이 아니라 사회적 시스템과 밀접한 관련을 가지고 있다. 따라서 기술적 오류의 가능성을 수정하고 사회적 시스템의 범주에서 기술적 접근이 이루어지도록 강제되어야 하며, 이를 위한 정책적 환경 조성이 필요하다.

이에 본 연구는 추천시스템이 가지고 있는 기본적 문제점의 실증적 규명, 사회적 시스템에서 문제가 될 수 있는 쟁점의 고찰, 공학적 측면과 비공학적 측면의 평가를 위해 필요한 연구방법론을 검토하고 실험을 통해 방법론적 타당성을 검토하고자 한다. 이를 위해 알고리즘 추천시스템에 필수적으로 적용되는 머신러닝의 유형과 각 유형의 장단점을 파악하고, 머신러닝 과정에서 편향이 나타나는 케이스와 보정방법을 살펴보는 단계를 통해 알고리즘의 부정적 결과를 보정하기 위한 방법을 탐색하고자 하는 것이다. 사회가 요구하는 최적의 결과는 알고리즘이 스스로의 학습만을 통해 제시할 수 있는 것이 아니라 사회적 환경을 반영하도록 개입함으로써 가능해질 것이다. 이러한 점에서 본 연구는 알고리즘의 문제점을 발견하고 보정을 적용하는 방법 도출을 중심으로 연구를 진행한 후 이를 지원하기 위한 제도와 정책에 관해 논의하고자 하였다.

나. 보고서의 구성

본 연구는 연구목적 달성을 위해 첫째, 추천시스템의 개념 및 추천시스템에 요구되는 쟁점 사항, 둘째, 공정성을 보장할 수 있는 알고리즘의 유형관련 연구, 셋째, 머신러닝 기반 추천시스템의 추천과정 및 편향성 보정 실증 연구를 통해 편향의 불가피성 및 보정의 방법을 실증, 넷째, 추천시스템의 공정성 확보를 위한 제도와 정책을 살펴본 후 다섯째, 제도와 정책 추진방법에 대한 전문가 조사를 바탕으로 추천시스템 편향 보정 및 공정성 보장을 위한 가이드라인을 제시하고자 한다.

제2장 추천시스템 알고리즘의 현황 및 문제점

제1절 알고리즘 기반의 추천시스템

1. 알고리즘 기반 추천시스템의 개념

전통적 관점에서의 알고리즘은 문제를 풀기 위한 계산법으로서 구조화된 절차나 방법, 계산 등을 실행하는 방식을 의미한다. 즉, 알고리즘은 입력된 정보와 출력 결과 사이의 과정에 해당되는 개념이다. 이러한 알고리즘은 입력 받은 정보를 바탕으로 사전에 규정된 절차를 거쳐 결과물을 출력해 내는데, 정보통신시스템에서의 알고리즘은 문제를 기계적 실행에 의해 해결하려는 목적을 가진다. 즉, 알고리즘은 입력된 정보와 출력 결과 사이의 과정에 초점을 맞추게 된다.

추천이 갖는 기본적인 의미는 이용자가 선호할 것으로 예측되는 무언가를 제안하는 것으로, 이를 데이터와 컴퓨터 연산을 적용하여 구축한 것이 추천시스템이다. 최근의 추천시스템은 보다 다양한 정보에 기반해 신속하고 정확도 높은 추천이 가능하도록 추천 알고리즘을 적용하고 있다. 예로, 넷플릭스(Netflix)와 같이 추천을 통해 콘텐츠 소비를 유도하는 서비스도 데이터의 분류 및 군집화에 근거한 알고리즘 기반 추천시스템을 운용중이다. 이러한 추천시스템은 이용자의 행동 및 성향정보에 근거하여 관련성과 참신성, 우연성⁴⁾을 갖춘 콘텐츠를 제안하는 데에 목적을 가지는데, 추천시스템이 제안하는 결과물은 이용자가 기존에 경험해보지 못한 것이라는 점에서 데이터에 근거한 완벽한 정답이 아닐 수 있다는 한계를 지닌다. 즉, 이미 도출된 결과에 대한 평가라는 확인절차를 통해 알고리즘의 정확도를 평가할 수 있다. 이로 인해 추천시스템 역시 머신러닝의 영역으로 편입되고 있다.

알고리즘의 근본적인 등장 배경은 인간이 수행할 과업을 대신하는 것으로, 알고리즘의 구성에 있어 인간의 판단과 사고방식이 필연적으로 개입된다. 즉, 입력과 출력 사이의 과정에 인간이 개입하여 그 처리방식을 규정하게 되는데, 이로 인해 알고리즘을 구성하는 사람의 세계관과 역사관, 윤리관 등이 출력 과정에 영향을 주어 출력 결과물의 중립성을 담보하기 어렵다는 점이 문제로 대두된다. 하지만 최근의 정보환경에서 알고리즘을 만드는 사람이 구조화 한 절차에서 발생할 수 있는 문제점과는 별개로 머신러닝과 딥러닝의 본질적 특성으로 인한 편향 발생 문제가 더욱 부각되고 있다.

Tom Michell(1997)⁵⁾은 머신러닝을 ‘특정한 T(Task)를 수행하는 과정에서 다양한 E(Experience)를 통해 P(Performance)의 변화를 학습하는 컴퓨터 프로그램’으로 정의하였다. 머신러닝은 E와 P가 존재하는 기계적 학습을 기반으로 한다. 앞서 살펴본 알고리즘은 ‘구조화된’ 절차로서 입력된 A에 B조건이 성립하면

4) 이용자에게 다소 의외이거나 놀라움을 주는 제안.

5) Mitchell, T. M. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 45(37), 870-877.

사전에 프로그래밍 된 절차를 통해 X를 동작시키지만, 머신러닝은 A라는 조건에서 결과가 X가 되는 조건 B를 찾도록 기계를 학습 시킨다. 머신러닝에서 조건 B와 프로그래밍 된 절차는 존재하지 않는다. 결과 X의 유무에 따라서 학습방법은 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 구분 되는데, 지도학습 방식은 분류(classification)와 예측(prediction) 알고리즘을 활용하여 예측모형을 개발한다. 이렇듯 지도학습을 바탕으로 예측모형을 개발하고 이를 활용하는 것을 머신러닝으로 한정하는 경우가 많은데, 이는 군집(clustering) 알고리즘을 활용하여 예측모형을 개발하는 비지도학습을 딥러닝으로 지칭하여 구분하고자 하는 접근에 따른 것이다.

결과적으로 알고리즘 기반 추천시스템은 인간이 인위적으로 구조화한 절차에 따라 출력을 하는 것이 아니라 머신러닝의 과정을 통해 인공지능 혹은 시스템이 분류와 예측 절차를 규정하고 이에 따라 추천이라는 결과를 하게 된다. 따라서 기존의 알고리즘과 현대의 알고리즘 기반 추천시스템에서의 알고리즘은 기본적인 개념과 작동방식에서 현저한 차이가 나타난다. 이러한 알고리즘 기반 추천시스템은 알고리즘을 만드는 사람의 세계관, 역사관, 윤리관이 배제되지만 학습 데이터에 내재된 편향이나 의도적 오류를 일반적인 절차로 이해할 수 있으며, 사회적 통합을 위해 의도적으로 조정하는 배려와 배제를 포함하지 않을 가능성이 높다. 또한 알고리즘이 도출하는 결과의 과정을 운용자도 알지 못하는, 이른바 ‘블랙박스’ 차원의 문제가 발생할 수 있다. 따라서 알고리즘 기반 추천시스템의 문제점 파악은 예측결과를 바탕으로 추천을 제공하는 머신러닝의 문제점을 해결함으로써 적절한 운용방향을 모색하는 목적뿐만 아니라, 딥러닝이 제시하는 예측 속에 내포될 수 있는 문제점의 제거 방안 모색이라는 의미를 포함하게 된다.

2. 알고리즘 기반 추천시스템의 사용 현황

가. 알고리즘 추천시스템의 현황과 발전과정

1) 추천시스템의 등장

ICT 기술이 발전함에 따라 대중들의 소비패턴이 빠르게 변화하고 있으며, 이에 맞춰 소비 시장 또한 급격한 변화를 맞고 있다. 과거 직접적인 구매 방식에서 벗어나 인터넷이나 홈쇼핑 등의 경로를 통해 간접적인 구매가 늘어나고 있을 뿐만 아니라 디지털 환경이 사회에 접목됨에 따라 등장하고 있는 각종 온라인 서비스 등 상품 유통의 본질적인 변화가 일어나고 있으며 그 영역 또한 지속적으로 확장되어가고 있다. 더불어 ICT 기술을 기반으로 함으로써 대중들이 접할 수 있는 상품들의 양이 급격하게 확대되고 있다. 따라서 대중들은 원하는 상품을 선택하는 데 어려움이 존재하게 되고, 상품을 제공하는 입장에서는 다양하고 시기각각 변화가는 대중들의 선호도를 추적하기 어려워졌다.

이에 대하여 소비 및 판매 양측의 어려움을 해소하고 모두 만족시킬 수 있도록 돕는 추천시스템이 등장

하였다. 초창기 가장 간단한 추천시스템으로는 높은 판매량 순으로 상품을 추천하는 베스트셀러 기반 추천 방식과 설문을 통해 선호도를 파악하고 상품을 추천하는 최소질의대상 상품결정 방식 등이 존재한다. 하지만 이러한 추천시스템은 개인화된 추천이 불가능하거나 신뢰성 저하 등의 문제점들이 존재한다. 따라서 최근에는 정보 필터링(Filtering) 방식과 연관성 분석(Association rule) 등 다양한 추천 기법들이 연구되고 있으며, 특히 정보 필터링 방식을 기반으로 하는 콘텐츠 기반, 협력 필터링 방식의 추천시스템과 두 가지 방식을 섞은 하이브리드 추천시스템이 대두되고 있다. 추천시스템에 대한 자세한 설명은 이후 추천시스템 현황 및 발전과정에서 서술한다.

2) 추천시스템의 현황 및 발전과정

현재 주로 사용되는 추천시스템은 협업 필터링, 연관성 분석, 콘텐츠 기반 방식의 추천시스템들이다. 이를 추천의 타입으로 보면 크게 이용자 정보를 기반으로 하여 자동으로 아이템 리스트를 선별해주는 개인화 추천(Personalized recommender), 상품 순위나 선택 상품 기준의 순위를 통해 추천을 해주는 비개인화 추천(Non-personalized recommender) 그리고 아이템 즉, 상품 자체가 보유하고 있는 정보(Contents)를 활용하여 추천하는 속성기반 추천(Attribute-based recommender)로 구분할 수 있다.

이러한 추천시스템의 초기 단계에서는 대부분 개인화된 추천을 제공하더라도 신뢰성에 저하를 보이는 등의 문제점들이 존재하였다. 이후 추천시스템의 접근방식과 알고리즘에 대한 연구가 계속되면서 연관성 분석(Association rule), 콘텐츠 기반, 협업필터링이라는 크게 3가지 접근방식으로 발전하게 된다.

먼저 연관성 분석은 장바구니 분석으로도 불리며 연관데이터 간의 연관 법칙을 찾는 방법으로 이용자의 행동 데이터를 기반으로 아이템 간의 연관관계를 만들고 이를 통하여 아이템을 추천하는 방식이다. 이용자가 한 아이템에 대하여 선호를 하였을 때, 연관성이 존재하는 다른 아이템에 대하여 선호한다는 조건부 확률을 기반으로 추천을 진행한다. 추천항목의 우선순위가 정해지기 때문에 기존의 기억기반 협업필터링 방식에서 발생하는 희소성과 Gray sheep⁶⁾문제를 해결할 수 있지만, 이용자나 아이템의 수가 많아질 경우, 데이터 희소성 문제가 등장할 수 있다.

이에 대한 해결방안으로 지지도와 신뢰도 계산 시 임계값을 설정하여 특정 항목집단에 대해서만 분석을 진행하는 방법이나 RFM⁷⁾기법을 이용해 이용자 그룹을 세분화하여 그룹 내에서 연관성 분석을 수행하는 방법 등이 연구되어 더욱 발전된 연관성 분석 추천시스템들이 등장하고 있다. 이러한 연관성 분석 기반 추천시스템은 유통업이나 홈쇼핑, 상품 진열 및 패키징, 번들링 등의 영역에서 주로 사용되고 있다.

다음으로 가장 널리 알려져 있고 대부분의 추천시스템에서 사용되는 협업필터링 기반 모델의 알고리즘은 현재까지의 추천시스템의 알고리즘 중에서 가장 우수한 성능을 보이고 있다. 협업필터링은 기존의 이용

6) 비슷한 선호도를 갖는 이용자의 선호 아이템을 대상 이용자에게 추천하는 협업 필터링 방식에서 비슷한 선호도를 갖지 않는 즉, 특이 선호도를 갖는 이용자에게는 알맞은 추천이 불가능한 문제를 말함.

7) RFM 분석은 Recency, Frequency, Monetary 기준을 바탕으로 고객을 분류하는 방법 중 하나임.

자 및 아이템 선호에 대한 데이터를 학습하여 이를 바탕으로 개인화된 추천을 제공하며 넓은 범위의 아이템 즉, 추천 아이템의 다양성을 보장한다. 이러한 협업필터링은 Goldberg et al.(1992)에 의해 처음으로 개념이 소개되었으며 크게 기억기반 협력필터링과 모델기반 협력필터링으로 나뉜다. 기존의 협업필터링에 해당하는 기억기반 협력필터링은 이용자가 입력한 선호도나 특정 아이템에 대한 평가를 기반으로 유사한 이용자들에게 추천을 제공하는 방식이다. 따라서 별도의 모델을 구축하지 않고 추천이 필요할 때마다 휴리스틱 기법을 이용해 결과를 도출하는 ‘Lazy learning’⁸⁾에 해당한다. 이는 데이터의 희소성이나 확정성 그리고 Gray sheep과 같은 문제점으로 이어진다. (손지은 외, 2015⁹⁾)

이에 대한 해결책 중 하나로 등장한 것이 바로 모델기반 협력필터링이다. 모델기반 협력필터링은 기계학습과 데이터마이닝 방식을 활용하여 학습데이터를 통해 모델을 구축한 후 사용자에게 아이템을 추천하는 방식이며 성능이 우수하여 널리 쓰이고 있다. 이러한 협력필터링은 등장 초기에 미네소타 대학의 Group-Lens를 시작으로 Amazon.com, CD Now.com, MovieFinder 등에서 사용되기 시작하였으며, 현재 교보문고나 아마존과 같은 온라인 서점을 비롯하여 각종 포털 및 언론사의 뉴스 서비스, Netflix, 멜론등과 같이 음악 및 영화 스트리밍 서비스 등의 영역에서 주로 활용되고 있다.

앞선 방식들보다 비교적 뒤늦게 등장한 콘텐츠 기반 추천시스템은 아이템이 보유하는 콘텐츠 자체를 직접 분석하여 아이템 간 혹은 아이템과 이용자 선호도 간의 유사성을 토대로 추천을 해주는 방식이다. 분석이 용이할 뿐만 아니라 메타데이터의 가용성으로 인해 영화, 음악, 도서뿐만 아니라 텍스트 기반의 뉴스나 인터넷 기사 등을 추천하는데 널리 쓰이고 있다. 협업 필터링에서는 추천 대상 이용자와 선호도가 유사한 이웃 이용자를 찾기 위해 여러 이용자의 평가점수 데이터가 필요한 반면 콘텐츠기반 접근방식에서는 추천 대상 이용자의 선호도를 파악하기 위해 해당 이용자만의 과거 데이터나 프로필 정보가 이용된다. 추천 대상이 되는 이용자에 대한 독립적인 정보만을 필요로 하기 때문에 다른 이용자 정보가 부족하더라도 추천이 가능하며 새로운 아이템이 추가되어도 First rater¹⁰⁾문제가 발생하지 않는다. 하지만 전제 조건인 과거 데이터와 프로필 정보가 존재하지 않는다면 활용이 불가능하고, 기존의 콘텐츠 기반 추천시스템은 이용자의 과거 데이터를 바탕으로 유사 아이템을 추천해주기 때문에 다른 이용자의 선호를 반영하지 못하여 비슷한 아이템들만 추천되고 새로운 속성의 아이템을 추천하지 못하는 과도한 특수화 문제가 존재한다.

이에 대한 해결책으로 추천시스템에 유전자 알고리즘을 적용함으로써 무작위 요소를 추가하거나 돌연변이 방식을 사용하는 기법이 연구되었고 이를 기반으로 만들어지는 새로운 콘텐츠 기반 추천시스템은 기존 콘텐츠기반 접근방식보다 우수한 성능을 보이며 발전 중이다.

대표적인 3가지 추천시스템 접근방식 이후 각각 장단점을 보유하고 있는 협업 필터링과 콘텐츠기반 추천 시스템을 조합하여 장점을 살리고 단점을 보완하는 하이브리드 추천시스템에 대한 연구가 활발히 이루어

8) 머신러닝에서 학습 데이터를 간편하게 저장하고 테스트 데이터가 도착할 때까지 대기하는 형태의 학습방법을 말함.

9) 손지은·김성범·김현중·조성준(2015). 추천시스템 기법 연구동향 분석. 대한산업공학회지, 41(2). 185-208

10) 협업 필터링에서 어느 사용자에게도 평가받지 못한 새로운 아이템이 있을 경우 아이템 간의 의미 상호작용이 생성될 수 없어 추천이 불가능한 문제를 말하며 Cold-start 문제와 유사함.

지고 있다. 하이브리드 추천시스템은 한 방식을 다른 방식에 더하거나, 융합하여 하나의 모델로 만들어 기존의 두 추천시스템 방식에서 흔히 존재하는 Cold start나 희소성 문제 등을 해결하고자 하는데, 여러 기술의 추천시스템을 융합하는 방식으로는 Weighted, Switching, Mixed, Feature Combination, Feature Augmentation, Cascade, Meta-level의 총 7가지 대표적인 하이브리드화 기법들이 존재한다. 이러한 하이브리드 추천시스템 방식은 Netflix, 빅인사이트, NewsDude 등 다양한 분야에서 기존의 추천시스템의 문제점들을 해결하며 추천시스템의 성능을 높이고 있다.

마지막으로 각각 추천시스템들은 장단점을 지니고 있기 때문에 필히 새로운 추천시스템이 더 나은 추천시스템이라고 할 수 없다. 다만 추천시스템의 발전 추세는 점점 머신러닝이나 딥러닝, 빅데이터 기술 등을 활용하여 알고리즘 모델을 형성하고 이를 기반으로 추천을 제공하는 방식으로 나아가고 있다. 하지만 이러한 추천시스템의 성능을 높이는 것과 별개로 자동화된 추천시스템으로 인해 발생하는 각종 사회적/기술적 문제들이 제기되고 있다. 관련하여, 본 연구는 추천시스템이 사회 전반에 적용되는 상황에서 이용자에게 영향을 미칠 수 있는 알고리즘 편향을 분석하고 공정성을 확보할 수 있는 방안을 제안하고자 한다.

나. 알고리즘 기반 추천시스템의 유형

알고리즘 기반 추천시스템은 추천하는 방식에 따라 크게 협업필터링(collaborative filtering), 콘텐츠 기반 필터링(content-based filtering), 딥러닝(deep learning) 기반 추천시스템, 이용자선택기반 추천시스템으로 구분할 수 있다.

1) 협업 필터링(collaborative filtering)

협업 필터링¹¹⁾은 기존 사용자들의 행동정보만을 이용하여 해당 사용자와 비슷한 성향의 사용자들이 선호하는 아이템(user-item recommendation), 또는 특정 아이템을 선호하는 사용자들이 선호하는 다른 아이템(item-item recommendation)을 추천하는 방식이다. 이러한 협업 필터링의 기본 가정은 ‘특정 아이템에 대해서 선호도가 비슷한 사용자들은 다른 아이템에서도 비슷한 선호도를 보인다.’라는 것이다(손지은 외, 2015)¹²⁾. 협업 필터링의 핵심 기술은 행렬분해(matrix factorization)¹³⁾로 사용자-아이템 행렬(user-item matrix) 또는 아이템-아이템 행렬(item-item matrix)을 분해하고 차원을 축소시킨 후 복원하는 과정을 통하여 사용자의 아이템에 대한 선호도 또는 아이템간의 유사도를 계산하는 기술이다.

11) Sarwar, B. et al. 2001. Item-based collaborative filtering recommendation algorithms. WWW: Proceedings of the International World Wide Web Conference. 10, (May 2001), 285-295.

12) 손지은·김성범·김현중·조성준(2015). 추천시스템 기법 연구동향 분석. 대한산업공학회지, 41(2). 185-208

13) Koren, Y. et al. 2009. Matrix Factorization Techniques for Recommender Systems. Computer. 42, 8 (Jul. 2009), 30-37.

기본적인 협업 필터링 모델은 이용자와 상품의 특성에 대한 어떤 정보도 활용하지 않는다. 때문에 기계 학습의 가장 핵심적인 공정성의 정의 중 하나인 “비자각에서 오는 공정성(fairness through unawareness)”¹⁴⁾¹⁵⁾을 충족한다. 비자각에서 오는 공정성이란 성별, 인종과 같은 보호되어야 할 이용자 정보가 알고리즘에 사용되지 않음으로써 성립되는 공정성이다. 더 강력하게는, 보호되어야 할 이용자 정보를 유추할 수 있는 정보(노동자 색을 통해 인종을 유추하거나 거주지를 통해 소득수준을 유추하는 경우)까지 사용되지 않음으로써 성립될 수 있다. 알고리즘에서 나타나는 편향을 완화하기 위해 성별, 나이, 장애와 같은 보호속성(protected class variables)과 대리변수를 학습데이터에서 삭제시킴으로써 비자각에서 오는 공정성을 충족하고자 하지만, 이러한 방법은 다른 측면에서 누락편향성이 존재할 수 있으며, 다른 대리변수들이 등장하여 이용자 정보를 유추할 수 있다는 점에서 한계가 있다.

또한 협업 필터링은 다른 공정성 문제를 갖는다. 바로 ‘콜드 스타트 문제(cold start problem)’¹⁶⁾이다. 협업 필터링은 기존의 사용자 행동기록에 기반하여 아이템을 추천하기 때문에 기존의 행동기록이 없는 신규 이용자는 좋은 추천을 받기 어려우며, 마찬가지로 신규 아이템은 좋은 품질을 갖추고 있더라도 추천되기 어렵다. 즉 소비자와 생산자 양측 모두의 입장에서 신규 항목에 대한 공정성 문제가 발생할 수 있다. 그리고 행렬분해를 기초로 사용하게 되는 협업필터링은 계산 과정이 복잡하고 많은 알고리즘에 해당되어 사용자수가 늘어나게 되면 효율성이 떨어진다는 단점이 있다(서봉원, 2016¹⁷⁾).

2) 콘텐츠 기반 필터링(Content-based filtering)

콘텐츠 기반 필터링은 아이템 항목의 특성(종류, 브랜드, 가격대 등)을 분석한 프로파일(item profile)과 이용자의 특성(성별, 인종, 지역, 나이 등)을 분석한 프로파일(user profile)을 추출하여 추천에 활용하는 방식을 말한다. 협업필터링 방식은 이용자의 행동기록을 사용하는 반면, 콘텐츠 기반 필터링은 항목자체를 분석하여 유사하거나 추천하는 콘텐츠를 구현하게 된다. 구체적으로 추천하는 방식을 살펴보면 항목에 대한 프로파일(item profile)과 사용자의 선호도를 추출한 프로파일(user profile)의 유사성을 계산하여 사용자가 선호할만한 콘텐츠를 추천한다(서봉원, 2016). 콘텐츠 내용을 분석하기 위한 알고리즘으로 군집분석(Clustering analysis), 인공신경망(Artificial neural network) 등의 기술이 사용된다. 콘텐츠 기반 필터링은 고전적인 추천 방식이었지만, 최근 기계학습 기술 등의 발전에 힘입어, 협업 필터링의 단점을 보완하기 위

14) Gajane, P. and Pechenizkiy, M. 2018. On Formalizing Fairness in Prediction with Machine Learning. FATML: Workshop on Fairness, Accountability, and Transparency in Machine Learning. (May 2018), 1-6

15) Chen, J. et al. 2019. Fairness Under Unawareness. FAT*: ACM Conference on Fairness, Accountability, and Transparency. (2019), 339-348.

16) Schein, A.I. et al. 2002. Methods and metrics for cold-start recommendations. SIGIR: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. 25, (2002), 253.

17) 서봉원(2016). 콘텐츠 추천 알고리즘의 진화. 한국콘텐츠진흥원. 방송트렌드&인사이트 5권

한 방안으로 다시 활용되고 있다.

콘텐츠 기반 필터링은 내용 자체를 분석하므로 협업 필터링에서 발생하는 콜드 스타트 문제를 해결할 수 있다. 하지만 다양한 형식의 항목을 추천하기 어렵다는 단점이 있으며 무엇보다 프로파일을 구축할 때 다양한 정보들을 가공하여 활용하므로 비자각에서 오는 공정성을 만족할 수 없게 된다.

이와 같이 추천시스템은 작동방식에 따라 필연적으로 비자각에서 오는 공정성과 신규 항목에 대한 공정성(콜드스타트 문제) 중 하나를 충족할 수 없게 된다. 행동기록에만 기반하는 협업필터링은 신규 항목에 대한 공정성을 충족하지 못하며, 이를 해결하기 위해 프로파일 데이터를 구축하게 되면 직간접적으로 비자각에서 오는 공정성을 충족할 수 없게 된다. 협업 필터링과 콘텐츠 기반 필터링에서의 공정성 문제를 요약하면 <표 II -1>과 같다.

<표 II -1> 대표적인 추천시스템에서의 공정성 문제

	협업 필터링	콘텐츠 기반 필터링
비자각에서 오는 공정성 (fairness through unawareness)	O	X
신규 항목에 대한 공정성 (cold start problem)	X	O

3) 딥러닝(deep learning) 기반 추천시스템

최근에는 다양한 분야에 걸쳐 딥러닝 기술이 적용되고 있으며 추천시스템도 예외는 아니다. 하지만 추천시스템에서의 딥러닝의 적용¹⁸⁾은 그것을 활용하여 콘텐츠 기반 필터링의 성능을 끌어올리는 것¹⁹⁾, 혹은 협업 필터링의 핵심 기술인 행렬 분해를 더 효과적으로 수행하는 방법을 개발하는 것²⁰⁾에 있다. 예를 들어, 음악 스트리밍 서비스업체 스포티파이(Spotify)는 기존의 협업필터링 기반의 음원 추천시스템에 딥러닝 기술을 적용해서 인공지능을 활용한 오디오 분석을 활용해 이용자의 취향에 적절한 콘텐츠를 추천하고 있다. 이는 기존의 패러다임을 더 잘 수행하는 방법론의 문제일 뿐이며, 딥러닝이 추천시스템 분야에 새로운 패러다임을 제시하는 것은 아니다.

18) Zhang, S. et al. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. ACM Computing Surveys. 52, 1 (Feb. 2019), 1-38.

19) van den Oord, A. et al. 2013. Deep content-based music recommendation. Advances in Neural Information Processing Systems. 26, (2013), 1-9.

20) He, X. et al. 2017. Neural Collaborative Filtering. WWW: Proceedings of the International World Wide Web Conference. (Apr. 2017), 173-182.

4) 이용자 선택 기반 추천시스템

최근에는 추천시스템을 적용하는 거대 플랫폼들이 공정성 문제에 관한 책임을 요구받는 상황이다. 이에 플랫폼 사업자들은 복잡한 알고리즘이 갖는 불확실성이 없도록 사용자의 결정에 기반한 추천의 비중을 높이고 있다. 예를 들면, 국내 최대 포털 사이트인 네이버의 메인 페이지에 게시되는 뉴스는 모두 ‘이용자가 구독하고 있는 언론사’의 기사가 ‘언론사가 배열한 순서’로 채워지며, 유튜브의 동영상 추천 중 상당수는 ‘이용자가 구독하고 있는 채널’의 최신 동영상으로 채워진다. 이러한 상황에서는 추천시스템이 UI 측면의 역할만을 담당하기 때문에 직접적으로 공정성에 관한 문제를 야기하지는 않는다. 하지만 (사용자가 특정 콘텐츠 제공자를 구독하는 경로로 기능하는) 플랫폼 검색엔진에서의 검색결과 랭킹에 대한 공정성 문제²¹⁾가 논의될 수 있다.

다. 알고리즘 추천시스템의 적용분야

추천 알고리즘은 각종 의사결정 상황에서 도움을 주는 도구로서, 시장에서 강력한 경쟁적 우위를 확보할 수 있는 도구로 여러 분야에서 활용되고 있다. 2015년 Jie Lu와 동료들²²⁾은 추천시스템이 사용되거나 연구되고 있는 적용(application) 분야를 8개 영역(e-거버넌스, e-비즈니스, e-커머스/e-쇼핑, e-라이브러리, e-러닝, e-투어리즘, e-리소스 서비스, e-집단 활동)으로 분류하여 각 영역의 사례와 선행연구에 대한 탐구를 진행한 바 있다. 해당 분류에 기반하여 실제 추천시스템이 적용되거나 연구되는 사례들을 살펴보면 다음과 같다.

1) e-거버넌스

e-거버넌스는 정부가 시민 및 기업에게 정보 전달을 위해 인터넷 등 정보통신 기술을 이용하는 것을 말한다. e-거버넌스의 활성화에 따른 정보의 폭발적 증가(overload)로 시민과 기업들이 필요한 정보를 찾기 어려워지는 문제가 발생했으며, 이에 따라 추천시스템이 이러한 문제의 해결책으로서 e-거버넌스 시스템에 도입되게 되었다²³⁾. e-거버넌스에서의 추천시스템 사용은 정부-시민(G2C, Government-to-citizen) 서비스와 정부-비즈니스(G2B, Government-to-Business) 서비스로 구분할 수 있다. 먼저 정부-시민 서비스(G2C)에서 사용되는 추천시스템에 대해서는 2008년 De Meo 등²⁴⁾이 제시한 바 있으며, 사용자의 개인정보

21) Castillo, C. 2019. Fairness and Transparency in Ranking. ACM SIGIR Forum. 52, 2 (Jan. 2019), 64-71.

22) Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: a survey. Decision Support Systems, 74, 12-32.

23) Guo, X., & Lu, J. (2007). Intelligent e-government services with personalized recommendation techniques. International journal of intelligent systems, 22(5), 401-417.

와 사용 디바이스에 기반해 가장 흥미를 가질만한 서비스를 제안하는 방식이 적용되었다. Luis Terán와 Andreas Meier(2010)²⁵⁾은 퍼지 클러스터링(fuzzy clustering) 기법을 통해 e-투표 과정에서 유권자들의 선호에 가장 가까운 후보를 추천해주는 추천시스템을 제안한 바 있다. 정부-비즈니스 서비스(G2B) 분야에서는 정부가 개인 비즈니스 사업자에게 적절한 비즈니스 파트너를 제안해주기 위해 개발된 BizSeeker²⁶⁾ 시스템을 예시로 들 수 있다. BizSeeker에서는 비즈니스 영역의 이용자들은 의미적 유사성을 사용한 아이템-기반 협업 필터링 모델로 추천된 비즈니스 파트너 후보들의 목록을 제공받게 된다.

2) e-비즈니스

e-비즈니스 영역은 비즈니스-비즈니스 서비스(B2B, Business-to-Business)에 해당한다. e-비즈니스의 추천시스템은 해당 분야의 일정 수준 이상의 지식이 필요하며, 이에 따라 협업 필터링 또는 콘텐츠 기반 필터링 추천 방법에 지식 기반(Knowledge-based) 필터링²⁷⁾이 결합하는 경우가 많다. B2B 영역에서는 주로 디지털 비즈니스 생태계 안에서 사업자 간 컴퓨팅 기반 추천 중재(negotiation)를 진행하는 시스템이 제안되어왔다. 고객 관계 관리(CRM, Customer-Relationship Management) 분야에 있어서, 2013년 Zui Zhang 등²⁸⁾이 통신사의 고객 관계 관리를 지원하는 추천시스템을 제안한 바 있다. FTCP-RS(Fuzzy-based telecom product recommender)라고 하는 이 소프트웨어 시스템은 소비자에게 적절한 서비스 플랜과 패키지를 추천하고 설명을 제공한다.

3) e-커머스/e-쇼핑

e-커머스 및 쇼핑은 비즈니스-소비자 서비스(B2C, Business-to-Consumer)에 해당하며, 추천시스템의 중요성이 크게 부각 되어온 분야라고 할 수 있다. 이러한 B2C e-커머스 웹사이트에서는 판매자, 구매자 정보, 구매 행동 기록, 점수매김(Rating) 등 다양한 데이터를 통해 소비자의 미래 구매 행동을 예측한다. e-커

24) De Meo, P., Quattrone, G., & Ursino, D. (2008). A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario. *Data & Knowledge Engineering*, 67(1), 161-184.

25) Terán, L., & Meier, A. (2010, August). A fuzzy recommender system for eElections. In *International Conference on Electronic Government and the Information Systems Perspective* (pp. 62-76). Springer, Berlin, Heidelberg.

26) Lu, J., Shambour, Q., Xu, Y., Lin, Q., & Zhang, G. (2010). BizSeeker: a hybrid semantic recommendation system for personalized government-to-business e-services. *Internet Research: Electronic Networking Applications and Policy*, 20(3), 342-365.

27) 지식 기반 추천(Knowledge-based recommendation)이란 특정 분야에 대한 전문적인 지식 구조(Ontology)를 먼저 구축한 뒤 이를 활용하는 방식의 필터링을 말한다.

28) Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., & Lu, J. (2013). A hybrid fuzzy-based personalized recommender system for telecom products/services. *Information Sciences*, 235, 117-129.

머스 분야에서 추천시스템을 사용한 대표적인 사례로는 아마존과 이베이가 있는데, 이들은 추천시스템을 통해 소비자가 구매할 제품을 찾는 과정을 지원한다²⁹⁾. 아마존(Amazon.com)의 경우 전체 판매의 약 35%가 추천시스템 'A9'을 통해 이루어지고 있다고 알려져 있다. 아마존의 추천시스템은 기본적으로 아이템 기반 협업 필터링에 기반하며, 먼저 아이템-아이템 행렬을 만들고 상품들 간의 유사성을 계산한 후, 사용자의 구매 상품과 관련된 또 다른 상품을 추천해주는 방식으로 이루어진다³⁰⁾.

4) e-라이브러리

디지털 라이브러리는 디지털 상으로 존재하는 지식 또는 객체를 이용자에게 제공한다. 이때 추천시스템은 이용자가 정보와 지식 소스를 찾고 선별할 수 있도록 도와주는 역할을 한다. 스탠포드 디지털 라이브러리 프로젝트의 일부인 Fab³¹⁾의 경우, 협업 필터링과 콘텐츠 기반 추천 기술이 혼합된 추천시스템을 제공한다. 그리고 Porcel 등(2010)³²⁾은 대학의 디지털 라이브러리(UDL)에서의 정보 탐색을 지원하는 추천시스템을 제안한 바 있다. 이러한 e-라이브러리 추천시스템에서는 콘텐츠 기반 필터링, 협업 필터링, 그리고 지식 기반 필터링이 혼합된 접근이 주로 사용된다.

5) e-러닝

e-러닝에 대한 추천시스템은 전통적 e-러닝 시스템이 개발된 2000년대 초반부터 지속적으로 많은 관심을 받아왔다. 주로 이용자가 선호하는 학습 주제와 방식, 이용자의 학습 행동 등의 정보에 기반해 적절한 학습 코스를 추천해주는 서비스가 제시된다. 이러한 e-러닝 관련 추천시스템에서는 이용자 행동 정보의 수집 및 추천이 중요하게 다루어진다. Zaiane(2002)³³⁾은 상관 규칙 마이닝 등의 기술을 비롯한 데이터 마이닝 기술을 제안하여 적절한 교육 콘텐츠를 추천하는 시스템을 제안하였다. Farzan과 Brusilovsky(2006)³⁴⁾

29) Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data mining and knowledge discovery*, 5(1-2), 115-153.

30) Greg, L. (2003). Amazon. comrecommendations item-to-item collaborative filtering. *Internet Computing*.

31) Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3), 66-72.

32) Porcel, C., & Herrera-Viedma, E. (2010). Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries. *Knowledge-Based Systems*, 23(1), 32-39.

33) Zaiane, O. R. (2002, December). Building a recommender agent for e-learning systems. In *International Conference on Computers in Education, 2002. Proceedings.* (pp. 55-59). IEEE.

34) Farzan, R., & Brusilovsky, P. (2006, June). Social navigation support in a course recommendation system. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 91-100). Springer, Berlin, Heidelberg.

가 개발한 CoureAgent는 학생들의 피드백 정보를 수집하여 학생들에 대한 정보를 구축한다.

6) e-투어리즘

인터넷과 모바일 기기의 발달은 관광 산업에도 큰 변화를 가져왔다. e-투어리즘과 관련한 추천시스템은 이용자에게 적절한 코스나 명소, 식당, 교통 정보 등을 추천한다. 식당 추천시스템 REJA(REstaurants of JAén)³⁵⁾의 경우 협업 필터링과 콘텐츠 기반 필터링을 혼합한 필터링 방법의 사용이 제안되었다. Tuukka Ruotsalo 등(2013)³⁶⁾은 모바일 기반의 장소 추천시스템을 제안한 바 있다. 해당 시스템은 온톨로지(Ontology)에 기반하여 다양한 맥락 정보를 필터링하는 프레임워크가 적용되었으며, 이러한 맥락 정보에는 디바이스에서 수집하는 정보까지 포함되었다. 이러한 e-투어리즘 관련 추천시스템에 있어서는 추천하는 아이템에 대해서 요구되는 복잡성 수준에 따라 적용되는 기술이 달라지게 된다. 예를 들어 식당과 같은 간단한 아이템에 대한 추천에 있어서는 콘텐츠 기반 필터링과 협업 필터링 알고리즘이 사용되지만, 여행계획 및 스케줄과 관련한 추천에 있어서는 관련 지식구조를 동원한 혼합적 추천 기술이 필요하다.

7) e-리소스 서비스

e-리소스란 이용자에 의해 업로드되는 동영상, 음악, 문서 등의 콘텐츠를 말한다. 이러한 리소스 서비스 추천시스템은 이용자들이 정보를 찾거나 공유하는 데에 낭비되는 시간과 에너지를 방지하는 것을 목적으로 한다. e-리소스 추천에 있어서는 굉장히 다양한 영역이 개입되며, 이에 따라 적용되는 추천 알고리즘 기술 역시 다양하다. 협업 필터링에 기반해 이용자의 태깅(Tagging)을 돕는 기술(Folksomomies)³⁷⁾, 협업 필터링과 콘텐츠 기반 접근에 기반해 관심을 가질 만한 TV 프로그램을 추천해주는 프로그램 추천시스템(PRS)³⁸⁾, 관심을 가지는 웹페이지나 뉴스를 추천해주는 시스템, 영화나 동영상 등의 콘텐츠를 추천해주는 시스템 등이 제안되어왔다.

35) L. Martinez, R. M. Rodriguez and M. Espinilla, "REJA: A Georeferenced Hybrid Recommender System for Restaurants," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy, 2009, pp. 187-190, doi: 10.1109/WI-IAT.2009.259.

36) Ruotsalo, T., Haav, K., Stoyanov, A., Roche, S., Fani, E., Deliai, R., ... & Hyvönen, E. (2013). SMARTMUSEUM: A mobile recommender system for the Web of Data. *Journal of Web Semantics*, 20, 50-67.

37) Zheng, N., & Li, Q. (2011). A recommender system based on tag and time information for social tagging systems. *Expert systems with Applications*, 38(4), 4575-4587.

38) Zimmerman, J., Kauapati, K., Buczak, A. L., Schaffer, D., Gutta, S., & Martino, J. (2004). TV personalization system. In *Personalized digital television* (pp. 27-51). Springer, Dordrecht.

8) e-집단 활동

각 개인이 아니라 집단으로 행동하는 사람들에게 추천 서비스를 제공하는 경우도 있다. 이를 위한 집단 추천시스템(GRS, Group Recommender System)으로는 온라인 커뮤니티를 대상으로 한 도서 추천시스템³⁹⁾, 소규모 관심 집단에 대한 영화 추천시스템⁴⁰⁾ 등의 적용이 제안된 바가 있다. 이들은 기본적으로 개인에 대한 추천 기술을 응용하는 형태로 구성되어 앞선 시스템들과 큰 차이가 없는 것처럼 보이기 쉬우나, 개별 이용자를 군집화하는 방법에 있어 차이를 보인다.

제2절. 추천시스템의 공정성 개념과 탐지방법

1. 추천시스템 공정성의 개념

가. 공정성의 기본 개념과 접근

추천시스템은 시스템이 데이터를 통해 학습할 수 있게 하는 AI의 한 형태로서 머신러닝을 기본으로 한다. 머신러닝 공정성(ML fairness)은 시스템의 알고리즘에서 나타날 수 있는 편향성 감소를 위한 방안을 제시하는 것에서 출발한다. 여기에서의 공정성은 추천시스템에서 고려해야 할 공정성과 맥락적으로 일치한다. Alice Xiang and Inioluwa Raji(2019)⁴¹⁾은 머신러닝, 추천시스템 등 알고리즘 관련 연구에서 제기하는 공정성이 법에서 다루는 공정성과 다른 접근을 함으로서 사회적으로 받아들여지지 않을 가능성을 제시하기도 하였다. 이는 법적으로 알고리즘의 공정성이 받아들여지는 것은 차별금지법과 같은 사회적 약속이 있기 때문이지만 알고리즘 관련 공정성 연구의 결과가 차별의 현상을 논할 뿐 연구결과에서 나타난 차별이 법적으로 어떻게 조치되어야 하며, 정책입안자나 법률 관계인을 설득할 수 있는지에 대한 관심이 부족하다고 본 것이기 때문이다. 이처럼 해외의 연구는 공정성을 현행 법률 시스템과 연관시켜 공정성의 개념을 사회적 맥락과 일치시키고자 노력하고 있다.

39) Kim, J. K., Kim, H. K., Oh, H. Y., & Ryu, Y. U. (2010). A group recommendation system for online communities. *International journal of information management*, 30(3), 212-219.

40) O'Connor, M., Cosley, D., Konstan, J. A., & Riedl, J. (2001). PolyLens: a recommender system for groups of users. In *ECSCW 2001* (pp. 199-218). Springer, Dordrecht.

41) Xiang, A., & Raji, I. D. (2019). On the Legal Compatibility of Fairness Definitions. *arXiv preprint arXiv:1912.00761*.

1) 법에서의 차별에 대한 관점

공정성은 어떠한 사유로 인해 차별받지 않아야 한다는 것을 기본으로 한다. 하지만 차별은 하나의 법률에서 논하는 논리가 아니다. 공정성의 전제가 되는 ‘차별’은 구조적 오류로 볼 수 있지만 이는 법적으로 사용되는 차별의 정의에 비해 단순한 경향이 있다. 미국의 경우 1964년부터의 법은 다양한 법률에서 차별 금지를 규정하는데, 여기에는 동기, 의도, 인과관계가 포함된다.

또한 차별은 매우 다양한 차원을 가진다. Mehrabi et al.(2019)⁴²⁾은 차별의 유형이 다양하다고 제시한다. 이들에 의하면 차별의 유형은 첫째, 인종 차별과 같이 유형이 개인의 보호된 속성이 자신들에게 비호감적인 결과를 초래할 때 발생하는 직접차별, 둘째, 개인은 겉으로 보기에 중립적이고 보호되지 않는 속성에 기초하여 대우받는 것처럼 보이지만 암묵적 효과의 결과로 부당한 대우를 받게 되는, 예를 들어 거주지역으로 인해 받는 차별과 같은 간접 차별, 셋째, 개인적 선호가 반영되어 비체계적으로 이루어지는 구조적 차별, 넷째, 통계적 결과를 바탕으로 개인을 미리 예측하는 통계적 차별, 다섯째, 외생변수를 통제하지 않음으로 인해 발생하는 비체계적 차별 등이 있다. 공정성은 이러한 차별의 배제에서 출발한다.

2) 공정성에 컴퓨터공학적 접근 방식

앞서 살펴본 Mehrabi et al.(2019)의 다섯 가지 차별의 유형은 인간 혹은 사회적 시스템이 의도적으로 불공정한 결정을 하는 경우를 구분한 것이다. 알고리즘 연구에서는 이와 같은 불공정이 나타날 수 있는 이유를 크게 두 가지 관점으로 본다. 첫째는 데이터의 문제이다. 사회적 시스템에서 인간의 선택은 스스로가 학습하는 것뿐만 아니라 이미 마련되어 있는 사회적 시스템이나 이미 이루어진 선택의 결과에 영향을 받게 된다. 알고리즘의 경우 스스로 학습한다는 점에서 알고리즘 공정성은 광범위한 그룹으로부터 훈련 데이터를 수집하고, 데이터 클리닝 과정에 충분한 다양성을 유지하도록 보장되어야 한다. 다만 다양한 훈련 데이터의 사용은 알고리즘 모델이 충분한 성능을 가지고 있다는 전제하에서 가능하다. 둘째는 비지각 공정성(fairness through unawareness) 문제이다. 인간이 처리했던 결과를 바탕으로 하는 데이터에는 이미 과거에 이루어졌던 의도적 불공정이 포함되어 있어서 특정한 변수를 체계적으로 배제할 수 있다. 이 때문에 알고리즘 공정성은 보호속성(예: 성별, 나이, 장애)을 의사결정의 수단으로 사용하지 않아야 한다고 권고한다. 하지만 보호속성을 의사결정의 수단으로 사용하지 않는 것은 보호속성으로 인해 차별받은 의도적 불공정을 파악하여 보정할 수 없게 할 수 있기 때문에 공정성을 담보하는 것은 아닐 수 있다.

머신러닝 학습과정이나 추천과정에서 공정성을 측정할 수 있다는 점에서 알고리즘의 공정성 측정방법도 최근 관심을 받고 있다. 개념적으로 알고리즘의 공정성의 측정방법은 그룹 공정성과 개별 공정성으로 구분할 수 있다. 그룹 공정성은 통계적 동등성을 추구한다. 이 방법에 따르면, 보호대상 그룹에서 선택된 개인의

42) Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

비율은 전체 사회에서의 보호대상 비율과 거의 같아야 한다. 예를 들어 한 사회에서 장애인의 비율이 5%라면 면접에 선발된 20명의 지원자에 장애인이 1명 포함되어야 한다는 접근이다. 하지만 보호대상이나, 그 외에도 비율을 참고할 사항이 많기 때문에 효과적이지 못하다. 이에 제시된 대안이 개별 공정성으로서, 개별 공정성은 유사한 개인이 유사한 결과를 가져야 한다는 접근이다. 이 접근방식은 매트릭스를 사용하여 당면한 작업에 대해 유사한 두 개인이 얼마나 유사한지를 정의한다(Dwork et al., 2012)⁴³⁾. 그러나 이러한 접근법은 공정한 유사성 측정기준을 정의해야 하는 문제와 측정기준에 비공정 정보를 포함하는지 여부 및 어떻게 포함시키는지에 대한 문제도 제기된다.

3) 공정성 연구에서 고려해야 할 법률 사항

(1) 상호교차성

알고리즘 연구에서는 공정성을 단일 계층 측면에서 접근하는 경우가 많고, 다양한 속성의 상호적 속성을 다루는 경향이 적다. 이에 상호교차성이 적용될 필요가 있다. '상호교차성'이라는 용어는 1989년에 미국의 비판적 법 이론가 이론(critical legal race) 연구자 킴벌리 윌리엄즈 크렌쇼(Kimberlé Williams Crenshaw, 1989)⁴⁴⁾에 의해 고안되고 체계화 되었다. 상호교차성 이론은 한 사람의 사회적 정체성을 규정하는 범주는 단일하지 않으며 젠더, 인종, 사회 계급 등 다양한 측면이 상호교차적으로 작용한 결과라는 점을 강조한다. 또한, 한 사람에게 작용하는 억압, 지배구조, 차별을 올바르게 이해하기 위해서는 다양한 측면들 사이에서 일어나는 상호작용과 창발적 속성을 체계적으로 분석할 필요가 있다고 본다.

(2) 절차적 공정성

알고리즘 분석에서 절차적 공정성은 모델 예측 과정으로 이해되지만 이는 법률적 측면의 절차적 공정성과 다른 개념이다. 사회적으로 절차적 공정성은 공정한 결과를 목표로 하지 않고 의사결정 과정에 집중하여 이를 통해 적절한 결과를 도출하려고 한다. 따라서 시스템 분석에 있어서 알고리즘 자체보다 알고리즘을 둘러싼 시스템, 예를 들어 사회적 시스템이나 법적 규제 등이 분석되어야 한다.

4) 공정성 결여와 차별의 결과로서 편향

공정성의 적용은 차별을 배제하는 것에서 출발하며 그 결과는 편향의 제거로 완성된다. Mehrabi et al.(2019)은 편향이 다음과 같이 매우 다양하게 나타난다고 분석하였다(<표 II -2> 참조).

43) Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference 214-226.

44) Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In University of Chicago Legal Forum, 140, 139-67.

<표 II -2> 편향 개념

유형	내용
역사적 편향	여성의 사회참여 제한과 같은 역사적 편향
대표성 편향	제도나 체제에서 서양국가들의 선호가 주로 반영되는 것과 같은 편향
측정 편향	미국의 위험평가 알고리즘인 COMPAS에서 나타나듯이 측정 방법의 오류로 인한 편향
평가 편향	어플리케이션 모델링 과정에서 발생하는 편향
집단 편향	소수 관찰자를 대상으로 한 결과를 전체집단에 잘못 적용함으로 인해 나타나는 편향
모집단 편향	데이터 집합의 모집단이 전체 모집단의 분포와 다를때 나타나는 편향
심슨의 역설	각각의 변수에 신경 쓰지 않고 전체 통계 결과를 유추하면서 발생하는 편향
종단 데이터 오류	횡단데이터를 종단으로 취급하면서 나타나는 오류
표본추출 편향	하위집단의 랜덤하지 않은 표본추출로 인해 발생하는 오류
행동적 편향	상이한 문화적 코드로 인한 행동의 차이와 반응의 분석오류
콘텐츠 편향	어휘, 의미 등 콘텐츠의 구조의 기본적 차이에서 발생할 수 있는 편향
링크 편향	실제 행동을 관찰하지 않고 링크의 노드만을 분석함으로써 나타나는 편향
일시적 편향	측정 시점의 차이에 따라 나타나는 편향
인기 편향	가짜리뷰나 소셜 봇 등의 순위조작 등과 같이 데이터 조작으로 인한 편향
알고리즘 편향	입력된 데이터와 알고리즘에 의해 발생하는 편향
사용자 상호작용 편향	사용자가 의도적으로 선택한 편향된 행동과 상호작용으로 인해 촉발되는 편향
제시 편향	정보가 제시되는 방법에 의해 선택의 결과가 달라질 수 있는 오류
순위 편향	상위권의 정보가 상대적으로 과다하게 선택되는 결과

나. 공정성 확보의 전제조건

1) 분류의 공정성

이론적으로 알고리즘이 추천하는 시스템은 연령, 성별, 젠더, 인종, 계급 등의 기본적 인구사회학적 속성과 무관한 추천이나 선택이 이루어져야 한다. 이에 추천시스템은 첫째, 기본적인 인구사회학적 속성이 예측을 위한 변수로서 사용되지 않아야 하고, 둘째, 기본적인 인구사회학적 속성과 무관하게 예측의 결과가 도출되어야 하고, 셋째, 유사한 속성의 개인은 통계적, 확률적으로 유사한 결과가 나타나야하고 넷째, 기본적 인구사회학적 속성 또는 기타 유사한 속성을 가진 개인들은 유사한 결과가 나타나야 한다. 이에 기본적 인구사회학적 속성은 공정성과 관련하여 보호되어야 할 보호 속성(protected attribute)으로서 기본적 인구사회학적 속성으로 인해 부당한 대우를 받지 않고 공정성(fairness)을 담보하는 알고리즘을 설계하는 것이 머신러닝, 딥러닝, 추천시스템 등 알고리즘 기반의 시스템에서 전제되어야 한다.

2) 학습 데이터

인간의 행동 데이터를 기반으로 하는 추천시스템에서 입력 데이터는 공정성 학습의 기초이다. 하지만 축적된 인간의 편향이 입력 데이터로서 활용됨으로 인해 문제가 나타날 수 있다. 사회적으로 우월한 집단에

속하는 집단의 특성에 반감이 있거나, 열등감을 느끼는 등의 집단적 차별(group discrimination)이 학습 데이터에 포함될 경우 알고리즘은 이를 정상적인 판단과정으로 이해할 수 있기 때문이다. 예를 들어 2016년 MS의 채팅봇 TAY는 공개되자마자 인종/종교 차별적 답변을 쏟아 내었는데, 이는 TAY의 학습과정에서 악의적인 백인 우월주의자, 여성 혐오자, 이슬람 혐오자들이 반사회적 문장을 지속적으로 주입함으로써 나타난 결과로 추정된다.

이에 알고리즘의 결과 자체를 그대로 믿는 것이 아니라 결과를 검토하는 과정이 필요하다. 가장 기본적인 방법은 인구사회학적 속성 집단의 대표성을 충분히 반영하고, 인구사회학적 속성 집단의 대표성을 적절하게 설정하여 편향성을 제거하며 유사한 항목은 대표적 속성으로 처리해야 한다.

2. 추천시스템 공정성 기존 연구

공정성은 추천을 받는 소비자, 추천받을 아이템을 제공하는 생산자의 관점에서 논의가 될 수 있다. 추천 받는 소비자 입장에서는 추천시스템의 추천이 소비자에게 적합한지(for Subjects)가 중요하고 아이템 제공자 입장에서는 제공되는 서비스나 상품이 적절하게 추천되는지가 중요하다.

가. 주체 관점의 공정성 연구경향

주체 관점의 공정성(Fairness for subjects) 연구는 크게 추천이 분류의 공정성을 만족하는지와 연관된다. Ning & Karypis(2011)는 2010년 이전 추천시스템 알고리즘의 주류방식인 매트릭스 요인화(Matrix Factorisation)의 개선을 위해 이용자의 구매 이력과 평가 이력을 합산해 Top-N 추천 리스트를 구성하는 Sparse Linear Method (이하 SLIM)이라는 알고리즘을 제안하였다. Yao & Huang(2017)은 SLIM 방식을 적용한 STEM(Science, Technology, Engineering, Math) 교육코스 추천시스템의 편향 검증 연구에서 평가(rating) 데이터가 이용 주체의 선호를 정확하게 반영하고 있음에도 불구하고 교육코스 편향이 발생하고 있음을 입증하였다. Burke et al., (2018)는 분류의 공정성에 따른 편향을 보정하는 방법으로 Balanced neighbourhood 방식을 제시하였다. 이들은 영화 장르를 추천해주는 Movie Lens를 이용하여 Ning & Karypis (2011)에서 제시한 추천시스템 알고리즘인 SLIM과 Balanced Neighbourhood 방식의 수정 과정을 거친 BN-SLIM-U를 적용한 결과 BN-SLIM-U의 공정성 점수가 높게 나타났다고 보고하였다.

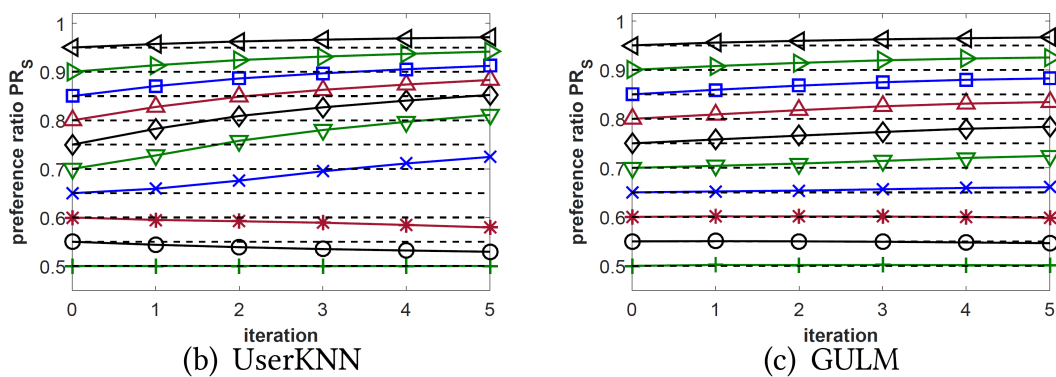
그러나 주체 관점의 공정성 연구들은 공정성의 문제를 입증하고 해결방식을 제시하지만 이러한 접근이 정확한 추천이라는 목적 달성을 담보하는 것은 아니라는 한계를 스스로 인정하고 있다. 이에 연구자들은 추천시스템의 주체(For Subjects) 관점의 연구에서 대상(For Objects)에 대한 연구도 고려되어야 한다고 주장하였다 (Burke et al., 2018; Yao & Huang, 2017).

나. 대상 관점의 공정성 연구경향

공정성 대상(Fairness for objects) 관점의 시각은 제공되는 서비스나 상품이 적절하게 추천되는지와 연관된다. 이는 추천시스템이 기본적인 문제점을 가지고 있기 때문이다. 예를 들어 전자상거래, 엔터테인먼트, 소셜미디어, 뉴스포털, 취업 등 광범위한 영역에서 적용되고 있는 추천시스템에서 기본적으로 정확성(accuracy)에 기반한 추천시스템은 사용자들이 관심이 적은 분야는 적게 표현되는 불균형 추천을 하면서 사용자의 관심영역이 점차 좁아지고 에코챔버나 필터버블과 같은 효과가 발생할 수 있다.

Steck (2018)는 사용자의 영화 선택 시나리오에 따라 Movie Lens의 추천시스템에 차이가 나타나는지 분석하였다. 분석결과 추천시스템이 트레이닝 데이터가 제한적이거나 노이즈가 많은 경우 불균형적인 추천을 발생시킬 수 있기 때문에 최적화된 측정(calibration)의 중요성을 제언하고 측정방안으로 다양성(diversity)을 반영하는 greedy algorithm (Ashkan et al, 2015)을 제언한 바 있다.

Tsintzou et al., (2020)은 제공되는 서비스나 상품의 추천에서도 분류의 공정성에 따른 추천의 편향이 발생한다고 보고하였다. 이들은 Movie Lens의 추천시스템 분석에서 성별에 따른 추천의 차이가 나타난다는 것을 밝혔는데, 유사한 속성을 가진 사용자와 비교하는 알고리즘인 UserKNN(User based K-Nearest-Neighbors) 대신 이를 보정하는 GULM(Group Utility Lose Minimization) 알고리즘의 필요성을 주장하였다.



[그림 II -1] UserKNN과 GULM

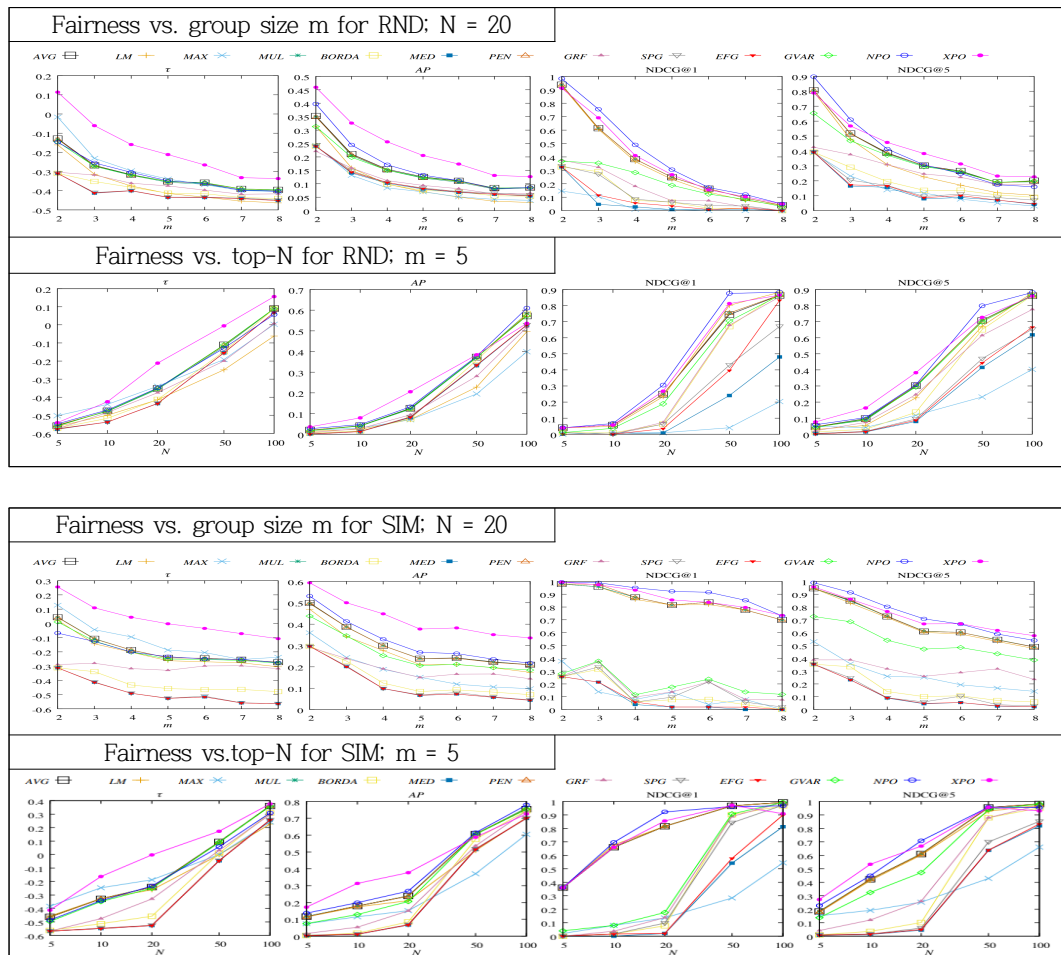
다. 집단 관점의 공정성 연구경향

집단 관점의 시각은 주체 관점 연구의 확장으로서 추천시스템이 집단구성원들의 개인별 선호를 시스템이 잘 반영하고 있는지에 관심을 가진다. Serbos 등(2017)은 음식, 쇼핑, 뷰티/스파, 헬스/의약품, 유흥을 추천하는 Yelp Challenge를 대상으로 하고 서비스의 연관성을 바탕으로 한 관계를 공정성 보정의 방법으로

사용한 알고리즘을 개발하여 적용한 결과 average rating 또는 least misery와 같은 접근법에 비해 공정성 측면에서 우수하다고 보고하였다. 이 연구에서 이용자 만족의 주요 요인은 추천된 아이템들 중 선호하는 아이템들로 인한 만족도이지만, 선호하지 않는 아이템들로 인한 불만족도 높은 영향을 미치는 것으로 나타났다.

Xiao 등(2017)은 Movie lens와 Movie Pilot의 데이터셋을 사용하고, 콘텐츠 간의 관련성을 중심으로 한 보정, 추천의 불균형 보정, 개인의 만족도를 결합하여 추천시스템의 공정성으로 제시함으로써 그룹 추천시스템에서 공정성을 전제하면서도 우수한 성능을 낼 수 있는 알고리즘 제시한 바 있다.

Sacharidis (2019)는 그룹 내 구성원들이 얻게 되는 ‘최소 효용’개념을 도입한 보정 알고리즘을 통해 추천시스템에서의 파레토 최적화 방법을 제안하였다. 이 연구에서는 공정성과 유용성을 기반으로 연구 모델과 알고리즘을 구성하고 Movie Lens 이용자를 대상으로 연구한 결과 파레토 최적화 방법을 적용한 top-N 추천시스템은 다른 알고리즘과 비교했을 때 효과적이라고 보고하였다.



3. 공정성 탐지 노력의 예: 순위의 공정성(fair ranking)

1) 순위 공정성의 특징

카드 회사의 신용카드 추천, 쇼핑 앱이나 배달 앱의 상품 및 서비스 추천 등은 어느 하나의 상품 및 서비스가 아닌 다수의 상품 및 서비스를 순차적 혹은 비순차적으로 제시한다. 이때 추천되는 대상은 너무나 많은 추천대상의 지극히 일부분이다. 조직 내에서의 서비스 제공이나 업무처리 등을 예로 들면 단순한 측정 기준으로 순위를 정하는 것은 공정성을 저해할 수 있다. 이에 순위공정성은 추천시스템에서 강조되는 영역이다. 순위공정성 확보를 위해서는 다음의 세 가지 특성이 요구된다. 첫째, 순위를 정할 수 있는 대상이 충분히 확보되어 사용자나 조직 구성원들에게 배분적/할당적 손해가 가는 것을 방지해야 한다. 둘째, 유사 아이템에 대한 일관된 처리로 개인적 공정성(individual fairness)을 확보해야 한다. 셋째, 보호되어야 하는 집단에 속한 아이템의 적절한 대표성으로 해당 집단 구성원들에게 대의적 손해(representational harm)가 가는 것을 방지해야 한다.

2) 공급자 측면과 사용자 측면에서의 추천시스템 공정성 이슈

앞에서는 추천시스템에서의 가장 대표적인 공정성의 개념인 ‘비자각에서 오는 공정성(fairness through unawareness)’과 ‘신규 항목에 대한 공정성’을 언급하였다. 그 외에도 다양한 공정성의 개념이 제안되어 있으며, 이는 크게 공급자 측면(content provider, object)에서의 공정성과 사용자 측면(user, subject)에서의 공정성으로 구분된다.

공급자 측면에서의 공정성은 아이템 자체 혹은 아이템의 제공자가 추천시스템에 의하여 차별 없이 대우되는지에 대한 문제이다. 공급자 측면에서의 공정성의 핵심은 아이템의 ‘품질’에 합당하게 추천하는 것이지만, 품질에 대한 정의의 불분명함, 품질 추론 기법에 있어서의 또 다른 공정성 문제가 얹혀있어 이에 대한 선행연구는 매우 제한적이다. 대신에 학계는 사용자별 입력 데이터에 대한 아이템 속성 비율의 편차를 줄이는 방식으로 공급자 측면에서의 공정성을 논의하고 있다.

사용자 측면에서의 공정성은 모든 사용자들이 속성에 관계없이 동등한 수준의 추천을 받는가에 대한 문제이다. 전산학적으로는 개별 사용자 혹은 사용자 그룹(예를 들면 남/녀)에 상관없이 아이템에 대한 선호도의 예측 오차의 수준을 균일하게 맞추는 것으로 사용자 측면의 공정성을 추구한다.

3) 순위 공정성 측정방법

공정성 측정의 대표적 방법은 관심 기반(attention-based) 측정방법과 확률 기반(probability-based) 측정방법이 있다. 관심 기반 측정방법은 상이한 집단에서 노출된 상태와 노출로 인한 효용성의 함수가 동일

해야 한다는 접근이다. 예를 들어 상이한 포털에서 노출된 상품의 만족도는 노출된 빈도에 비례해야 할 것이다. 또한 동일한 포털에서 상이한 집단에 노출되더라도 만족도는 유사해야 할 것이다. 즉, 주어진 랭킹 P에서 상이한 집단(G) A, B에 노출되고 그로 인한 효용성 평균이 U라고 하면

$$\frac{Exposure(G_A|P)}{U(G_A)} = \frac{Exposure(G_B|P)}{U(G_B)}$$

를 만족해야 한다.

확률 기반(probability-based) 측정은 랭킹이 확률적 과정을 통해 생성된다고 가정하며 랭킹의 기대된 특징과 그 관측의 편차를 측정하는 방법을 사용한다.

4) 공정한 순위 생성방법

순위 공정성 확보의 방법은 편향이나 문제점을 제거하는 시점에 따라 구분된다. 머신러닝 기반의 추천시스템은 데이터가 치우치지 않아야 하는 것을 전제로 하는데, 기본적인 머신러닝의 문제점으로서 초기 데이터 부족으로 Cold Start까지 감안하면 편향이나 문제점 제거는 훈련 데이터 전처리 단계에서 공정성을 확보해 편향에 의한 잠재적 영향을 줄이는 전처리(Pre-processing) 방법과 후처리가 보완적으로 적용되어야 한다. 후처리(Post-processing) 방법은 특정한 제한(constraint)에 따라 측정 대상을 재구성하는 방법으로 공정한 순위 시스템 (fair ranking system)의 주류방식이다. 도중 처리(In-processing) 방법은 동시에 발생할 수 있는 조건이나 현상을 통제하는 방법이다.

5) 미국 NIST (National Institute of Standards and Technology) Fair Ranking Track, 순위공정성 탐지노력

NIST (National Institute of Standards and Technology)는 최근 데이터 기반 사회의 만연한 문제인 알고리즘의 공정성 부족을 해결하기 위해 페어 랭킹 트랙 (Fair Ranking Track)⁴⁵⁾을 운영했다. NIST는 추천 알고리즘이 데이터가 어떤 방식으로든 편향된 경우 해당 편향을 반영하는 결정을 내린다는 점에 우려를 표하고, 그 사례로 스트리밍 서비스가 제안하는 음악 아티스트와 취업 면접 추천 여부에 영향을 미치는 실제 결과를 제시했다.

45) <https://fair-trec.github.io/>



TREC 2020 Fair Ranking Track

The TREC Fair Ranking track evaluates systems according to how well they *fairly* rank documents. The 2020 focuses on scholarly search and fairly ranking academic abstracts and papers from authors belonging to different groups.

You can find all the information about the 2019 TREC Fair Ranking track [here](#).

TIMELINE

- **May 19, 2020:** guidelines released.
- **July 1, 2020:** training queries and corpus released
- **August 17, 2020:** evaluation queries released
- **August 28, 2020:** submissions due
- **late September, 2020:** evaluated submissions returned

DOWNLOADS

- [TREC 2020 Fair Ranking Track Participant Instructions](#): guidelines, experimentation protocol, and data format description (updated May 19, 2020).
- [TREC 2020 Fair Ranking Track training data](#): training queries, and a subset of OpenCorpus containing a superset of the training query candidate documents (retrieve from this pool). **Note: by downloading the data sample contained in this package, you agree to the [Semantic Scholar Dataset License Agreement](#).**
- [TREC 2020 Fair Ranking Track evaluation data](#): eval queries are in the TREC-Fair-Ranking-eval-sample-no-rel.json file, while the eval sequences are in TREC-Fair-Ranking-eval-seq.csv. For the retrieval task, please retrieve the results from the subset of OpenCorpus provided with the training data. **Note: by downloading the data samples contained in this package, you agree to the [Semantic Scholar Dataset License Agreement](#).**
- [validate-run-rerank.py](#): rerank task submission validation script and example [submission](#). (updated August 19, 2020).
- [validate-run-retrieve.py](#): retrieval task submission validation script and example [submission](#). (updated August 19, 2020).

RESOURCES

- [Google Group](#)
- [Twitter](#)
- [Slack channel](#)
- [2019 TREC Fair Ranking \(archive\)](#)

CITATION

General references to the TREC Fair Ranking track should use the following citation,

[그림 II -2] TREC 2020 Fair Ranking Track

추천시스템이 편견을 학습하지 않도록 하기 위한 노력의 일환으로 2020년 미국 메릴랜드 주 NIST의 게이 더스 버그 캠퍼스에서 열리는 장기 실행 텍스트 검색 컨퍼런스(TREC⁴⁶⁾)의 일환으로 올해 페이 랭킹 트랙 트랙을 시작했다. Microsoft, Boise State University 및 NIST의 연구자들이 제안하고 조직 한 이 트랙 (기본적으로 새로운 연구 분야의 인큐베이터)은 공정성에 대한 연구를 조정하는 것을 목표로 한다. 데이터 및 검색 기술의 편향 정도를 측정하는 적절한 방법을 찾아 이를 제거하기 위한 전략을 얻기 위한 방법이다.

이를 위해 특정 검색 도구에서 사용하는 특정 데이터 세트를 선택 했다. 비영리 단체인 Allen Institute for Artificial Intelligence⁴⁷⁾에서 개발한 Semantic Scholar⁴⁸⁾ 검색엔진 은 학자들이 해당 분야와 관련된

46) <https://trec.nist.gov/>

논문을 검색하는 데 도움이 되는데, 이 검색엔진의 편향성을 측정하는 실험을 진행했다. Allen Institute는 TREC에 Semantic Scholar에 대한 쿼리의 400GB 데이터베이스, 반환된 답변 목록, 검색자에 대한 각 답변의 관련성 측정값으로 각 답변에 대한 클릭 수를 제공했다.

또한 상용 검색엔진 결과와 학술 연구논문 검색 모두에서 자주 발생하는 문제에 집중하기로 결정하고, 특정 검색어를 실행할 때마다 목록 상단에 동일한 답변의 편향을 분석했다. 학술 검색의 한 가지 문제점은 상위 기관에서 가장 잘 알려진 연구자가 작성한 논문 목록을 자주 반환한다는 점이다.

연구 트랙의 측정에 따르면, 공정한 알고리즘은 질의에 대한 응답으로 공정한 노출 점유율을 중요하게 고려한다. 이는 더 눈에 띄는 기사가 여전히 더 자주 표시되고 덜 눈에 띄는 기사는 덜 표시 될 수 있음을 의미한다. 물론 이것이 공정성을 정의하는 유일한 방법은 아니지만, 공정성 문제를 정량화하는 것이 적절한 첫 단계로 평가하고 있다.

4. 추천시스템 공정성 탐지방법

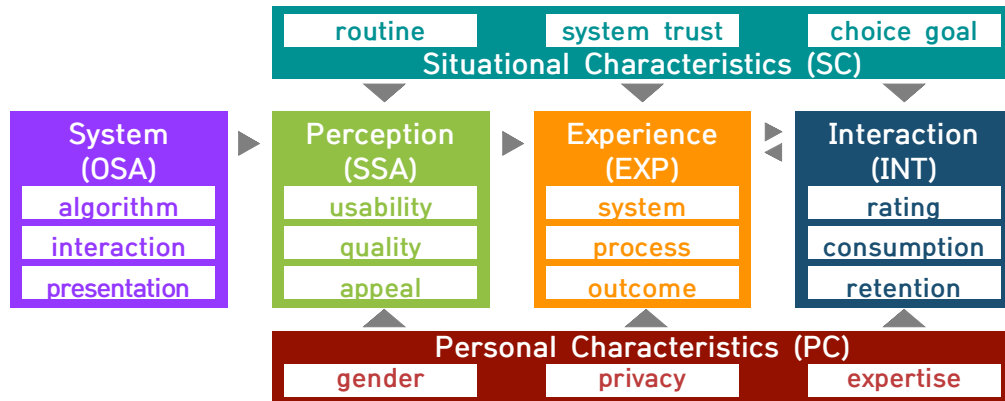
가. 전통적 분석 방법 검토

시스템 중심 접근(System-centric approaches)과 이용자 중심 접근(User-centric approaches)은 추천시스템 분석에서 가장 선호된 전통적 분석방법이다. 시스템 중심 접근은 미리 수집한 이용자들의 취향 데이터셋(preferences dataset)과 추천시스템이 예측한 이용자들의 예측 취향 데이터셋이 얼마나 정확히 일치하는가 여부에 따라 머신러닝(machine learning) 기반의 추천시스템 알고리즘의 성능을 측정하는 공학적 방법을 이용한다. 반면 이용자 중심 접근 방법은 전통적인 형태의 실험이나 서베이를 통해 사용자 만족을 측정하는 방법을 이용한다.

시스템 중심 접근은 과거의 선택·선호를 담고 있는 데이터를 바탕으로 작동하기 때문에 플랫폼을 처음 이용하는 이용자에서 나타나는 상태인 콜드스타트, 기존의 선택이 과하게 선택되는 문제점이 보고되고 있다. 특히 객관적 오차 지표가 실제 추천 서비스 이용자들의 이용 경험 만족도를 충실히 반영하지 못한다는 단점이 거론된다. 이용자 중심 접근은 측정방법이나 변수 설정 등에 따라 일관된 결과를 보여주지 못하는 단점이 있다. 이에 닌넨버그와 윌렘센(Knijnenburg & Willemsen, 2015)은 추천 서비스 이용자 연구의 이론적 방법론적 공백을 메우기 위하여 추천시스템의 이용자 중심 평가 모델(User-Centric Evaluation Framework: UCEF)을 제안하였다.

47) <https://allenai.org/>

48) <https://www.semanticscholar.org/>

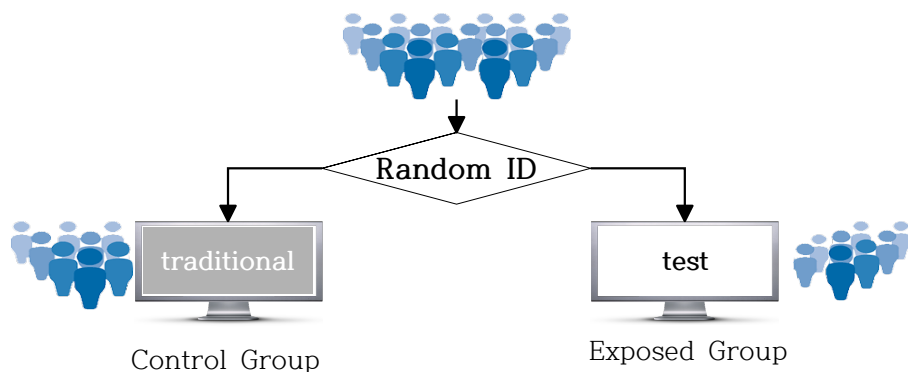


[그림 II -3] 난빈바스와 윌람센(2015)의 사용자 중심 평가 모델(User-Centric Evaluation Framework: UCEF)

UCEF는 추천시스템 연구자들이 통일적인 방법론적인 틀에서 연구를 진행하게 하여 연구 결과에 대한 교차 비교가 가능하도록 하고 사용자 연구의 신뢰성과 일관성을 구축했지만 사용자 중심 연구의 전통인 실험실 연구(lab experiments)와 서베이 방법의 틀 안에서 제시되고 있다는 한계가 있다.

나. 추천시스템의 사회적 영향 연구방법 검토

기존 분석방법의 문제점 대안으로 제시된 연구방법 중 대규모 온라인 통제실험방법은 기본적으로 기존의 UI·UX 연구에서 사용되어왔던 전통적인 A/B 테스트를 온라인에서 대규모로 수행하는 방식이다 (Kohavi et al., 2009). 이러한 방식은 관심 분야에 대한 전문적인 지식과 실험설계에 관한 기본 지식만 가지고 있으면 얼마든지 대규모 온라인 통제실험을 수행하는 것이 가능하다. 대규모 온라인 통제실험방법은 온라인 방문고객을 무작위로 A, B 두 그룹에 배치, 이때 A 그룹은 통제그룹으로서 기존의 방식에 노출되고, B 그룹은 대조그룹(실험그룹)으로서 새로운 방식에 노출된다.



[그림 II -4] 대규모 온라인 통제실험방법

이러한 방법은 기존의 추천 서비스 이용자 실험 연구가 가지고 있던 방법론적 단점인 소규모 샘플, 제한된 변수로 인해 각 연구 결과와 방법론상의 일관성 부재(온라인 실험실 연구의 경우), 그리고 통제의 결여로 인한 변수들 간의 인과관계의 내적 타당 저하(온라인 필드 실험의 경우) 문제를 동시에 극복할 수 있는 방법이다.

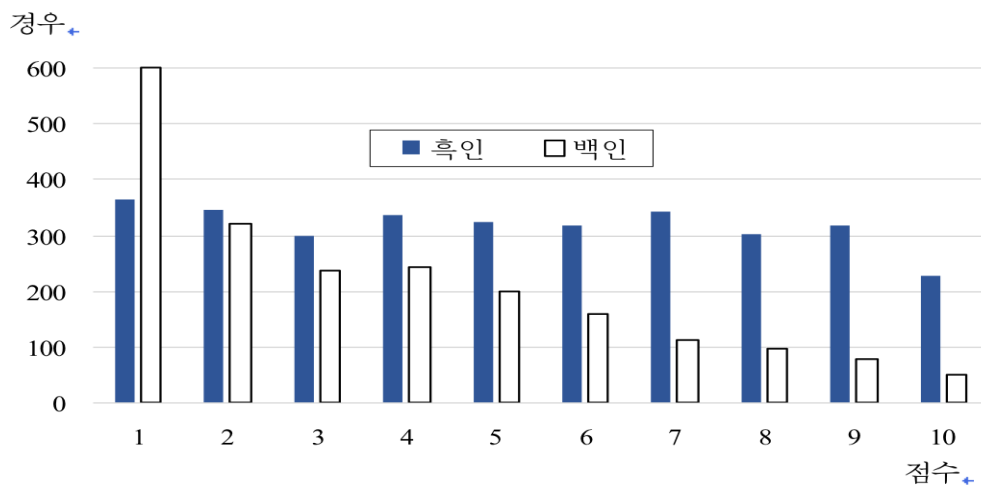
루기에리, 페드레티, 튜리니(2010)는 온라인 통제 실험 방법을 통해 데이터 마이닝 기술에서 알고리즘의 차별을 발견하는 기법을 제시하였다. 예를 들어 전미 고용기회 평등 위원회(US Equal Employment Opportunity Commission: EEDC)의 차별적 효과 여부 유무 원칙은 보호 속성(성별, 인종, 등: X)을, 양가성(binary)의 기준변인(예측되는 변인, 예를 들어 고용여부: C)을 사용하여 다른 변인들을 통제된 상태에서 보호 속성을 포함한 집단의 C에 대한 비율이 0.8 이하인 경우 해당 알고리즘에서 차별적 효과가 발생했다고 판정하게 된다.

$$\frac{PR(C=YES|X=0)}{PR(C=YES|X=1)} \leq \tau = 0.8$$

다. 추천시스템의 분석 결과에 대한 논쟁과 해석의 어려움

알고리즘의 편향과 관련하여 가장 많은 논란과 논쟁을 일으킨 사건은 약 20개 주의 미국 법원에서 사용하는 Northpointe라는 사설회사에서 개발한 위험 평가(risk assessment) 알고리즘 컴파스(COMPAS)이다.

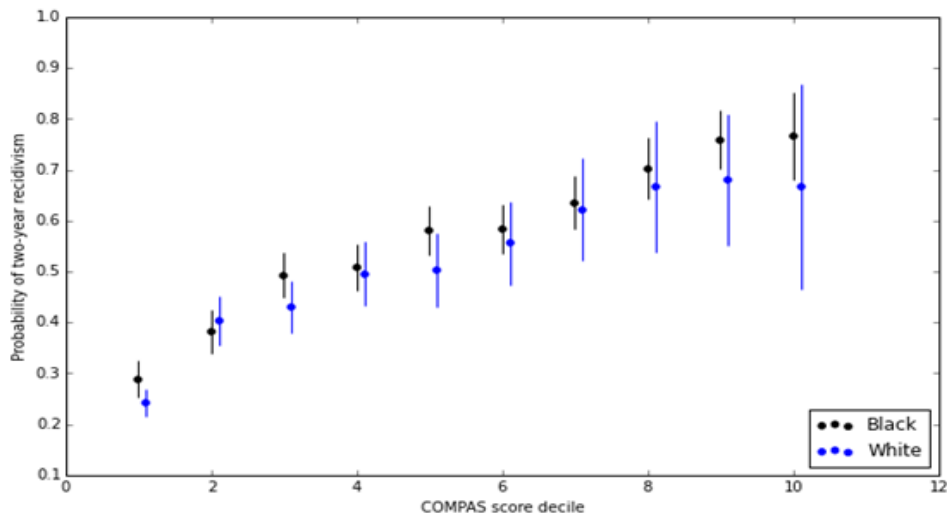
프로퍼블리카(2016)는 미래의 범죄자로 낙인찍힐 위험은 흑인 피고인이 백인 피고인에 비해 두 배에 달하고 백인 피고인들은 흑인 피고인들에 비해 종종 낮은 위험을 가진 것으로 잘못 판정되고 있다고 지적한 바 있는데([그림 II -5] 참조)⁴⁹⁾, 만일 COMPAS 알고리즘이 인종 편향을 가졌다면, 사실은 범죄 위험이 큰(high risk) 백인들이 위험이 낮은(low risk)것으로 판정되어 가석방 받을 확률이 흑인들보다 훨씬 클 것이고 따라서 실제 재범확률(recidivism probability)은 백인이 흑인보다 훨씬 높을 것으로 예상된다.



[그림 II -5] COMPASS에 의한 백인과 흑인의 위험평가 점수분포

49) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

그러나 아베 공(Abe Gong, 2016)의 분석에 따르면, 실제 위험평가에 따른 재범비율은 백인과 흑인 사이에 차이가 거의 나타나지 않았다([그림 II -6] 참조)⁵⁰⁾.



[그림 II -6] COMPASS의 스코어에 따른 인종별 재범확률

이러한 모순은 알고리즘 그 자체의 편향에서 기인하기보다는 기본적으로 흑인들의 재범률(baseline recidivism rate)이 백인보다 높기 때문에 발생하는 것이다.

따라서 알고리즘이 기계적인 공정에 그치지 않고 차별적 효과(disparate impact)까지 고려한다면 제 1종 오류(false positive) 비율까지 고려해야 한다. 높은 위험자로 판정된 사람 중 재범을 저지르지 않은 자를 고려하여 비율상의 인종 간 차이가 없도록 해야 한다. 그러나 코버트 데이비스 등(Corbett-Davies et al., 2017)의 연구에 따르면 알고리즘이 차별적 효과까지 고려한다면 흑인의 경우 백인의 경우보다 가석방의 기준을 더 낮추는 수밖에 없다. 결국 새로운 차별을 도입하는 것인 동시에 차별적 효과까지 고려한다면 알고리즘의 정확성과 효율성은 어느 정도 희생할 수밖에 없다.

공정성 확보를 위한 추천시스템의 정확성에 대한 희생을 우려하는 의견과 달리, 추천시스템의 알고리즘이 ‘공정성’을 최우선으로 고려해야 한다는 견해도 존재한다. Forbes는 자사의 아티클 ‘What Does Fairness in AI Mean?’을 통해 인공지능 설계에 있어 반드시 지양해야 할 점으로 ‘불공정한 편향의 형성 (creation of unfair bias)’을 강조한 바 있다.

‘공정성의 추구’가 아닌 ‘불공정한 편향이라는 지양할 대상’을 제시한 것은, 해당 아티클이 프린스턴 대학의 Arvind Narayanan 조교수가 제시한 ‘⁵¹⁾공정성에 대한 21가지 정의’를 언급하며 공정성이 하나의 절대적인 개념이 아닌 상황의존적인 개념임을 함께 강조하고 있기 때문이다. 공정성에 대한 이 같은 접근 방식은 과거 국내 포털사이트 네이버의 뉴스 서비스 알고리즘의 공정성이 ‘공정성에 대한 정의 부재’로 인한 비판을 받은 것과 상반된다⁵²⁾.

50) <https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84>

51) 2018년도 프린스턴 대학 강의 (링크: <https://www.youtube.com/watch?v=jIXluYdnyyk>)

따라서 Forbes가 정의 내린 추천시스템 속 알고리즘의 공정성은 각 알고리즘이 적용되는 고유의 상황 속에서 ‘불공정한 편향을 발생시키지 않는 것’이다. 더 나아가 Forbes는 인공지능을 설계하고 완성해나가는 과정 속 편향이 발생할 가능성이 편재함에 따라 다음과 같이 세 가지 편향 발생 경로에 대한 재고를 통해 미시적 관점에서 알고리즘 내 편향 발생을 억제할 필요가 있음을 제시하고 있다.

첫째로, 표본 대표성이 결여된 데이터를 통한 편향 발생 가능성이 있다. 알고리즘의 학습에 사용된 데이터가 고려 대상이 되는 모든 집단을 고루 내포하지 못할 경우 알고리즘의 학습 결과가 편향될 수 있다는 것이다. 이에 대한 사례로 ⁵³⁾Buolamwini & Gebru (2018)의 상업용 성별 판별 알고리즘 속 특정 성별 및 피부색에 대한 차별 실험 연구를 들고 있다.

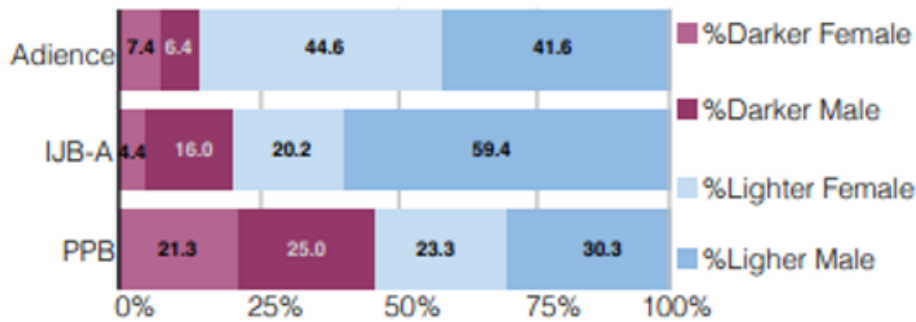
<표 II -3>각 성별 구분 알고리즘의 피부색과 성별에 따른 인식률(eg., DF=Darker Female)

Classifier	Metric	DF	DM	LF	LM
MSFT	PPV(%)	76.2	100	100	100
	Error Rate(%)	23.8	0.0	0.0	0.0
	TPR(%)	100	84.2	100	100
	FPR(%)	15.8	0.0	0.0	0.0
Face++	PPV(%)	64.0	99.5	100	100
	Error Rate(%)	36.0	0.5	0.0	0.0
	TPR(%)	99.0	77.8	100	96.9
	FPR(%)	22.2	1.03	3.08	0.0
IBM	PPV(%)	66.9	94.3	100	98.4
	Error Rate(%)	33.1	5.7	0.0	1.6
	TPR(%)	90.4	78.0	96.4	100
	FPR(%)	22.0	9.7	0.0	3.6

해당 연구에서는 MSFT와 Face++, IBM 등 세 가지 상업용 성별 분류 알고리즘의 피부색과 성별에 따른 인식률을 측정하였다. 그 결과 대체적으로 밝은 피부의 남성 (LM, Lighter Male)일수록 높은 인식률을 보이는 반면, 어두운 피부의 여성 (DM, Darker Female)일수록 낮은 인식률을 보이는 것을 알 수 있었다. 이에 대한 원인으로 알고리즘의 인풋 데이터로 사용된 인물 데이터셋 (i.e. Adience, IJB-A, PPB)에 특정 피부색을 지닌 성별에 대한 표본 수가 부족한 것을 지적한 바 있다.

52) http://news.khan.co.kr/kh_news/khan_art_view.html?art_id=201811291703001

53) Buolamwini, J. & Gebru, T.. (2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91 (Access at: <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>)



[그림 II -7] 인물 데이터세트별 표본의 성별 및 피부색

둘째, 사회적 편견에 오염된 데이터를 통한 학습으로 알고리즘에 편향이 발생할 수 있다. 이는 학습 데이터의 출처에 대한 문제로, 언어 데이터와 같이 학습 데이터를 형성한 주체가 인간일 경우, 그 데이터 속에는 인간사회의 관행과 해당 언어권의 문화가 뿌리 깊게 반영되어 있다는 것이다. 예로, ‘아빠는 간호사다’보다 ‘엄마는 간호사다’라는 텍스트 데이터의 등장 빈도가 더 높을 경우 해당 데이터는 필연적으로 성 편향을 지니게 된다.

셋째, 머신러닝 모델링 과정에서 잘못된 목적 함수 (Objective Function), 즉 알고리즘의 우선순위 설정 과정에서 편향이 발생할 수 있다. 이에 대한 사례로 ⁵⁴⁾Obermeyer et. al.(2019)의 헬스케어 분야의 알고리즘 속 편향 도출 연구를 들고 있다. 해당 연구는 의료 서비스를 제공하는 데 있어 우선순위를 설정하는 기준을 환자의 ‘아픈 정도’로 설정하더라도, 환자의 아픔에 대한 알고리즘의 판단 척도를 어떻게 설정하느냐에 따라 예기치 못한 편향이 발생할 수 있음을 시사하고 있다. 예로, 아픔의 정도에 대한 판단 척도로 ‘의료비 지출액’을 설정하였을 때 예상과는 달리 의료보험 등에 대한 지출이 상대적으로 낮은 흑인 인구가 덜 아픈 환자로 분류되어 의료 서비스 제공의 우선순위에서 제외되는 결과를 갖는 등 편향 발생의 소지를 보인바 있다.

이와 같이 추천시스템 알고리즘의 공정성을 강조하는 견해는 공정성이라는 개념이 지니는 모호함과 상황특수성으로 인해 각 알고리즘의 고유한 목적에 따른 공정함의 틀을 명확히 하고, 이를 해결하기 위한 미시적인 관점에서의 노력이 필요함을 제시하고 있다.

54) Obermeyer, Ziad & Powers, Brian & Vogeli, Christine & Mullainathan, Sendhil. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. 366. 447-453.
10.1126/science.aax2342. (Access at:
<https://www.ehdc.org/sites/default/files/resources/files/Dissecting%20racial%20bias%20in%20an%20algorithm%20used%20to%20manage%20the%20health%20of%20populations.pdf>)

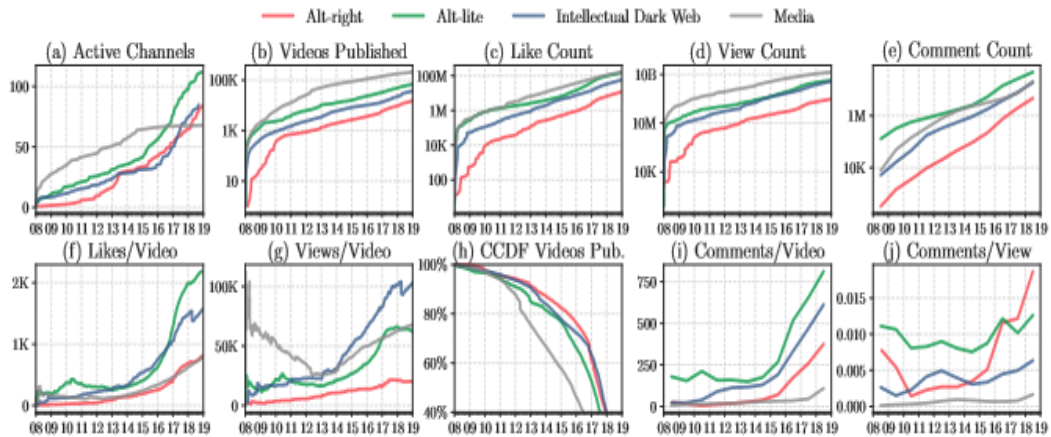
5. 추천시스템 공정성 평가에 대한 기존 접근

가. 온라인 동영상 공유 플랫폼(유튜브 등)에 제기된 추천시스템의 부작용

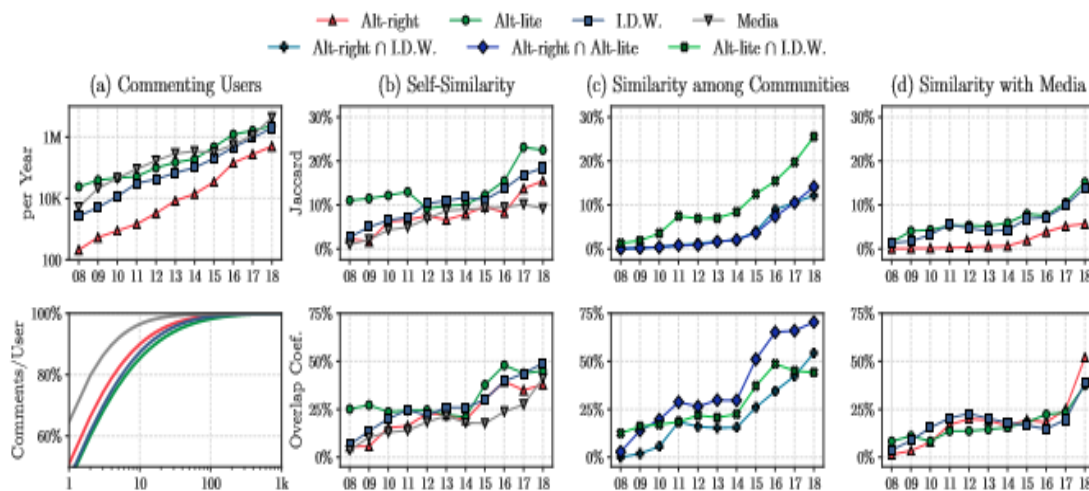
유튜브는 팔로우를 중심으로 뉴스나 콘텐츠를 소비하는 다른 소셜 미디어 서비스와는 달리, 직접 검색 또는 관심사를 기반으로 한 유튜브의 추천 알고리즘으로 인한 소비가 큰 편으로(Nielsen, 2019), 유튜브 이용자의 시청 시간 중 70%가 추천된 영상으로부터 나온다(마정미, 2020). 하지만 추천 알고리즘의 추천 정보만 접한 이용자는, 관심이 없거나 본인의 의견과 다른 콘텐츠를 보지 못한 채 ‘필터 버블(Filter bubble)’ 속에 갇힐 수 있다.

우리사회에서 가장 흔하게 볼 수 있는 추천시스템의 부작용은 정치적으로 편향된 콘텐츠의 추천이다. 추천받은 사람은 편향된 추천을 바탕으로 자신의 편향에 무감각해질 수 있다. 만약 추천받은 시스템의 콘텐츠와 일반적인 언론의 콘텐츠에서 차이가 나타날 경우 언론의 편향을 의심하게 된다. 이는 큰 문제점을 발생시키는데, 미디어 체계의 구조와 작동방식은 해당 사회의 정치 체계 구조와 유기적으로 연관되어 있기 때문에(Hallin & Mancini, 2004), 언론이 편향돼 있다고 지각할 때 언론 신뢰뿐만 아니라 민주주의에 대한 신뢰도 손상될 수 있다(Tsfati & Cohen, 2005). 따라서 추천된 영상 위주의 유튜브 뉴스 시청 등에서 편향된 정보를 지속적으로 접하게 된다면 사회갈등의 요인으로 작용하거나 사회 불신의 요인이 될 수 있다. 이에 Roose (2019)는 유튜브 동영상 추천 알고리즘의 편향성(bias)과 극화(polarization)에 대한 문제를 제기한바 있다.

유튜브 추천 동영상 극화에 관한 최신 연구인 스위스 로잔 연방 공과대(Eidgenössische Technische Hochschule Zürich)와 브라질 미나스제라이스 연방대학(Universidade Federal de Minas Gerais)의 연구(Ribeiro et al., 2020)는 채널 추천 스냅샷들로부터 추천 그래프를 구성한 후 랜덤 워크(random walk) 시뮬레이션을 통하여 특정 조건에서 정치적 동영상 소비 극화의 가능성을 보여주었다. 이들은 사용자들의 채널과 콘텐츠에 대한 선택과 변화를 이해하기 위해 대규모 데이터 연구 필요. 대규모 관련 채널의 풀을 수집하고 채널에 대한 유튜브의 추천들과 데이터 수집, 유튜브 채널에 게재된 동영상을 기존 연구의 유형론에 입각하여 프로그래밍한 시스템으로 분석하여 채널 유형화 하였다. 그리고 유튜브의 추천 알고리즘이 특정 키워드를 검색하는 채널 사용자로 하여금 더욱 극단적이며 편향적인 콘텐츠를 지향하도록 하는데 어떠한 역할을 하는지 검증하기 위해 시뮬레이션을 통해 극단화의 경로 측정하고, 커뮤니티 및 미디어 채널의 활동량 증가는 채널 제작, 영상 출시 및 댓글의 증가 등 사용자 반응을 통해 확인하였다. 이를 통해 이들은 서로 다른 채널과 집단의 사용자들 사이의 교차점 분석을 통해 댓글을 다는 사용자의 특성을 파악하고, 콘텐츠 제작자들과 댓글을 다는 사람들의 논조의 진화를 추적하며, 어조의 변화를 연구하였다.

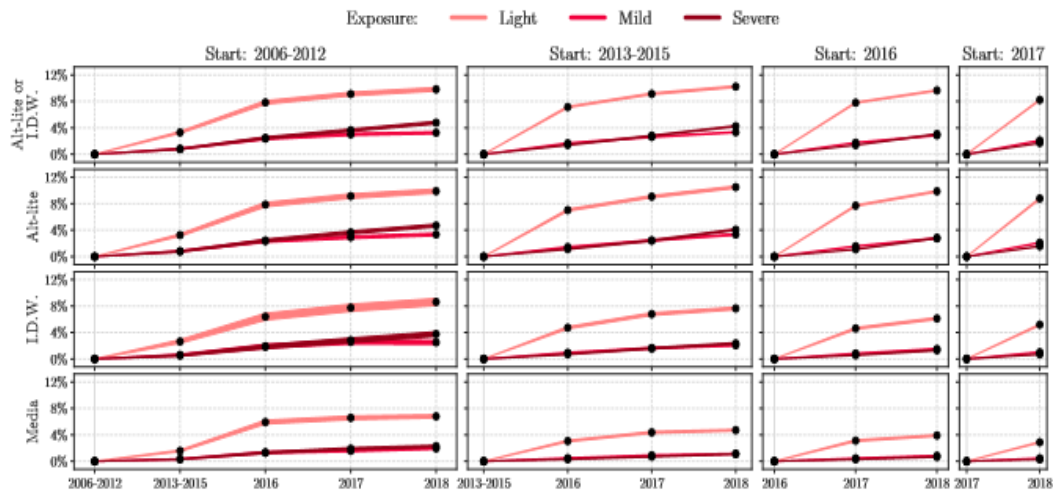


[그림 II-8] 커뮤니티와 미디어 채널에 대한 활동량 증가

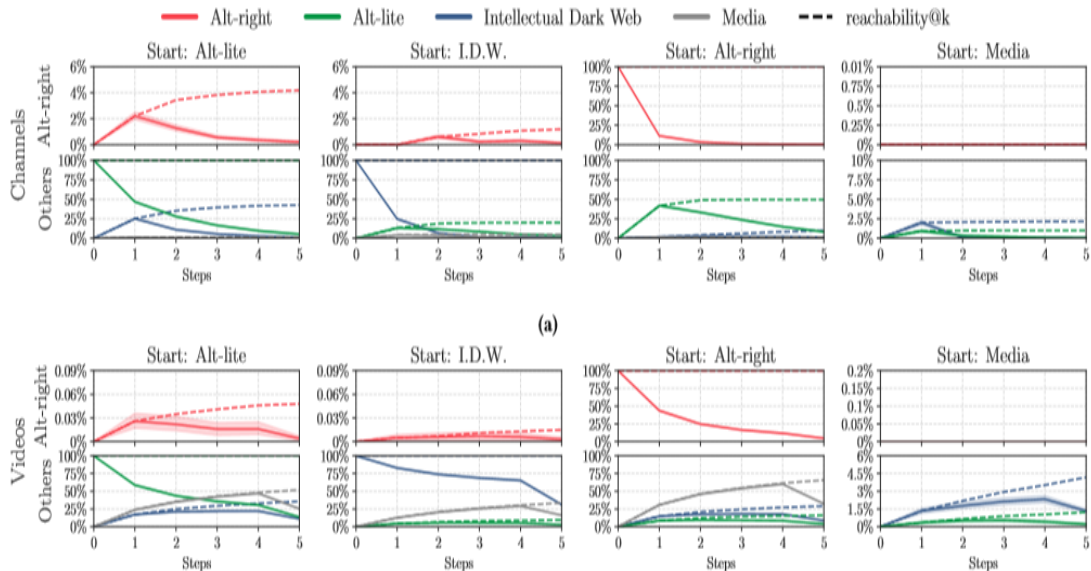


[그림 II-9] 사용자 교차점 분석

또한 사용자 이동 추적, 사용자의 논조의 진화를 추적하고 그들의 이야기가 불러일으킨 현상, 어조의 변화 연구, 문제가 되는 채널이 동일한 사용자를 기반으로 하고 있는지, 이들 이용자들이 어떠한 유형의 콘텐츠로 이동하는지를 확인하였으며, 정보극화를 측정하기 위해 댓글을 다는 사용자를 활용, 사용자의 궤적에 초점을 맞추어 사용자 의견을 통한 추론 및 그래프 시뮬레이션으로 극단화 측정하였다.



[그림 II -10] 사용자의 콘텐츠 이동양상



[그림 II -11] 채널과 영상에 대한 임의 행로(random-walks)시뮬레이션 결과

나. 자동화된 소프트웨어 기반 개인화 알고리즘 편향 및 극화 추적 시도와 한계

로잔 연방 공과대와 미나스제라이스 연방대학의 연구는 33만 건이 넘는 동영상에 대한 방대한 양적분석을 수행(Ribeiro et al., 2020)하였음에도, 유튜브 전체 추천 내역(history)을 얻는 것은 불가능하기 때문에, 계속 변화하는 블랙박스 알고리즘을 한정된 순간(a time slice)에 대해서만 분석할 수밖에 없었던 점을 지적하였다. 즉, 개인화된 결과물의 차이를 추적하지는 못하였고 대신 개인화되지 않은 추천 스냅샷들로부터 그래프를 구성한 후 랜덤 워크(random walk) 시뮬레이션을 통하여 실험을 진행한 것을 연구의 한계로 들었다.

Ledwich & Zaitsev (2019)는 YouTube의 800개 이상의 채널에 대한 추천내역을 분석하여 네 가지 타입의 급진화(radicalization)를 검증하였다: (1) Radical Bubble: 급진적인 동영상을 보았을 때 비슷한 영상을 추천하는 정도가 그렇지 않은 동영상을 보았을 때보다 높다. (2) Right-wing Advantage: 유튜브의 추천 알고리즘이 우파 콘텐츠를 더 많이 추천한다. (3) Radicalization Influence: 이용자들이 추천 알고리즘을 통해 더 극단적인 콘텐츠에 노출된다. (4) Right-wing Radicalization Pathway: 유튜브 알고리즘이 메인스트림과 중도좌파 채널의 이용자들에게 극우 콘텐츠를 추천한다. 그 결과, (1)에 대해서만 부분적인 근거를 확인할 수 있었고, (2), (3), (4)에 대해서는 근거가 없다고 결론지었다. Ledwich & Zaitsev (2019)는 구글 알고리즘 개발진이 상용 비디오 서비스의 추천 알고리즘을 서술한 선행연구(Zhe Zhao, 2019)에 기반한다면, 개인의 과거 동영상 시청 내역이 추천 결과물에 큰 영향을 주지 않을 것이라고 분석하였다. Munger & Phillips (2019)는 유튜브 상 극화된 영상의 노출은 알고리즘보다 이용자에 있다고 주장하였는데, 이는 약 천만 명의 페이스북 유저 데이터를 분석하여 이용자들의 편향적 뉴스 소비 행동은 알고리즘으로 인한 랭킹보다 개인의 선택이 더 크게 기인함을 보였던 Bakshy et al., (2015)의 연구와도 일맥상통한다.

다. 개인화 추천 알고리즘 역분석의 어려움

추천 알고리즘은 블랙박스 형식으로 작동하며 자체적으로 끊임없이 개선되고 있어 추적이 어렵다. 또한 네이버, 다음 등 국내 포털 사이트의 경우 주민번호 당 계정 생성 수를 3개로 제한하고 있으며, 주요 하나의 IP에서 다량의 계정을 생성하는 것을 차단하는 메커니즘을 갖추고 있어 개인화 실험을 위한 가상 계정을 생성하여 분석하는 것이 불가능하다. 비디오 스트리밍 서비스의 경우 잘못된 광고비 집행 방지 등을 위하여 봇 탐지 연구들이 활발히 진행되었으며(Chen et al., 2015; Marciel et al., 2016), 이러한 메커니즘으로 인하여 로봇 에이전트의 구현이 제한될 수 있다는 장점이 있다.

제3장 추천시스템 공정성 확보를 위한 정책 및 제도적 논의

추천알고리즘은 앞서 살펴본 바와 같이 데이터 기반의 자동화된 의사결정(automated individual decision-making)을 말한다. 알고리즘은 추천시스템 뿐만 아니라 다양한 영역에서 반영될 수 있다. 현재까지는 추천알고리즘만을 한정하여 공정성을 확보하고자 노력한 사례는 드물다. 그러므로 이번 장에서는 알고리즘이 반영되는 여러 가지 영역에서 공정성을 확보하기 위한 정책, 제도적 노력들을 살펴보고 실제 사례들을 논의하고자 한다.

제1절 개발단계에서의 투명성 확보 방안

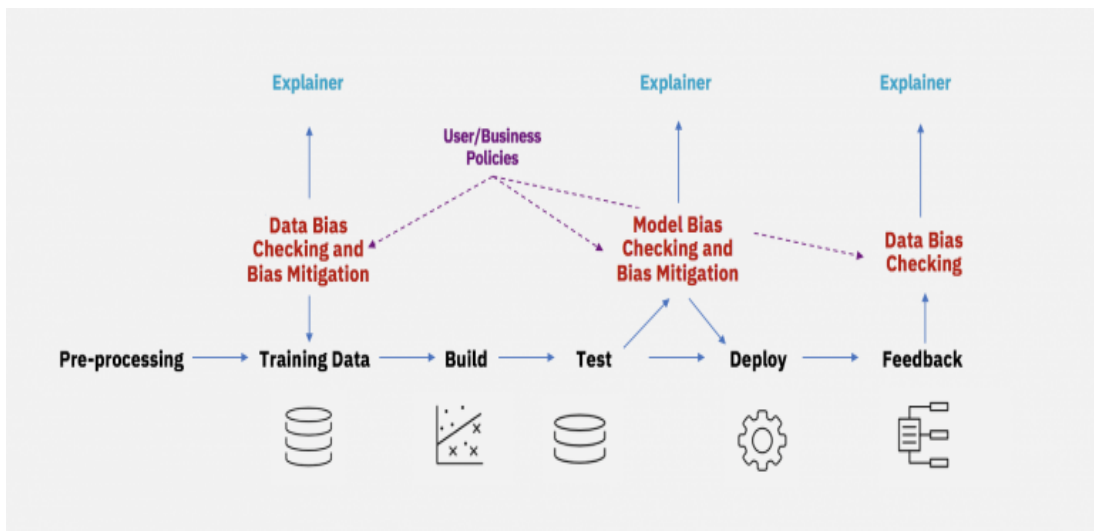
1. 개발자를 위한 오픈소스 툴킷

컴퓨팅 기술의 발전과 확장으로 인해 머신러닝을 이용한 알고리즘의 개발 및 사용이 쉬워지고, 빠르게 사용영역이 확장되고 있다. 이와 동시에 알고리즘으로 인한 편향성 문제가 지속적으로 발생 및 제기되고 있으며, 이를 해결하고자 하는 노력 또한 정부, 민간, 기업 등 여러 주체가 참여하여 다양하게 이루어지고 있다. 하지만 현재 머신러닝 알고리즘의 사용 측면과 달리 편향성으로 발생하는 공정성의 문제는 개인이나 중소기업 등의 일반적인 영역에서는 비용, 인력 등의 이유로 다뤄지기 어렵다.

반면, 리소스가 풍부한 구글, IBM과 같은 대기업이나 연구실들은 이러한 문제를 다룰 수 있는 여건을 충족하고 있다. 따라서 최근 머신러닝 모델 개발 및 사용과정에서 지속적으로 발생하는 알고리즘의 편향성을 측정하고 알고리즘 공정성을 개선할 수 있는 다양한 ‘공정성 툴킷’들을 개발하고, 이를 공정성 문제를 실질적으로 다루기 어려운 영역에서도 공적으로 활용할 수 있도록 오픈소스 형태로 제공하는 노력들이 이루어지고 있다.

IBM에서 2018년에 개발하여 오픈소스 서비스를 제공하는 Github을 통해 공개한 ‘AI Fairness 360’은 머신러닝 모델내에서 그리고 AI 애플리케이션 수명주기 전반에 걸쳐 알고리즘 편향성의 원인이 되는 편견과 차별을 탐색하여 모델 개발자에게 보고해줄 뿐만 아니라 완화해주는 툴킷이다. 해당 오픈소스 툴킷에는 70개 이상의 연구용 공정성 매트릭과 11개의 고유한 편향 완화 알고리즘이 포함되어 있으며 금융, 인적자본 관리, 의료 및 교육을 포함한 실제 산업 전반의 적용사례로 변환될 수 있도록 설계 및 개발되었다. 현재는 2020년 06월을 기준으로 업그레이드되어 scikit-learn 데이터 과학 라이브러리 및 R과의 호환성을 추가하는 등 지속적으로 기능 및 성능 업데이트가 이루어지고 있다⁵⁵⁾. AI Fairness 360은 기존의 IBM이 갖고있는 뛰어난 기술력과 보유하고 있는 방대한 데이터를 통해 높은 수준으로 개발된 오픈소스 툴킷이기 때문에

기능적으로 뛰어날 뿐만 아니라 다양한 기능을 제공한다. 오픈소스로 제공되는 파이썬 패키지에는 넓은 범위의 알고리즘 공정성 연구 커뮤니티에서 개발한 9가지의 알고리즘이 포함되어 있어 불가피하게 발생하는 편향을 완화시켜준다. 또한 대중성을 위하여 scikit-learn의 모델 형성 및 예측 기능인 fit/predict의 경향성에 맞춰 개발함으로써 연구자들의 협업을 가능케하고, IBM 자체의 통합적인 연구결과들을 다양한 산업분야에서 종사하고 있는 데이터 과학자 및 공학자, 개발자들에게 그들이 이해할 수 있도록 전달이 가능하다. 이와 더불어 3가지 튜토리얼들을 통해 연구자뿐만 아니라 실무에 종사하는 사람들에게 높은 접근성을 제공하고 있다. 이를 통해 AI Fairness 360은 기존의 오픈소스 툴킷들과 다르게 단순 매트릭스가 아닌 편향 완화에 집중되어 보다 직접적으로 알고리즘 공정성 문제에 기여하고, 연구분야뿐만 아니라 실질적으로 알고리즘들이 적용되고 사용되는 산업영역에서의 사용성 그리고 소프트웨어 설계파트까지 고려하여 알고리즘 공정성 문제까지 직접적으로 고려하기 어려운 연구자 및 개발자들에게 현실적인 도움을 제공한다.

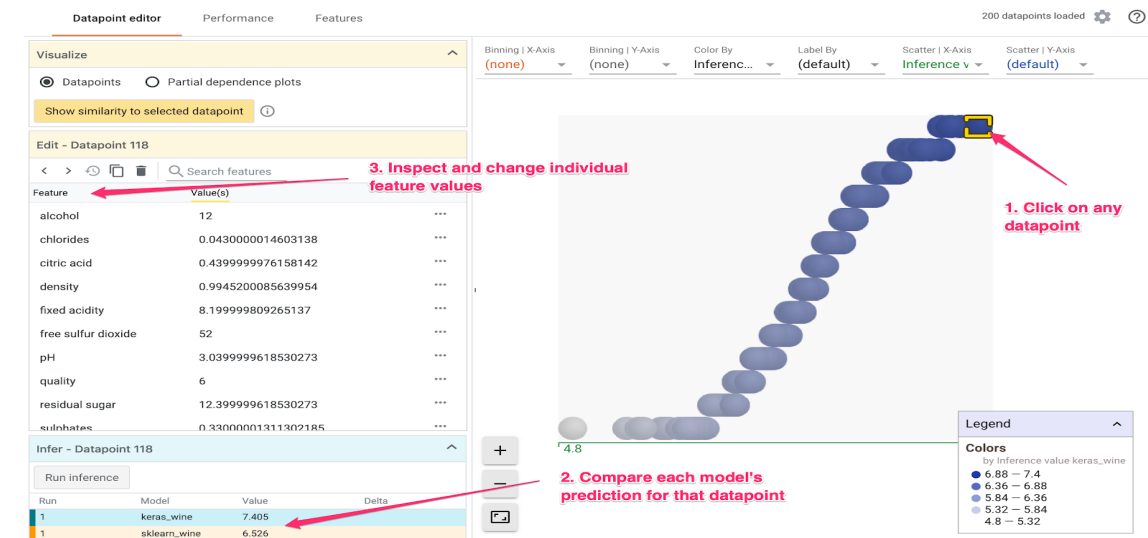


[그림 III-1] <AI Fairness 360>의 AI 모델 사이클 전반에 대한 툴킷 작동원리

대표적인 플랫폼 기업이자 테크 자이언트인 Google에서 개발한 ‘What-if’ 툴킷은 Google의 데이터관련 프로그래밍을 위한 오픈소스 소프트웨어 라이브러리인 TensorFlow 추가 및 내장되어 사용자로 하여금 머신러닝 모델의 공정성을 쉽게 분석할 수 있도록 돕고 있다. 개발자로 하여금 머신러닝 모델의 분류과정에서의 블랙박스화 회귀적 머신러닝 모델에 대한 이해를 확장시킬 수 있도록 대량의 샘플 데이터 셋과 알고리즘 모델에 대한 추론을 가능케 해주며, 즉각적으로 추론 결과를 다양한 방식으로 시각화하는 기능을 수행한다. 현존하는 오픈소스 툴킷 중 시각화 기능이 뛰어난 편에 속하여 머신러닝 알고리즘에 대한 손쉬운 이해를 제공함으로써 알고리즘 공정성 개선에 기여하고 있다. 더불어 샘플들을 직접 편집하거나 프로그램에 맞춰 재실행하는 기능을 통해 모델의 변화 결과를 보여주어 모델의 성능과 공정성을 모두 측정할 수 있도록 돕는다.⁵⁵⁾ What-if 툴킷은 현재 머신러닝 개발에 있어 가장 대중적으로 사용되는 오픈소스인

55) <https://aif360.mybluemix.net/>

TensorFlow에 알고리즘 공정성 툴킷을 코드 한 줄 없이 적용할 수 있도록 웹 애플리케이션 형태로 개발되었는데, 이는 그동안 코드를 능숙하게 다루는 숙련자만의 영역이었던 모델 분석 및 평가를 TensorFlow를 이용하는 모든 사람들이 접근할 수 있도록 만들어 줌으로써 알고리즘 공정성 이슈를 보다 그라운드 레벨로 현실화했다. 더불어 주요 기능인 인퍼런스(Inference) 결과를 시각적으로 보여주는 것, 데이터 포인트를 선택해 피처(Feature) 값을 자유롭게 편집할 수 있는 것, 개별 피처가 어떤 영향을 끼치는지에 대한 결과를 시각적으로 제공하는 등 강점인 시각화 기능을 통해 인터랙티브한 방식으로 간편하게 머신러닝 모델을 분석할 수 있게 해주어 연구자 및 개발자들이 모델의 편향성을 보다 높은 효율로 완화할 수 있게 해준다.



[그림 III-2] <what-if> 공정성 툴킷의 시각화 기능

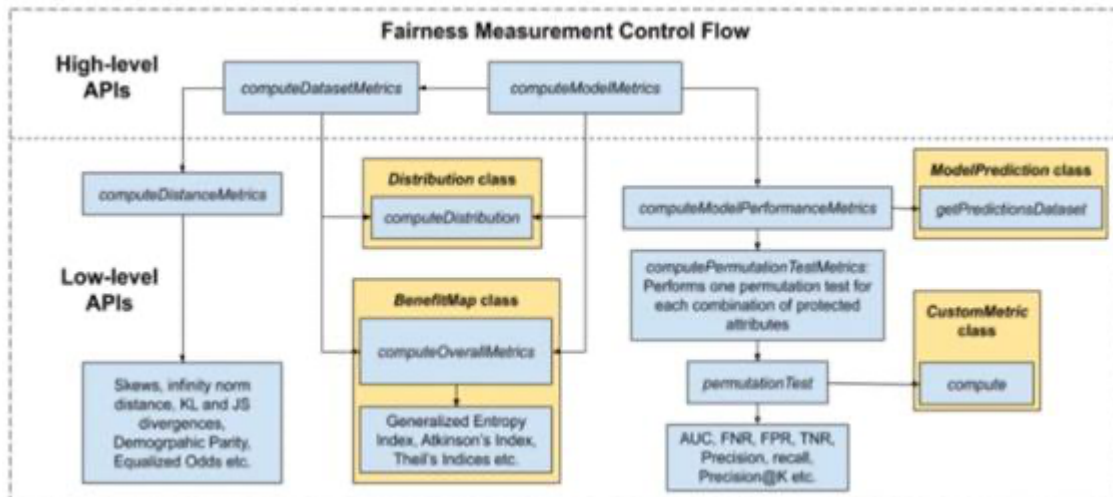
또 다른 오픈소스 툴킷인 'Fairlearn'은 Microsoft에서 개발된 툴킷으로써 IBM의 공정성 툴킷과 마찬가지로 Github를 통해 오픈소스로 제공 중에 있다. 툴킷을 개발한 Microsoft에서는 알고리즘 공정성을 우선적으로 해결되어야 할 사회공학적 도전과제로 보고 사회적 그리고 기술적 문제가 복합적으로 얽혀 발생하는 편향성 문제를 완화하기 위해 Fairlearn을 개발하였다. 해당 툴킷은 IBM의 오픈소스 툴킷들과 마찬가지로 공정성 매트릭과 최신 편향성 완화 알고리즘에 대하여 넓은 선택지를 제공하고 있으며 여러 모델과 동시에 비교분석이 가능한 상호작용적 시각화 기능을 제공하여 알고리즘 모델의 편향성 완화와 공정성 개선에 도움을 주고 있다.⁵⁷⁾ Fairlearn 오픈소스 패키지는 궁극적으로 모델이 갖고 있는 혹은 영향받는 여러 요소들과 이로 인해 발생할 수 있는 편향을 완화할 수 있는 전략을 평가할 수 있도록 도와준다. 이를 통해 알고리즘 공정성 문제에 대하여 연구자 및 개발자들이 인공지능 및 머신러닝 모델을 개발하는 데에 있어 필요한 시나리오의 장단점을 제공해줌으로써 주도적인 모델 개발을 가능케 해준다.

비즈니스에 특화되어 특정 업계 사람들이 서로 구인 및 구직, 정보파악 등의 서비스를 제공하는 SNS기

56) <https://pair-code.github.io/what-if-tool/>

57) <https://fairlearn.github.io/>

업인 LinkedIn은 'LiFT'라는 오픈소스 공정성 툴킷을 개발하여 제공하고 있다. 해당 툴킷은 머신러닝 모델 학습 도중 배치되어 사용될 수 있으며, 학습 데이터셋의 편향성에 대한 점수를 매기고, 모델의 공정성을 평가함과 동시에 학습모델의 서브그룹들에 대한 성능에 차이를 탐지할 수 있다. 또한 다양한 모델들에 대하여 가변적으로 적용이 가능하고 재사용할 수 있으며, 다양한 부가기능들을 통해 높은 편의성을 제공하여 알고리즘 공정성 개선에 기여하고 있다. 특히 취업과 관련된 추천시스템을 활용하고 있는 회사인만큼 내부적으로 툴킷을 적용 및 확대하여 사람 및 알고리즘에 의한 편향을 완화하고 공정한 취업기회 제공을 목표로 하고 있다.⁵⁸⁾ LiFT의 경우 소개하고 있는 여러 오픈소스 툴킷과 마찬가지로 편의성과 높은 성능을 통해 알고리즘 공정성을 보다 다양한 영역과 여러 레벨에서 다룰 수 있도록 돕고 있다. 또한 개발업체인 LinkedIn의 특성이 반영되어 사회적 영역에서 중요한 이슈인 구인구직의 문제를 다룸으로써 알고리즘이 인간사회에 미칠 수 있는 영향과 이로 인해 발생하는 편향에 대하여 보다 직접적으로 다루고 있다는 점에서 의미를 갖는다.



[그림 III-3] <LiFT>의 공정성 측정 프레임워크 플로우 차트

앞선 오픈소스 툴킷의 사례들이 대기업 중심이라면 'Ethics and Algorithms Toolkit'은 데이터 커뮤니티인 DC, 샌프란시스코 정부 관계자, 존 홉킨스 및 하버드 대학 연구자들이 모여 개발한 오픈소스 공정성 툴킷이다. 이들은 정부 차원에서의 알고리즘 이용이 불가피함과 동시에 모든 사람과 알고리즘이 편향성을 갖는다는 점을 중심으로 정부의 데이터 활용과 알고리즘 개발 및 활용이 갖는 의미를 이해하도록 돕고, 잠재적 리스크들을 명료하게 만들고, 리스크들을 완화시킬 수 있는 방법들을 제공하기 위해 해당 툴킷을 개발하였다. 특히 도시들과 같은 정부의 데이터 및 알고리즘 사용에 중점을 두어 정부 당국과 관계자들에게 민간, 언론, 학회 등으로부터 제기되는 정부의 알고리즘 편향성 문제와 공정성 개선에 대한 실용적인 툴킷을 제공한다는 점이 특징이다.⁵⁹⁾ Ethics and Algorithms 툴킷은 알고리즘을 가장 민감하고 넓은 범위에 적용

58) <https://venturebeat.com/2020/08/25/linkedin-open-sources-toolkit-to-measure-ai-model-fairness/>

하고 있는 주체인 정부관계자들이 개발에 참여함으로써 정부기관과 같은 공공환경에서 사용하기 적합하도록 목표를 설정하여 개발하기 때문에 민간영역에서 개발한 오픈소스 툴킷과 달리 사회공공 측면에서 알고리즘의 편향을 보다 효과적으로 완화해준다.

또 다른 오픈소스 공정성 툴킷인 ‘Deon’은 데이터 사이언스 프로젝트를 진행하면서 연구진 혹은 개발자로 하여금 일종의 윤리적인 체크리스트를 손쉽게 추가할 수 있도록 도와주는 툴킷으로써 Driven Data 연구소에서 개발 및 배포되었다. 고대 그리스어로 Duty, 즉 의무를 의미하는 단어인 Deon이라는 이름처럼 해당 오픈소스 툴킷은 데이터 사이언스뿐만 아니라 머신러닝 및 AI의 개발과정에서 발생하고 지나칠 수 있는 윤리적 문제점들을 개발자들이 지속적으로 체크하고 인지하여 그 과정에 영향을 미칠 수 있도록 현실적이고 적용 가능한 리마인더를 제공하는 것을 목표로 한다.⁶⁰⁾ 이외에도 알고리즘 공정성을 위한 오픈소스 툴킷에는 워싱턴 대학에서 개발하여 제공하는 ‘Lime’, 시카고 대학 데이터 사이언스 및 공공정책 센터에서 개발한 ‘Aequitas: A Bias and Fairness Audit Toolkit’ 등이 존재하고 있으며 알고리즘을 사용하는 모든 영역에서 알고리즘 편향성을 인지하고 완화할 수 있도록 하여 공정성을 개선할 수 있도록 돕고 있다. 이처럼 다양한 영역이 주체가 되어 개발하는 오픈소스 툴킷은 알고리즘 공정성이라는 불가피한 문제를 보다 다양하고 넓은 영역, 그리고 다양한 레벨에서도 다룰 수 있도록 돕고 있다. 간편하고 강력한 도구인 알고리즘은 이미 사회 전반 곳곳에 침투하여 있기 때문에 이러한 노력은 알고리즘 공정성 문제를 표면적으로 다루는 것이 아닌, 알고리즘 편향 문제 개선에 보다 실질적인 역할을 수행하는 중이라고 할 수 있다.

제2절 자율규제활동과 공정성

1. 기업차원의 자율활동

가. 해외기업

1) 구글

구글은 ‘책임 있는 AI(Responsible AI)⁶¹⁾’ 프레임워크에 기반해 사회와 공익에 대해 책임 있는 AI개발에 대한 연구결과와 도구, 실행지침을 공유하고 있다. 구글이 AI 책임에 대해 관심을 가지는 주제는 크게 공정성(Fairness), 해석가능성(Interpretability), 프라이버시(Privacy), 보안성(Security) 등으로 그 주제를 나눌

59) <https://ethicstoolkit.ai/>

60) <https://deon.drivendata.org/>

61) Responsible AI Practices. URL :

<https://ai.google/responsibilities/responsible-ai-practices/?category=fairness>

수 있으며 권장사례와 세부 실행방안 등을 확인할 수 있다. 공정성의 경우, 구글은 AI의 영향력이 사회 전반에 걸쳐서 증대되고 있는 상황에서 누구에게나 공정하고 포용적인 시스템을 지향하는 것이 중요하다고 강조한다. 그러나 구글은 기존 데이터를 기반으로 학습한 알고리즘은 기존 데이터가 가지고 있는 문제나 편향(bias)을 학습하거나 확대할 수 있으며, 여러 테스트를 거치더라도 현실적으로 완전한 공정성을 구축하기 어려우며 셋째, 공정성에 대한 표준적인 정의가 없어 공정성을 확보하는 것은 매우 어려운 문제임을 강조한다. 구글은 공정성을 확보하기 위한 노력으로 다양성(Diversity) 측면에서 접근하여 공정성 문제에 대한 연구를 진행하고 있으며, AI 및 알고리즘 개발 시의 공정성 확보를 위해 네 가지 실천방안을 추천하고 있다.

첫째, 공정성과 포용에 대한 구체적인 목표를 사용해 모델을 디자인하는 것이다. 사회과학·인류학 등 다양한 관련 전문 영역의 관점을 고려하고 구체적인 공정성의 목표를 계획해 이를 달성하기 위한 객관적인 알고리즘을 디자인해야 한다. 둘째, 모델을 훈련하고 검증하기 위한 적합한 데이터셋을 사용하는 것이다. 개발자는 자신의 데이터셋 자체의 공정성을 평가하여야 하는데, 데이터셋 자체가 편향 및 한계가 존재할 수 있기 때문이다. 데이터셋에 대한 평가 이후, 자신의 시스템의 목적과 관점에 따라 적합한 데이터셋으로 모델을 구축해야 한다. 셋째, 불공정 편향 확인을 위해 시스템을 체크해야 한다. 의도적으로 편향 상황을 설계한 데이터를 통한 지속적 테스트를 통해 시스템의 편향을 확인하고, 이를 지속적으로 수정하는 피드백 루프(feedback loop)를 구성해야 한다. 마지막으로, 퍼포먼스를 분석하는 것이다. 척도나 데이터의 서브그룹을 조정하는 방식으로 모델의 결과물에 대해 다양하게 분석함으로써 모델이 실제 세계의 문제에 잘 적용되는지 확인해야 한다.

공정성 외에 구글은 해석가능성(Interpretability), 프라이버시(Privacy), 보안(Security)의 측면에서 AI의 사회적책임 실천방안들을 제시하고 있다. 구체적으로 AI시스템에 대해서 의문을 제기하고 이해할 수 있는 해석가능성을 확보하는 것이 중요하며 이에 대한 실천방안으로 학습모델을 해석가능하게 디자인할 것, 훈련된 모델을 이해할 수 있을 것 등을 제시하고 있다. 또한 구글은 AI가 법적으로 프라이버시를 보호할 뿐만 아니라 사회의 일반적 인식과 개인적인 수준에서 프라이버시도 함께 고려해야함을 강조하고 있다. 마지막으로 구글은 AI시스템이 외부 간섭이나 공격에 관계없이 의도한대로 작동해야하는 보안적인 측면을 강조하고 있으며, 세부 권장 사례와 실행지침을 제공하고 있다. 이와 같은 실천방안을 추천하는 것 외에도 구글은 AI 및 알고리즘의 공정성과 관련하여 지속적으로 다양한 연구를 수행하고 있으며, 해당 결과 및 오픈소스 툴킷을 공유하고 있다. 이들의 주제는 크게 ‘편향 측정’, ‘공정한 머신러닝 모델과 데이터셋 구축’, ‘불공정성 경감을 위한 테크닉’, ‘이용자와 사회에 대한 이해’로 분류될 수 있으며 관련 페이지에서 바로 연구논문을 다운로드하거나 분석 실습을 진행해볼 수 있다.

2) IBM

글로벌 IT기업인 IBM의 IBM연구소(IBM Research)에서는 <2020년 AI분야에 대한 5가지 전망>을 발표하며 ‘신뢰할 수 있는 AI(Trusted AI⁶²⁾)’를 만들어내기 위한 노력을 늘려갈 것을 강조했다. IBM은 ‘신뢰할 수 있는 AI’라는 명목 아래 ‘공정성(Fairness)’, ‘가치정렬(Value Alignment)’, ‘강건성(Robustness)’, ‘설명가능성(Explainability)’, ‘투명성과 책임감(Transparency and Accountability)’ 등 다섯 가지 개념을 자사 인공지능의 핵심 역량이자 규율로 채택, 강조하고 있다.

먼저, ‘공정성(Fairness)’를 통해 인공지능을 통해 도출된 결과물이 지닐 수 있는 편향 가능성에 대한 지속적인 경계를 강조하고 있다. 공정성을 확보할 수 있게 적절하게 보정된 AI는 사용자의 선택을 지원하고, 빠른 결정과 평가를 할 수 있도록 기회를 제공할 수 있다. 인공지능이 적용되는 모든 영역에 대해 각기 다른 편향의 유형과 데이터 타입에 대응할 수 있도록 AI Fairness 360 툴킷 등의 활용을 제시한다. 그 밖에 ‘가치정렬(Value Alignment)’을 통한 인공지능 기반 의사결정의 정확성과 효율성을, ‘강건성(Robustness)’를 통한 인공지능 모델의 지속가능함과 강건함을, ‘설명가능성(Explainability)’를 통해 알고리즘의 설명가능성을, ‘투명성과 책임감(Transparency and Accountability)’를 통해 알고리즘의 투명성과 알고리즘에 대한 인간의 책임감이 필요함을 기술하고 있다.

이뿐만 아니라 IBM에서는 AI에 대한 투명성과 책임성을 강조하며 정책적인 측면에서 AI에 대한 정밀규제를 찬성한 바 있다. 구체적으로 IBM 정책연구소(IBM Policy Lab)에서는 AI에 대한 5가지 규제 가이드를 제시했다. 첫 번째는 AI윤리책임자를 지정하여 AI시스템에서 발생할 수 있는 위험을 완화하고, 내부 지침과 규정을 관리하고 책임질 수 있게 해야 한다는 것이다. 두 번째는 사업분야와 업무에 따라 적용되는 AI가 다양하고 잠재적 위험요소도 다르기 때문에 사업분야별로 위험에 따른 맞춤 규정을 수립해야한다는 것이다. 세 번째는 기업의 영업비밀에 해당하는 소스코드까지는 공개하지 않더라도 AI가 어떤 목적을 가지고 시스템에 적용되는지 사용자에게 투명하게 공개해야 한다는 것이다. 네 번째는 AI가 내린 결정에 대하여 의사결정 과정과 이유에 대해서 설명할 수 있어야 한다는 점이다. 마지막으로 AI시스템을 적용하는 것에 앞서 기업 내에서 공정성, 편향성, 보완성 등에 대한 사전 테스트가 이뤄져야 한다는 것이다. 이 테스트는 기업과 정부가 함께 참여하는 ‘공동규제’방식이어야 한다는 점을 제안했다. 마지막으로 IBM에서는 “AI 오픈스케일(OpenScale)”이라는 AI편향 검증 플랫폼을 선보인바 있다. AI오픈스케일을 통해 제안되는 가치는 5가지로 첫째, AI의 의사결정 과정을 이해하기 쉬운 용어로 설명하는 것, 둘째 자동화된 편향제거기술로 편향성을 탐지하는 것, 셋째 GDPR등 법률과 규정을 준수할 수 있도록 기업이 모든 데이터, 기록을 보존하는 것 등이 해당된다.

62) <https://www.research.ibm.com/artificial-intelligence/trusted-ai/>

3) MS(Microsoft) Responsible AI⁶³⁾

마이크로소프트 (i.e. MS)는 ORA (the Office for Responsible AI) 설립을 통해 자사 인공지능 시스템의 책임감에 대한 정책을 펼치고 있다. 구체적으로 마이크로소프트는 ‘공정성(Fairness)’, ‘신뢰성&안전성(Reliability & Safety)’, ‘프라이버시&보안(Privacy & Security)’, ‘포괄성(Inclusiveness)’, ‘투명성(Transparency)’, ‘책임감(Accountability)’ 등 여섯가지 원칙을 바탕으로 책임감 있는 인공지능을 위한 기틀을 다지고 있다. 사람을 공정하게 다뤄야 한다는 ‘공정성(Fairness)’와 더불어 인공지능 시스템의 안정성을 논하는 ‘신뢰성&안전성(Reliability & Safety)’, 보안 및 프라이버시를 이야기하는 ‘프라이버시&보안(Privacy & Security)’, 다양한 사람들이 널리 이롭게 사용할 수 있어야 한다는 포괄성의 ‘포괄성(Inclusiveness)’, 이용자가 이해할 수 있어야 한다는 ‘투명성(Transparency)’, 그리고 알고리즘에 대한 인간의 책임감인 ‘책임성(Accountability)’를 통해 ‘Responsible AI(책임감있는 AI)’를 구성하려는 모습을 볼 수 있다. MS사는 이러한 6가지 윤리적인 원칙이 정부, 학계, 기업, 시민사회 등에서 AI를 개발하고 활용할 때 적용되어야 함을 강조했다.

이뿐만 아니라 MS에서는 내부윤리위원회를 두고 “책임지는 AI : 대화형 AI개발자를 위한 10가지 지침 (Responsible bot : 10 guidelines for developers of conversational AI)”를 발표한 바 있다. 해당 가이드라인의 구체적인 세부 지침을 살펴보면, AI가 적용되는 사례에 목적을 명시할 것, 제품 또는 서비스 일부로 AI가 사용됨을 명확하게 들어낼 것, AI가 적절한 문화규범을 존중하고 오용을 방지하도록 설계할 것, AI가 사람들에게 공정하게 적용될 수 있게 설계할 것, AI가 프라이버시를 존중할 수 있도록 할 것 등의 내용이 해당된다. 마지막 지침에서는 “책임을 받아들이는 것”이라는 규정으로 AI로 인해서 발생하는 윤리적 이슈에 대해서 책임을 강조하는 모습을 찾아볼 수 있다. 또한 MS사에서는 AI기술이 활용되는 방향을 논의하기 위해서 사내에 엔지니어, 변호사 등 다양한 전문가들이 포함됨 거버넌스 위원회를 운영중이다. 이러한 모습을 통해 MS사가 기술적인 요구사항에 앞서서 책임과 윤리가 선결되기 위해 여러가지 노력을 하는 것을 유추할 수 있다.

나. 국내기업

1) 카카오의 알고리즘 윤리 헌장

국내 IT기업 카카오 <알고리즘 윤리 헌장> 발표를 통해 국내에서는 최초로 인공지능 윤리헌장을 제정하였으며, 자사가 생각하는 AI 윤리의 기준점들을 제시한 바 있다. 최근에는 먼저 발표된 여섯 가지 원칙에 더해 사회적으로 화두가 되고 있는 아동과 청소년에 대한 새로운 조항을 신설했다. 알고리즘의 윤리에 대

63) <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar6>

해서 알고리즘 개발자가 의도적으로 개입하여 차별이 발생할 가능성, 데이터 수집과 관리측면에서 사생활 침해와 같은 윤리성의 부재, 알고리즘의 원리와 작동방식에 대한 정보가 투명하게 공개되지 않을 경우의 문제점 등 주요 이슈를 중심으로 카카오의 윤리현장이 수립됐다(국민권익위원회, 2019⁶⁴)

카카오는 첫째, ‘알고리즘 기본 원칙’을 통해 윤리현장을 도입한 목적을 설명하고 있다. 카카오는 알고리즘과 관련된 모든 사안은 우리 사회의 윤리의 틀을 벗어나지 않으며, 인간의 행복과 편익을 추구함을 강조한다. 이어서 둘째, ‘차별에 대한 경계’를 통해 알고리즘의 결과물로 인한 의도적인 사회적 차별이 일어나지 않도록 함을 이야기하고, 셋째, ‘학습 데이터 운영’을 통해 알고리즘 학습에 사용되는 데이터가 반드시 사회 윤리에 근거하여 수집 및 분석, 활용됨을 강조한다. 넷째, ‘알고리즘의 독립성’을 통해 알고리즘이 타인에 의해 자의적으로 훼손되거나 영향받지 않도록 엄중한 관리가 필요함을 제안하고, 다섯째, ‘알고리즘에 대한 설명’을 통해 이용자와의 신뢰관계 및 기업 경쟁력을 훼손하지 않는 범위에서 알고리즘에 대해 성실히 설명할 것을 약속하고 있다. 여섯째, ‘기술의 포용성’을 통해 알고리즘을 기반으로 제공되는 서비스가 우리 사회 전반을 포용할 수 있도록 명시하며, 마지막으로 최근 추가된 ‘아동과 청소년에 대한 보호’를 통해 아동 및 청소년이 알고리즘으로 인해 부적절한 정보와 위험에 노출되지 않도록 함을 강조하고 있다.

카카오측 알고리즘 윤리현장은 국내 기업 중 최초로 인공지능 알고리즘 설계 시 준수해야 할 사항을 명문화했으며 민간기업 차원에서 인공지능 윤리기준을 공표했다는데 의의가 있다(과학기술정책연구원, 2019⁶⁵). 특히 마지막 ‘기술의 포용성’을 추가함으로써 의도치 않게 발생할 수 있는 사회적 소외를 방지하기 위한 의지를 드러냈다. 그러나 한편으로는 카카오 윤리현장 내에서 이용자가 일반 사용자 뿐 만 아니라 알고리즘 서비스와 산업에서의 직간접적인 이해관계자를 모두 포함하고 있지만 실제 윤리의 주요 행위자로서의 이용자 역할이 잘 드러나지 않는다는 한계가 지적된다(이원태, 2018⁶⁶). 또한 알고리즘 관련 기술 및 서비스에 따라서 다양하게 성격이 구분될 수 있으므로 세부적이고 구체적인 행위 원칙이 제시될 필요가 있다(이원태, 2018).

2) 기타 사례 : (주)코난테크놀로지, 한국인공지능윤리협회

카카오에서 ‘알고리즘 윤리현장’은 발표 이후, IT기업을 포함한 산업계에서는 관련된 기업들은 알고리즘 윤리와 관련하여 다양한 시도를 진행하고 있다. 첫째, 국내 인공지능기업인 (주)코난테크놀로지는 인공지능 모델과 학습데이터의 편향성을 보정하기 위한 연구를 2022년까지 45개월 간 진행할 것이라고 발표했다. (주)코난테크놀로지는 자연어처리(NLP)와 자연어이해(NLU)를 기반으로 지속적으로 성장하고 있다. 메신저 혹은 정부나 기관의 대민서비스 부문에 활용되는 ‘KONAN BOT’, 대용량의 정형 및 비정형 데이터를 분석하는 ‘Konan Analytics’, 딥러닝 알고리즘 기술을 바탕으로 추천, 매칭 솔루션을 제공하는

64) 국민권익위원회(2019) 기업윤리 브리프스 : 네오에텍스 - 인공지능AI와 기업윤리. 2019년1호

65) 과학기술정책연구원(2019). 인공지능 기술 전망과 혁신정책 방향(2차년도)

66) 이원태(2018). 카카오 알고리즘 윤리현장의 의미 : 개발자 자율규제 의지 천명... ‘이용자 윤리’ 보완 필요. 「신문과방송」. 통권 568호. 52~55

‘SmartMatch’ 등 여러가지 서비스를 제공하고 있다. (주)코난테크놀로지와 KAIST, 연세대학교, 페이스북, 마이크로소프트 등이 참여한 연구에서는 크게 세 가지 목표를 가지고 진행된다. 첫째, 인공지능 모델의 편향성을 분석하고 공정성 개념을 확립하고자 하며 둘째, 데이터 및 인공지능 모델의 편향성을 탐지하고, 완화하고, 제거하는 알고리즘을 연구하고 마지막으로 편향성 분석을 위한 시각화 도구를 산업영역에 적용하고자 하는 것이다. 이를 위해 구체적으로 편향성을 분석하기 위해 ‘공정성’에 대한 개념 융합연구를 수행할 예정이며, 정성적 평가를 통해 산업영역에 적용된 결과를 분석한다. 그 후, 편향성을 분석하고 시각화한 후, 실제 다양한 데이터를 활용해서 기존의 측도와 개발된 측도를 비교분석 할 예정이다. 코난테크놀로지는 이 연구를 통해 학습데이터와 AI모형의 편향성 측도를 개발하고, 이를 탐지-완화-제거하기 위한 프레임워크를 개발하여 인공지능이나 학습모델의 편향성과 공정성의 윤리적 문제를 해소하고자 하는 목표를 가진다고 발표했다. 이런 시도는 산학연계로 진행되는 알고리즘 편향보정 관련 연구라는 점에서 의의를 찾아볼 수 있으나 현재 진행 중인 연구로 결과를 선불리 판단하기는 어려운 상태이다.

그 외에도 한국인공지능윤리협회(Korea Artificial Intelligence Ethics Association)에서는 2019년 10월 23일에 인공지능 윤리헌장(The AI Ethics Charter)⁶⁷⁾을 제정하여 공표한 바 있다. 이 윤리헌장의 서문에서는 인공지능이 생활 깊숙이 자리잡고 있으며, 이로 인한 파급력이 막대해져 인공지능의 역작용과 위험성을 인지하고 대응하는 방안을 마련하는 게 중요하다고 강조한다. 이 윤리헌장은 크게 5가지 장에서 37가지 세부 원칙으로 구분되며 정부, 기업, 단체, 개인에게 공통으로 적용된다(<표VI-1>참고). 첫째, 인간과 인공지능 관계에 있어서 세부적으로 인공지능은 인간을 행복하게 하기 위한 도구라는 점, 해를 끼쳐서는 안 된다는 점, 인간의 보편가치를 존중해야하는 점, 인공지능의 어떤 의사결정도 인간에 의해 수정이나 취소될 수 있다는 원칙을 포함한다. 인간과 인공지능 관계와 관련된 세부 원칙들은 인공지능이 ‘도구’로서 활용되어야 하며 ‘인간의 가치’를 침범해서는 안 된다는 점을 강조하고 있다. 둘째, 인공지능의 목적과 관련되어 ‘선하고 안전한 인공지능’원칙을 서술하고 있다. 인공지능은 불순한 목적으로 만들어져서는 안되며, 비상시 중단할 수 있는 방안에 대한 세부 원칙으로 구성된다. 구체적으로 인공지능 기술과 알고리즘에 불안정한 요소가 발견될 시 즉각 수정되어야 하며, 알고리즘 기록은 문서화에 의해 투명하게 관리되어야 하며, 보안 체계를 유지해야 한다는 원칙 등을 포함하고 있다. 셋째, 인공지능 개발자의 윤리들과 관련된 13가지 원칙을 포함한다. 구체적으로 인공지능 개발자가 윤리적 책임의식을 가져야 하며, 합의된 안전 개발 지침에 의거해 인공지능 제품과 서비스를 만들어야 한다는 원칙을 포함한다. 넷째, 인공지능의 소비자 관점에서의 원칙이 제시되고 있는데 인공지능 제품과 서비스를 범죄의 목적으로 사용해서는 안 되며 올바른 방법으로 사용해야 한다는 세부원칙을 포함한다. 마지막으로 공동의 책임과 이익공유의 측면에서 사회적 가치를 추구해야 한다는 점을 강조하고 있다. 세부 실행지침으로는 인공지능의 편익과 혜택을 온 인류가 공정하게 누리고 향유할 수 있어야 하는 등의 내용을 포함하고 있다. 이러한 인공지능 윤리헌장은 ‘선한 인공지능(Good AI)’ 개념을 담고 있으면서 개발자와 소비자를 위한 실제적이고 구체적인 지침들이 마련됐다는데 의미를 찾을 수 있다. 한 편으로는 기술발전에 따라 윤리헌장은 지속적으로 수정되어 대중적인 확산을 도모

67) 한국인공지능윤리협회(2019.10.23.). 인공지능 윤리헌장. URL : <https://kaiea.org/aicharter>

할 필요가 있다. 뿐만 아니라 선언적 의미의 공포내용을 업무에 반영하고 오류를 수정하는 구체적 방법과 감시기능의 적용이 구체화 되어야 한다.

2. 정부차원의 제도적 활동

가. EU의 GDPR(General Data Protection Regulation)의 알고리즘 투명성

GDPR은 2018년 5월 25일부터 시행되고 있는 EU의 개인정보보호법령으로 방대한 분량의 전문(Recital)과 99개 조항으로 이루어진 본문으로 구성되어 있으며 정보로 인해서 차별받지 않을 권리를 규정하고 있다⁶⁸⁾. 1995년 EU는 개인정보보호규범인 ‘데이터 보호 지침(DPD : Data Protection Directive)’을 제정했으며 인공지능 및 알고리즘과 같은 새로운 기술이 확산되고 개인정보 환경이 급변하면서 2016년 GDPR체제로 대체하여 2018년부터 발효되어 회원국에 적용되어져왔다(이원태, 2016⁶⁹⁾). GDPR에서는 알고리즘 투명성(algorithmic transparency)과 관련해서 ‘공정하고 투명한 정보처리(fair and transparent processing)의 원칙’을 선언했으며 자동화된 의사결정에 대해서 ‘설명을 요구할 권리(right to explanation)’을 제안하고 있다.

GDPR 제22조에서는 프로파일링⁷⁰⁾을 포함해서 자동화된 의사결정을 제한하고 있으며, 제13조, 제14조, 제15조에서는 정보주체의 권리를 명시하고 있다. 제22조에서는 예측을 목적으로 이뤄지는 개인정보의 자동화된 모든 형태의 정보를 어떻게 수집하고, 저장할 것인지 다루고 있다. 구체적으로 정보주체의 명시적 동의가 선행되어야 하며, 권리보호를 위해서 이익을 제기할 수 있는 권리를 제공해야 한다.

제13조, 제14조, 제15조에서는 관리자(controller)가 정보를 수집, 처리하는 과정에서 정보주체에게 프로파일링과 같은 자동화된 의사결정이 이뤄지고 있다는 점, 의미있는 정보와 정보를 처리함에 있어서 예상되는 결과를 정보주체에게 설명해야 한다고 명시하고 있다. 즉, 관리자에게 설명을 제공할 의무를 규정한 것으로 굿맨과 플렉스먼(2016)⁷¹⁾은 이를 ‘설명을 요구할 권리(right to explanation)’이라고 GDPR을 해석했다. GDPR의 전문에서는 투명하고 공정한 정보처리를 위해서 정보주체에게 정보가 제공될 수 있을 것, 정보주체에게서 발생하는 위험을 최소화할 것을 강조한다. 인공지능과 알고리즘이 작동하는 원리가 복잡하고, 이해할 수 없는 블랙박스(black box)인 만큼 이용자에게 작동원리 혹은 관련된 설명을 제공함으로써 투명성을 제고해야 한다는 것이다.

68) KISA에서는 GDPR대응지원센터를 운영중에 있음. URL : <https://gdpr.kisa.or.kr/index.do>

69) 이원태(2016). EU의 알고리즘 규제 이슈와 정책적 시사점. KISDI Premium Report. 16(12)

70) 프로파일링은 자연인의 업무성과, 경제적 상황, 건강, 개인적선호, 관심사, 신뢰도, 행태, 위치 등에 관한 분석을 예측하기 위해서 행해지는 경우로 자연인에 관한 개인적 특징을 평가하기 위해 개인정보를 사용해 이뤄지는 모든 형태의 자동화된 개인정보 처리라고 정의함(이선구(0019). <알고리즘의 투명성과 설명가능성 : GDPR을 중심으로>; 재인용)

71) Goodman, Bryce & Flaxman, Seth(2016). “European Union regulations on algorithmicdecision-making and a ‘right to explanation’”. 2016 ICML Workshopon Human Interpretability in Machine Learning

나. 미국 FTC의 User Artificial Intelligence and Algorithms⁷²⁾

미국 연방거래위원회(Federal Trade Commission, 이하 FTC)에서는 기계적 알고리즘이 편향을 가질 수 있으며 차별을 발생시키고 심화시킬 수 있다는 문제를 인식하고 이에 대응하기 위한 하나의 방안으로 2020년 4월 <AI와 알고리즘 이용에 대한 지침(User Artificial Intelligence and Algorithms)>을 발표한 바 있다. 해당 지침은 기업이 알고리즘이 이용자의 권리를 보장하고 권익을 향상시키기 위해 어떻게 알고리즘을 관리해야 하는지 구체적이고 명확하게 방안들을 제시하고 있다. FTC는 기업이 AI 및 알고리즘을 사용하고자 할 때, ①투명해야 하며(Be Transparent) ②소비자에게 설명 가능해야 하며(Explain your decision to the consumer) ③공정해야 하며 (Ensure that your decisions are fair), ④견고성과 실증가능성을 갖춰야 하며(Ensure that your data and models are robust and empirically sound) ⑤책임성을 갖춰야 한다(Hold yourself accountable for compliance, ethics, fairness, and nondiscrimination)고 강조했다.

1) 투명성을 제고할 것(Be Transparent)

가장 먼저 FTC는 사용자와 상호작용하는 자동화된 알고리즘과 같은 툴을 사용함에 있어서 이용자를 기만하지 않아야 한다고 강조했다. 그리고 대규모 데이터를 모을수록 알고리즘의 기능이 향상되고 이용자를 위한 혜택도 늘어나지만 데이터셋을 수집하는 방법에 대해서 더욱 주의해야 한다고 강조했다. 오디오 및 비디오와 같은 민감한 정보를 포함한 데이터를 수집하는 것을 주의해야 하며, 데이터 수집과정이 투명해야 함을 강조했다. 마지막으로 신용평가사가 제 3자 공급업체의 정보를 기반으로 해서 자동화된 결정을 내리는 경우에는 소비자에게 해당 정보를 투명하게 알려야 하며 소비자가 이를 확인하고 부정확한 정보를 수정할 수 있는 권리를 가질 수 있게 해야 한다.

2) AI 및 알고리즘의 의사결정에 대해 소비자에게 설명할 수 있을 것(Explain your decision to the consumer)

두 번째로는 AI 및 알고리즘에 의해 수행된 의사결정에 대해서 소비자에게 설명해야 한다는 점을 강조했다. 세부 실행지침으로는 알고리즘에 의해서 소비자에게 불리한 결정이 내려졌을 경우, 이에 대해서 소비자에게 명확하게 설명해야 한다는 것이다. 또한, 알고리즘을 사용해서 어떤 결정을 내릴 경우에 결정에 영향을 준 요인을 투명하게 공개하고, 알고리즘에 의해서 거래조건이 변경되는 경우에도 소비자에게 반드시 알려야 한다는 세부지침을 제시하고 있다.

72) FEDERAL TRADE COMMISSION(2020.04.08.) <Using Artificial Intelligence and Algorithms>. URL : <https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms>

3) 결과의 공정성에 대해 보장할 것(Ensure that your decisions are fair)

세 번째로 공정성(Fairness)을 확보하기 위한 세부 지침을 제시했다. AI는 자칫하면 사회적 약자(취약계층)에 대한 차별을 일으킬 수 있기 때문에 이에 대해서 주의를 기울이는 것이 필요하다. 알고리즘을 사용하기 전과 사용한 후를 주기적으로 테스트하여서 사회적 취약계층에 대해서 차별이 발생하지 않도록 주의해야 한다. 또한 알고리즘에 입력되는 데이터에 보호속성에 대한 변수 혹은 대리변수가 투입되지 않는 것도 중요하지만 이와 별개로 도출되는 결과에서도 차별이 발생하지 않는지를 확인하고 관리할 필요가 있다는 점을 강조한다. 마지막으로 의사결정에 사용되는 정보를 수정하거나 접근할 수 있는 기회를 이용자에게 보장해야 한다.

4) 데이터와 모델에 대해서 강건성을 갖추고, 실증적인 타당성을 보장해야 할 것 (Ensure that your data and models are robust and empirically sound)

네 번째 실행지침에는 고객의 데이터를 자동화된 의사결정에 활용하는 제3자에게 제공하는 경우에 데이터의 정확성을 알려야 한다는 내용이 포함되었다. 또한 AI모델이나 알고리즘이 개발자가 의도한 방향으로 작동하고 혹시 차별적인 요소가 없는지 통계적인 원칙과 방법론을 사용하여 확인하고 재검증할 필요가 있다고 강조한다.

5) 법령준수, 윤리, 공정성 및 비차별성에 대한 책임을 갖추고 있을 것(Hold yourself accountable for compliance, ethics, fairness, and nondiscrimination)

마지막으로 알고리즘의 편향으로 인해서 이용자에게 피해가 발생하는 것을 방지하기 위해 알고리즘 운영자는 데이터가 대표성을 가지고 있는지, 편향성을 보정할 수 있는지, 빅데이터를 기반으로 한 예측이 얼마나 정확한지, 빅데이터에 의존했을 때 윤리적 혹은 공정성의 문제가 발생하지 않는지와 같은 자가질문을 수행해야 한다고 강조했다. 또한 AI나 알고리즘에 대한 무단사용/오용/남용을 방지해야 하며 책임을 수행할 수 있는 방법을 고려해야한다.

다. 국내 지능정보사회 윤리가이드라인⁷³⁾

기계적 알고리즘이 인간의 규범과 행위에 큰 영향을 미치게 되면서 기존의 규범체계의 근본적인 변화가 예상된다. 이러한 변화에 대응하기 위해 국내에서는 과학기술정보통신부와 한국정보화진흥원에서 AI와 같

73) 과학기술정보통신부·한국정보화진흥원(2020). <지능정보사회 윤리가이드라인>

은 지능정보기술에 대한 윤리가이드라인을 발표한 바 있다. 이 윤리가이드라인의 특징은 지능정보사회의 행위주체를 개발자, 공급자, 이용자로 구분하고 행위주체별로 세부 규범내용을 포함하고 있다는 점이다. 우선 지능정보기술이 지켜야할 공통원칙을 제시하고 있는데 공공성(Publicness), 책무성(Accountability), 통제성(Controllability), 투명성(Transparency)이 해당된다. 공공성이란 지능정보기술이 가능한 다수에게 도움을 주어야 한다는 개념이며, 책무성이란 지능정보기술로 인한 사고 등의 책임분배를 명확하게 해야 한다는 개념이다. 통제성은 오작동에 대한 대비책을 마련하고, 이용자의 이용선택권을 최대한 보장해야 한다는 의미이며 마지막으로 투명성은 지능정보기술에 이용자, 소비자, 시민의 의견이 반영되어야 하며 이용단계에서 관련된 정보를 공개하고 공유되어야 한다는 점이다. 4가지 기본원칙을 바탕으로 하여 행위주체별로 세부지침을 제시하고 있다.

주체별로 세부 지침들을 살펴보면, 개발자는 공공성을 확보하기 위해 기술개발 시 사회적 차별요소를 배제하고 사회적 약자나 취약계층을 위한 접근성을 보장해야 한다. 그리고 윤리적 절차를 충실히 이행하여 연구하거나 개발하고, 전 과정에 대한 책임을 공유하여 책무성을 확보해야 한다. 통제성을 갖추기 위해서는 지속적인 품질관리와 기술적 제어장치를 마련해야 하고, 투명성을 확보하기 위해서 위급상황 시 필요데이터를 제공하거나, 은닉기능을 개발하지 않아야 한다. 공급자는 공공의 이익에 부합하는 제품을 공급하여 공공성을 확보할 수 있도록 해야 하며, 이용자의 권리를 보장하고 지능정보서비스의 의사결정의 조건과 범위를 확립하여 책무성을 가질 수 있도록 해야 한다. 또한 지능정보기술을 활용한 제품 혹은 서비스에 대해서 검증기준과 통제조치를 마련하고 이용자의 선택권을 보장하여 통제성을 갖출 수 있도록 해야 하며 마지막으로 이용자 정보를 부당하게 이용하지 않으며, 제품이나 서비스가 미치는 부정적인 영향을 사전에 파악하여 반영하여 투명성을 확보해야 한다. 마지막으로 이용자는 공공성의 측면에서 악의적 이용을 하지 말아야 하며, 책무성의 측면에서 이용자의 윤리책임을 숙지하고 있어야 하고, 통제성의 측면에서 지능정보기술에 대한 이용역량을 강화해야 하고, 투명성의 측면에서 개인정보활용에 대한 감시하는 역할을 수행해야 한다. 해당 윤리가이드라인은 정부가 포함된 민관협력차원의 윤리적 대응접근이라는 점과 행위주체별로 세부적 규범내용을 포함한 종합적인 접근이라는데 의의가 있다.

3. 알고리즘 영향성 평가

알고리즘 영향성 평가(AIA, Algorithmic Impact Assessment)는 EU 미래과학기술 분과의 라이스만과 동료들이 2018년 New York University의 AI Now Institute 보고서에서 제안한 프레임워크다(Reisman et al., 2018)⁷⁴⁾. 이는 Public Agency에 대해 알고리즘 설명가능성 및 책무성 부과의 기준과 거버넌스를 제시할 알고리즘 영향평가의 필요성을 주장한다.

74) Reisman, D., Schultz, J., Crawford, K., & Whittaker, M.(2018.4. Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability. AI Now.

이는 정부활동에 따른 사회 및 관련 공중에 대한 영향을 평가하고 점수를 매기는 영향성 평가(Impact Assessment) 프레임워크에 기반하고 있다. 여타 영향성 평가 도구가 공중 및 외부 연구자들에게 공유되지 않는 반면, AIA는 다양한 Public Agency와 사람들을 평가 절차 안에 적극 참여시키고 검토 및 토의하게 하여 광범위한 개인, 커뮤니티, 연구자, 정책입안자들이 책무성 확보에 기여할 수 있게 한다는 데에 의의가 있다.

알고리즘 영향성 평가 프레임워크의 수행과정은 다음과 같이 크게 다섯 단계로 이루어진다: ①시행 전 리뷰 단계(Pre-acquisition review) ②초기 기관정보 공개 요구 단계(Initial agency disclosure requirements) ③비평 단계(Comment period) ④법적 절차 검토 단계(Due process challenge period) ⑤ AIA 갱신 단계(Renewing AIAs).

먼저 시행 전 리뷰 단계에서는 알고리즘 규제 범위를 정부기관과 개발업자와의 사전협의를 통하여 결정하게 된다. 구체적으로, 기관과 공중 양측에게 사전에 자동화된 의사결정 시스템의 적용에 대해 평가할 수 있게 해줌으로써 기관과 공중이 시스템 도입에 대해 협상해야 하는 부분을 파악할 수 있게 해주는 것이다. 이후 알고리즘 실행 이전 단계에서 정부 측 전문가와 알고리즘 개발 주체가 만나 해당 알고리즘의 사용 목적, 데이터 수집의 범위와 방식, 설계 디자인, 결과산출 방식, 잠재적인 위협요인(편향이나 오류의 가능성) 및 예상되는 혜택 등에 대해 논의하고 자체 평가를 통하여 규제 대상 여부를 결정하게 된다. 세 번째 비평 단계에서는 문서화된 형태로 시민사회에 논의 결과를 공유하고 일정 기간 시민사회의 외부 전문가들로부터 자체 평가에 대한 피드백을 받는다. 논의된 사항 이외의 알고리즘의 심각한 위해 요소가 발견되는 경우 개발업자에게 개선안을 요구하게 된다. 네 번째 법적 절차 검토 단계에서는 정부기관의 주도 아래 개발업자와 시민사회와의 절충점을 모색하게 되며, 마지막 AIA 갱신 단계를 통해 평가 과정을 주기적으로 이루어 알고리즘 일정 부분의 변경으로 발생할 수 있는 잠재적 위협에도 대처할 수 있게 된다.

이와 같은 프레임워크를 차용한 대표적인 기관은 캐나다 정부의 사례에서 찾을 수 있다. 캐나다 정부는 자동화된 의사결정(Automated Decision-Making)의 공공 및 일상 부문 적용에 대비하여 알고리즘의 투명성, 책무성, 입법성, 절차적 공정성 등에 대한 정책적 원칙(Directive on Automated Decision-Making)을 가지고 있는데, 이 원칙을 지지하는 도구 중 하나로 AIA 측정 도구를 제안하고 있다.

Algorithmic Impact Assessment v0.8

Page 3 of 13

Risk Profile
Is the project within an area of intense public scrutiny (e.g. because of privacy concerns) and/or frequent litigation?

☒ Yes
☐ No

Are clients in this line of business particularly vulnerable?

☒ Yes
☐ No

Are stakes of the decisions very high?

☐ Yes
☒ No

Will this project have major impacts on staff, either in terms of their numbers or their roles?

☒ Yes
☐ No

Previous Next Complete

Impact Level: 1 Current Score: 9 Raw Impact Score: 9 Mitigation Score: 0

[그림 III-4] 캐나다의 알고리즘 영향성 평가(Algorithmic Impact Assessment)

2020년 8월 31일 기준으로 캐나다 정부 홈페이지에서 AIA의 베타 버전⁷⁵⁾이 오픈소스 형태로 공개되어 있어 누구나 테스트에 참여할 수 있을 뿐 아니라, 공개된 Github⁷⁶⁾를 통해 개발 및 오류사항 보고 등에 참여할 수도 있다. 35분 이내의 시간 동안 자신이 진행하고 있는 프로젝트에 대한 60여개의 질문에 대해 답을 하고, 그 결과에 따라 알고리즘 영향성 관련 점수를 도출한다.

75) open.canada.ca/aia-eia-js/?lang=en

76) github.com/canada-ca/aia-eia-js/issues

제4장 시뮬레이션 기반 실험방법론의 적용과 공정성 평가

제1절 추천시스템 공정성 관련 선행 실험

추천시스템의 공정성에 관한 실증적 연구는 극히 제한적이다. 이유는 학계에서는 실제 추천 알고리즘이나 추천 내역에 대한 정보를 얻는 것이 불가능에 가깝기 때문이다. 그렇기 때문에 추천시스템의 공정성에 대한 기존 연구들은 주로 합성 데이터(synthetic data)를 이용한 협업 필터링의 시뮬레이션을 통해 진행되어왔다.⁷⁷⁾⁷⁸⁾

콘텐츠 기반 필터링의 경우 사용할 수 있는 데이터 필드와 그를 가공하는 방식이 다양하여 표준적인 방식이 없다. 그에 비하여, 협업 필터링의 작동 방식은 대부분 행렬 분해에 기반하고 있어 알고리즘이 정형화된 편이며, 시뮬레이션 실험을 위한 합성 데이터를 구축하기 용이하다.

특히 협업 필터링의 공정성에 대한 시뮬레이션 연구들은 공급자 측면보다는 사용자 측면에서의 공정성을 주로 다루었는데, 공급자 측면의 공정성은 품질을 반영해야 하기 때문에 기계적 공정성을 적용하기 어려운 반면, 사용자 측면의 공정성은 (오차수준의 균일화 같은) 기계적 공정성의 적용이 용이하기 때문이다.

본 연구에서는 가상 데이터를 이용하여 ‘전처리 단계에서의 입력 데이터 불균형 보정을 통한 추천 결과물 편향성 보정 방법’을 제안하며, 이를 실제 MovieLens 데이터에 적용하여 평가할 것이다. 본 실험연구에서 사용되는 편향성에 관한 정량적 정의는 다음의 주요 관련 연구에 기반하고 있다. 아래의 주요연구 2는 추천 알고리즘의 최적화 방법 변경, 주요연구 3과 4는 후처리 과정을 통해서 편향성을 보정하고자 하였으나, 본 연구는 전처리 단계에서의 작업을 통해 편향성을 보정하는 점이 차별점이다.

1. (주요연구1) Gajane, P. and Pechenizkiy, M.(2018) 연구

<On Formalizing Fairness in Prediction with Machine Learning. FATML: Workshop on Fairness, Accountability, and Transparency>

딥러닝/기계학습 분야에서의 ‘공정성’은 개인이나 집단을 차별하지 않는 것을 의미한다. 특히 대상이 어떤 분류에 속하는지 예측하는 분류 문제는 예측을 받는 대상에 의해 공정성이 정의된다. 약자들의 법으로 보호받아야 하는 특징을 보호속성(protected attributes)이라고 정의한다. 여기에는 성별, 인종, 종교, 지역, 장애, 가족의 상태 등이 있을 수 있다.

77) Tsintzou, V. et al. 2018. Bias Disparity in Recommendation Systems. DAB: Workshop on Data and Algorithm Bias. (Nov. 2018), 1-4.

78) Mansoury, M. et al. 2019. Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison. RMSE: Workshop on Recommendation in Multistakeholder Environments. (Aug. 2019), 1-7.

‘공정성’은 하나로 정의할 수 없으며, 어떤 시스템에 어떤 대상을 위하여 어떻게 정의하나에 따라 여러가지 정의를 내릴 수있다. Gajane과 Pechenizkiy는 기계학습모델의 예측에 있어서 ‘공정성’의 기존 4가지의 개념을 소개한다. 첫째, 모델이 예측 과정에서 보호속성(protected attributes)을 명백하게 사용하지 않으면 그 모델은 ‘비자각에서 오는 공정성(fairness through unawareness)’을 가진다고 한다. 둘째, 여러 대상들의 보호속성만 다르게 변경했을 때, 같은 예측결과가 나온다면 이를 ‘반사실적으로 공정(counter-factually fairness)’하다고 한다. 셋째, ‘집단 공정성(group fairness)’는 같은 그룹 내에서는 예측 모델이 거의 동등한 확률로 결과를 예측해야 한다는 것을 의미한다. 넷째는 ‘개인 공정성(individual fairness)’로 유사한 개인은 예측 결과값도 유사해야 한다.

추가로 기존 기계학습 분야에서 잘 고려하지 않았던 2가지의 공정성 개념을 제안하였다. 첫째는, ‘자원의 균등성(equality of resources)’으로 특정 개인들의 의도적인 결정과 행동에서 나오는 결과일 때만, 사회적 혜택의 불균등한 분배가 공정하다. 즉, 각 개인이 자신들의 이익을 위해 의도적으로 나오는 야망과 선택에는 자원의 균등성이 적용될 수 있으나, 개인들이 선택하지 않은 선천적으로 타고난 재능 같은 경우에는 고려하지 않아도 된다. 둘째는, ‘기능 가능성의 균등성(equality of capability of functioning)’이다. 이는 존재하거나 행하는 상태에 대해 사회적으로 타고난 특징들은(예. 나이, 성, 인종, 계층 등) 같은 기회에 있어서 동일하지 않은 힘을 가지도 있다고 해도 모든 사람들이 만족해야 한다는 것을 의미한다.

2. (주요연구2) Yao, S. and Huang, B(2017) 연구

<Beyond Parity: Fairness Objectives for Collaborative Filtering. NIPS: Advances in Neural Information Processing Systems>

머신러닝 분야 최고 학회 중 하나인 NIPS에 발표된 논문으로, 협업 필터링 기반 추천 알고리즘의 공정성을 연구하였다. 기존 공정성 정의 및 측정법이 충분하지 않음을 밝히고, 4개의 새로운 공정성 정의를 제안한다. 이들을 이용하여 보호해야하는 약자 혹은 소수 집단의 예측 오차 값(prediction error)이 그렇지 않은 집단의 예측 오차값과 비슷할 때 공정하다고 정의한다. 이러한 공정성 정의를 기계학습 모델의 학습 목표(learning objective)에 추가하면 불공정성(unfairness)을 최소화하는 추천 알고리즘을 학습시킬 수 있다. 또한 학습목표에 공정성 정의를 새롭게 추가함으로써 과대평가와 과소평가를 줄일 수 있다는 것을 보여주었다. 실험은 가상의 데이터와 실제 데이터(MovieLens 1M dataset)⁷⁹⁾ 모두를 사용하였으며, 이들이 새롭게 제안하는 공정성 정의가 기존 공정성 정의보다 좋으며, 모델을 더 공정하게 학습시킬 수 있다고 주장한다.

79) MovieLens라는 사이트에서 6,040명의 사용자와 3,883개의 영화에 대해 사용자 별로 선호하는 영화를 수집한 데이터로 <https://grouplens.org/datasets/movielens/1m/>에 공개되어있다.

3. (주요연구3) Tsintzou, V 외 (2018) 연구

<Bias Disparity in Recommendation Systems. DAB: Workshop on Data and Algorithm Bias>

엔터테인먼트, 상업, 고용 등의 다양한 분야에 사용되는 추천시스템에서의 ‘공정성’ 문제도 중요해지고 있다. 추천시스템은 사용자의 취향을 고려하여 가장 좋아할 만한 항목을 추천해주는 시스템이다.

기존 연구와는 달리 단순히 학습된 모델의 불균형만 고려하는 것이 아니라, 입력 데이터의 편향 불균형을 고려하는 연구도 진행되었다. 입력 데이터로 사용되는 사용자의 취향에 이미 편향 불균형이 있다면, 추천시스템의 결과물에도 불균형이 있을 수 있기 때문이다.

가상 데이터와 간단한 추천 모델을 활용하여 여러 조건들 아래에서의 주어진 데이터의 불균형과, 추천시스템 결과물의 불균형 사이의 상관관계를 관찰하였다. 이 연구에서 강조하는 것은 단순히 입력 데이터의 불균형이 그대로 결과의 편향 불균형으로 나타난다는 것이 아니라, 사용자의 불균형을 추천시스템이 더 증폭시킬 수 있다는 점이다. 더해서 거대한 편향값이 추천시스템의 입력 데이터로 사용되면 이는 강화 효과(reinforcing effect)를 일으켜, 장기적으로 사용자를 극단화(polarize)하는 부정적인 효과를 가져올 수 있다는 점도 강조한다.

추가로 이 연구에서는 간단히 편향 불균형을 줄일 수 있는 후처리 랭킹 알고리즘(re-ranking algorithm)을 제안하였다. 또한 MovieLens 1M data를 이용해 실제 데이터에서의 불균형이 더 큰 불균형을 일으킨다는 것을 관찰하였다.

4. (주요연구4) Mansoury, M. 외 (2019) 연구

<Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison. RMSE: Workshop on Recommendation in Multistakeholder Environments>

이전의 추천시스템에서의 편향 불균형 연구⁸⁰⁾에서 더 나아가, 여러 알고리즘의 편향 불균형(bias disparity)과 추천 성능(ranking quality)의 트레이드 오프(tradeoff)를 측정해본 연구가 진행되었다. 기존 연구에서 간단한 추천 모델을 이용한 것에 반해, 이번 연구는 비슷한 이웃기반(neighborhood-based), 모델 기반(model-based), 신뢰 기반(trust-aware) 추천 알고리즘의 총 3가지 최신 알고리즘 종류에 포함되는 13가지 추천 알고리즘을 이용하여 실험하였다. 데이터 셋으로는 Yelp 데이터셋⁸¹⁾을 이용하였다.

실험은 총 두 단계로 진행되었다. 10개의 가장 선호하는 항목 카테고리에 대한 추천 결과의 편향 불균형을 보았고, 각 추천 알고리즘에 대해 모든 카테고리에 대한 평균 편향 불균형 정도를 알아보았다. 그

80) Tsintzou, V. et al. 2018. Bias Disparity in Recommendation Systems. arXiv.

81) Yelp Challenge의 데이터 셋으로 209,393개의 사업체에 대해 1,968,703명의 사용자의 8,021,122개의 리뷰를 포함하고 있다. 여러 사업들에 대한 사용자의 평점 및 평가 및 사용자 정보 등을 수집하였다. <https://www.yelp.com/dataset>에서 확인할 수 있다.

결과 이웃 기반 알고리즘이 다른 알고리즘에 비해 현저히 좋은 결과를 보였다. 또한 대부분의 경우 평점 데이터 이외에 다른 정보를 사용하는 것이 추천시스템의 성능을 향상시킨다는 것을 보였다.

제2절 추천시스템 공정성 실험설계와 절차

1. 연구문제 및 연구방법

가. 시뮬레이션 실험연구의 목적

본 연구진은 협업필터링 기반의 추천시스템이 이용자 측면에서 갖는 편향성과 공정성에 초점을 두었다. 본 연구에서는 Tsintzou와 동료들(2018) 및 Mansoury와 동료들(2019)이 사용한 정의를 따라 추천시스템의 편향성을 이용자그룹별 입력 데이터에서의 아이템 카테고리의 비율과 추천 결과물에서의 해당 비율의 차이(bias disparity, 자세한 정량적 정의는 후술)로 규정한다. 또한, 이에 기반하여, 입력 데이터에서 이용자 그룹 규모의 불균형으로 인해 발생하는 추천시스템의 편향성을 보정하는 방법을 제안하고 그 효과를 측정한다.

나. 연구문제

해당 연구에서는 협업필터링 기반의 추천시스템이 이용자 측면에서 가지는 편향성과 공정성을 파악하기 위해 아래와 같은 연구문제를 구성했다.

연구문제 1. 가상데이터를 이용한 실험환경에서, 협업 필터링 기반의 추천시스템은 두 이용자 그룹에 대하여 이용자 수의 불균형 정도에 따라 추천결과물의 편향성에 차이를 보이는가?

연구문제 2. 두 이용자 그룹의 규모에 따른 추천결과물의 편향성의 차이가 존재할 경우, 인위적인 이용자 데이터 증대 방법을 통해 편향성의 차이를 줄일 수 있는가?

연구문제 3. 제안된 데이터 증대 방법의 추천시스템 편향성 보정 효과는 가상 데이터를 이용한 실험환경과 실제 데이터에 적용한 실험환경에서 어떻게 다른가?

다. 연구방법

본 실험연구에서는 가상으로 생성한 행동기록 데이터(synthetic data generation)를 이용해 입력데이터의 전처리 과정에서 두 이용자 그룹의 불균형도를 보정하는 것이 추천결과물의 편향성의 차이에 끼치는 영향

을 관찰하며, 해당 방법론을 MovieLens 실제 데이터에 적용한다. 대략적인 연구방법은 아래와 같이 요약할 수 있으며, 세부적인 사항은 제3절을 통해 서술할 것이다.

- 1) 가상 데이터의 구축 : Tsintzou와 동료들(2018)이 제안한 데이터 합성 방법에 기반하여, 두 사용자 그룹 G_1 , G_2 와 두 아이템 그룹 C_1 , C_2 를 갖는 가상의 사용자-아이템 행동기록 행렬(user-item matrix)를 만든다. 이때, 두 사용자 그룹의 규모 $|G_1|:|G_2|$ 를 1:19, 2:18, 3:17, ..., 19:1의 19 단계로 조절하여 사용자 수의 불균형 정도를 조작한다.
- 2) 추천결과물 편향성 측정 : 위의 과정에서 만든 각각의 행렬에 대해 가장 기본적인 협업필터링 기반 추천 알고리즘인 User-kNN 알고리즘과 SVD 기반 알고리즘을 적용하여 추천시스템의 편향성을 측정한다.
- 3) 데이터 증대 방법을 통한 편향성 보정 효과 측정 : 두 사용자 그룹의 규모 $|G_1|:|G_2|$ 를 보정하는 4가지의 방법을 제안하고, 각 방법을 적용했을 때, User-kNN 알고리즘과 SVD 기반 알고리즘에서 편향성을 측정한 후, 2)에서 측정된 편향성과 비교한다.
- 4) 실제 데이터에 적용 : 가상데이터를 통해 실험한 데이터 증대 방법을 MovieLens 데이터에 적용하여 편향성 보정 효과를 측정한다.

제3절 실험결과와 함의

1. 정의 및 모델

가. 선호도 및 편향성 정의

n 명의 사용자 집단을 U , m 개의 추천 항목 집합을 I 라고 정의한다. 이를 통해 $n \times m$ 크기의 연관 행렬 S 에 대해 사용자 u 가 항목 i 를 선호했으면 $S(u,i)=1$, 그렇지 않으면 0인 행렬을 만들 수 있다. 여기서 선호한다는 것은 사용자 u 가 어떤 포스팅 i 에 좋아요를 클릭함, 어떤 물건 i 를 구매함, 영상 i 를 시청함 등이 될 수 있다.

이 사용자들이 각자 A_u 라는 특징을 지닌다고 가정한다. 예를 들면 이 특징은 성별 혹은 나이가 될 수 있다. 그리고 같은 특징을 지닌 사용자들을 같은 집단으로 묶어서, 전체 사용자 집단을 여러 집단으로 나눌 수 있다. 이 실험에서는 간단히 하기 위해 사용자를 2개의 집단으로 나눈다. 한 집단은 보호집단/소수집단(protected group)이 된다. 이와 같은 방식으로 항목들 또한 각각 특징 A_i 를 지닌다고 가정하고, 이 특징을 기반으로 같은 특징을 지니는 항목들끼리 분류하여 여러 카테고리로 만든다. 즉 한 카테고리 내에는 같은 특징을 지니는 항목들이 존재한다. 예를 들면, 특징 A_i 를 영화의 장르라고 했을 때, 로맨스와 액션 영화 카테고리

리로 나눌 수 있다.

주어진 연관 행렬 S 에 대해, 우리는 사용자 집단 G 의 카테고리 C 에 대한 선호도(preference ratio) $PR_S(G, C)$ 를 다음과 같이 정의할 수 있다. 이는 사용자 집단 G 의 전체 카테고리의 항목에 대한 선호도 정도에 대한 사용자 집단 G 의 카테고리 C 에 대한 선호도 비율이다.

$$PR_S(G, C) = \frac{\sum_{u \in G} \sum_{i \in C} S(u, i)}{\sum_{u \in G} \sum_{i \in I} S(u, i)}$$

또한 이 선호도 $PR_S(G, C)$ 를 $P(C) = |C|/m$, 즉 모든 항목 중에서 임의로 선택했을 때 카테고리 C 에 속하는 항목을 선택할 확률과 비교하면 편향성을 파악할 수 있다. 사용자 집단 G 의 카테고리 C 에 대한 편향성 $B_S(G, C)$ 는 다음과 같은 식으로 나타낼 수 있다.

$$B_S(G, C) = \frac{PR_S(G, C)}{P(C)}$$

이는 임의로 카테고리 C 에 속하는 항목을 고를 확률에 대한 사용자 집단 G 가 카테고리 C 의 항목을 고를 확률이다. 이 값이 1보다 작으면, 음성 편향(negative bias)으로 해당 집단 G 는 평균적으로 카테고리 C 의 항목을 선호할 경향이 적다는 것을 의미한다. 반면에 이 값이 1보다 크면, 양성 편향(positive bias)으로 해당 집단 G 가 평균적으로 카테고리 C 의 크기와 상관없이 카테고리 C 의 항목을 선호할 경향이 크다는 것을 의미한다.

추천 알고리즘의 결과물로 사용자 u 의 r 개의 추천 항목을 지니는 행렬을 정의할 수 있다. 모든 사용자에게 대한 추천 결과의 집합은 이진 행렬인 R 로 나타낼 수 있다. 행렬 R 은 사용자 u 에 대해 항목 i 를 추천하면 $R(u, i) = 1$ 로 표기하고, 그렇지 않으면 0값을 지닌다. 행렬 R 을 이용하여 추천 알고리즘 결과물의 선호도와 편향성인 $PR_R(G, C)$ 와 $B_R(G, C)$ 를 계산할 수 있다.

입력데이터와 추천된 결과물에 대한 사용자 집단 G 의 카테고리 C 의 편향성 차이를 비교하기 위해서, 기존 연구[4]에서 사용되었던 편향성 차이(bias disparity) 정의를 사용한다. 이는 편향성 값에 대한 상대적 변화 정도로, 다음과 같이 계산할 수 있다. 편향성 차이는 0에 가까울수록 해당 추천시스템이 편향성이 없다는 것을 의미하며, 추천시스템의 편향성 정도는 편향성 차이의 절댓값으로 그 정도를 비교할 수 있다. 편향성 차이는 양수 혹은 음수가 나올 수도 있으며, 양성은 기존 사용자의 편향성 방향으로 증폭시킨 것을 의미하며, 음성인 경우 해당 추천시스템이 사용자의 편향을 감소하는 방향으로 추천을 해주었다는 것을 의미한다.

$$BD(G, C) = \frac{B_R(G, C) - B_S(G, C)}{B_S(G, C)}$$

위의 선호도 및 편향성 정의는 모두 기존 추천시스템에서의 편향성 차이 연구에서 자주 사용된 정의[4]를 채택하였다.

나. 추천 알고리즘 정의

이번 실험에서는 기존 연구에서 사용되었던 User-KNN 알고리즘 외에 추가로 SVD 알고리즘을 사용하여 진행되었다. 이는 하나의 특수한 추천 알고리즘에만 적용되는 것이 아니라 다른 추천 알고리즘에도 실험 결과가 적용된다는 것을 보여준다.

1) User-KNN 알고리즘(K-Nearest-Neighbors Algorithm)

User-KNN 알고리즘은 사용자 기반 추천 알고리즘으로 K-최근접 이웃 알고리즘이라고 불린다. 이를 간단히 설명하자면 사용자 u에게 항목을 추천해주기 위해, 사용자 u와 비슷한, 즉 과거에 비슷한 항목에 대해 선호도를 보인, k명의 이웃 사용자들을 고르고, 이 k명의 이웃 사용자들이 선호했던 항목들을 추천해주는 것이다. 이를 식으로 나타내면 다음과 같다.

$$V(u,i) = \frac{\sum_{n \in N_k(u)} JSim(u,n)S(n,i)}{\sum_{n \in N_k(u)} JSim(u,n)}$$

각 사용자 u에 대해, k명의 가장 유사한 사용자를 찾고 이 사용자의 이웃 집단을 $N_k(u)$ 라고 정의한다. 유사도 계산 방법에는 여러 가지가 있지만, 이번 실험에서는 가장 많이 사용되는 자카드 유사도(Jaccard similarity)를 사용한다. 이 자카드 유사도는 행렬 S를 이용해 쉽게 계산할 수 있다. $V(u,i)$ 는 유틸리티 값으로 사용자 u의 k명의 최근접 이웃들이 해당 항목 i를 골랐을 확률이다. 사용자 u가 기존에 선호한다고 하지 않았던 항목들에 대하여 유틸리티 값이 높은 r개의 항목을 사용자 u에게 추천한다.

2) SVD 알고리즘 (Singular Value Decomposition Algorithm)

SVD 알고리즘은 행렬 인수분해(matrix factorization) 방법 중의 하나로, 행렬 인수분해는 추천 알고리즘의 협력 필터링(collaborative filtering) 방법 중 하나이다. 행렬 인수분해는 크기가 큰 행렬을 작은 행렬들의 곱으로 유사하게 나타내어 계산과 저장을 용이하게 한다. 또한 작은 행렬들의 곱을 통해 원래의 행렬과 유사한 행렬로 복귀하고, 이 복귀한 행렬을 이용해 항목들을 추천할 수 있다.

SVD 알고리즘은 원래의 행렬 $m \times n$ S를 3개의 행렬의 곱인 $U\Sigma V^T$ 로 나타낸다. 여기서 U는 AAT를 고유값 분해(eigendecomposition)해서 얻은 $m \times m$ 크기의 직교행렬, V는 ATA를 고유값 분해(eigendecomposition)해서 얻은 $n \times n$ 크기의 직교행렬, Σ 는 AAT, ATA를 고유값 분해해서 나오는 고유값(eigenvalue)들의 제곱근을 대각 원소로 $m \times n$ 크기의 대각행렬이다. 여기서 Σ 의 대각원소 중에 작은 값을

제거하여 행렬이 $k \times k$ 크기가 되도록 줄이고, U 와 V 도 해당하는 행과 열을 제거하면 각각 Σ' , U' , V' 을 얻을 수 있다. 이 새롭게 얻어진 행렬을 다시 곱해 $U'\Sigma'V'T$ 를 구하면 기존 S 와 유사한 행렬 S' 을 얻을 수 있다. 여기서 얻어진 사용자 u 에 대한 항목 i 값 $S'(u,i)$ 를 이용하여, 각 사용자 u 에 대해 기존에 선호하지 않는다고 했던 모든 항목 i 에 대해 $S(u,i)$ 값이 높은 r 개의 항목을 추천할 수 있다. 이 실험에서는 $k=100$ 값을 사용한다.

2. 가상의 데이터를 이용한 편향성 차이 실험 결과

가. 가상 데이터 생성

전체 사용자 집단을 n_1 명의 사용자를 가지는 G_1 집단과 n_2 명의 사용자를 지니는 G_2 집단으로 나눈다. 이와 유사하게 전체 항목 집단을 m_1 개의 항목을 가지는 C_1 카테고리와 m_2 개의 항목을 지니는 C_2 카테고리로 나눈다. 이 실험에서 사용자 집단 G_1 은 C_1 카테고리의 항목을 선호하고, 사용자 집단 G_2 은 C_2 카테고리의 항목을 선호한다고 가정한다. 이 선호도를 정량화하기 위해 데이터 생성 과정에서 ρ_1 , ρ_2 의 두 개의 매개변수를 추가로 사용한다. ρ_i 는 사용자 집단 G_i 의 카테고리 C_i 에 대한 선호도 $PR_S(G_i, C_i)$ 를 의미한다. 예를 들어, $\rho_1 = 0.7$ 은 G_1 의 70%의 선호했던 항목이 C_1 에 속한다는 뜻이다.

각 실험에는 총 1000명의 사용자와 1000개의 항목을 생성했다. 또한 각 사용자 별로 1000개의 항목의 5%에 해당하는 50개의 항목을 과거에 선호한 항목 데이터로 설정했으며, 추천시스템을 통해 $r=10$ 개의 항목을 새롭게 추천해준다. 3의 실험은 모두 User-KNN 알고리즘을 이용해 추천 알고리즘을 구현하였으며, $k=10, 20, 50$ 으로 User-KNN에도 변화를 주어 총 3가지 알고리즘을 비교한다. 모든 실험은 총 3번의 실험 결과의 평균치 값이다.

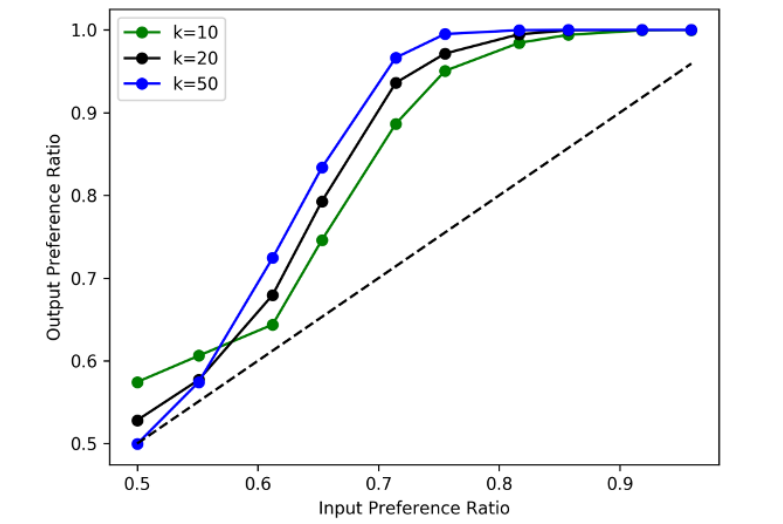
입력 데이터 값인 사용자의 선호 데이터가 편향되어 있을 때, 추천시스템을 통해 추천된 항목들은 더 심하게 편향된다는 것을 보이기 위해 크게 두 가지 실험을 진행한다. 첫 번째 실험은 선호도의 역할을 파악하기 위함이고, 두 번째 실험은 사용자 집단이나 카테고리 크기의 역할을 파악하기 위해 진행하였다. 여기서 집단이나 카테고리의 크기는 해당 집합에 속하는 개체의 수를 의미한다.

나. 선호도 변화에 따른 실험 결과

이 실험에서는 동일한 크기를 가지는 두 개의 집단 G_1 , G_2 과 동일한 크기를 가지는 두 개의 카테고리 C_1 , C_2 를 가지는 데이터를 생성하여 실험을 진행한다. 각 그룹의 선호도를 변화시켜 추천시스템을 통해 나오는 결과물의 선호도와 비교해본다. 또한 실험은 두 그룹이 동일한 선호도를 가지는 경우와 그렇지 않은 두 경우를 나눠 진행한다.

1) 두 사용자 집단이 대칭적 선호도를 가지는 경우

이 실험에서 두 집단인 G_1 , G_2 가 동일한 선호도 $\rho_1 = \rho_2 = \rho$ 를 지닌다고 가정한다. ρ 는 0.5이상 1 이하의 값으로 0.05의 간격을 가진다. [그림IV-1]은 사용자의 선호도 ρ 에 따른 추천 결과물의 선호도 $PR_R(G_1, C_1)$ 그래프이다. 이 실험의 경우 정의에 의해 선호도를 2로 나눈 값이 편향성과 동일하므로, 이해를 쉽게 하기 위해 그래프에서는 편향성 대신에 선호도를 사용하였다. [그림IV-1]에서 점선은 입력 데이터의 선호도 및 편향성이 추천시스템 결과물의 선호도 및 편향성과 동일할 경우를 나타내므로, 편향성 차이가 없는 경우를 나타낸다.



[그림IV-1] 사용자의 선호도에 대한 추천 결과물 선호도

입력데이터의 편향성이 작은 경우($PR_S \leq 0.6$), 추천결과물의 편향성은 감소하거나 입력 데이터의 편향성과 유사한 값을 가진다. 이 경우에, User-KNN 알고리즘에서 항목 추천을 받는 과정에서 사용자들은 두 그룹 모두에 이웃을 가지고 있을 것이다. 반면에 입력 데이터의 편향성이 큰 경우 ($PR_S > 0.6$), 추천결과물의 편향성은 가파르게 상승하여, $PR_S = 0.8$ 인 경우 최대값을 가진다. 이를 통해 입력 데이터의 편향성이 큰 경우, 추천 시스템은 두 집단을 양극화하여 추천 항목을 각자의 선호하는 카테고리에서만 추천해주는 것을 확인할 수 있다.

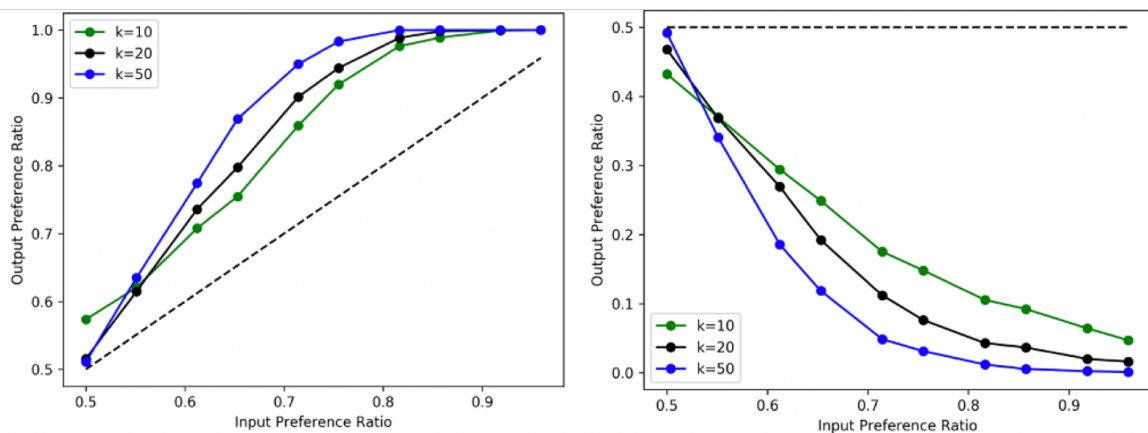
또한 이 실험에서 사용된 User-KNN 알고리즘에서 고려하는 최근접 이웃 수인 k 를 증가시키면 추천 결과물의 편향성이 증가한다. 많은 이웃들을 고려할수록, 추천하는 항목들의 편향성의 강도를 더 커지게 하며, 알고리즘은 다른 카테고리의 항목들을 더 잘 구별할 수 있게 된다.

2) 두 사용자 집단이 비대칭적 선호도를 가지는 경우

이 실험에서는 한 사용자 집단은 편향성을 가지도록, 나머지 하나의 집단은 편향성을 가지지 않도록 하여, 추천 결과물의 편향성을 비교해보는 실험이다. 즉 한 사용자 집단 G_2 의 선호도는 0.5로 고정시키고, 나머지 하나의 사용자 집단 G_1 이 선호도 ρ_1 를 0.5에서 1 사이 값을 가지도록 바꾸어, 추천 결과물의 편향성을 비

교해본다. [그림IV-2]의 왼쪽 그래프에서 ρ_1 에 따른 추천 결과물의 G_1 의 C_1 에 대한 선호도를 파악할 수 있고, 오른쪽 그래프에서 ρ_2 에 따른 추천 결과물의 G_2 의 C_2 에 대한 선호도를 파악할 수 있다.

이 실험을 통해 집단 모두 편향된 상태보다, 한 집단만 편향된 경우 그 집단의 추천 결과물의 편향성이 크게 나타남을 관찰할 수 있다. 놀랍게도 편향되지 않은 집단 G_2 가 전체 편향성 증폭 정도를 완화하는 것이 아니라 오히려 편향된 집단 G_1 의 편향성을 증폭시킨다는 점이다. 더해서 [그림IV-2]에서 볼 수 있듯이 편향되지 않은 집단은 편향된 집단의 편향성을 반영한다. 이는 추천 알고리즘이 편향되지 않은 집단 G_2 가 C_2 를 선호한다고 했던 실험보다 C_1 을 선호하는 것으로 인식하여 카테고리 C_1 에 대한 전체적인 선호도 및 편향성을 증가시키기 때문이다.



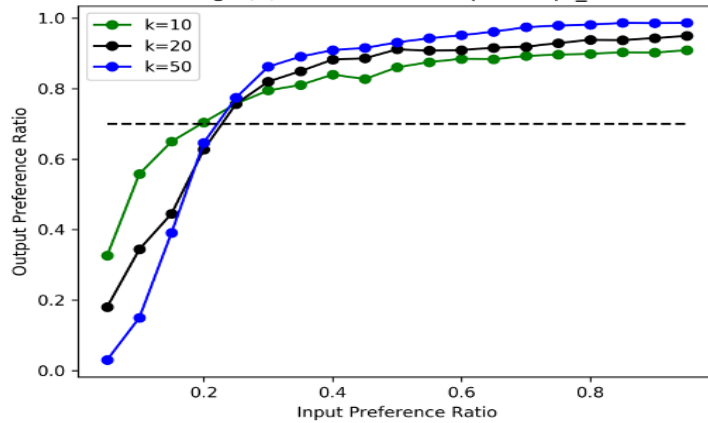
[그림IV-2] 비대칭적 선호도를 가지는 경우, 두 집단의 사용자 선호도에 따른 추천 결과물 선호도.

* (왼쪽) 사용자 집단이 0.5에서 1.0까지의 선호도를 가지는 경우
(오른쪽) 사용자 집단이 0.5의 고정된 선호도를 가지는 경우

다. 사용자 집단 및 카테고리 크기 변화에 따른 실험 결과

1) 사용자 집단 크기 변화

두 개의 집단 G_1, G_2 의 크기가 다를 때, 즉 포함된 사용자 수가 서로 다를 때에 대해 실험해본다. n_1 명의 사용자를 가지는 G_1 집단에서의 n_1 을 전체 사용자 수인 n 의 ϕ 배라고 정의한다. n_1 은 n 의 5%에서 95%까지의 값을 가지며 그 간격은 10%로 설정한다. 아래의 실험에서 두 집단의 선호도는 동일한 $\rho_1 = \rho_2 = 0.7$ 값으로 고정한다. [그림IV-3]은 집단 크기 비율인 ϕ 에 따른 추천 결과물의 선호도 비율 $PR_R(G_1, C_1)$ 을 나타낸 것이다. $PR_R(G_2, C_2)$ 은 아래 그래프와 거울 대칭인 그래프이기 때문에 생략한다. 아래 실험에서 k 는 User-KNN 알고리즘에서의 최근접 이웃을 k 명 선택했음을 의미한다.

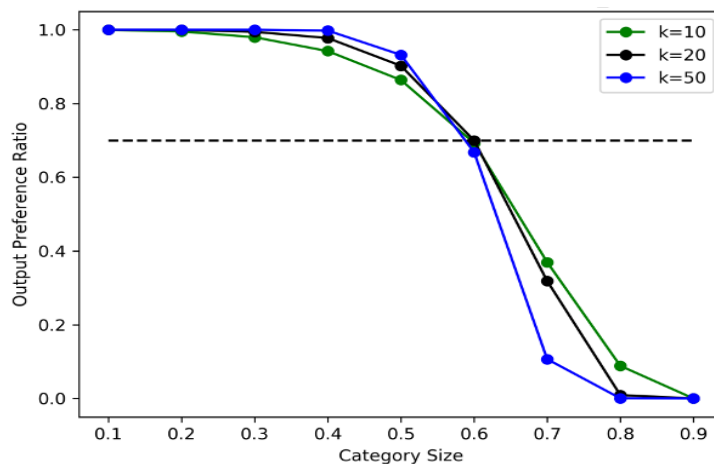


[그림IV-3] 두 집단의 크기 비율에 따른 추천 결과물 선호도

이 실험을 통해 우리는 집단크기비율 $\phi \leq 0.3$ 에 대해 $PR_R(G_1, C_1) < 0.7$ 이므로, G_1 집단은 음성 편향성 차이를 지님을 알 수 있다. 이는 소수 집단의 선호도가 다수 집단에 의해 묵살된다는 것을 의미한다. 집단크기 비율 ϕ 이 0.35에서 0.5까지의 중간 값을 가질 때는, 두 그룹간의 크기 차이가 있음에도 불구하고 두 그룹 모두의 편향성이 극대화된다. 하지만 크기가 큰 집단의 편향성 증가 정도가 크기가 작은 소수 집단에 비해 더 크다.

2) 카테고리 크기 변화

두 항목 카테고리의 크기가 다를 때 추천 결과물의 선호도는 어떻게 변화하는 지 알아보기 위해 진행하였다. C_1 의 크기인 m_1 을 전체 항목 개수인 m 의 θ 배라고 정의한다. θ 는 0.1에서 0.9까지의 값이 될 수 있다. 아래의 실험에서 두 집단의 선호도는 동일한 $\rho_1 = \rho_2 = 0.7$ 값으로 고정한다. 아래 [그림IV-4]는 θ 에 따른 추천 결과물의 선호도 $PR_R(G_1, C_1)$ 를 나타낸 것이다. $PR_R(G_2, C_2)$ 은 아래 그래프와 거울 대칭인 그래프이기 때문에 생략한다.



[그림IV-4] 두 카테고리의 크기 비율에 따른 추천 결과물 선호도

$\theta \leq 0.7$ 에 대해 G_1 의 C_1 에 대한 편향성은 ρ_1/θ 과 동일하므로 양성 편향성을 지닌다. 하지만 C_1 의 크기가 커질수록 G_1 의 C_1 에 대한 편향성은 줄어든다. 카테고리 크기가 크지 않을 경우($\theta \leq 0.5$), 추천 결과물의 편향성은 카테고리 크기와 상관없이 증폭된다. $\theta > 0.7$ 인 경우에 G_1 은 원래의 선호도와 다르게 C_2 를 선호하는 방향으로 편향된다. 흥미로운 점은 θ 가 0.6에서 0.7 사이의 값을 가지는 경우인데, 이 경우 G_1 은 C_1 을 선호하는 방향으로 편향되나 그 정도가 미미하여, 더 편향된 그룹에 의해 추천 결과가 C_2 에 편향되게 된다.

3. 데이터 증대 방법을 통한 편향성 보정

가. 제안하는 편향성 보정 방법의 정의

편향성 차이(bias disparity)의 문제를 해결하기 위해, 우리는 데이터가 추천시스템에 입력되기 전에 간단히 데이터 전처리(pre-processing)를 통해 편향성 차이를 줄일 수 있는 방법을 제안한다. 실제로 다양한 시스템 및 플랫폼에서 사용되는 추천시스템은 하나의 알고리즘을 사용하지 않고, 여러 알고리즘을 결합하여 매우 복잡하게 구현되어 있다. 따라서 모든 추천 알고리즘의 편향성 차이를 줄이는 단 하나의 방법을 제안하는 것은 불가능에 가깝다.

따라서 이번 실험에서는 추천 알고리즘에 입력 값으로 들어가는 사용자의 평가 데이터 값을 수정함으로써 편향성 차이를 줄이는 것을 목적으로 한다. 아래는 이번 실험에서 제안하는 편향성 보정을 위한 네 가지 데이터 증대 방법이다.

방법1 : 우리가 제안하는 첫 번째 데이터 증대 방법은 소수 집단의 크기를 다수 집단과 동일하도록 임의의 유저를 여러 명 선택하고 복제해 늘리는 것이다. 여기서 소수집단은 그 집단에 포함된 사용자 수가 전체 사용자 수의 절반보다 작은 경우로, 다수 집단은 그 집단에 포함된 사용자 수가 전체 사용자 수의 절반보다 작은 경우로 정의한다. 소수 집단을 G_1 , 다수집단을 G_2 이라고 가정하고, 소수집단의 크기를 n_1 , 다수집단의 크기를 n_2 라고 했을 때, G_1 내의 사용자를 임의로 $(n_2 - n_1)$ 번 선택하고 해당 사용자의 항목 선호 벡터를 복제해 G_1 에 추가한다.

방법2 : 두 번째 데이터 증대 방법은 소수 집단의 크기가 다수 집단의 크기의 0.25배 보다 작을 때, 소수 집단의 크기가 다수 집단의 크기의 0.25배가 되도록($0.25 * n_2 = n_1$)조정하는 방법이다. 이 과정은 첫 번째 방법과 유사하게 동일한 소수집단 내의 임의의 유저 벡터를 복제하여 추가하는 방식이며, 대신($0.25 * n_2 - n_1$)번 추가한다.

방법3 : 세 번째로 제안하는 데이터 증대 방법은, 소수집단에서 임의로 두 유저를 고른 후에 두 유저의 선호하는 항목을 더해서 새로운 유저 값을 만들어 소수 집단에 추가하는 방식이다. 소수 집단과 다수 집단의 크기가 같을 때까지 이 과정을 반복하여 추가한다.

방법4 : 마지막으로 제안하는 데이터 증대 방법은 소수 집단과 다수 집단의 크기를 동일하게 맞추기 위해, 두 명의 사용자를 임의로 뽑고 그 값에서 임의로 항목들을 선택하는 방법이다. 두 명의 사용자를 임의

로 뽑은 뒤, 두 사용자가 선택한 항목들에 대해서 임의로 10개의 항목들을 뽑고 이를 선택한 항목으로 가지는 가상의 사용자 값을 소수 집단에 추가한다. 이 과정을 소수 집단과 다수 집단의 크기가 같아질 때까지 반복한다.

나. 제안하는 편향성 보정 방법의 장단점 및 기존 방법론과의 비교

이번 실험에서 제안하는 편향성 보정 방법은 데이터 입력과 추천 알고리즘 사이에 데이터 증진 단계를 추가하는 방법으로 편향성을 보정한다. 이는 추천 알고리즘 자체를 수정하지는 않으므로, 추천 알고리즘을 수정하는 것에 비해 매우 적은 비용으로 적용할 수 있다는 장점이 있다. 하지만 이는 시스템에 쓰이는 추천 알고리즘이 어떻게 작동하는지는 고려하지 않았기 때문에, 모든 추천 알고리즘에 적용된다고 보기는 어렵다. 다음은 추천 알고리즘에서 편향성을 줄이는 기존 방법론과 이번 실험에서 제안하는 편향성 보정과의 비교이다. 추천 알고리즘에서의 편향성 연구는 시작 단계이기 때문에, 비교할 편향성 보정 방법의 수가 다양하지 않다.

2017년에 진행된 협업 필터링(collaborative filtering) 기반 추천 알고리즘의 공정성 연구⁸²⁾에서 새로운 공정성 정의를 제안하며, 새로운 공정성 정의를 기계학습 모델의 학습목표(learning objective)로 설정하여 모델을 새롭게 학습시켜 공정성을 향상시키는 방법을 제안한다. 이 방법은 데이터와 추천 알고리즘 모두를 고려하여 불공정성을 최소화하는 추천 알고리즘 모델을 학습시킬 수 있다는 장점이 있다. 또한 이 연구에서 공정성 정의를 학습목표에 추가하는 것을 통해 과대적합과 과소적합을 줄였다는 것을 보였다. 반면에 우리가 제안하는 방법과는 달리 기존 추천 알고리즘 모델을 그대로 사용하는 것이 아니라 추천 알고리즘 자체를 새롭게 학습시켜야 하기 때문에 비용이 매우 많이 든다는 단점이 있다.

2018년에 진행된 사용자 데이터의 편향성을 추천시스템이 더 강화한다는 연구⁸³⁾에서는 이러한 편향 불균형을 줄이기 위해 추천 알고리즘(User-KNN) 끝에 한 단계를 덧붙이는 후처리(post-processing) 방법인 GULM(Group Utility Loss Minimization) 알고리즘을 제안했다. 탐욕 알고리즘(greedy algorithm)을 이용해 항목을 다른 카테고리 변경시켜, 이 추천되는 항목들에 편향성 차이가 없으면서 동시에 추천 점수(utility)를 손실을 최소화하고자 하였다. 사용자의 선호도가 뚜렷하지 않은 경우(선호도 0.65 이하)는 좋은 결과를 나타냈으나, 선호도가 뚜렷한 경우(선호도 0.65 이상)에는 여전히 편향성 차이가 있음을 확인할 수 있었다. 반면에 우리가 제안하는 방법은 선호도가 뚜렷한 경우에도 편향성을 줄일 수 있다는 것을 실험을 통해 확인하였다. 또한 GULM 알고리즘은 탐욕 알고리즘을 이용하기 때문에 단순히 데이터를 복제하는 우리가 제안하는 방법보다는 비용이 많이 소모된다는 단점이 있다. 반면에 우리는 추천 점수의 손실을 고려

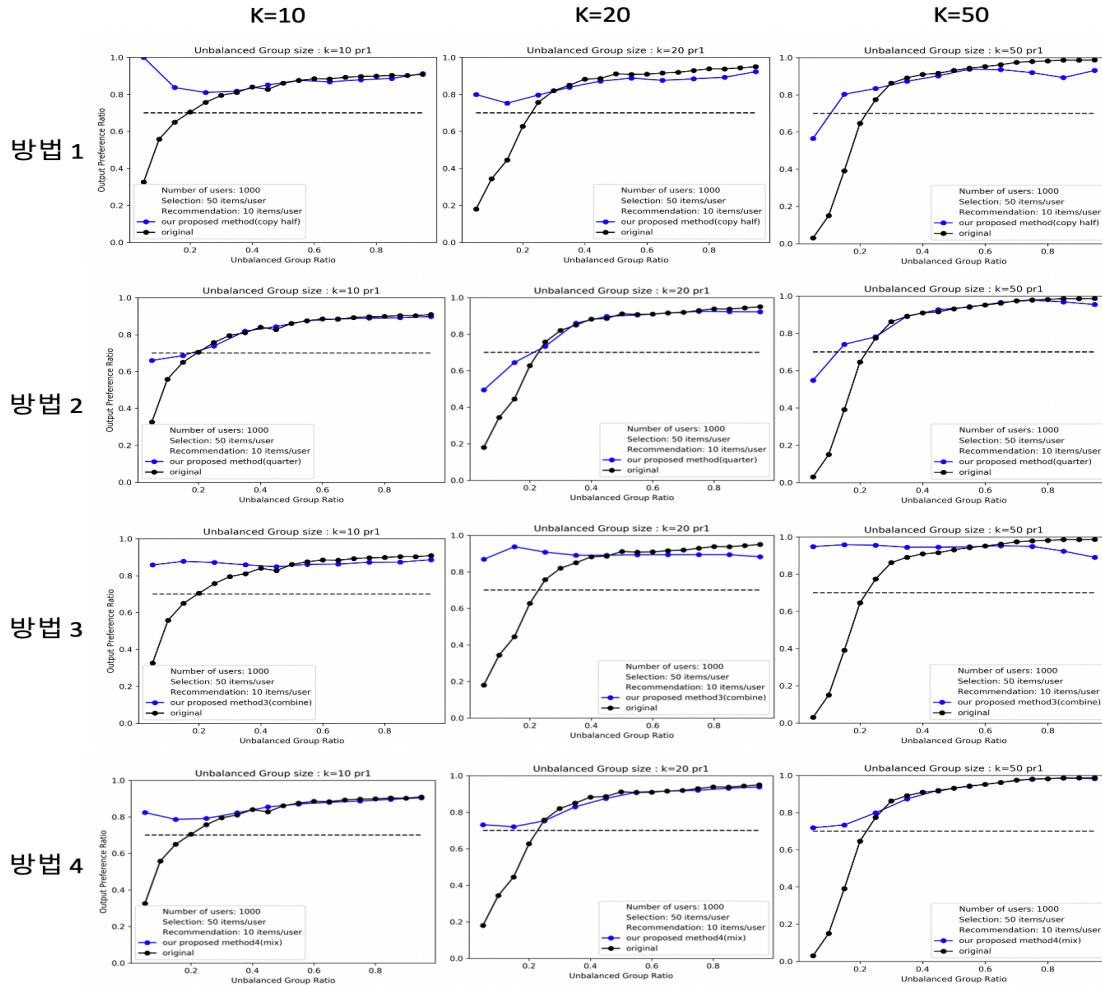
82) Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems* (pp. 2921-2930).

83) Tsintzou, V., Pitoura, E., & Tsaparas, P. (2018). Bias disparity in recommendation systems. *arXiv preprint arXiv:1811.01461*.

하지 않았으므로 추천점수를 고려한 GULM 알고리즘의 추천 결과물이 사용자 만족도는 더 높아, 만족도가 중요한 실제 추천시스템에 쓰이기 적합하다.

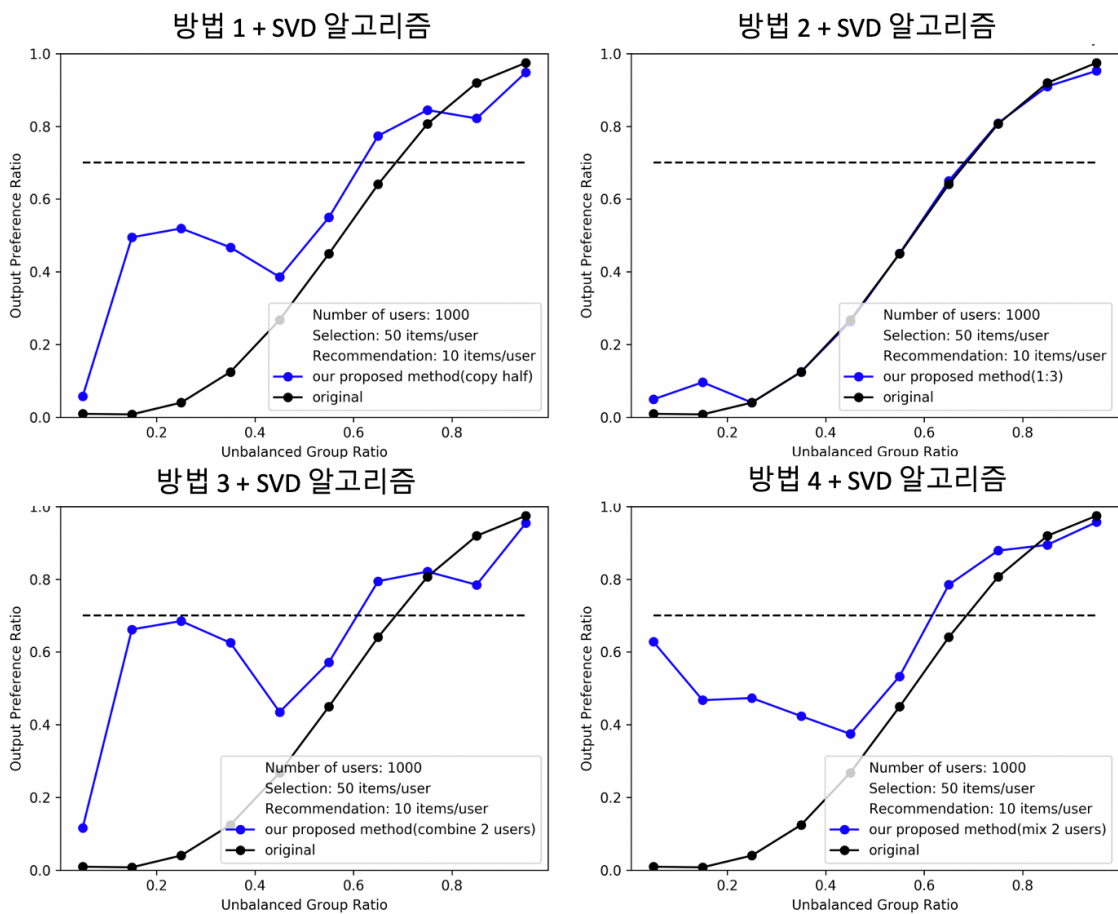
다. 가상 데이터를 이용한 편향성 보정 실험

G_1 에 해당하는 사용자 수 n_1 가 이 전체 사용자 수 n 에서 차지하는 비율을 ϕ 라고 정의한다. n_1 은 n 의 5%에서 95%까지의 값을 가지며 그 간격은 10%로 설정한다. 실험 결과 그래프의 x축은 G_1 에 해당하는 사용자 수가 이 전체 사용자 수에서 차지하는 비율 ϕ 이며, y축은 추천 알고리즘을 통해 추천된 결과물에서의 G_1 의 C_1 에 대한 선호도 $PR_R(G_1, C_1)$ 를 나타낸다. 그래프에서 점선은 입력 데이터에서의 선호도 $PR_S(G_1, C_1)$ 를 나타내며, 검은 실선은 기존 데이터를 이용해 추천한 결과물의 선호도, 그리고 파란 실선은 우리가 제안한 데이터 증대 방법을 이용하여 추천할 때의 결과물의 선호도를 나타낸다. [그림IV-5]에서 실선이 점선에 가까운 값을 가질수록 편향성 차이가 작다는 것을 나타낸다. 따라서 아래 시험한 모든 경우에 대해 파란색 실선이 검은색 실선에 비해 점선에 가까우므로, 우리가 제안하는 방법을 통해 편향성 차이를 줄일 수 있다.



[그림IV-5] 데이터 증대 방법을 통한 추천알고리즘(User-KNN알고리즘)의 결과물 선호도 보정결과

제안하는 방법이 비단 User-KNN알고리즘에만 적용되는 것이 아니라, 다른 알고리즘에도 적용된다는 것을 보여주기 위해, 위에서 사용되었던 User-KNN 알고리즘 외에도 SVD 알고리즘을 사용해 추가 실험을 진행하였다. [그림IV-5]와 [그림IV-6]에서 볼 수 있듯이, User-KNN과 SVD알고리즘을 이용한 추천시스템의 경우, 제안하는 모든 데이터 증대 방법이 기존 데이터를 사용하여 추천하는 것보다 더 적은 편향성 차이를 보임을 관찰할 수 있다. 특히, 방법 3의 경우 G_1 의 크기가 적당히 작은 전체 사용자 수의 15%에서 35%에 해당할 경우, 구간 평균 편향성 차이가 -0.6437 에서 -0.0627 로 크게 감소한다. 또한 방법 4의 경우 G_1 의 크기가 매우 작은 전체 사용자 수의 5%에 해당하는 경우에 유의미한 편향성 차이 감소 효과를 보임을 확인할 수 있다.



[그림IV-6] 데이터 증대방법을 통한 추천알고리즘(SVD알고리즘)의 결과물 선호도 보정 결과

- * 이번 연구에서 제안하는 네 가지 방법의 데이터 증대 방법을 통한 추천 알고리즘(SVD 알고리즘)의 결과물의 선호도 보정 결과를 나타낸 그래프이다. 점선이 기존 입력 데이터 값이 가지고 있는 선호도이며 검정색 실선이 원래 데이터를 이용했을 때의 결과이며, 파란색 실선이 우리가 제안하는 방법을 통해 얻은 추천 결과물의 선호도이다. X축은 소수집단의 집단 크기 비율이고, y축은 추천 알고리즘을 통해 추천된 결과물의 선호도 및 편향성을 나타낸다.

나. 실제 데이터를 이용한 편향성 차이 보정 실험

이 실험에서는 실제 사용자와 선호 데이터를 이용해 실험하기 위해 MovieLens 1M Dataset (MovieLens 기업에서 제공하는 공개된 데이터셋으로 6,000명의 실제사용자가 4,000개의 영화에 대해 총 100만개의 별점 평가를 포함하는 데이터셋이다. <https://grouplens.org/datasets/movielens/1m/>)를 사용하였다. 이 데이터셋에서 액션과 로맨스 카테고리에 포함된 영화, 각각 468과 463개의 영화만 고려하였다. 또한 전체 사용자 중에서 적어도 90개 이상의 별점 평가를 한 사용자들만 고려하여 총 1,259명의 사용자를 데이터로 사용하였다. 이 사용자들은 981명의 남성과 278명의 여성으로 구성되어있다.

표1은 User-KNN(k=10)와 SVD 알고리즘을 이용해 구현한 추천시스템에 대해 입력 데이터의 편향성과 추천 결과물의 편향성 그리고 편향성 차이를 나타낸 것이다. 편향성 차이를 줄이려면 기존 입력 데이터의 편향성과 추천 결과물의 편향성이 유사해야 한다. 아래 표에서 보면 User-KNN의 경우 액션영화에 대한 남성의 편향성은 1.3925였으나, 추천 알고리즘을 통해 산출된 데이터의 편향성은 1.6786으로 편향성이 증대되었음을 확인할 수 있다. 반면에 로맨스 영화에 대한 남성의 편향성은 0.5787이었으나 추천결과물의 편향성은 0.2716으로 선호하지 않은 정도가 더 강화되었음을 확인할 수 있다.

<표 IV-1> 실제 데이터에 대한 User-KNN(위)과 SVD(아래)의 편향성 차이

	액션 영화			로맨스 영화		
	입력 데이터 편향성	추천 결과물 편향성	편향성 차이	입력 데이터 편향성	추천 결과물 편향성	편향성 차이
남성	1.3925	1.6786	0.2054	0.5787	0.2716	-0.5307
여성	0.9688	1.1486	0.1855	1.0335	0.8405	-0.1867

	액션 영화			로맨스 영화		
	입력 데이터 편향성	추천 결과물 편향성	편향성 차이	입력 데이터 편향성	추천 결과물 편향성	편향성 차이
남성	1.3925	1.0623	-0.2371	0.5787	0.6126	-0.6126
여성	0.9688	1.4800	0.5276	1.0335	0.4848	-0.5309

이 실험에서는 위의 실제 MovieLens 데이터에 대해 우리가 제안한 4가지 데이터 증대 방법을 적용하여, 편향성 차이의 변화를 확인해보고자 하였다. 아래 표2는 User-KNN(k=10)알고리즘과 SVD 알고리즘을 이용한 추천시스템에 4가지 방법에 대한 각 남성, 여성 사용자 집단의 액션, 로맨스 카테고리에 대한 편향성 차이(BD)를 나타낸 것이다. 여기서 편향성 차이는 절댓값으로 크기를 비교하며 0에 가까울수록 추천시스템의 편향성, 즉 기존 사용자가 가진 선호도를 증폭하거나 감쇠하는 정도가 적다. 가상의 데이터에 적용된 것과는 달리 실제 데이터에 적용해 본 경우, 모든 집단과 모든 카테고리에 대해 편향성 감소를 보이지는 않았고, 오히려 편향성이 증가한 경우도 있었다.

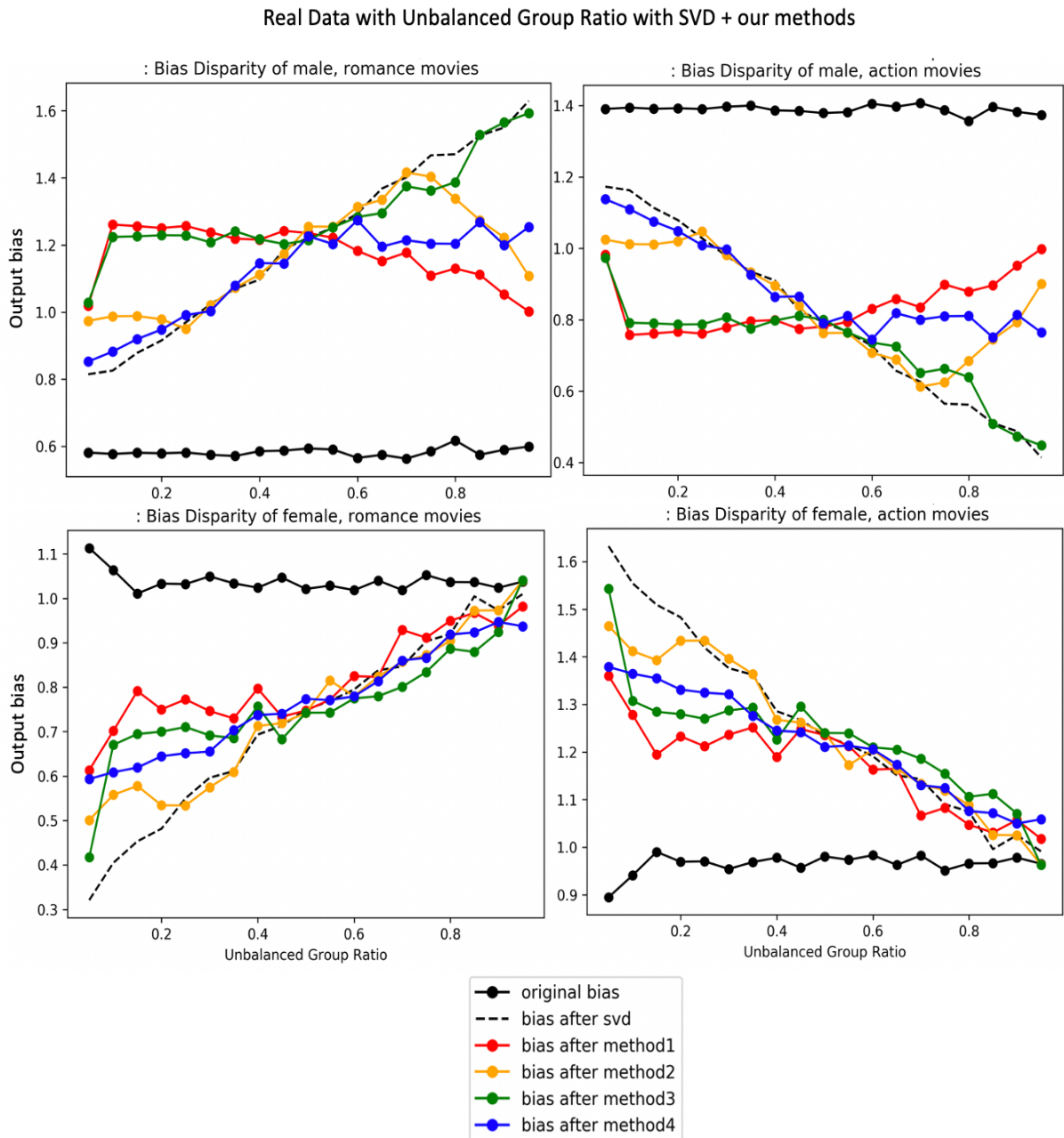
방법3을 활용한 User-KNN 추천시스템의 경우 모든 사용자집단과 항목 카테고리의 편향성 차이가 기존 데이터를 이용한 추천시스템보다 감소하였다. 하지만 동일한 방법 3을 SVD 추천시스템에 적용한 경우 여성 집단에 대해서만 편향성 차이 감소효과를 보였다. 또한 방법 1과 방법 4의 경우 User-KNN 추천시스템에 적용할 경우, 오히려 편향성 차이를 증가하는 효과를 보였다. 하지만 이 방법을 SVD 추천시스템에 적용할 경우, 여성 집단에 대해서는 편향성 감소를 보였다.

<표 IV-2> 실제 데이터를 이용한 User-KNN(위)와 SVD(아래)과 우리가 제안한 4가지 방법을 적용한 추천시스템의 편향성 차이 비교

	BD(남성,액션)	BD(여성,액션)	BD(남성,로맨스)	BD(여성,로맨스)
기존 데이터	0.2054	0.1855	-0.5307	-0.1867
방법1	0.2143	0.5438	-0.5537	-0.5607
방법2	0.2032	-0.5248	0.1796	-0.1765
방법3	0.1804	-0.0360	-0.4660	0.0369
방법4	0.2054	0.3000	-0.5307	-0.3026

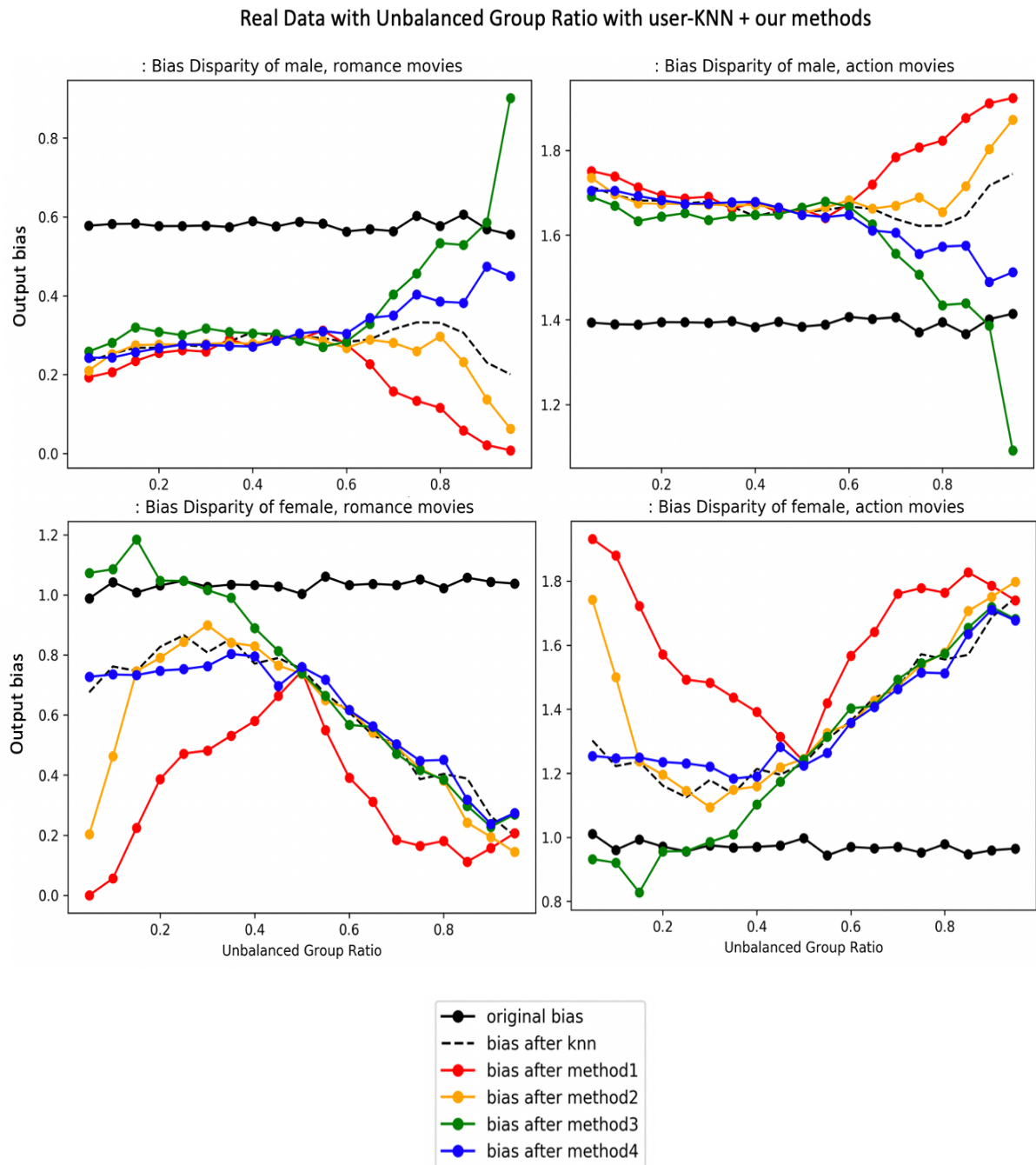
	BD(남성,액션)	BD(여성,액션)	BD(남성,로맨스)	BD(여성,로맨스)
기존 데이터	-0.2371	0.5276	0.6126	-0.5309
방법1	-0.4525	0.2297	1.1688	-0.2334
방법2	-0.2630	0.4991	0.6794	-0.4938
방법3	-0.4269	0.2874	1.1027	-0.2935
방법4	-0.2598	0.3813	0.6710	-0.3856

[그림IV-7]과 [그림IV-8]은 실제 데이터의 사용자 집단 비율을 조절하여 기존 추천 알고리즘의 편향성과 이번 실험에서 제안하는 방법을 적용한 추천 알고리즘의 편향성 차이를 비교하는 그래프이다. 각 그래프는 각 집단의 카테고리별 그래프를 나타낸 것이다. 검은색 실선은 기존 입력 데이터의 편향성을 의미하며, 점선은 추천 알고리즘만 이용했을 경우의 편향성, 유색 실선은 각각 이번 실험에서 제안하는 네 가지 방법을 적용한 결과물의 편향성을 나타낸 것이다. 모든 그래프에서 기존 입력 데이터의 편향성(검은 실선)보다 추천 알고리즘 결과물의 편향성(점선)이 차이가 있다는 것을 확인할 수 있다. 추천 알고리즘이 기존 데이터의 편향성 정도를 그대로 반영하는 것이 아니라, 기존 데이터의 편향성보다 더 큰 편향성을 가지도록 추천을 해주는 것임을 확인할 수 있다. 반면에 이번 실험에서 제안하는 여러 방법들을 적용한 추천 알고리즘의 결과(유색 실선) 대부분은 추천 알고리즘만 적용한 결과물의 편향성보다 더 입력 데이터에 가까운 편향성을 가짐을 확인할 수 있다. 하지만 제안하는 모든 방법들이 항상 더 좋은 결과를 나타내는 것은 아니다. 예를 들어 User-KNN을 이용한 남성의 로맨스 영화에 대한 편향성 그래프([그림IV-8]의 첫 번째 그래프)에서 방법3과 방법4를 적용한 경우, 단순 User-KNN만 이용한 경우보다 입력 데이터와 비교적 적은 편향성 차이를 가지나, 방법 1과 방법2는 오히려 더 큰 편향성 차이를 가짐을 확인할 수 있다.



[그림IV-7] 실제 데이터 사용자 집단 비율을 조절한 데이터 SVD 추천알고리즘 이용한 편향성 차이

* 실제 데이터의 사용자 집단 비율을 조절한 데이터의 SVD 추천 알고리즘을 이용한 편향성 차이 결과 그래프이다. 각 그래프는 순서대로 남성의 로맨스 영화에 대한 편향성 차이, 남성의 액션 영화에 대한 편향성 차이, 여성의 로맨스 영화에 대한 편향성 차이, 여성의 액션 영화에 대한 편향성 차이를 나타낸다. 검은 실선은 입력 데이터의 편향성 차이를 의미한다. 검은 점선은 SVD 추천 알고리즘만 이용한 결과물의 편향성이며, 유색 선 그래프는 각각 이 실험에서 제안한 네 가지 방법을 적용한 SVD 추천 알고리즘의 결과의 편향성이다.



[그림IV-8] 실제 데이터 사용자 집단 비율을 조절한 데이터 User-KNN 편향성 차이 결과

* 실제 데이터의 사용자 집단 비율을 조절한 데이터의 User-KNN 편향성 차이 결과 그래프이다. 각 그래프는 순서대로 남성의 로맨스 영화에 대한 편향성 차이, 남성의 액션 영화에 대한 편향성 차이, 여성의 로맨스 영화에 대한 편향성 차이, 여성의 액션 영화에 대한 편향성 차이를 나타낸다. 검은 실선은 입력 데이터의 편향성 차이를 의미한다. 검은 점선은 User-KNN 추천 알고리즘만 이용한 결과물의 편향성이며, 유색 선 그래프는 각각 이 실험에서 제안한 네 가지 방법을 적용한 User-KNN 추천 알고리즘의 결과의 편향성이다.

따라서 이 연구에서 진행한 가상의 데이터를 이용한 실험에서 우리가 제안한 방법이 편향성 차이를 줄이는 결과를 보였다고 하더라도, 실제 데이터에 적용하는 경우 이 방법들이 모두 편향성 차이를 줄인다고 보장하긴 어렵다. 또한 이 실험에서는 User-KNN과 SVD 의 두 가지 알고리즘을 이용해 추천시스템을 구현

하였는데, 실제 데이터를 이용한 경우 사용하는 추천 알고리즘에 따라 결과가 다를 수 있다. 추천 시스템에 사용하는 알고리즘은 이외에도 다양하기 때문에, 모든 알고리즘에 데이터 증대 편향성 보정 방법을 적용했을 때의 결과는 보장할 수 없다. 실제 추천시스템에 사용되는 알고리즘은 하나로 구현되지 않으며 매우 복잡하게 구성되어 있기 때문에, 하나의 알고리즘에 적용하여 편향성 감소 효과를 보았어도 여러 알고리즘을 결합해 사용한 추천시스템에도 같은 효과를 보인다고 말하기 어렵다

4. 결과 및 함의

온라인을 통해 영상 추천, 쇼핑 아이템 추천, 교육, 검색, 온라인 구인구직 등의 다양한 분야에서 추천시스템을 사용한다. 추천시스템에 사용된 알고리즘에는 데이터 상의 편향 등의 이유로 알고리즘 편향이 있을 수 있다. 알고리즘 편향은 알고리즘의 의사결정 과정이 특정 인구 집단에게 부정적인 결과를 야기할 수 있는 것을 의미한다.

이번 실험에서 가상의 사용자와 그들의 평가 정보 데이터를 만들고 이를 이용해 User-KNN과 SVD 알고리즘을 이용한 추천시스템을 구현해 실험을 진행하였다. 실험을 통해 User-KNN을 이용한 추천시스템의 경우, 기존의 사용자가 선호한다고 했던 정도보다 더 증폭된 정도로 항목들을 추천해주는 것, 즉 알고리즘 편향을 확인할 수 있었다.

추가로 알고리즘의 수정을 최소화 하기 위해 알고리즘에 입력하기 전에, 입력 데이터를 전처리하는 방법으로 편향성 차이를 줄일 수 있는 네 가지 방법을 제안하였다. 이 방법은 가상의 데이터를 이용한 경우 편향성 차이감소 효과를 극명하게 보였으나, 실제 데이터를 이용한 실험에서는 방법이나 집단에 따라 효과가 다르게 나타났다.

이번 실험에서 사용한 User-KNN과 SVD는 사용자 정보를 이용한 협업 필터링(collaborative filtering)에 포함되는 알고리즘이다. 하지만 실제 추천시스템에는 협업 필터링뿐만 아니라 항목 자체의 특성을 고려하는 콘텐츠 기반 필터링(contents-based filtering)의 알고리즘 또한 사용되며, 하나의 알고리즘만 사용되지 않고 여러 알고리즘을 결합하는 방식으로 사용된다. 따라서 우리가 제안하는 데이터 증대 방법이 모든 알고리즘과 모든 추천시스템에 편향성 차이를 감소시키는 효과를 불러일으킬 것이라고는 보기 어렵다. 하지만 이 실험을 통해, 전체 알고리즘을 고치지 않고, 간단히 데이터만 수정함으로써 편향성 차이를 줄임을 보였다.

추후 연구로는 이 연구에서 사용된 추천알고리즘 외에 다양한 알고리즘을 이용해서 추천시스템의 편향성을 보이고, 이 추천시스템에도 간단한 전처리 과정을 통해 편향성 차이 감소 효과를 보이는지 확인해볼 수 있다. 또한 단순히 하나의 알고리즘만 사용하는 것이 아니라 여러 알고리즘을 결합해서 추천시스템을 구현할 수 있으며, 편향성 차이 뿐만 아니라 추천 정확도 등의 평가방법을 추가할 수 있다. 또한 이 연구에서 사용된 편향성 및 편향성 차이 정의는 기존 관련 연구에서 자주 쓰이는 정의를 채택하였으므로, 이를 이용하여 여러 플랫폼들의 추천시스템의 편향성 차이를 실험을 통해 계산해볼 수 있을 것이다. 또한 이러한 추천시스템에도 데이터 전처리 과정이 편향성 감소에 효과를 미치는지 확인해 볼 수 있다.

제5장 추천시스템 공정성을 평가 및 보정 방법론에 대한 전문가의견

제1절. 전문가조사 방법 및 연구문제

1. 전문가 심층조사 목적

학계와 산업계 각 분야의 전문가들을 대상으로 심층조사를 진행함으로써 알고리즘 공정성에 대한 개념과 이를 보정하기 위한 방법론적 쟁점들을 파악하고, 쟁점별 이슈 및 해결책을 모색하여 효율적인 정책 입안에 기틀을 제공하는데 목적을 둔다. 특히 이번 전문가 심층조사는 크게 개념적인 접근, 방법론적인 접근, 정책적인 접근인 세 가지 측면으로 알고리즘 공정성 보장에 대해 살펴봄으로써 알고리즘 공정성을 보장하기 위한 실용적이고 유의미한 결과를 도출하는 데 목적이 있다. 구체적인 연구목적으로 첫째, 알고리즘 공정성에 대한 보편적 정의가 가능한지, 가능하다면 전문가들이 합의하는 공정성 개념이 존재하는지를 살펴보고자 한다. 둘째, 추천시스템에서 사용되는 알고리즘의 공정성을 확보하기 위한 가장 적합한 분석방법을 확인하고 편향 보정을 위한 대안을 모색하고자 한다. 셋째, 라이스만과 동료들(Reisman et al., 2018)이 알고리즘 설명가능성 및 책무성 부과의 기준과 거버넌스를 제시한 알고리즘 영향평가(AIA) 방법론을 검토해서 추천시스템 알고리즘의 사용 목적, 데이터 수집의 범위와 방식, 설계 디자인, 결과산출 방식, 잠재적인 위험요인(편향이나 오류의 가능성) 및 예상되는 혜택 등을 평가할 방법론적 틀과 방향을 제시하고자 한다. 특히 이번 조사에서는 방법론적인 측면에 주목하여 알고리즘의 공정성을 보장할 수 있는 분석방법을 다양하게 제시하고, 이에 대한 전문가 의견을 구체적으로 탐색하여 알고리즘 공정성을 확보할 수 있는 방향성을 제안하고자 한다.

2. 연구문제

연구문제 1. (개념적 접근)알고리즘 공정성에 관한 보편적 정의가 가능한가?

연구문제 2. (방법론적 접근)알고리즘 공정성을 분석하기 위한 가장 적합한 분석방법은 무엇인가?

연구문제 3. (제도적 접근)알고리즘 공정성 보장방안을 위한 영향성 평가 제도가 필요한가?

3. 연구방법

추천시스템의 설명가능성을 높이고 편향을 보정하기 위해 공정성에 대한 개념적 논의, 보장 방안에 대한 방법론적 논의, 제도적 장치와 다양한 대안과 수단을 탐색하여 대안을 제시하기 위해 전문가 심층조사를

실시했다. 연구 대상은 방법론적, 시스템 및 법적 문제점을 복합적으로 분석할 수 있도록 학계, 산업계, 법조계 등 다양한 계층별 전문가 10명으로 구성됐다. 연구문제를 바탕으로 알고리즘 공정성 개념, 방법론적인 보안방안, 제도적인 영향성평가 등에 관련된 질문지를 구성했으며 전문가들은 질문지를 바탕으로 하여 다양한 의견을 제시했다. 조사는 총 일주일간 진행되었으며, e-mail을 통해 반구조화된 형태의 설문으로 진행했다.

<표 V-1> 전문가조사 리스트

no	전문가	소속	전공분야
1	전문가A	공공기관	언론학
2	전문가B	학계	언론학
3	전문가C	산업체	정책학
4	전문가D	학계	법학
5	전문가E	학계	경영학
6	전문가F	산업체	정책학
7	전문가G	산업체	정책학
8	전문가H	학계	전산학
9	전문가I	산업체	전산학
10	전문가J	학계	전산학

전문가 조사를 진행한 내용은 크게 네 가지로 구분된다. 첫째, 추천시스템의 알고리즘 공정성 개념에 대한 기초적인 의견을 수집했다. 둘째, 이를 보장할 수 있는 방법을 제안하고 쟁점별 이슈 및 해결책에 대한 다학제적인 종합적 방법론을 탐색했다. 셋째, 제도적인 측면에서 알고리즘 영향성 평가제도를 파악하고 다양한 의견을 수집했다. 마지막으로 알고리즘 공정성과 관련된 최근 연구 동향 및 추가적인 개선의견들을 수집했다.

제2절. 전문가조사 결과

1. (연구문제1:개념적 접근) 알고리즘 공정성에 관한 보편적 정의가 가능한가?

추천시스템 등에서의 ‘알고리즘 공정성’에 대한 보편적이고 합의된 정의가 없는 점에 대해서 전문가들에게 ‘알고리즘 공정성’에 대한 보편적 정의가 가능한지, 가능하다면 그 핵심개념은 무엇인지 의견조사를 진행했다. 또한, ‘알고리즘 공정성’에 대한 보편적 정의가 어렵다면 그 이유와 대체가능한 개념에 대해서 무엇인지에 대해서 조사했다. 이에 대해서 전문가들은 굉장히 다양한 의견을 제시했다. 전문가들 사이에서도 ‘알고리즘 공정성’에 대한 정의가 가능하다와 불가능하지 않다고 의견이 나뉘었으며 세부적인 정의에 대해서는 다양한 의견이 제안됐다.

전문가들은 추천알고리즘의 공정성을 ‘추천 결과의 편향성’ 정도로 정의했다. 이에 대한 구체적인 예시로

포털에서의 상위 뉴스콘텐츠 기사를 들었다. 뉴스기사들을 정치적인 관점을 고려하여 이진분류로 나누어 구분하면 정량적인 평가가 가능할 것이라고 제안했다. 다만 이 편향성이 사용자들에 의한 실제 행위가 적용된 것인지, 포털에서 고의적으로 조정된 것인지 제 3자 입장에서 알기 어려운 한계가 있으며, 도메인에 따라 편향성 정의에 필요한 기준의 수가 증가할수록 정량적 측정 및 측정된 수치에 대한 정확도를 보장하기 어렵다고 한계를 들었다. 전문가J는 ‘보편인권에 벗어나지 않음’으로 알고리즘 공정성을 정의하고 있었다. 구체적으로 전문가J는 “1) 보편인권에서 벗어나 많은 사람들이 가지고 있는 차별적 편견을 결과로 내뱉지 않아야 하고, 2) 그 결과가 사회로 흘러가서 다시 알고리즘의 학습 및 논리 소스에 활용되는 경로를 끊을 수 있어야 하며, 3) 알고리즘 특성상 선형(prior) 활용도가 높은 경우 그것을 빠르게 수정할 수 있는 장치가 필요하다.”라고 자세하게 제안했다. 전문가D는 설명적 관점에서 ‘알고리즘 공정성’에 대한 일반적인 정의는 가능하지만 알고리즘의 사용이 보편화되고 사례가 축적됨에 따라 분야별 알고리즘 공정성의 의미가 구체화 될 수 있음을 강조했다. 또한 전문가D는 공정성에 대한 핵심개념으로 ‘공정성을 실제로 접근할 것이냐, 절차적으로 접근할 것이냐’에 따라 달라진다고 의견을 제시하며 실제로 접근한다면 공정성 핵심에 ‘차별과 편견의 배제’가 중요한 요소가 되고 절차적으로 접근하면 ‘투명성’이 핵심개념으로 등장하게 될 것이라고 주장했다.

반면 ‘알고리즘 공정성’ 개념을 보편적으로 정의내리기 어렵다고 응답한 전문가H는 “ ‘공정성’ 자체에 대한 정의조차 논쟁의 대상이 된다고 생각하기 때문에 알고리즘에 적용되는 공정성을 개념화하는 것은 더욱 어렵다.” 라고 말했다. 추가로 전문가H는 알고리즘 혹은 인공지능은 비도덕(amoral)이라고 하는데 이는 도덕(moral)도 부도덕(immoral)도 아닌 옳고 그름에 대한 판단 가치 없이 학습 및 행동 하는 것을 일컫고 모든 기술에 대한 양면적인 사용(dual use)을 강조했다. 즉, 핵분열 기술을 통해 발전소를 만드든지, 핵폭탄을 만드느지로 갈리는 것이다. 따라서 ‘알고리즘 공정성’이 보편적으로 정의내리기는 힘들다고 전문가H는 재차 강조했다. 대체가능한 개념으로는 ‘사람들이 알고리즘 결과에 대해 얼마나 만족 혹은 납득하였는가’가 있다고 응답했다. 공정성의 측면에서 사람들에게 알고리즘의 결과가 납득 가능한지 조사하고 이에 대해서 다수가 만족한다면 ‘알고리즘 공정성’의 대리지표로 사용될 수 있음을 확인했다. 전문가C도 공정성이란 분배적, 절차적, 상호작용적 관점을 모두 포함해야 하며 기계학습모형이 이해할 수 있는 공정성은 수학적으로 정의되어 정량평가가 가능해야 하는데 알고리즘에서 활용한 공정성 개념을 수학적으로 정의내리는 것은 어렵다고 지적했다. 즉 “알고리즘 공정성에 대한 단일한 최상의 정의를 내리기 어려우며, 분배의 관점에서 구분된 통계적 개념들조차 동시에 만족되기 어렵다”라고 응답했다.

비슷한 맥락으로 전문가F도 알고리즘의 공정성 개념이 보편적으로 적용되기 어렵다고 응답했다. 전문가F는 국내외에서 알고리즘 공정성에 대한 논의가 되고 있고, 원칙적인 수준에서의 경계와 우려는 있지만 아직 구체적인 합의나 정의가 있다고 보기 어렵다고 지적했다. 특히 전문가F는 알고리즘이 편향되지 않고 독립적이게 제공하는 것을 정의내리기 매우 어렵다고 재차 강조하며, 설사 독립적으로 제공된다고 하더라도 그것을 공정하다고 보기는 어렵다고 응답했다. 그러나 전문가F는 ‘편향’에 대해서는 고려해야 한다고 강조했다. 즉 “공정성의 반대인 편향에 대한 우려는 존재하고, 이는 사회적이고 국가적으로 논의가 필요해 보인

다. 공정성에 대한 절대적인 개념을 정립하는 것은 어려운 일이지만 그 개념의 극단적 대적점에 있는 부분을 고려하는 것은 공감대를 구할 가능성이 높다.” 라고 응답했다. ‘공정성’에 대한 개념을 정의하여 공정하지 못한 것을 불공정하고 편향적이라고 하기보다는, 사회적으로 문제가 되는 불공정하고 편향적인 부분에 대한 합의가 우선적으로 선행되어야 한다고 강조했다.

전문가A는 ‘공정성’이라는 개념이 역사적으로 항상 논란이 되어왔으며, 알고리즘 영역의 공정성 개념을 도입하게 되었을 때 논란이 다시 발생하게 될 것이라고 지적했다. 특히 계량적 관점에서 공정성 연구결과를 분석한 Narayanan(2018)의 사례를 들며 수학적으로 공정성에 대한 정의는 30여 개 이상 정의되어 왔지만 완벽한 공정성에 대한 개념이 없으며 상당수가 서로 상충하고 있음을 지적했다. 즉, 공정성이라는 개념이 수학적으로 정의할 수 있는 개념이 아니고 당시의 사회수준과 맥락에 따라 변화할 수 있다고 강조했다. 공정성과 마찬가지로 편향과 차별이라는 개념 또한 시대와 사회적인 맥락인 문화가 반영되기 때문에 정의가 어렵다고 응답했다. 그러므로 전문가A는 ‘공정성’이라는 개념보다는 데이터의 투명성과 그에 따르는 다양성 개념이 더 중요할 수 있다고 강조했다. 추천시스템에서 활용하는 데이터는 연산 가능한 데이터로 직관적으로 문제가 있음이 보이지만 직관을 데이터에 반영하기 어려운 현실이라고 응답했다. 그러므로 데이터 투명성이 전제된다면 학습이 이루어지기 전에 문제를 사전에 파악하고, 이후 데이터를 보완할 수 있는 방향을 찾을 수 있기 때문에 유의미할 것이라고 주장했다.

전문가E는 학문적인 공정성 정의와 일반 사람이 인지하는 공정성의 정의가 우선적으로 일치되어야 하지만, ‘알고리즘의 공정성’에 대해서 합의된 사회적 정의를 도출하는 것은 현실적으로 불가능하다고 지적했다. 왜냐하면 사람이 내리는 의사결정을 두고도 ‘공정성(fairness)’에 대한 정의가 모두 다른데, 이에 대한 정의가 확립되기 어렵다는 것이다. 전문가E는 “출신대학이라는 변수를 채용 알고리즘에 예측변수로 투입하는 곳이 공정한지, 공정하지 않은지에 대해서는 완벽하게 합치된 사회적 동의를 얻기 어렵다.”라고 쉽게 예시로 설명하기도 한다. 그러나 그럼에도 불구하고 알고리즘 공정성에 대한 정의를 확립하려는 노력은 지속되어야한다고 강조한다. 공정성에 대한 보편적 정의를 확립하는 노력이 지속됨으로써 시대에 따라, 국가적 특성에 따라 ‘다수’가 생각하는 공정성에 기반해 알고리즘의 공정성을 확보해나갈 수 있다고 주장한다. 또한 전문가B는 개인화의 중심에는 개인 데이터가 존재하기 때문에 알고리즘이 개인마다 서로 다른 콘텐츠를 추천하는 상황에서 공정성에 대한 보편적 정의는 내리기 어렵다고 응답했다.

추가적인 의견으로 전문가G는 알고리즘 공정성이 입력(input) 데이터의 객관성 x 수집정당성 x 합리성 x 투명성에 기반해 생성되는 목적 개념이므로 알고리즘의 공정성 이전에 활용 데이터의 공정성에 대한 정의가 우선적으로 선행되어야 한다고 응답했다. 전문가들의 ‘알고리즘 공정성’에 대한 개념적 접근은 다양한 관점들이 있었지만 ‘알고리즘 공정성’에 대한 개념과 현재 발생한 격차를 메우기 위해서는 더 많은 연구와 투자가 이루어져야 한다는 점은 의견이 동일했다.

<표 V-2> 개념적 측면 전문가 의견

분류	전문가 의견
알고리즘 공정성 개념정의가 가능하다	- (전문가I) ‘추천결과의 편향성’으로 정의 가능함 - (전문가J) ‘보편인권에 벗어나지 않음’으로 정의가능함. 차별적 편견을 결과로 내뱉지 않고, 그 결과가 활용되는 경로를 끊을 수 있고, 알고리즘을 수정할 수 있는 장치가 필요함 - (전문가D) 설명적 관점에서 일반적 정의는 가능하지만, 알고리즘이 보편화됨에 따라 분야별 알고리즘 공정성 의미가 구체화될 수 있음. 공정성을 실체적으로 접근한다면 ‘차별과 편견의 배제’가 핵심개념이지만 절차적으로 접근한다면 ‘투명성’이 핵심개념이 될 수 있음
알고리즘 공정성 개념정의가 불가능하다	- (전문가H) 알고리즘이라는 것은 비도덕(amoral)으로 사용방식에 따라 도덕, 부도덕 여부가 달라지기 때문에 ‘알고리즘 공정성’을 개념화하는 것은 어려움. 대체가능한 개념으로 ‘사람들이 알고리즘 결과에 대해 얼마나 만족 혹은 납득하였는가’를 사용할 수 있음 - (전문가C) 공정성은 정량평가가 가능해야하는데 알고리즘에서 활용한 공정성 개념을 수학적으로 정의내리는 것은 어려움 - (전문가F) 알고리즘이 편향되지 않고 중립적이게 제공하는 것을 정의내리기 어려우며 중립적이라고 하더라도 그것이 공정하다고 보기는 어려움 - (전문가A) 공정성, 편향, 차별이라는 개념은 수학적으로 정의하기 어려우며, 사회 수준과 맥락(문화)에 따라 변화할 수 있는 개념임. 데이터의 투명성과 다양성 개념이 보다 중요함 - (전문가E) 알고리즘 공정성에 대한 사회적으로 합의된 정의를 도출하는 것은 어렵지만 보편적 정의를 확립하려는 노력은 지속될 필요가 있음 - (전문가B) 개인화가 고도화 된 사회에서 알고리즘 공정성에 대한 보편적 정의는 어려움
추가 의견	- (전문가G) 알고리즘 공정성 이전 활용 데이터의 공정성에 대한 정의가 우선적으로 선행되어야함

2. (연구문제2:방법론적 접근) 알고리즘 공정성을 분석하기 위한 가장 적합한 분석방법은 무엇인가?

알고리즘 공정성 분석을 위해 1) 이용자기반의 실험, 2) 대규모 온라인 패널기반 실험, 3) 시스템 기반의 자료를 바탕으로 한 연구, 4) 시뮬레이션 기반 실험 등 다양한 분석방법을 제시하고 이에 대한 전문가 의견을 수집했다. 전문가 조사 결과, 알고리즘 공정성 보장을 위한 방법론이 하나의 의견으로 수렴되기보다는 다양한 의견이 제시됐다. 전문가F는 연구진이 제시한 모든 방법에 장단점이 존재하기 때문에 한 가지 방식으로 추천하는 것이 어려운 일이고 바람직하지 않다고 응답했다. 전문가C는 “공정성에 대한 지표와 개념, 분석방법은 분야와 상황에 따라 달리 적용되고 있으며 산업계 차원의 합의가 이뤄지지 않았다. 현재로서는 어느 특정 방법을 사용하기보다는 다양한 분석방법을 모두 적용하고, 결과를 검증하고 이를 통해 합의를 도출하기 위한 노력이 필요한 단계이다.”라고 강조했다. 특히 전문가C는 IBM의 AI Fairness 360 오픈소스 툴킷 사례와 구글이 2019년 11월에 발표한 공정성지표 서비스(베타) 사례에서 글로벌 IT기업들이 ‘알고리즘 공정성’에 대해 자체적으로 연구개발 투자를 진행하고 있다는 점을 강조했다. 구체적으로 IBM의 AI Fairness 경우 70개 이상의 공정성 지표와 10개의 최신 편향 완화 알고리즘을 포함하고 있는데 이는 개발자들이 공개적으로 사용할 수 있으며 구글의 공정성지표 서비스(베타)는 텐서플로 모델 분석, 공정성 지표,

텐서플로 데이터 검증, What if Tool (WIT) 등을 포함하고 있으며 Github에 공개된 점을 소개했다.

전문가들은 1) 이용자기반 실험, 2) 대규모 온라인 패널기반 실험, 3) 시스템 기반의 자료를 바탕으로 한 연구, 4) 시뮬레이션 기반 실험에 대해서 각각의 장단점을 설명했다. 1) 이용자기반 실험에 대해 전문가B는 실제 추천시스템을 특정 사이트에서 구현하고, 추천시스템의 공정성을 바꿀 수 있는 parameter(예측 변수 구성 혹은 최대화하고자 하는 objective function에 공정성 요소가 포함되는가)를 조정하며 A/B 테스트를 실시하고, 상품 구매 여부와 더불어 이용자에게 직접 경험한 알고리즘의 공정성(perceived fairness)을 물어보는 방법론이 의미 있을 것이라고 강조했다. 또한 이 방법은 실제 추천시스템이 작동하는 일상적인 상황에서의 공정성 인식을 이용자에게 직접 물어볼 수 있다는 장점이 있지만, A/B 테스트를 구현하고 시행하는 데 인력적인 비용이 클 수 있다는 단점을 지적했다. 전문가D는 실증 연구 방법론을 개발 활용하는 것이 중요한데 제안된 연구방법론 중 이용자 기반 실험이 가장 적절하다고 응답했다.

2) 대규모 온라인 패널 기반 실험의 경우 전문가E는 가상의 추천시스템 시나리오에 대해서 온라인 패널 설문방식으로 시도할 수 있는데 이는 충분한 모수를 확보할 수 있는 장점이 있지만 필드 실험방식에 대해서 외적타당성이 떨어질 수 있다는 문제를 지적했다. 전문가H는 ‘알고리즘 공정성’의 대리지표로 알고리즘 결과에 대한 납득으로 본다면 많은 사람들의 다양한 생각을 가늠할 수 있는 대규모 온라인 패널 기반 실험이 적절하다고 의견을 제시했다.

3) 시스템 기반의 자료를 바탕으로 한 연구에 대해서 전문가F는 이용자의 실제 행동 데이터를 분석하고 미충족된 욕구(unmet needs) 등을 도출할 수 있다는 장점이 있지만 이용자들이 지각하는 공정과 편향에 대한 생각을 알기 어렵다고 강조했다. 또한 시스템 기반의 자료를 바탕으로 분석할 경우 데이터를 얻고 분석하는 것이 현실적으로 어렵다는 점을 지적했다.

전문가E는 4) 시뮬레이션 기반 실험의 경우 여러 버전의 추천시스템을 돌려보면서 추천 정확도(accuracy)와 알고리즘 공정성(fairness) 지표를 비교할 수 있다는 장점이 있다고 강조했다. 다만 이때 공정성은 사용자가 느끼는 인지적 공정성이 아니라 통계적으로 정의된 공정성이 되기 때문에 장점은 통계적 공정성을 테스트할 수 있지만 반면 통계적 공정성과 인지된 공정성이 다를 수 있다는 단점이 동시에 존재한다고 응답했다. 한편, 전문가G는 공정성 분석을 위해 추천시스템 이용자의 피드백 정보가 필수인데 가상의 이용자를 생성하기 어렵고, 생성해서 시뮬레이션 한다고 하더라도 그 결과의 신뢰성을 담보하기 어렵기 때문에 현실적으로 분석하기 어렵다고 지적했다.

알고리즘 공정성을 분석하기 위한 적합한 분석방법으로 전문가들은 각 방식들을 보완 적용하거나 복합적으로 사용하는 것이 적절하다고 응답하는 경우가 많았다. 특히 전문가E는 각각의 장단점이 다르므로 복합적으로 사용하는 것이 가장 타당하다고 응답했다. 구체적으로 이용자기반 실험, 대규모 온라인 패널 기반 실험, 시뮬레이션 기반 실험을 보완적으로 사용해야 한다고 제안했다. 전문가D도 알고리즘 및 모델 개선과 개발에 있어서는 3번과 4번 방식이 필연적이지만 실제 적용을 위해서는 1번과 2번 방식으로 이어져야 한다고 다양한 방

식의 보완과 적용이 필요하다고 응답했다. 전문가F 또한 데이터 리소스가 충분하다면 한 가지 방식을 채택하기보다 다양한 방식의 연구를 분석해서, 공정하지 못한 부분에 대해서 극복하는 것이 더 현명한 방식이라고 제안했다. 그럼에도 불구하고 우선순위를 정한다면 3) 시스템기반의 행동데이터 분석에 이용자의 실제 요구나 목소리를 들을 수 있는 방식의 조합이 적절하다고 의견을 더했다. 전문가의 경우, 불특정 다수의 이용자 모집과 피드백 정보수집의 용이성 등을 고려하여 2) 대규모 온라인 패널 기반 실험이 가장 적합하다고 제안했다. 전문가A 또한 시스템 기반의 자료를 바탕으로 연구한 결과를 토대로 시뮬레이션 기반 실험을 진행하고, 대규모 온라인 패널 실험으로 나아가는 것이 적절하다고 응답했다.

추가적인 의견에 대해서 전문가D는 Ali et al(2019)에서 볼 수 있는 사례처럼 연구대상 플랫폼에서 실제 광고 등을 하여 노출 결과를 분석하는 방법도 시도할만한 연구방법이라고 제안했다. Ali et al(2019)는 페이스북 광고를 통해 게시된 채용공고가 직업 유형에 따라 특정 인종 또는 성별이 더 많이 노출된다는 연구결과를 발표했다. 예를 들어, 별목공 채용공고에 노출된 사람의 90% 이상은 남성이며, 70% 이상이 백인이었고 청소부 채용공고는 65% 이상 여성이며 75% 이상이 흑인인 것으로 나타난 사례를 소개했다.

<표 V-3> 방법론적 측면 전문가 의견

분류	전문가 의견
1) 이용자 기반 실험	- (전문가E) 실제 이용자에게 공정성 인식을 물어볼 수 있다는 장점이 있지만, A/B 테스트 구현에 발생하는 인력적 비용 우려 - (전문가D) 실증 연구방법론의 중요성에 기반, 가장 적합한 방법론
2) 대규모 온라인 패널 기반 실험	- (전문가E) 충분한 모수 확보가 가능하나 외적 타당성 저하의 문제 발생 우려 - (전문가H) ‘알고리즘 결과에 대한 납득’을 알고리즘 공정성의 대리지표로 본다면 적절
3) 시스템 기반의 자료를 바탕으로 한 연구	- (전문가F) 미충족된 욕구 등의 도출에 용이하나, 이용자들의 공정·편향에 대한 지각 파악에 어려움
4) 시뮬레이션 기반 실험	- (전문가E) ‘추천 정확도’와 ‘알고리즘 공정성’ 지표 비교 가능하나 통계적 공정성과 인지적 공정성의 차이 고려 필요 - (전문가H) 가상의 이용자 생성이 어려우며 이에 따른 결과의 신뢰성 저하 문제 우려
알고리즘 공정성에 대한 최적의 분석방법	- (전문가E) 1,2,4번 방식의 보완적 사용 - (전문가H) 알고리즘 모델 개발에 있어서는 3,4번 방식 활용이 필요하며 실제 적용을 위해서는 1, 2번 방식 활용 필요하며, 단일 방법론 선정 시 가장 적합한 방법은 2번 - (전문가F) 이용자의 실제 요구나 목소리를 파악 가능한 3번 방식이 가장 적합 - (전문가A) 3번 결과를 바탕으로 4번을 시행하고, 최종적으로 2번 방식을 활용

3. (연구문제3:제도적 접근) 알고리즘 공정성 보장방안을 위한 영향성 평가 제도가 필요한가?

알고리즘이 사회제도가나 시장에 적용되면서 추천시스템의 알고리즘 공정성 보장방안으로 제시되고 있는 ‘알고리즘 영향성 평가제도’에 대한 필요성에 대한 전문가 의견을 수집했다. 구체적으로 캐나다의 영향성평가 사례를 제시하고 평가주체와 평가대상에 대한 의견을 확인했다. 조사 결과, 대부분의 전문가들은 알고리즘 영향성 제도가 필요하다는 의견에 동의했으나 영향성 평가의 수행방식, 주체 등에 대해서는 사회적 합

의가 필요하다고 강조했다. 전문가D는 추천시스템 알고리즘의 영향성 평가가 리스크 관리의 중요한 수단으로 인공지능 시스템에 관한 법 규범이 확립되지 않은 현실에서 필요한 제도라고 응답했다. 그러나 영향성 평가를 수행하는데 있어서의 방식, 주체, 평가의 범위, 평가결과에 대한 대응방식은 우선적으로 선행 연구가 수행되어 그 연구결과를 기반으로 사회적 합의를 도출하고 관련 제도 및 거버넌스 구축이 필요하다고 응답했다. 특히, 전문가D는 법적인 관점에서 현재 시행되고 있는 영향성 평가 제도에 대한 한계를 제시했다. 즉, 국내에서는 기술영향평가, 개인정보영향평가, 환경영향평가, 교통영향평가 등 다양한 영향평가제도가 각종 법률에 도입되고 있지만 평가대상, 수행방식, 수행역량이 부족하다고 지적했습니다. 다른 전문가H는 자동화된 의사결정이 인간에게 미치는 영향에 대한 고려 없이 개발된다면 심각한 문제를 초래할 수 있으므로 알고리즘 영향성평가 제도가 반드시 필요하다고 강조했다. 전문가C는 AI 오남용이나 악의적 활용을 막기 위해 사전적 대응(precautionary approach)이 반드시 필요하기 때문에 알고리즘 영향성평가 제도가 반드시 필요하다고 강조했다. 특히 위험도에 기반 한 규제(risk based regulation)로서 적절한 조치라고 판단했다. 전문가A 또한 사전에 소스코드를 검사하고 규제하는 방식은 현실적 적용이 어렵기 때문에 캐나다와 같이 서비스 제공자의 투명한 답변을 전제로 하여 사회적 영향력을 평가하는 것이 유용한 방식이 될 것이라고 응답했다. 전문가A는 알고리즘 영향성 평가제도가 도입됐을 경우 구체적인 실행방식에 대해서 제안했다. 특히 ‘투명성’이 확보되는 것이 중요하다고 재차 강조하며, 투명하게 답변하고 그 결과를 시민에게 공개하는 방식으로 제공해야 한다고 제안했다. 구체적으로 “영향평가결과, 영향을 미칠 수 있는 수준이 어느 정도인지 서비스 제공 시 반드시 공개하도록 하고, 어떤 노력을 통해 영향 수준을 낮출 수 있었다는 기록도 함께 공개하는 방식이 가능할 것 같다. 시민들로 하여금 믿을 수 있는 서비스임을 인증하는 방식으로 운영하는 것이다.” 라고 응답했다.

전문가들은 알고리즘 영향성 평가제도가 수립되는 것에 대해서는 동의하고 있었지만 그 방향성과 실행 방향에 대해서는 다양한 의견을 제시했다. 전문가B 는 알고리즘 영향성제도가 필수적이지만 평가 주체가 시민사회가 되어야 한다고 주장했다. 전문가D는 알고리즘 영향성 평가제도는 독립성과 전문성을 갖춘 제3의 기관에서 수행하는 것이 바람직하고 타당하며, 추천시스템의 종류와 사업 범위 등을 고려해서 자율규제 형식으로도 설계가 가능할 것으로 판단했다. 이와 비슷한 의견으로 전문가H는 IRB(Institutional Review Board)처럼 전문가로 구성된 위원회가 맡아야 한다는 의견을 제시했다. 그리고 평가대상에는 실제적으로 적정 규모 이상의 단체나 회사가 받아야 한다고 응답했다. 전문가G 또한 영향성 평가를 진행한다면 공공영역에서 적용되어야 하고, 프로파일링 금지권 등 법제도 개선이 병행되어야 한다고 응답하며 영향평가의 주체는 별도의 기관을 설립해서 중립적인 견지에서 진행되어야 할 것을 강조했다. 전문가A는 알고리즘 영향평가를 공공영역에서 우선 실시한 뒤, 민간영역으로 전파 가능할 것이라고 봤다. 공공이 주체가 되더라도 그 과정과 결과, 사용된 방식이 모두 투명하게 공개되어야 한다고 강조했다. 특히, 평가대상이 민간영역으로 확산될 경우 미국 상원에서 발의된 “알고리즘 책무성 법안(Algorithm accountability bill)”의 규정을 활용가능하다고 응답했다. 전문가A는 알고리즘 책무성 법안에서 규정하고 있는 대상을 구체적으로 들며 이처럼 공공영역에서 출발한 후 확대하는 것이 적절하다고 강조했다. 구체적으로 알고리즘 책무성 법안은 그

대상을 (A) 최근 회계연도 기준 3년 간 평균 연 총 수입액이 5천만 달러를 초과한 기업, (B) (i)100만 명 이상의 소비자, 혹은 (ii)100만 개 이상의 소비자 기기에 대한 개인 정보를 소유하거나 통제하는 기업, (C) 위의 (A) 또는 (B) 요건을 충족하는 개인이나 파트너십, 혹은 법인 기업에 의해 실질적으로 소유, 운영, 통제되는 기업, (D) 고객이나 고용인이 아닌 개인에 대한 정보를 판매 및 거래하거나 제 3자에게 정보에 대한 접근권을 제공하기 위해 사업의 상당한 부분으로서 데이터를 수집, 축적, 유지하는 데이터 중개자 혹은 기타 상업적인 기업으로 규정하고 있다고 덧붙였다.

이와는 조금 다른 의견으로 국가 및 공공기관이 평가 주체가 되어야 한다는 의견도 제시됐다. 대표적으로 전문가C도 “지능정보화 기본법 제56조에 지능정보 서비스 등의 사회적 영향평가를 규정하고 있으며 이에 따라 사회적 영향평가의 주체를 국가 및 지방자치단체로 규정하고 있기 때문에 정부산하의 전문 공공기관에 해당 업무가 위임위탁 될 것으로 예상된다”라고 응답했다. 평가 주체는 공공기관이지만 평가 과정에서는 다양한 배경의 업계, 학계, 시민단체의 전문가들이 참여해서 평가가 이루어지도록 해야 한다고 응답했다. 평가대상에 대해서는 지능정보화 기본법 제56조에서 “국민의 생활에 파급력이 큰 지능정보서비스”로 분류하고 있기 때문에 모든 자동화 추천시스템이 아니라 국민 생활에 파급력이 큰 인공지능을 대상으로 해서 그것을 도입한 국가기관, 공공기관, 민간기업이 모두 대상이 되어야 한다고 응답했다. 이와 비슷하게 전문가E는 공정성 및 영향성 평가의 주체가 국가 및 공공기관이 되어야 하며 추천시스템을 구현하여 사용하는 모든 민간기업 및 공공기업이 평가의 대상이 되어야 한다고 강조했다. 특히, 사기업의 경우 상품/서비스 추천 알고리즘 중에서도 금융이나 교육, 의료와 관련된 서비스나 상품은 대중의 일상생활에 큰 영향을 미치므로 평가의 대상이 되어야 한다고 강조했다. 구체적으로 전문가E는 1차적으로는 국가 차원에서 알고리즘 공정성에 대한 자체 평가 가이드라인을 제공하며, 2차적으로는 적극적인 감사방식을 통해 공공부분 알고리즘의 공정성을 관리해야 한다고 응답했다. 특히 추천시스템 개발자가 해당 알고리즘의 편향성에 대해 가장 잘 알고 있고, 외부 감사자가 완벽히 코드를 이해하고 편향성을 파악하는 것이 어려울 수 있으므로 자체 평가 가이드라인이 중요하다고 강조했다.

알고리즘 영향성 제도가 필요하지 않다고 응답한 전문가들은 알고리즘 영향성 평가제도가 특정 분야에 적용되어야 한다는 의견을 제시했다. 전문가 F는 대부분의 알고리즘이 적용되는 분야에서는 알고리즘 영향성 평가제도가 무의미할 것이라고 지적했다. 특히 알고리즘을 통해 나타나는 추천 결과는 빅데이터와 정보의 홍수 속에서 효율적으로 이용자 개인에게 적합한 정보를 제공하기 위한 수단 중 하나이기 때문에 영향성 평가 제도가 어떤 분야에서 필요한지, 필요한 정도의 피해가 있는지 등에 대한 연구결과와 사회적 합의가 우선 선행되어야 한다고 강조했다. 예를 들어, 연구진이 제시한 캐나다의 사례의 경우 알고리즘이 적용되는 공공분야가 영향력이 크다고 판단하여 영향성 제도를 도입한 것으로 판단되지만 이 사례만으로는 공공부분이라고 영향성 평가가 다 필요하다고 동의하기 어렵기 때문에 합리적인 접근이 필요하다고 응답했다. 특히, 공공부분의 경우 사회적 합의를 통해 반드시 필요하다고 생각되는 분야이면서 영향성 평가가 중립성을 지킬 수 있고 전문성을 보유한 곳에서 실시되는 것이 합리적이라고 강조했다. 민간과 기업의 경우 영향성 평가 제도가 오히려 인공지능과 빅데이터 분석의 경쟁력 저하를 막을 가능성이 있기 때문에 영

향력 평가는 가급적으로 자율에 맡겨야 한다는 의견을 제시했다. 또 다른 전문가 I는 추천시스템이 적용되는 도메인에 따라 구분될 필요가 있다는 의견을 제시했다. 예를 들어, 뉴스콘텐츠의 경우 사회적으로 중요한 결정을 해야 하는 주제에 대해서 특정 목적을 가진 뉴스들만 편향적으로 뉴스 소비자에게 전달된다면 왜곡된 진실이 발생하여 사회적 비용이 발생하기 때문에 뉴스콘텐츠와 같이 도메인에 대한 추천시스템의 편향성은 공공기관이 관리해야 할 필요성이 있다고 강조했다. 그러나 그 외에 상품이나 제품에 대한 추천은 본래 목적이 이윤창출로 자본주의 시장원리에 따라 편향이 발생하는 것이 자연스럽다는 의견을 제시했다.

<표 V-4> 제도적 측면 영향성 평가제도에 대한 전문가 의견

분류	전문가 의견
영향성평가제도가 필요하다	- (전문가D) 리스크 관리의 중요한 수단으로 인공지능 시스템에 관한 법규범이 확립되지 않은 현실에서 필요한 제도임. 그러나 세부적인 시행방식에 대해서 선행연구가 수행되고, 사회적 합의가 도출되어 제도 및 거버넌스 구축이 필요함. 영향성 평가 제도는 독립성과 전문성을 갖춘 제3의 기관에서 수행하는 것이 바람직함 - (전문가H) 자동화된 의사결정이 인간에 미치는 영향에 대한 고려 없이 개발된다면 심각한 문제를 초래할 수 있으므로 영향성 평가제도가 반드시 필요함. 전문가로 구성된 위원회가 맡아야 하며 평가대상은 적정규모 이상의 단체나 회사가 받아야 함 - (전문가C) AI오남용이나 악의적 활용을 막기 위해 사전적 대응(precautionary approach)이 반드시 필요함. 정부산하의 전문 공공기관에 업무가 위임위탁되나, 평가과정은 업계, 학계, 시민단체 전문가들이 참여해야 함 - (전문가A) 사전에 소스코드 검사하고 규제하는 방식이 어렵기 때문에 캐나다와 같이 서비스 제공자의 투명한 답변을 전제로 한 영향성 평가제도가 유용할 수 있음 - (전문가G) 공공영역에서 우선 적용되어야 하며, 영향평가의 주체는 별도의 기관을 설립해야 함 - (전문가E) 공공성 및 영향성 평가의 주체가 국가 및 공공기관이 되어야 하며, 추천시스템을 구현하는 모든 민간기업 및 공공기업이 평가대상이 되어야 함
영향성평가제도가 필요하지 않다	- (전문가F) 알고리즘 적용되는 분야에서 알고리즘 영향성 평가가 무의미할 것. 영향성 평가제도에 대한 연구결과와 사회적 합의가 우선 선행되어야 함 - (전문가I) 추천시스템이 적용되는 도메인에 따라 구분될 필요가 있음. 예를 들어 뉴스콘텐츠를 제공하는 경우에는 필요하지만 상품이나 제품에 대한 추천은 자본주의 시장원리에 따라야 하므로 영향성 평가제도가 적합하지 않음

4. 추가적인 전문가의견

추가적으로 앞으로 사회에 영향을 미칠 수 있는 자동화된 추천알고리즘에 대해서 전문가 의견조사를 진행했다. 다양한 알고리즘 사례들이 제시되었는데 전문가B는 정부와 공공기관에서 도입할 자동화된 추천시스템 알고리즘이 시민사회에 큰 영향을 미칠 것이라고 응답했다. 구체적으로 행정안전부가 입법예고한 전자정부법 일부개정법률안에서 인공지능을 활용한 지능형 행정처리 근거와 이에 대한 행정적, 기술적, 재정적 지원에 대한 법적 근거를 마련하고 있어 챗봇, 스마트시티 등 정부의 각 영역에 인공지능이 도입되고 시

민생활에 큰 영향을 끼칠 것이라고 응답했다. 전문가C는 “뉴스·미디어와 문화콘텐츠 추천시스템은 민주시민으로의 가치관과 역량 형성에 있어서 영향을 미치고, 상품, 금융서비스에 관한 추천시스템은 개인의 삶의 질과 경제생활에 영향을 미치고, 직장 및 노동, 인재채용 채용시스템은 근로 방식과 환경에 영향을 미치기 때문에 우선적으로 주목하여야 한다.”고 응답했다. 추가로 복수의 전문가들은 딥러닝을 활용한 추천시스템 알고리즘을 주목해야 한다고 응답했다. 전문가E는 CNN, RNN, LSTM 등 초기 방식보다 연합학습(Federated Learning)이나 GAN(Generated Adversarial Networks)가 주목할만하다고 응답했다. 특히 연합학습은 개인의 행태를 개인화하고, 다른 수많은 사용자 정보를 업데이트한 내용으로 다시 업데이트하고 이 과정을 무한반복하는 방식으로 다른 이용자의 특성을 반영한 개인화 추천으로 주목할만하다고 강조했다. 또한 전문가D는 딥러닝을 활용해 개별 고객이 개별상품 및 콘텐츠를 소비할 확률을 직접 예측하는 방식이 주목받을 것이라고 예측했으며 구체적으로 아래와 같이 설명했다.

“과거에는 추천시스템만을 위하여 설계된 CF 등이 주를 이루었다면, 최근에는 수집되는 고객 데이터의 양과 질(granularity와 accuracy)이 모두 향상됨에 따라 고객의 web/app log를 포함한 digital footprint를 예측변수로 하여 개별 고객-상품 조합에 대하여 구매 확률을 직접 예측하는 방식이 각광받지 않을까 생각된다. Input data로는 앞서 언급한 고객의 web log가 중요한 역할을 할 것이며, 그중에서도 사용자가 해당 사이트에서 보이는 행동 패턴이 중요한 지표가 될 것이다(전문가D).”

반면 전문가G는 성공적인 추천시스템은 인공지능모델인 딥러닝(deep learning)을 적용한 기술들이 많이 있는데, 이는 필연적으로 딥러닝의 단점을 같이 가져간다고 지적했다. 대표적으로 판단에 대한 이유를 드는 것이 부족하다는 점, 학습데이터가 많이 필요하다는 점, 학습에 있어서 시행착오가 요구된다는 점이 있다고 응답했다. 그래서 전문가G는 “판단에 대한 이유 설명가능성 향상 알고리즘”에 대해 주목하고 있다고 응답했다. 구체적으로 “판단은 추천시스템에 최종적인 추천 결과로, 이 추천 결과가 왜 나타났는지 이유를 알아야 잘못되거나 위험한 추천 결과를 찾아내고 이를 수정 혹은 삭제 가능하다. 이러한 이유를 통해서 사용자에게 대한 신뢰를 향상시킬 수 있다.”라고 주장했다. 전문가B는 협업필터링이 딥러닝 알고리즘과 결합하거나 Item2Vec과 같은 알고리즘과 결합하기도 한다고 응답했다.

전문가들은 인공지능과 제도적인 개선에 관해서 다양한 의견을 제시했다. 전문가A는 ‘시민과의 협업’이 중요하다고 재차 강조했다. 그 예로 미디어 스타트업인 ‘더 마크업(The Markup)’의 최근 두 가지 프로젝트를 소개했다. 구체적으로 전문가A는 아래와 같이 소개했다. 아래 사례를 통해 시민들과의 협력을 통해 알고리즘의 공정성을 보장할 수 있도록 나아가야 할 필요가 있다고 제안했다.

“‘더 마크업’이 최근 만든 블랙라이트(Blacklight, <https://themarkup.org/blacklight>)는 해당 사이트가 내 개인정보를 얼마나 수집하는지를 분석해서 보여준다. 응답해당 사이트에 애드 트래커가 있는지, 써드파티 쿠키가 있는지, 쿠키 추적을 하는지, 나의 접속 동안 행위(세션)를 기록하는지, 키보드 후킹을 하는지, 페이스북 픽셀이 설치돼 있는지, 구글 애널리틱스의 ‘remarketing audiences’가 활성화돼 있는지 등을 30초~1분 사이에 분석해 제공한다. 이렇게 쌓인 정보들은 다시 새로운 사이트 분석에 활용됩니다. 시민들의 자발적 협력을 유도하기 위해 블랙라이트를 만든 것이다. 또, ‘더 마크업’이 최근 시작한 ‘시민 브라우저 프로젝트(The Citizen Browser Project)’는 시민들과의 협업을 최우선에 두고 있다. 페이스북과 유튜브 등 소셜 미디어를

통해 허위정보가 어떻게 확산되는지를 분석하는 것을 목적으로 추진 중이며, 1,200명의 시민들에게 ‘더 마크업’이 만든 브라우저를 사용하게 해 이들의 브라우저에서 페이스북과 유튜브가 수집하는 정보들과 계정 활동 내역을 실시간으로 수집해 분석하는 방식이다. 시민 패널들에게 보상비를 주고, 표본 할당을 통해 구성하며 물론 개인 식별 정보는 삭제한다. 이를 통해 플랫폼이 어떤 콘텐츠를 확산하고 있는지를 추적해 밝히는 것이다 (전문가A).”

전문가들은 알고리즘 공정성에 대해서 매우 많은 관심이 필요하다고 재차 강조했다. 전문가C는 알고리즘의 기술과 서비스의 궁극적인 목적이 인간의 좋음(betterness)에 기여하는 것으로, 알고리즘 공정성이 학계, 산업계, 정치/행정, 시민 등 각계 각층이 관심을 가지고 지속적인 연구가 필요한 영역이라고 응답했다. 또한 전문가H는 ‘알고리즘 공정성’에 대한 다학제적 접근과 융합연구가 필요하다고 응답했으며 전문가I는 추천시스템 알고리즘의 공정성이 요구되어야 하는 응용도메인과 그렇지 않은 도메인을 구분하는 것이 옳은지 아닌지가 논쟁중이지만 이 연구뿐만 아니라 다양한 연구가 결실을 맺어 사회적 공공성이 요구되는 편향을 제거하여 투명하고 공정한 사회를 이루는데 기여해야 한다고 응답했다. 전문가G는 신뢰기반의 AI생태계를 위해 국가와 기관별 정책 마련의 필요성을 촉구했다. 신뢰기반 AI생태계를 확보하기 위해 전방위적 제도개선이 병행되어야 하며, 한편으로는 개발자 대상의 윤리강령을 보편화하고 이를 기업문화로 정착시키는 것이 중요하다고 강조했다. 전문가E는 공정성을 반영한 알고리즘을 크게 3가지로 분류하며 이에 대한 방법론을 구체적으로 아래와 같이 제안했다.

“일반적으로, 공정성을 반영한 알고리즘은 크게 3가지로 나뉜다. 이런 알고리즘을 bias mitigation algorithm이라 하고, 전처리 단계에서 적용되면 pre-processing bias mitigation algorithm, 모델 학습 도중 적용되면 in-processing bias mitigation algorithm, 모델 학습 후에 적용되면 post-processing bias mitigation algorithm이라고 한다. 이렇게 3가지 타입의 방법 중에서도 공정성 측면에서는 앞의 2가지 방법론이 좀 더 인지된 공정성이 좋을 수 있을 것 같다. 그 이유는 pre-processing bias mitigation이나 in-processing bias mitigation이 사후적으로 결괏값을 조정한다는 느낌이 적으므로 대중이나 사용자에게 공정성이 해쳐진다는 느낌을 덜 줄 수 있기 때문이다(전문가E).”

전문가F는 해당 연구와 관련하여 여러가지를 제안했는데 첫째, 알고리즘 추천에 의한 공정성 이전에 알고리즘을 제외한 추천 공정성에 대해서도 함께 고민이 필요하며 둘째, 공정성이 적용되는 상황을 구분할 필요가 있다고 제안했다. 예를 들어, 편향성이 용인되거나 사회적으로 감내할 수 있는 분야가 있는 반면 공정성에 대한 경계와 주의가 필요한 부분을 잘 정리해야 한다는 것이다. 셋째, 추천 결과에 대한 공정성에 초점을 맞추기 보다는 알고리즘을 구성하는 프로세스의 공정성에 대한 관심이 중요하며, 넷째 알고리즘 공정성에 대해서는 쉽게 정의내리기 어렵지만 어떻게 달성해야 하는 것도 어려운 문제이기 때문에 다양한 방식으로 확장하여 생각할 필요가 있다고 강조했다. 예를 들어, 다양한 방식의 추천 알고리즘을 제공하거나, 알고리즘이 아닌 다른 추천방식을 함께 제공하거나 선택할 수 있게 보완하는 방식 등을 고려할 수 있다. 마지막으로 공정성에 대한 논의에 있어서 알고리즘 추천방식에 대한 올바른 이해와 이용을 위한 교육과 리터러시를 개발하는 노력이 함께 병행될 필요가 있다고 강조했다.

제6장 결론 및 제언

제1절 주요 연구요약

1. 추천시스템 알고리즘의 현황 및 문제점

추천시스템 알고리즘은 데이터 기반의 자동화된 의사결정(automated individual decision-making)을 의미한다. 과거에는 추천행위가 인간이 개입하여 이뤄졌다면 기술이 빠르게 발전하고 데이터가 빠르게 쌓임에 따라 데이터를 기반으로 하여 인간의 의사결정에 도움을 주는 추천 알고리즘의 역할은 매우 커지고 있다. 알고리즘 기반 추천시스템은 인간이 인위적으로 구조화한 절차에 따라 결과를 출력하는 것이 아니라 머신러닝 과정을 통해서 시스템 내에서 분류를 하고 예측절차를 거쳐서 추천결과를 내놓게 된다. 디지털기술이 일상화됨에 따라 소비시장은 급격하게 확대되고 소비 및 판매 양측에 도움을 주는 추천시스템이 등장하게 됐다. 알고리즘 기반 추천시스템은 대표적으로 협업필터링, 콘텐츠기반, 딥러닝 기반 추천시스템, 이용자 선택기반 추천시스템 방식이 있다. 협업필터링은 기존 사용자의 행동정보를 이용해 사용자와 비슷한 성향의 사용자가 선호하는 아이템을 추천하는 방식으로 비자각에서 오는 공정성은 만족할 수 있으나 콜드스타트 문제가 발생한다. 콘텐츠기반 필터링은 아이템 항목의 특성(종류, 브랜드 등)을 분석한 프로파일과 이용자 특성을 분석한 프로파일을 추출해서 추천에 활용방식을 말한다. 내용 자체를 분석하기 때문에 콜드스타트 문제를 해결할 수 있지만 프로파일을 기반으로 수행되기 때문에 비자각에서 오는 공정성을 만족할 수 없다. 딥러닝 기반 추천시스템은 어떤 방법이 새롭게 있다기보다 딥러닝 방법론이 추천시스템에 적용되는 것이라고 할 수 있다. 예를 들어 딥러닝을 적용해서 콘텐츠 기반 필터링 성능을 끌어올리는 것, 협업필터링 핵심기술인 행렬분위를 효과적으로 수행하는 방법을 찾는 것 등을 들 수 있다. 마지막으로 이용자 선택 기반 추천시스템은 이용자가 선택한 내용들을 기반으로 추천이 이뤄지는 것을 뜻한다. 이러한 추천시스템의 알고리즘은 정부, 비즈니스, 쇼핑 및 상거래, 교육, 여행 등의 다양한 분야에서 적용되고 있다.

추천시스템의 공정성 개념을 파악하기 위해서는 우선 공정성 개념에 대해서 살펴볼 필요가 있다. 공정성이라는 것은 어떠한 사유로도 차별받지 않아야 한다는 것을 의미한다. 차별이라는 것은 보호적 속성에서 비롯되는 직접차별, 독립적이고 보호되지 않는 속성에 의한 간접차별, 비체계적으로 이뤄지는 구조적 차별, 통계적 결과를 바탕으로 개인을 미리 예측하는 통계적 차별, 외생변수를 통제하지 않아 발생하는 비체계적 차별 등이 있으며 공정성은 이러한 차별의 배제에서부터 시작하게 된다. 특히 알고리즘에서 편향과 차별 같은 불공정이 나타날 수 있는 이유를 두 가지 관점에서 볼 수 있다. 첫째는 사용하는 데이터의 문제로 발생할 수 있기 때문에 알고리즘 공정성을 확보하기 위해서는 광범위한 그룹으로부터 훈련 데이터를 수집하고 충분한 다양성을 확보하도록 해야 한다. 다른 이유로는 비자각에 대한 공정성이 발생할 수 있다. 인간이

처리했던 결과의 데이터가 이미 불공정이 포함되어 있을 수 있기 때문에 보호속성을 의사결정의 수단으로 사용하지 않아야 하지만 또 한편으로 그렇게 된다면 보호속성으로 인해서 차별받는 의도적 불공정을 보정할 수 없게 되어 공정성을 담보할 수 없게 된다. 공정성을 보장하기 위해서 법률적인 고려사항으로 상호교차성과 절차적 공정성을 고려할 수 있다. 추천시스템에서 공정성을 확보하기 위해서는 다음과 같은 전제조건이 필요하다. 첫째, 추천알고리즘은 인구사회적 속성과 무관한 추천이 이루어져 분류의 공정성을 전제하고 있어야 한다. 둘째, 알고리즘에 활용되는 학습데이터가 인간의 편향이 개입되어 있을 수 있으므로 알고리즘 결과 자체를 그대로 믿는 것이 아니라 결과를 검토하는 과정이 필요하다. 추천시스템 공정성과 관련된 기존 연구들을 살펴보면 주제관점의 공정성과 대상관점의 공정성 연구, 집단관점의 공정성 연구경향으로 살펴볼 수 있다.

추천시스템에서의 공정성을 탐지하기 위해서는 전통적으로 시스템 중심 접근과 이용자 중심 접근이 사용되고 있다. 시스템 중심 접근은 이용자 데이터셋과 추천시스템이 예측한 이용자들의 예측취향 데이터셋이 얼마나 정확히 일치하는지 살펴보는 것이다. 시스템 중심 접근방법은 플랫폼을 처음 이용하는 이용자에게서 나타나는 콜드스타트 문제, 실제 서비스 이용자들의 이용경험 만족도를 충실히 반영하지 못한다는 단점 등이 지목된다. 이용자 중심 접근은 측정방식은 전통적인 형태의 실험이나 서베이를 통해 사용자 만족을 측정하는 방법인데 측정방법이나 변수설정 등에 따라 일관된 결과를 보여주지 못한다는 한계가 있다. 이처럼 추천시스템 서비스 이용자 연구의 이론적이고 방법론적인 공백을 채우기 위해 닌넨버스와 윌렘센(Knijnenburg & Willemsen, 2015)는 ‘추천시스템의 이용자 중심 평가모델’을 제언한 바 있다. 해당 모델은 추천시스템 관련 연구자들이 통일적인 방법론 틀에서 연구를 진행하게 하여 연구의 신뢰성과 일관성을 확보하는데 기여했지만 이용자 중심 연구의 전통적인 실험실연구와 서베이 방법 틀 안에서 제시되고 있다는 점에서 한계가 있다. 추천시스템 공정성평가에 대한 기존 접근을 살펴보면 온라인 동영상 공유 플랫폼 등에서 정치적으로 편향된 콘텐츠가 추천되는 등 추천시스템의 부작용에 대해서 문제가 제기되고 관련 연구가 진행되고 있다. 한편으로는 자동화된 소프트웨어에 기반하여 알고리즘 편향이나 극화를 추적하려는 연구들이 진행되고 있다. 예를 들어, 리베이루(Ribeiro et al., 2020)는 33만건이 넘는 동영상에 대한 양적 분석을 진행했지만 계속 변화하는 알고리즘의 한정된 순간만 분석할 수 밖에 없다는 한계를 보였다. 유튜브의 800개 이상 채널에 추천내역을 분석해서 네 가지 타입의 급진화를 검증한 또 다른 연구에서는 개인의 과거 동영상 시청 내역이 추천 결과물에 큰 영향을 주지 않는 것으로 파악했다(Ledwich & Zaitsev, 2019). 이러한 연구에서 살펴볼 수 있듯이 추천알고리즘은 블랙박스 형식으로 작용하면서 끊임없이 변화하기 때문에 추적이 어렵다는 한계가 있다.

2. 추천시스템 공정성 확보를 위한 정책 제도적 방향

추천시스템의 공정성 확보를 위한 정책 제도적 방향을 수립하기 위해 두 가지 측면에서 연구가 진행됐다. 첫째는 알고리즘의 공정성을 확보하기 위한 정책 제도적인 실제 사례들을 조사한 것이다. 둘째는 전문가조사를 통해 알고리즘 공정성의 개념과 공정성을 확보할 수 있는 방법론에 대해 탐구한 것이다. 우선 알고리즘 공정성을 확보하기 위해서 국내외의 공공 및 민간 영역에서 다양한 사례들을 찾아볼 수 있었다. 개발단계에서의 투명성 확보를 위해 리소스가 풍부한 구글, IBM과 같은 대기업들은 알고리즘 공정성을 개선할 수 있는 다양한 ‘공정성 툴킷’을 개발하고 있다. 특히 IBM에서 2018년에 개발해 오픈소스 서비스로 공개적으로 제공되고 있는 “AI Fairness 360”이 대표적이다. 구글에서 개발한 ‘What-if’ 툴킷은 Google의 텐서플로에 추가 및 내장되어 사용자로 하여금 머신러닝 모델의 공정성을 쉽게 분석하도록 지원하고 있다. 이 외에도 Microsoft사에서 개발된 오픈소스툴킷인 ‘Fairlearn’, LinkedIn의 오픈소스 공정성 툴킷인 ‘LiFT’ 등의 사례를 들 수 있다. 이러한 ‘공정성 툴킷’은 오픈소스 형태로 제공되면서 개발자들이 알고리즘 공정성을 보장할 수 있도록 하는 데 도움을 주고 있다.

개발단계에서의 노력 외에도 기업들은 자율적인 규제활동을 통해서 공정성 보장에 노력하고 있다. 구글은 ‘책임있는 AI(Responsible AI)’의 프레임워크 내에서 사회적 가치를 수행하는 책임있는 AI개발에 대한 실행지침을 공유하고 있으며 IBM은 ‘신뢰할 수 있는 AI(Trusted AI)’를 만들기 위해 공정성, 가치정렬, 강건성, 설명가능성, 투명성과 책임감과 같은 다섯가지 개념을 핵심 역량으로 강조하고 있다. 국내 사례로는 카카오 알고리즘 윤리헌장을 들 수 있는데 특히 최근에는 개발자가 인위적으로 개입해 차별이 발생할 가능성이냐 아동과 청소년 같은 취약계층을 보호하는 강령을 신설한 바 있다. 이러한 알고리즘 윤리헌장은 국내 기업 최초로 인공지능 알고리즘 설계 시 고려해야 할 사항을 명문화했다는 데 의의가 있다. 이 외에도 한국인공지능윤리협회에서는 ‘인공지능 윤리헌장’을 통해 개발자가 지켜야 하는 세부 강령을 제안한 바 있다.

기업의 자율활동 외에도 정부 차원에서 제도적인 활동을 통해 알고리즘 공정성 확보를 위해 노력하고 있는데 가장 대표적으로 EU의 GDPR(General Data Protection Regulation)을 들 수 있다. GDPR에서는 알고리즘 투명성(algorithmic transparency)에서 ‘공정하고 투명한 정보처리 원칙’을 강조하고 있으며, 자동화된 의사결정에 대해서 ‘설명을 요구할 권리’를 명시하고 있다. 미국 FTC에서는 <AI와 알고리즘 이용에 대한 지침(User Artificial Intelligence and Algorithms)>을 발표하여 이용자의 권리를 보호하기 위해 알고리즘을 어떻게 관리해야 하는지 구체적으로 명시하고 있다. 국내에서는 지능정보사회 윤리가이드를 통해 지능정보기술이 지켜야 할 공통원칙을 공공성, 책무성, 통제성, 투명성으로 제시하고 있다. 마지막으로 알고리즘 영향성평가(Algorithmic Impact Assessment) 제도를 통해 알고리즘이 사회 및 공중에 미칠 수 있는 영향을 평가할 필요가 있다. 영향성 평가제도를 통해 정책입안자, 연구자, 개인, 개발자들이 책무성 확보에 기여하는 데 기여 할 수 있다. 알고리즘 영향성 평가제도는 1) 시행 전 검토단계, 2) 초기 기관정보 공개요구단계, 3) 비평단계, 4) 법적절차 검토단계, 5) AIA 갱신단계로 나눌 수 있다. 이러한 영향성 평가제도는 일상에 적용되는 알고리즘의 영향성을 평가하고 발생할 수 있는 편향에 대응할 수 있는 수단으로 활용될 수 있다.

3. 추천시스템 공정성 보정을 위한 시뮬레이션 기반 실험

협업필터링 기반의 추천시스템이 이용자 측면에서 가지는 편향성과 공정성에 주목하여 시뮬레이션 실험 연구를 진행했다. 본 연구에서는 추천시스템의 편향성을 이용자 그룹별 입력데이터의 아이템 카테고리의 비율과 추천 결과물에서의 해당 비율의 차이로 규정했다. 또한 입력데이터에서 이용자 그룹 규모의 불균형으로 인해서 발생하게 되는 편향성을 보정하는 방법을 제안하는 데 목적이 있다. 이를 위해 본 연구는 세 가지 연구문제를 규정했다. 첫째, 가상데이터를 이용한 실험환경에서 협업필터링 기반의 추천시스템이 두 이용자 그룹에 대해 이용자 수 불균형에 따라 추천결과물의 편향성에 차이를 보이는지, 둘째, 규모에 따른 편향성이 존재할 경우 인위적인 이용자 데이터 증대 방법을 통해 편향성의 차이를 줄일 수 있는지, 셋째, 제안된 데이터 증대방법의 추천시스템 보정효과는 실험환경과 실제 환경에서 어떤 차이를 보이는지다. 이 실험을 위해 가상으로 생성한 행동기록 데이터(synthetic data generation)를 이용하여 편향성의 차이를 확인했으며 해당 방법론을 MovieLens의 실제 데이터에 적용했다.

시뮬레이션 실험에서는 가상의 사용자와 그들의 평가 정보 데이터를 만들고 이를 이용해 User-KNN과 SVD알고리즘을 이용해 추천시스템을 구현해서 실험을 진행했다. User-KNN 알고리즘을 이용한 경우 기존의 사용자가 선호했던 정도보다 더 증폭된 정도로 항목을 추천해주는 것이 확인되어 알고리즘으로 인해 편향이 발생할 수 있음을 확인했다. 알고리즘 편향을 보정하기 위한 방안으로 해당 실험에서는 입력데이터를 전처리하여 편향성 차이를 줄일 수 있는 네 가지 방법을 제안했다. 네 가지 방법은 가상의 데이터를 이용한 경우에는 편향성 감소 효과를 극명하게 보이고 있었으나, 실제 데이터를 이용한 실험에서는 방법 혹은 집단에 따라 효과가 다르게 나타났다.

실제 추천시스템에서는 이번 연구에서 사용한 User-KNN과 SVD와 같은 협업필터링(collaborative filtering) 방법 외에도 콘텐츠 기반 필터링(content-based filtering) 알고리즘 등 다양한 방식으로 알고리즘이 사용된다. 그러므로 이번 연구에서 제안한 데이터 증대 방법이 모든 추천시스템에서 편향성 차이를 감소시킨다고 결론짓기는 어렵다. 그러나 이 연구는 편향성 및 편향성 차이의 정의를 기존 관련 연구에서 자주 쓰이는 정의를 채택하였으므로 이를 활용해 여러 플랫폼의 추천시스템 편향성 차이를 실험을 통해 확인하는데 기여할 수 있다. 추후 연구에서는 추천시스템 외에 다양한 알고리즘을 이용해서 편향성의 감소 효과를 보이는지 확인이 필요하다.

4. 추천시스템 공정성을 평가 및 보정 방법론에 대한 전문가의견

이후 이 연구에서는 학계와 산업계 각 분야들의 전문가들을 대상으로 심층조사를 진행함으로써 알고리즘 공정성에 대한 개념을 파악하고, 공정성을 보장하기 위한 방법론적 쟁점을 파악하고 공정성 보장을 위한 정책적인 방향성을 확인했다. 우선 ‘알고리즘 공정성’ 개념에 대해 보편적 정의가 가능하다고 판단한 전문가들은 ‘추천결과의 편향성’, ‘보편인권에 벗어나지 않음’ 등으로 정의하고 있었으며 핵심개념으로 ‘차별과 편향의 배제’, ‘투명성’ 등을 뽑았다. 반면 다수의 전문가들은 ‘알고리즘 공정성’을 보편적으로 개념화하

기 어렵다고 응답했다. 그 이유로 알고리즘 혹은 인공지능은 비도덕(amoral)의 차원이기 때문에, 공정성은 수학적으로 정량평가가 가능해야 하지만 알고리즘 공정성은 이러한 것이 어렵기 때문에, 알고리즘이 편향되지 않고 중립적이게 제공하는 것은 매우 어려우며 중립적이라고 하더라도 그것을 공정하다고 보기 어렵기 때문에, 공정성이라는 개념은 사회문화적인 배경에 따라 달라지기 때문 등 여러가지 이유를 설명했다. 이를 보완할 수 있는 대체개념으로 한 전문가는 ‘사람들이 알고리즘 결과에 대해 얼마나 만족 혹은 납득하였는가’를 사용할 수 있다고 응답했다. 그러나 모든 전문가들이 ‘알고리즘 공정성’에 대한 더 많은 연구와 투자가 이뤄져야 한다는 점을 강조했다. 전문가들을 대상으로 알고리즘 공정성을 분석하기 위한 가장 적합한 분석방법에 대해서 조사했을 때 하나의 의견으로 수렴되기보다는 다양한 의견이 제시됐다. 대부분의 전문가들은 이용자 기반의 실험, 대규모 온라인 패널 기반 실험, 시스템 기반의 자료를 바탕으로 한 연구, 시뮬레이션 기반 실험 등 다양한 분석방법론이 모두 장단점이 있으며 이를 통합적으로 사용하여 각 방법론의 단점을 보완하는 방향으로 나아갈 필요가 있다고 응답했다. 마지막으로 알고리즘 공정성을 보장하기 위한 방안으로 영향성 평가제도에 대한 전문가 의견을 조사했을 때, 반드시 필요하다는 의견과 무의미하다는 의견이 모두 존재했다. 영향성 평가제도가 필요하다고 응답한 전문가는 AI 오남용이나 악의적 활용을 막기 위해서는 사전적 대응이 반드시 필요하다고 강조했다. 다수의 전문가들이 영향성 평가제도에 대한 필요성에 대해서 동의했지만 그 방향성과 실행방향에 대해서는 다양한 의견을 제시했다. 알고리즘 영향성 평가제도의 주체는 독립성과 전문성을 갖춘 제3의 기관이나 전문가로 구성된 위원회, 시민단체, 정부기관 등 다양한 의견이 제시됐다. 그 평가대상에 대해서도 공공영역 우선 적용 후 확대돼야 한다는 의견, 어느정도 규모 이상의 기업이 모두 포함되어야 한다는 의견 등 다양했다. 그러나 영향성 평가제도가 필요없다고 응답한 전문가는 알고리즘 적용되는 분야에서 알고리즘 영향성 평가제도가 무의미하며, 알고리즘 영향성 평가제도에 대한 우선적으로 연구결과와 사회적 합의가 선행되어야 한다고 강조했다. 또 다른 전문가는 알고리즘이 적용되는 분야에 따라 영향성 평가제도의 적용 유무가 바뀌며, 시장경쟁에 기반하여 추천알고리즘이 적용되고 있는 분야에서는 적절하지 않다고 응답했다.

제2절 제언

1. 추천시스템 공정성 보정을 위한 방법론적 제언

추천시스템 알고리즘은 다양한 요인에 의해 편향이 발생한다. 편향은 1) 알고리즘 조작, 2) 알고리즘 기능·설계상의 오류, 3) 머신러닝의 바탕이 되는 학습데이터의 편향이 주된 이유이다. 추천시스템의 편향이 미치는 영향은 이미 페이스북의 감정전이 실험, 구글의 검색 알고리즘, Epstein 등의 편향효과 실험 등에서 이미 확인된 바 있다(e.g., Caliskan, 2018; Blodget et al., 2018; van Miltenburg, 2017).

추천시스템의 영향이 가시화됨에 따라 이용자 보호 관점에서 추천시스템 알고리즘의 편향을 줄이고 공

정성을 확보하기 위한 다양한 노력이 진행되고 있다. 이 주제가 다학문적으로 진행되고 있는 관계로, 추천 시스템 알고리즘의 공정성에 대한 합의된 개념 정의가 부재한 가운데, △공정성 개념의 조작적 정의의 문제, △ 탐지방법의 어려움 등이 한계로 제기되고 있다.

이 연구에서는 시뮬레이션 기반 실험연구를 통해 데이터 단계의 보정을 통해 추천시스템의 편향을 줄일 수 있다는 의미있는 결과를 도출했다. 이와 더불어 전문가 조사를 통해 공정성 평가 및 보정을 위한 방법론적 시도의 필요성과 의미를 확인했다는 점에서 의의를 둔다.

연구방법론과 관련해서, 향후 관련 연구 및 제도적 논의를 위해 제언을 하면 아래와 같다.

가. 이용자 보호 평가 차원에서 추천시스템 공정성의 구성 개념화(construct) 필요

현재까지 발표된 추천시스템의 공정성 관련 연구들은 대부분이 개념정의를 명확하게 하지 않고, ‘편향’과 ‘차별’의 반대개념으로 ‘공정성’ 개념을 사용하고 있다. 해당 추천시스템의 개발 목적에 따라 편향을 해석하는 성격이 달라질 수 있다는 점을 고려하더라도, 개념적 차원에서 추천시스템의 합의된 측정 기준들이 필요하다.

이를 위해서는 아래와 같은 후속 연구가 필요하다.

- ▷ 추천시스템 공정 연구에 대한 메타분석적 연구와 접근을 통해 개념 차원의 하위 요소들을 분류하고 개념 차원 분류와 기존 연구의 효과 등을 종합적으로 분석할 필요가 있다. 이를 통해 정책적 차원에서 필요한 공정성의 평가적 개념을 구성하는 것이 요구된다.
- ▷ 이용자 보호 속성이 반영되는 직접차별, 중립적이고 보호되지 않는 속성에 의한 간접차별, 비체계적으로 이뤄지는 구조적 차별, 통계 예측차별, 외생변수를 통제하지 않아 발생하는 비체계적 차별 등 발생원인에 따른 다양한 불공정 문제들 가운데 이용자 보호 관점에서 사회적 영향이 크고 제도화 논의가 필요한 영역을 구분하는 작업이 필요하다.

나. 추천시스템 공정성 평가를 위한 다양한 알고리즘별 특성 분석실험 필요

이번 연구에서 사용한 User-KNN과 SVD와 같은 협업필터링(collaborative filtering) 방법 외에도 콘텐츠 기반 필터링(content-based filtering) 알고리즘 등 다양한 방식으로 알고리즘이 사용된다. 최근에는 딥러닝을 활용한 추천시스템 알고리즘이 늘어나고 있다. 연합학습(Federated Learning)이나 GAN(Generated Adversarial Networks) 등의 방식 등이 적용되면서, 편향의 원인을 판단하는 데 어려움이 커지고 있다.

이번 연구의 확장적 차원에서, 알고리즘의 유형별로 편향의 차이가 어떻게 다른 패턴으로 나타나는가를 연구해서, 적용되고 있는 알고리즘의 특성에 맞추어 편향탐지의 방향을 예측하는 근거를 마련하는 것이 필요하다.

- ▷ 추천시스템에 적용되는 다양한 알고리즘의 효과를 확장적으로 비교할 수 있는 확대된 시뮬레이션 연구 방법과 실제 데이터 기반의 검증이 필요하다.

다. 통합적 실험연구를 위한 실험 플랫폼 구축과 대규모 데이터셋의 구축

추천시스템에서의 공정성 연구에서 사용된 시스템 중심 접근과 이용자 중심 접근은 각각의 한계를 모두 갖고 있다. 시스템 중심 접근은 이용자 데이터셋과 추천시스템이 예측 결과값을 비교하는 것이 주된 방법인데, 이는 추천시스템을 보유한 기업 등이 자기 보유 데이터를 활용해 사용해온 방법이다. 제3자가 특정 시스템에 접근해서 데이터셋을 수집하는 것의 어려움이 존재하기 때문에 다양한 방법을 통한 대규모 데이터 수집기법이 최근 활용되고 있다.

대표적인 사례는 프로퍼블리카가 페이스북의 정치광고의 편향성을 검증하기 위해 2020년 현재 진행 중인, 크라우드 소싱 방식의 개인 데이터 수집방법이 있다. 크라우드소싱이 어렵다면, 대규모 패널을 확보해서 추천 되는 내용을 자동화적인 방식으로 데이터를 수집하는 것을 고려할 수 있다.

또다른 방식은 추천의 결과물이 나타내는 ‘다양성의 축소’나 ‘분극화’와 같은 영향 또는 결과 개념을 조작적으로 정의내려서, 자동화된 에이전트를 통해 데이터를 시계열적으로 그리고 대량으로 수집해서 검증하는 방식이 고려 가능하다.

이용자에게 미치는 영향을 탐지하기 위해서는 Epsetin 등(2015, 2017)과 황용석 등(2018)이 사용한 이용자 기반 대규모 온라인 실험연구를 수행하는 것도 중요하다. 이 연구는 알고리즘에 의한 불공정성인 추천이 가져오는 이용자의 지각 및 선택행동에 미치는 영향을 계량화할 수 있다는 점에서 중요하다.

- ▷ 대규모 데이터를 확보해서 시스템 기반적인 실험을 구현하기 위해서, 제3자 영역에서의 데이터 수집이 필요하며, 이를 위한 데이터 수집 플랫폼의 구축이 요구된다.
- ▷ 추천시스템의 편향이 가져오는 지각 및 행동적 영향을 탐지하기 위한 대규모 통제된 온라인 실험연구가 병행될 필요가 있다. 이 연구결과는 추천시스템 알고리즘의 직접 및 간접적 효과를 통계적으로 증명하는 데 도움이 된다.

2. 추천시스템의 공정성 보정을 위한 정책적 제언

추천시스템 알고리즘의 공정성 문제는 서비스 제공자의 책무(accountability)의 영역이다. 즉, 추천시스템 제공자가 시스템이 사회적 가치를 훼손하지 않는지를 모니터링 할 필요가 있으며, 이를 위한 윤리적 가이드라인 등을 마련하는 것이 필요하다.

현재 다양한 차원의 노력들이 전개 중이며, 그 가운데 향후 이용자 보호 정책으로서 필요한 노력을 제시하면 다음과 같다.

가. 추천시스템의 영향성 평가를 공공분야에 적용

추천시스템의 영향성평가제도가 캐나다 등에서 공공분야에 적용되고 있다. 이 제도는 추천시스템의 설계 단계부터 합의된 공정성 개념을 기반으로 추천시스템이 작동되도록 가이드하는 내용과 시스템의 영향성을 평가하는 차원을 모두 담고 있다. 추천시스템은 민간영역에서 기업 등의 경쟁력을 좌우하는 지적 재산으로서, 제도를 통해 공정성과 같은 사회적 가치를 강제하기는 어렵다. 그러나 공적 시스템에 한해서는 그 정당성이 확보됨으로써, 이를 우선적으로 적용할 필요가 있다.

영향성 평가방법을 고안하게 되면, 민간의 영향력 있는 추천시스템이 이용자 및 시장 보호를 침해하는 사건이 발생했을 때, 그 영향의 크기나 방향을 가늠할 수 있는 평가적 개념과 도구로 활용할 수 있다. 이에 이를 위한 방법론적 및 제도적 연구가 필요하다.

- ▷ 공공분야 추천시스템에 대한 영향성 평가를 개발 사전단계와 사후 단계로 구분하여 그 방법을 제시해야 하며, 이를 위한 평가적 방법론 정립이 필요하다.
- ▷ 민간 추천시스템에 의한 이용자 보호 침해 사례가 발생했을 때, 그 영향의 크기나 방향을 가늠할 수 있는 평가적 방법론으로 활용 가능하다.

나. 추천시스템 알고리즘의 투명성 강화를 위한 자율노력 지원

추천시스템의 개발단계에서부터, 운영 단계까지 개발주체가 자율적으로 투명성을 강화하고 편향을 최소화하기 위한 노력들이 진행 중이나 아직까지 제한적이다. 기업의 자율규제 활동의 하나로 추천시스템의 투명성을 강화 및 이용자 보호 노력을 지속하기 위한 자율규약 및 자율적 모니터링과 실험결과의 공표 등의 노력을 진작시키는 것이 필요하다.

이를 위해 자율규제를 잘 수행하는 기업 등에 인센티브를 제공하는 방안을 고민해야 하며, 자율노력의 확산을 위한 다양한 협의체 등의 구성이 필요하다.

- ▷ 추천시스템의 공정성을 확보하고 운영의 투명성을 높이기 위한 기업의 자율규제활동을 진작하기 위해 개발 및 운영단계별 투명성 가이드라인을 개발하는데 지원할 필요가 있다.
- ▷ 자율활동의 성과를 공유하고 확산시키기 위한, 협의체 구성 등 공동의 협력 플랫폼이 필요하다.
- ▷ 이용자보호제도의 하나로, 추천 알고리즘의 투명성을 위한 자율활동을 포함하고 이를 지원하는 제도적 노력이 고려될 수 있다.

다. 이용자의 경감심 고취와 자율능력을 기르기 위한 정책

추천시스템 알고리즘이 갖는 편향의 문제를 이용자들이 지각하고 이를 합리적으로 이용할 수 있도록 다양한 캠페인 활동과 교육사업 등이 확대될 필요가 있다. 디지털리터러시 교육에 알고리즘의 효과성과 관련된 내용을 포함하는 작업이 필요하다.

- ▷ 이용자를 대상으로 추천시스템의 작동원리와 영향을 이해시키기 위한 교육 프로그램을 강화하고, 기존의 디지털리터러시 교육에 이를 포함시키는 것이 필요하다.
- ▷ 이용자의 인식제고를 위한 다양한 캠페인 활동이 필요하다.

라. 추천시스템의 작동방식 등에 대한 이용자 대상 고지 노력

황용석 등(2018)은 알고리즘이 작동하는 설명방식에 따라서 이용자들이 알고리즘에 능동적으로 대처하는가를 실험한 바 있다. 즉, 추천 알고리즘의 각기 다른 설명방식(화이트박스과 블랙박스를 포괄하는)이 사용자들이 검색결과를 평가하는 데에 미치는 영향을 검증하였다. 이는 앱스타인 등(Epstein, 2015;2018) 등의 연구자에서 시도된 사용자가 편향된 검색결과를 인지하였을 때 정보의 평가에 미치는 영향을 설명방식과 연관지어 구체화한 실험이었다. 이 연구에서는 검색 알고리즘의 설명이 사용자에게 다르게 제공되었을 때 사용자들은 검색 알고리즘의 인지성(Awareness), 정확성(Correctness), 해석가능성(Interpretability)과 인지된 책무성(Accountability)을 평가하는 데에 차이가 있는지를 실험을 통해 검증하였다.

연구결과, 알고리즘의 설명방식이 다르게 제공되었을 때 사용자들이 인식하는 인지성, 정확성, 해석가능성, 인지된 책무성에 통계적으로 유의미한 차이가 있음을 발견하였다. 검색 방식에 아무런 설명이 제공되지 않은 참여자들은 알고리즘의 네 가지 영역 모두 가장 높은 점수를 준 반면 알고리즘의 존재, 블랙박스의 영역(Why), 화이트박스의 영역(How) 중 하나의 정보만을 제공받은 참여자들은 보다 낮은 점수를 주었다. 마지막으로, 알고리즘에 대해 복합적인 설명을 들은 사용자들은 검색 알고리즘에 대해 가장 비판적인 평가를 내렸음을 발견했다.

황용석 등(2018)년 연구 결과는 검색 알고리즘에 대한 설명방식이 복잡적이고 다양할수록 사용자들은 검색 결과의 정확성과 신뢰성에 대해 보다 명확히 판단할 수 있음을 보여주며 이는 알고리즘이 적용된 검색 시스템 이용자들의 권익을 위해 검색 알고리즘이 상세하고 복합적으로 제공되어야 한다고 제언하고 있다.

이 연구를 고려했을 때, 추천시스템이 이용자에게 그 작동원리 등 결과값이 미치는 요인을 고지하고 설명하는 것이 중요하다는 것을 알 수 있다. 따라서 이와 관련한 고지 노력에 관심을 기울일 필요가 있다.

- ▷ 추천시스템의 작동방식 등을 이용자에게 고지하는 등 이용자보호에 필요한 정책적 수단을 강구할 필요가 있다.

참 고 문 헌

국내 문헌

- 마정미 (2020). 유튜브 저널리즘과 공론장(public sphere)에 관한 연구, 『한국소통학보』, 19(1), 217 ~ 246.
- 황용석·김기태 (2019). 개인화 서비스 진전에 따른 자동추천시스템 연구 동향과 방법론적 특성 연구. 『사이버커뮤니케이션학보』, 36(2), 221 ~ 253.
- 황용석·김기태 (2020). 알고리즘 기반 자동화된 의사결정의 설명 가능성에 대한 연구. 『언론정보연구』, 57(3), 41 ~ 80.
- 황용석·이현주·정재관·김기태(2018). 《알고리즘 이용자 보호 이슈의 방법론적 접근》, 정보통신정책연구원 위탁연구보고서 18-20.

해외 문헌

- Ali et al(2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes, <<https://doi.org/10.1145/3359301>>
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). "Exposure to ideologically diverse news and opinion on Facebook." *Science*, 348(6239), 1130~1132.
- Burke, R., Sonboli, N., & Ordonez-Gauger, A. (2018, January). "Balanced neighborhoods for multi-sided fairness in recommendation." In *Conference on Fairness, Accountability and Transparency*, 202~214.
- Chen, L., Zhou, Y., & Chiu, D. M. (2015). "Analysis and detection of fake views in online video services. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2s), 1~20.
- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *University of Chicago Legal Forum*, Vol. 140, 139~67.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214-226.
- Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), E4512-E4521.
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM: Human-Computer Interaction*, 1, 42.
- Hallin, D. C., & Mancini, P. (2004). "Comparing media systems: Three models of media and politics." Cambridge, UK: Cambridge University Press.

- Kaminski, M. E., & Malgieri, G. (2019). Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. Available at SSRN 3456224.
- Kevin Roose (2019). "The Making of a YouTube Radical." NY Times
(<https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>)
- Knijnenburg, B. P., & Willemsen, M. C. (2015). "Evaluating recommender systems with user experiments." In *Recommender Systems Handbook*, 309~352. Springer, Boston, MA.
- Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). "Controlled experiments on the web: survey and practical guide." *Data mining and knowledge discovery*, 18(1), 140~181.
- Ledwich, M., & Zaitsev, A. (2019). "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization." *arXiv preprint arXiv:1912.11211*.
- Marciel, M., Cuevas, R., Banchs, A., González, R., Traverso, S., Ahmed, M., & Azcorra, A. (2016, April). "Understanding the detection of view fraud in video content portals." In *Proceedings of the 25th International Conference on World Wide Web*, 357~368.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
- Munger, K., & Phillips, J. (2019). "A supply and demand framework for YouTube politics." Retrieved from <https://osf.io/73jys>
- Nielsen (2019). "News media report: fresh news ecosystem." YouTube Journalism. Nielsen Korea.
- Ning, X., & Karypis, G. (2011, December). "Slim: Sparse linear methods for top-n recommender systems." In *2011 IEEE 11th International Conference on Data Mining*, 497~506. IEEE.
- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute, 1~22.
- Ribeiro, M.H. , Ottoni, R., West, R., Almeida, V.A.F., & Meira, W. (2020). "Auditing radicalization pathways on YouTube," *FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp 131~141.
- Roose, K. (2019). "The making of a YouTube radical." The New York Times. Retrieved from <https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>
- Ruggieri, S., Pedreschi, D., & Turini, F. (2010). "Data mining for discrimination discovery." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2), 1~40.
- Sacharidis, D. (2019, April). "Top-N group recommendations with fairness". In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 1663~1670.
- Serbos, D., Qi, S., Mamoulis, N., Pitoura, E., & Tsaparas, P. (2017, April). "Fairness in package-to-group recommendations." In *Proceedings of the 26th International Conference on World Wide Web*, 371~379.

- Steck, H. (2018, September). "Calibrated recommendations." In Proceedings of the 12th ACM conference on recommender systems, 154~162.
- Troy, D. (2019, 5, 21). "Netflix rules in customer satisfaction survey, followed by vue." Streaming Media. Retrieved from <https://www.streamingmedia.com/Articles/News/Online-Video-News/Netflix-Rules-in-Customer-Satisfaction-Survey-Followed-By-Vue-131889.aspx>
- Tsfati, Y., & Cappella, J. N. (2003). "Do people watch what they do not trust? Exploring the association between news media skepticism and exposure." *Communication Research*, 30(5), 504~529.
- Tsintzou, V., Pitoura, E., & Tsaparas, P. (2018). "Bias disparity in recommendation systems." arXiv preprint arXiv:1811.01461.
- Xiang, A., & Raji, I. D. (2019). "On the Legal Compatibility of Fairness Definitions." arXiv preprint arXiv:1912.00761.
- Xiao, L., Min, Z., Yongfeng, Z., Zhaoquan, G., Yiqun, L., & Shaoping, M. (2017, August). "Fairness-aware group recommendation with pareto-efficiency." In Proceedings of the Eleventh ACM Conference on Recommender Systems, 107~115.
- Yao, S., & Huang, B. (2017). "Beyond parity: Fairness objectives for collaborative filtering." In Advances in Neural Information Processing Systems, 2921~2930.
- Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., ... & Chi, E. (2019, September). "Recommending what video to watch next: a multitask ranking system." In Proceedings of the 13th ACM Conference on Recommender Systems, 43~51.

[부록]

추천시스템의 편향보정 및 공정성 보장방안 연구	ID					
바쁘신 가운데 시간을 전문가 조사에 시간을 내어 주셔서 감사드립니다. 건국대 디지털커뮤니케이션연구센터 와 KAIST 전산학부 공동 연구진은 정보통신정책연구원의 협동과제인 "추천시스템의 알고리즘 편향 보정에 관한 연구: 방법론적 접근"을 수행하고 있습니다. 이 연구에서는 추천시스템의 공정성 개념과 기업의 관련 책무활동 등을 탐구 하고 있으면, 가상의 기계학습 기반 추천 알고리즘에 대한 시뮬레이션 실험을 통해 공정성 보정 또는 편향감소에 필요한 데이터 처리방안 등을 연구하고 있습니다. 추천시스템의 공정성 또는 편향감소를 위한 방법론적 연구과제로서, 관련 주제에 대한 전문가 분들의 의견을 얻고자 합니다. 해당 주제가 최신성이 있고 주제의 복잡성이 있어 최대한 포괄적인 질문형태로 문항을 구성했습니다. 답변이 가능하신 문항에 대해 의견을 주시면 연구에 큰 도움이 되겠습니다. 황용석 교수(건국대 미디어커뮤니케이션학과), 문수복 교수(KAIST 전산학부) 외 참여 연구진 일동 ※ 조사관련 문의 : konkukdcrc@gmail.com						

성 함		소 속	
-----	--	-----	--

Q1. 선생님께서는 자동화된 추천시스템 알고리즘 중 앞으로 영향을 미치거나 주목할 만한 알고리즘은 무엇입니까?

Q2. 추천시스템 등 '알고리즘 공정성'에 대한 다양한 정의가 있지만, 아직 합의된 정의는 없습니다. 선생님께서는 '알고리즘 공정성'에 대한 보편적 정의내리기가 가능하다고 보십니까? 그렇다면 그 핵심 개념은 무엇이라고 생각하십니까? 알고리즘 공정성에 대한 개념이 보편적으로 적용하기 어렵다면, 그 이유는 무엇이며 대체 가능한 개념은 무엇이라고 생각하십니까?

Q3. 알고리즘 공정성 분석을 위한 방법론 중 전통적인 1) 이용자 기반의 실험, 2) 대규모 온라인 패널 기반 실험, 3)시스템 기반의 자료를 바탕으로 한 연구, 4) 시뮬레이션 기반 실험 등 알고리즘 공정성에 대한 다양한 분석방법이 적용되고 있습니다. 선생님께서 생각하시기에, 적합하거나 추천하고 싶은 연구방법은 무엇입니까?

Q4. 현재 사용 또는 적용되고 있는 자동화된 추천시스템 알고리즘 중 앞으로 영향을 미치거나 주목할 만한 알고리즘은 무엇입니까?

Q5. 알고리즘의 사회제도가 시장에 적용되면서, 자동화된 의사결정이 늘어나고 있습니다. 추천시스템의 알고리즘 공정성에 대한 '영향성 평가' 논의가 있습니다. 현재 캐나다에서는 공공부분에 적용되는 알고리즘을 대상으로 이를 적용하고 있습니다. 선생님께서는 알고리즘 영향성 평가가 필요하다고 생각하십니까? 그렇다면 그 평가 주체는 누가 되어야하며, 누구(공공 또는 민간, 영향력 있는 기업 또는 전체 자동화 추천시스템 등)를 대상으로 해야 할지 의견을 주십시오. 필요하지 않다고 생각하시면 그 이유를 적어주십시오.

Q6. 최근 알고리즘 공정성 분석과 관련해 주목할만한 연구결과가 있으시다면 자료나 내용을 공유해 주시면 감사하겠습니다.

Q7. 연구진에게 추가적으로 전하고 싶은 의견이나 정보가 있으시면 적어주십시오.

다시 한 번 고견에 감사드립니다.

2020년 11월

연구진 일동

*연락처 : konkukdcrc@gmail.com

1. 본 연구보고서는 정보통신정책연구원의 지원으로 수행한 건국대학교 산학협력단의 연구결과물입니다.
2. 본 연구보고서의 저작권은 정보통신정책연구원에 있으며, 본 연구내용을 발표할 때에는 반드시 정보통신정책연구원의 지원으로 건국대학교 산학협력단에서 수행한 연구결과임을 밝혀야 합니다.