15기 추천시스템 세미나

ToBig's 15기 강의자

류채은

추천시스템과 윤리적 도전

Recommender Systems and their Ethical Challenges

Contents

Unit 01 | Introduction & Abstract

- 6가지 윤리적 도전(Ethical Challenges) 제시
- 추천 시스템 사용자 이외의 다른 이해당사자(stakeholders) 고려
- 추천 시스템은 개개인의 디지털 환경과 사회적 상호작용에 관여하기에 윤리 고려가 중요

- 추천 시스템의 윤리적 논의 현재 배경:
  - 유아기(infancy)
  - 분열(fragmentation)

- 윤리적 도전을 추천시스템의 설계(design), 배치(deployment), 사용(use), 트레이드 오프(trade-off)를 기준으로 다뤄야 함

Unit 02 | A working Definition of Recommender Systems

- Level of abstraction(LoA): 추상성 수준; 개인에게 무엇이 보이고 무엇이 모호하거나 안 보이는지

- LoA에 의존한 시험적 정의(Working Definition)의 3가지 한도(parameters)
  - 1. 옵션의 범위(Space of options)
  - 2. 좋은 추천(good recommendation)
  - 3. 추천시스템의 평가 척도

- 의사 결정 보조(Decision support)
- 다중 이해당사자 환경(multi-stakeholder environment): 사용자, 공급자, 시스템 관리자
- 고려 대상: content-based, collaborative filtering, and combination of both

Unit 03 | How to Map the Ethical Challenges Posed by RS

- 고려대상: RS의 작동(behavior)과 영향(impact)
- 윤리적 이슈의 2가지 차원:
  - 1. 작동의 영향(Impact of Operation)
    - 1) 이해당사자의 실용성(utility)에 미치는 악영향
    - 2) 권리 침해
  - 2. 즉각적 손해(harm)과 관련된 자들의 위험에 대한 노출 및 권리 침해

	Immediate Harm	**Exposure to Risk**
Utility	e.g. inaccurate recommendations	e.g. A/B testing (see section 4.1)
Rights	e.g. unfair treatment	e.g. leaking of sensitive information

Unit 04 | 추천시스템의 윤리적 도전

추천시스템의 윤리적 도전(Ethical Challenges)

- 1. 윤리적 내용 Ethical Content
- 2. 정보 보호 Privacy
- 3. 자율성과 정체성 Autonomy and Personal Identity
- 4. 불투명함 Opacity
- 5. 공정성 Fairness
- 6. 양극화와 사회적 조작가능성 Polarization and manipulability

Unit 4.1 | 윤리적 내용 Ethical Content

- 추천 받는 내용(content)이 윤리적이도록 문화적·윤리적 선호도 기반으로 필터링하는 방법

- **사용자 중심 접근 방식(User-centered approach)**
  : 사용자의 유용성(utility)와 권리에 미치는 악영향 최소화

- 방법: 사용자가 추천시스템이 개인 정보를 다루는 방법을 직접 조절하도록 해 마케팅 편향(marketing bias)과 내용 검열(content censorship)을 필터링하고 온라인 실험을 면하도록 함.

- Short-comings:
  - 필터의 존재를 통한 사용자의 민감한 정보 폭로 가능성
  - 사용자에게 과도한 책임 전가
  - 문제가 되는 활용 분야

Unit 4.2 | 정보 보호 Privacy

- 정보 침해 위험(Privacy Risks)의 4가지 유형:
  - 1) 사용자 동의 없이 정보가 수집되고 공유될 때
  - 2) 수집된 정보가 외부인에게 노출될 때
  - 3) 데이터를 통해 사용자에 대한 추론할 때
  - 4) 다른 사용자들의 상호작용을 통해 사용자에 대한 모델을 구성할 때
- 해결 방안
  - 1) 설계적 방안 architectures approach
  - 2) 알고리즘 측면 방안algorithmic approach
  - 3) 정책적 방안 policy approach (ex. GDPR 법안)

- 사용자 중심 접근 방법의 위험성

Unit 4.3 | 자율성과 정체성 Autonomy and Personal Identity

- 자율성 침해: 사용자에게 특정한 방향으로 추천을 하거나 선택의 폭을 제한함으로써 특정한 종류의 내용에 중독되게 할 수 있음
  - 범위: 순수한(benign) 추천/설득/강요
  - Captology(인간들이 특정 행동을 하도록 유도하는 기술)

- 정체성 침해의 두 가지 이유:
  - 1) 사용자에 대한 모델의 연속적 구조 변경(reconfiguration)
  - 2) 추천시스템의 사용자 분류(categorization)와 사용자 스스로 본인에 대해 내린 분류의 불일치

Unit 4.4 | 불투명함 Opacity

- 사용자에게 추천시스템이 어떻게 '생각'하는지에 대한 **설명**을 제공하는 것은 자율성 침해를 최소화하는데 도움을 줌.

- '좋은 설명' = 사실적 설명(Factual explanations)
  : 어떤 결과가 일어나기 위한 이전의 필수적 조건/상황이 무엇이었는지에 대한 설명

- 사실적 설명을 제공하기 매우 어려움(불투명함)
  - 이유1) 사용자의 다른 사용자의 정체에 대한 접근 불가
  - 이유2) 모델 연산과정의 복잡함
  - 이유3) winner-takes-all senarios

Unit 4.5 | 공정성 Fairnes

- 사회적 편향(Social biases) 초래 가능성
  - 1) Observation bias: 피드백 수집의 연속적 수행을 통해 생긴 특정 그룹에 대한 편향
  - 2) Population imbalance: 이미 존재하는 사회적 편향/편견을 수집된 데이터가 반영

Unit 4.6 | 양극화와 사회적 조작가능성 Polarization and social manipulability

- Filter Bubble: 다양한 관점으로부터 사용자 격리
  - 사회적 논의 감소
  - 공정 숙의(public deliberation) 침해
  - 민주주의적 기관 활동 약화

- 사회적 조작 가능성
  - 정치적 선전(propaganda)

- 사용자에 대한 관련성(relevance)과 다양성(diversity) 사이의 Trade-off
  - 뜻밖의 재미(Serendipity)의 제안

Unit 05 | 결론

- Ethical challenges 분류 결과

	Immediate Harm	Exposure to Risk
Utility	Biased recommendations (4.1)	Opacity (4.4) Questionable content (4.1)
Rights	Unfair recommendations (4.5) Encroachment on individual autonomy and identity (4.3)	Privacy (4.2) Social manipulability and Polarisation (4.6)

- 추천 시스템의 사용자(receiver)외 제공자(provider), 사회 등 제 3자에 대한 논의가 중요

들어주셔서 감사합니다.