웹 로그인 아이디와 사용자 성별 및 연령대 관계 분석과 예측 모형

2015. 9. 14.

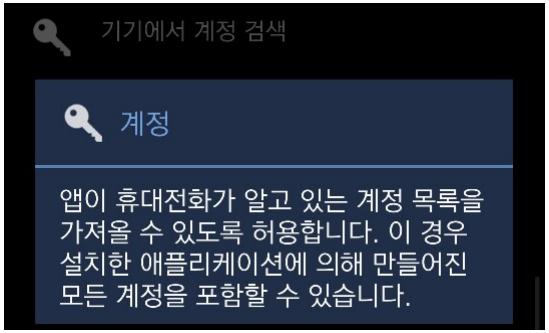
박희웅

서울대학교 산업공학과

연구 배경

- **웹 사용자의 인구통계적 정보를 추론하는 문제는 꾸준한 관심의 대상**
  - 검색 쿼리, 사이트 방문 로그, 소셜 네트워크 분석, 작성한 텍스트 분석 등 이용
  - Jaech and Ostendorf (2015)에서 사용자이름으로 성별 예측. 연령대는 안함

- **로그인 아이디를 이용해 사용자의 성별 및 연령대를 예측의 장점**
  - 날이 갈수록 개인정보 유출 문제가 심각해짐에 따라 사이트 가입 시 제한된 개인정보만을 입력 받는 추세
  - 사용자의 이용 내역이 존재하지 않는 콜드 스타트(cold start) 문제 해결
  - 모바일 어플리케이션의 경우, 기기 내 계정 수집 권한을 받으면 로그인 아이디 수집 가능



연구 방향

- **데이터 탐색과 분석을 통해 기존 사회과학적 연구와 비교**
  - 사람이 아이디를 보고 남녀를 판별할 수 있는지, 구분 지을 수 있는 요소를 추출해낼 수 있는지 실험 (Cornetto and Nowak, 2006)
  - 아이디를 보고 어떤 인상을 형성되는지 실험 (Pelletier, 2014)

- **성별 및 연령대 예측 모형 개발**
  - 우리나라 사용자에게도 Jaech and Ostendorf (2015)의 성별 예측 방법이 적절한지, 혹은 더 발전시킬 수 있는지 조사
  - 연령대 예측 모형 개발

- **주요 피쳐 추출 혹은 예측 모형 경량화**
  - 아이디 만으로는 예측 모형의 한계가 예상되며, 다른 소스의 정보들과 함께 예측 모형을 만들 수 있도록 핵심 피쳐 벡터 추출
  - 모바일 기기에 탑재될 수 있도록 모형에 필요한 저장 공간과 계산 시간 최소화

데이터

- **성별 + 연령대**
  - 네이버

	10대	20대	30대	
남자	10921	11066		3051
여자	9425	5478		

- **성별**
  - Jaech and Ostendorf (2015)

남자	22101
여자	27043

- **semi-supervised learning 용도**
  - Lui and Baldwin (2012) – 150만 개
  - McCormick (2014) – 976141개

연구 일정 및 참고 문헌

연구 일정

- 다음 시간 - Jaech and Ostendorf (2015) 리뷰
- 9월 말 – 데이터 전처리 및 탐색
- 10월 초 – 데이터 분석 및 가설 검증
- 10월 말 – 베이스라인 예측 모형 구현
- 중간 발표
- 11월 초 – 예측 모형 성능 개선
- 11월 말 - 주요 피쳐 추출 혹은 모형 경량화
- 12월 초 – 최종 결과 발표 및 논문 준비
- 최종발표

참고 문헌

Jaech, A., & Ostendorf, M. (2015). What Your Username Says About You. *rXiv:1507.02045 [cs]*. Retrieved from http://arxiv.org/abs/1507.02045.

Cornetto, K. M., & Nowak, K. L. (2006). Utilizing usernames for sex categorization in computer-mediated communication: Examining perceptions and accuracy. *CyberPsychology & Behavior*, *9*(4), 377-387.

Pelletier, L. (2014). You've Got Mail: Identity Perceptions based on Email Usernames. *Journal of Undergraduate Research at Minnesota State University, Mankato*, *9*(1), 13.

Lui, M., & Baldwin, T. (2012, July). langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25-30). Association for Computational Linguistics.

Rich McCormick. 2014. 4.6 million snapchat phone numbers and usernames leaked, January.