웹 로그인 아이디와 사용자 성별 및 연령대 관계 분석과 예측 모형
- 문제 정의와 기존 방법 리뷰

2015. 9. 16.

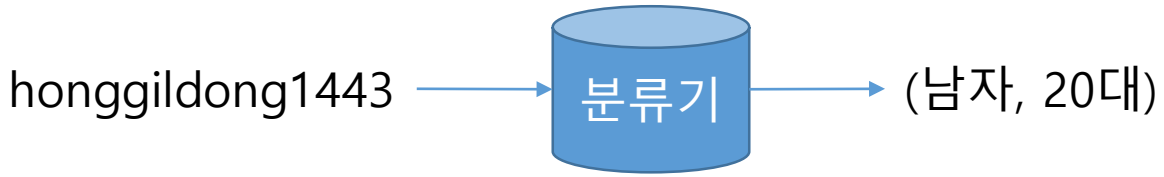
박희웅

서울대학교 산업공학과

문제 정의

- **웹 로그인 아이디**
  - 웹사이트에서 사용자가 서비스를 이용하기 위해 회원가입할 때, 사용자가 지정하는 고유 식별자. 영어로는 username 혹은 login name
  - 웹사이트마다 다르지만 대체적으로 6~20자 길이의 문자열로 이루어지며 알파벳, 숫자, 특수문자의 혼용이 허가되지만 띄어쓰기나 한글은 불가능. 대소문자를 구분하는 사이트도 있음
  - 예) honggildong1443

- **분류 문제**
  - 아이디가 주어지면, 해당 아이디의 사용자의 성별과 연령대를 맞추는 문제
  - 분류기를 학습하기 위해 label된 아이디 데이터 셋이 주어지며, semi-supervised 학습 용도로 많은 양의 label 없는 데이터를 활용할 수 있다고 가정



관련 기존 연구

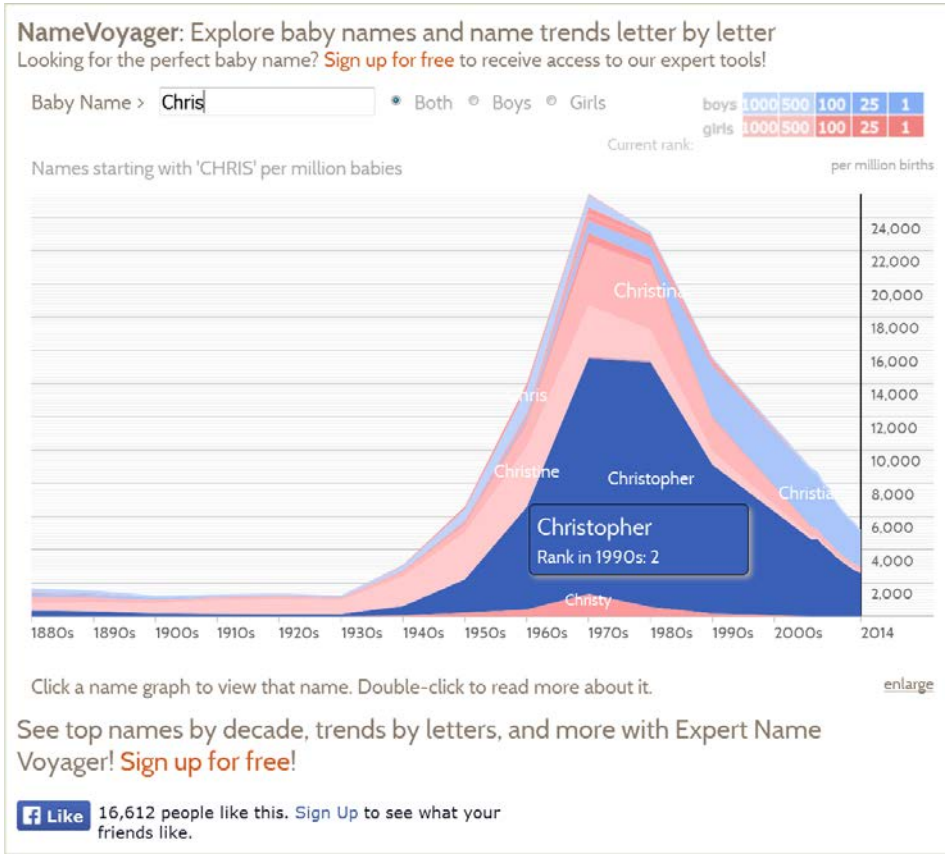
- **이름으로부터 성별 예측 문제와 관련성**
  - 우리나라 웹 사용자 중 상당수가 아이디에 자신의 이름 혹은 이니셜을 일부로 활용. 따라서 이름으로부터 성별을 예측할 수 있다면, 아이디에서도 성별을 예측할 가능성이 존재
  - 아이디에 이름이 들어간 예) gildong1443, honggd1443, hgd1443, ghdrlfehd1443

- **관련 논문**

제목	저자	게재지	연도	정확율
Discriminating gender on twitter	Burger et al.	In Proc. EMNLP	2011	89%
What's in a Name? Using First Names as Features for Gender Inference in Twitter	Liu and Ruths	In Proc. AAAI	2013	85%
미등록 이름 명사 인식 및 성별 구분	강유환 et al.	한국정보과학회	2004	94%
Determining the Gender of Korean Names for Pronoun Generation	Park, S. B. et al.	WASET	2007	81%
나이브 베이지안을 사용한 성명에 대한 성별 구분 연구	임명재 et al.	한국인터넷방송통신학회	2013	86%

관련 기존 연구

- 이름과 성별 데이터베이스



www.babynamewizard.com

순위	남자			여자		
	이름	인원수	백분율	이름	인원수	백분율
1	민준	12,145 명	0.5695 %	유진	29,092 명	0.9902 %
2	지훈	12,106 명	0.5677 %	지영	23,119 명	0.7869 %
3	현우	9,422 명	0.4418 %	지은	21,531 명	0.7328 %
4	주원	8,198 명	0.3844 %	수진	20,869 명	0.7103 %
5	시우	7,731 명	0.3625 %	지혜	20,332 명	0.692 %
6	현준	7,693 명	0.3607 %	지원	19,577 명	0.6663 %
7	우진	7,515 명	0.3524 %	민지	19,411 명	0.6607 %
8	지후	7,471 명	0.3503 %	민정	19,293 명	0.6567 %
9	승현	6,810 명	0.3193 %	혜진	18,905 명	0.6435 %
10	서준	6,804 명	0.3191 %	수연	18,782 명	0.6393 %
11	도현	6,661 명	0.3124 %	지연	18,681 명	0.6358 %
12	진우	6,600 명	0.3095 %	지현	18,649 명	0.6348 %
13	동현	6,532 명	0.3063 %	서연	18,506 명	0.6299 %
14	준영	6,472 명	0.3035 %	서현	16,680 명	0.5677 %
15	준호	6,447 명	0.3023 %	현정	15,644 명	0.5325 %
16	지원	6,401 명	0.3002 %	지윤	14,782 명	0.5031 %
17	지호	6,332 명	0.2969 %	은정	14,534 명	0.4947 %
18	건우	6,288 명	0.2949 %	은지	14,450 명	0.4918 %
19	성민	6,047 명	0.2836 %	소영	14,397 명	0.49 %
20	준혁	5,870 명	0.2753 %	수정	14,001 명	0.4766 %

www.erumy.com

관련 기존 연구

아이디로부터 성별 예측

제목	저자	게재지	연도	정확율
Discriminating gender on twitter	Burger et al.	In Proc. EMNLP	2011	77%
What Your Username Says About You	Jaech and Ostendorf		2015	74%

What Your Username Says About You

- Aaron Jaech and Mari Ostendorf, Dept. of Electrical Engineering, University of Wachinton
- 아이디로부터 성별을 맞추는 것과, 아이디로부터 모국어가 무엇인지 맞추는 두 가지 문제를 다룸

What Your Username Says About You

- **주요 개념**
  - morpheme: minimal linguistic unit with lexical or grammatical meaning. 우리말로는 형태소
  - u-morphs: 다른 사용자의 아이디와 공유할 수 있을 만큼 작은 단위의 문자열이지만 의미를 보존하고 있음. morpheme과 같은 개념이지만 일반적인 단어에서가 아닌 아이디 문자열으로부터 인코딩된 형태소

- **Morfessor**
  - Creutz and Lagus (2006)
  - Unsupervised morphology induction
  - Maximizing the likelihood of the data and the likelihood of the model라는 두 가지 상반된 목표를 최적화시키는 minimum description length (MDL) 목적함수를 갖고 있음
  - 아이디를 u-morphs로 변환하는 주 알고리즘으로 쓰임
  - 속도를 빠르게 하기 위해 알고리즘 과정에서 소문자와 대문자 사이에 구분자 토큰을 삽입. 예) HongGildong -> Hong$Gildong

What Your Username Says About You

- **단순 베이즈 분류기**
  - 아이디 u 는 형태소 $m_1, m_2, \ldots, m_n$으로 분리된 다음 확률이 가장 높은 클래스 c_i로 분류

$$\arg\max_i p_C(c_i) \prod_{k=1,\ldots,n} p(m_k|c_i)$$

Where $p_C(c_i)$ is the class prior and $p(m_k|c_i)$ is class-dependent

  - 스무딩 기법 이용하여 $p(m_k|c_i)$ 추정 (Frank and Bouchaert, 2006)

$$p(m_k|c_i) = \frac{1}{Z}\left(1 + \beta \cdot \frac{n(m_k, c_i)}{n(c_i)}\right)$$

Where n(*) indicates counts and β controls the strength of smoothing

What Your Username Says About You

- **결과**
  - U-morph를 통해 찾아진 구별력 있는 형태소들

	Male	Female
u-morph	guy, mike, matt, josh	girl, marie, lady, miss
trigram	guy, uy#, kev, joe	irl, gir, grl, emm

  - U-morph를 이용하면 실험 데이터에 대해 필요한 메모리 사이즈가 trigram과 비슷하고 4-gram에 비해 1/5-1/10 가량만 차지
  - 오분류율

Features	Error Rate	
	Supervised	Self-Training
3-gram	28.7%	32.0%
4-gram	28.7%	29.4%
u-morph	27.8%	25.8%