웹사이트 아이디 구조적 분석을 통한 사용자 추론
Inferring User Information from Login Name based on Structural Analysis

2015. 12. 7.

박희웅

서울대학교 산업공학과

서론

- 웹 사용자의 인구통계적 정보나 관심사를 추론하는 문제는 꾸준한 관심의 대상이 되어 왔다.
  - 검색 쿼리, 사이트 방문 로그, 소셜 네트워크 분석, 작성한 텍스트 분석 등 이용

- 웹 아이디는 웹사이트에서 사용자가 서비스를 이용하기 위해 회원 가입할 때, 사용자가 지정하는 고유 식별자로 영어로는 username 혹은 login name 이다.
  - 웹사이트마다 다르지만 대체적으로 6~20자 길이의 문자열로 이루어지며 알파벳, 숫자, 일부의 특수문자에 대해 혼용이 허가되지만 띄어쓰기나 한글은 불가능. 대소문자를 구분하는 사이트도 있다.
  - 예) honggildong1443
  - 웹 아이디는 사용자별로 고유해야 하고 스스로 기억하고 있어야 하므로, 사용자 신상정보나 관심사에 대한 단서를 포함하고 있을 가능성이 높다.

서론

- 웹 아이디는 웹사이트 가입과 동시에 얻을 수 있는 데이터로 사용자의 이용 내역이 존재하지 않는 콜드 스타트(cold start) 문제에 큰 도움이 된다.
  - 날이 갈수록 개인정보 유출 문제가 심각해짐에 따라 사이트 가입 시 제한된 개인정보만을 입력 받는 추세
  - 근래에는 다수의 웹사이트가 구글이나 페이스북 계정으로도 로그인을 가능하게 하여, 로그인 계정정보 외의 초기 개인정보 획득 기회가 줄어들고 있다.
  - 모바일 어플리케이션의 경우, 기기 내 계정 수집 권한을 받으면 아이디 수집 가능하다.

서론

- 기존 관련 연구로는 Burger et al. (2011)에서 트위터의 screen name을 보고 사용자의 성별을 예측했었다.
  - 트위터의 screen name은 로그인 아이디와는 별개로 화면에 나타나는 이름으로 아이디보다는 닉네임과 가깝다.
  - 닉네임은 사용자가 웹 서비스 이용시 반드시 기억하고 있어야 하지 않아도 되기에 웹 사이트의 분위기나 내용에 따라 작명하는 방식이 상이할 수 있다.
  - 3, 4-gram 단어의 남녀별 빈도수 차이를 이용해 성별을 구분했고, 77% 성별 정확도를 나타냈다.

- 최근에는 Jaech and Ostendorf (2015)이 유명 데이팅 사이트의 username을 이용해 사용자의 성별과 모국어를 추론했다.
  - 데이팅 사이트의 username은 로그인 시에도 이용되지만, 가입하면 자동으로 생성되는 자신의 미니 홈페이지 프로필에 큼지막하게 게재된다.
  - 데이팅 사이트인만큼 이성에 어필하기 위해 username에 성적 특징을 표현하는 단어들을 일반 웹아이디보다 많이 포함한다.
  - 단어로부터 형태소를 분해하듯이 의미 단위로 아이디를 토큰화하고 빈도수를 집계해 성별 확률을 도출했다.

서론

- 이번 연구에서는 기존 연구가 아이디를 토큰화하고 집단별로 토큰 수를 집계하여 집단을 구분하는 것에서 나아가, 품사 태깅과 같이 토큰화된 단위 요소 의미를 자동화된 알고리즘으로 알아내고자 한다.
  - 예를 들어, honggildong1443은 hong + ggildong + 1443 으로 분해될 수 있으며 각각은 사용자의 성씨, 이름, 출생년도를 의미한다.
  - 이러한 구조 분석을 통해 아이디로부터 성별뿐만 아니라, 이름이나 생일 등의 다양한 정보를 얻어낼 수 있을 것으로 기대된다.

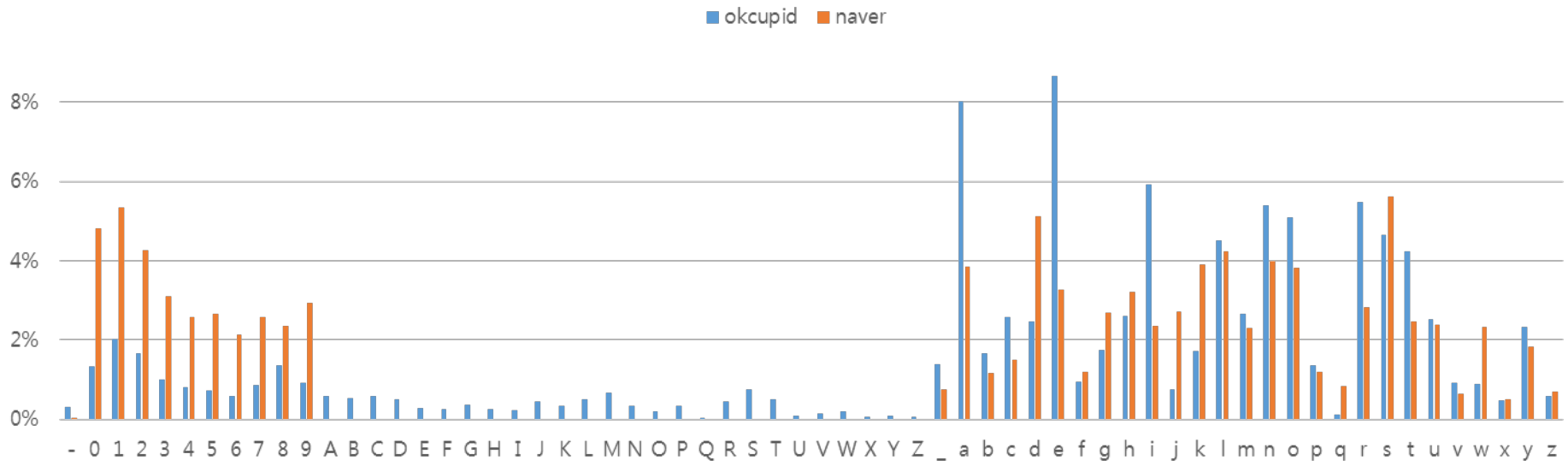
- 이 연구의 목적은 다음 세가지다.
  - 우리나라 사용자들의 웹 아이디를 수집하고 통계적인 분석을 통해 어떤 특징을 갖고 있는지 살펴본다.
  - 구조적 분석 모형을 이용해 아이디의 토큰화된 단위 요소 각각에 의미를 레이블링할 수 있는 방법을 제안한다.
  - 제안된 방법의 타당성을 검증하고 활용 가능성을 타진하기 위해, 아이디 구조 분석 결과를 이용해 사용자의 성별과 연령을 추론했을 때 분류 성능을 확인한다.

우리나라 사용자들의 웹 아이디 특성

- 웹 아이디를 분석하여 사용자 정보를 추론하기에 앞서, 우리나라 사용자들이 쓰는 웹 아이디의 특성을 살펴본다.
  - 국내 유명 포털 사이트 아이디 3만여 개를 수집했다.
  - 국내 사용자들의 아이디와 해외 사용자들 아이디의 특성을 비교하기 위해, Jaech and Ostendorf (2015)에서 이용한 okcupid.com 아이디 3만여 개도 분석했다.

우리나라 사용자들의 웹 아이디 특성

- 먼저 아이디를 구성하는 글자의 분포를 알아본다.
  - 아이디를 수집한 국내 포털 사이트는 대문자를 허용하지 않았으나 해외 사이트는 허용했다.
  - 우리나라 사람들이 해외 사용자보다 숫자를 더 많이 쓴다.
  - 해외 사용자들이 영어 모음인 a, e, i, o를 우리나라 사용자들보다 빈번히 쓴다.
  - 숫자 중에서는 0, 1, 2를 특히 많이 사용한다.



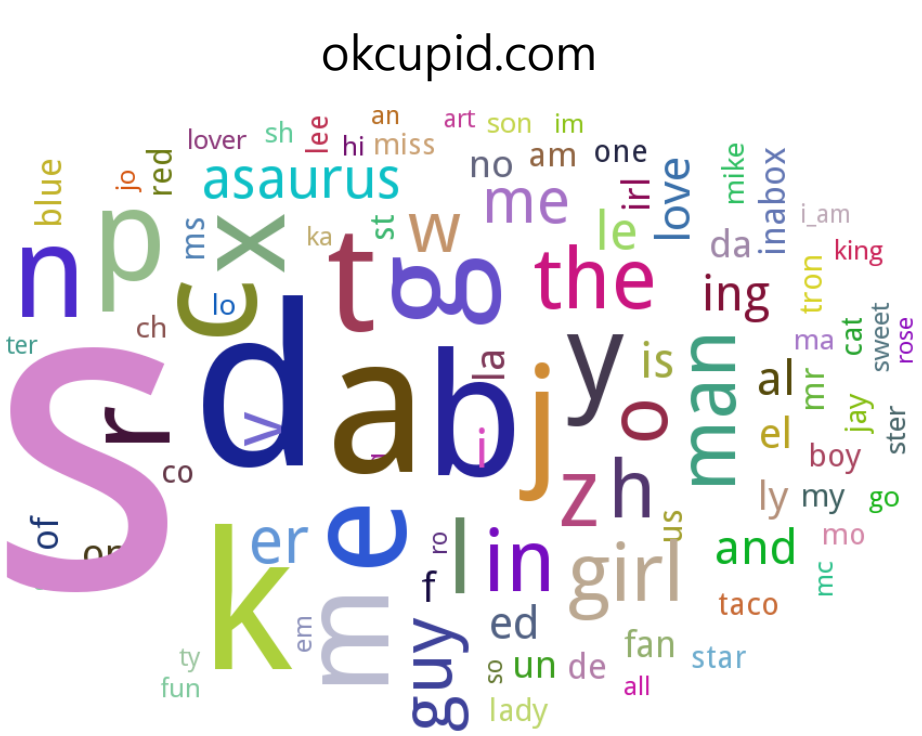
우리나라 사용자들의 웹 아이디 특성

- 띄어쓰기를 허용하지 않는 특성 때문에, 웹 아이디는 보통 고유한 의미를 갖는 더 작은 의미의 단위들로 쪼개질 수 있다.
  - Jaech and Ostendorf (2015)은 단어에서 형태소 단위로 나누는 것처럼 웹 아이디를 작은 의미 단위로 쪼개어 의미 단위 형태소의 빈도수를 분석했었다.
  - 이때 사용된 형태소 분해기인 Morfessor (2006) 알고리즘은 Unsupervised morphology induction으로, Maximizing the likelihood of the data and the likelihood of the model라는 두 가지 상반된 목표를 minimum description length (MDL) 목적함수로 최적화시킨다.

- 수집한 국내 포털 사이트 아이디와 해외 사용자 아이디 각각에 대해 Morfessor 알고리즘을 이용해 형태소를 분리하여 빈도수를 집계했다.
  - 알고리즘 수행 이전에 숫자 부분과 문자 부분을 분리하고, '-' 와 '_' 는 구분자로 취급하여 아이디에서 삭제하고 구분자를 기준으로 아이디를 분리했다.

우리나라 사용자들의 웹 아이디 특성

- 단어 구름을 이용해 분해된 형태소를 빈도수가 많을수록 크도록 나타냈다.
  - Okcupid 사용자들은 사전에 등장하는 평상시에도 자주 쓰는 단어들이 빈도수 상위 목록에 주로 포함되었다.
  - 반면 국내 사용자들은 자신의 이름을 아이디에 많이 썼는데, 소리 나는대로 영문으로 표기하기도 하지만 자판에서 한글 위치에 있는 알파벳을 그대로 쓰기도 한다.



우리나라 사용자들의 웹 아이디 특성

- 특정 연도나 날짜를 아이디에 많이 씀.

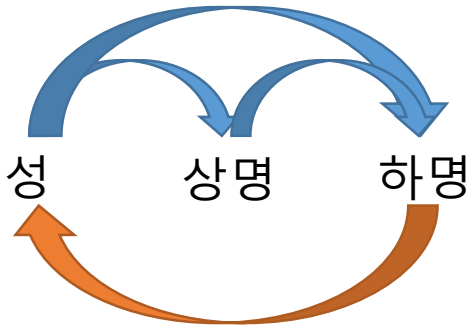
웹 아이디 구조 분석 모형

- 아이디를 형태소 단위로 분해한 뒤 각 요소에 의미적인 태그를 달아 분류 및 분석을 용이하게 할 수 있지 않을 것이다.
  - 우리나라 사용자들은 아이디에 자신의 이름이나 특정 날짜를 많이 쓴다는 것을 확인했으므로 이름과 날짜를 구조 분석을 통해 알고자하는 요소로 삼았다.
  - 이름자는 성, 이름 그리고 이름은 다시 상명자와 하명자로 구성된다.
  - 또한 한글 자판을 그대로 친 것인지, 영어 발음으로 쓴 것인지로 구별 지어진다.
  - 날짜는 연도, 월, 일로 나눠진다.

- 품사 태깅에 많이 쓰였던 은닉 마코프 모형(HMM)을 이용하여 아이디 의미 요소를 태깅한다.
  - 은닉 마코프 모형의 초기 상태별 확률과 전이 행렬, 그리고 출력 확률을 적절히 제어해 줌으로써 각각의 의미 요소를 모형의 은닉 상태로 대응시킨다.
  - 웹 아이디를 문자 부분와 숫자 부분으로 분리한 뒤, 각각에 형태소 분해한 결과에 대해 문자 부분 HMM과 숫자 부분 HMM으로 적합한다.
  - 아이디는 평균 3개의 형태소로 구성되므로, 은닉 마코프 모형에서 상태 전이 시에 직전 상태 정보만 이용하더라도 충분하다.

웹 아이디 구조 분석 모형

- 문자 부분 HMM



parkheewoong
parkwoong
heewoongpark

transition matrix 초기값

	한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
한글 성	0	1	1	0	0	0	1	1
한글 상명	0	0	1	0	0	0	1	1
한글 하명	1	0	0	0	0	0	1	1
영문 성	0	0	0	0	1	1	1	1
영문 상명	0	0	0	0	0	1	1	1
영문 하명	0	0	0	1	0	0	1	1
그 외	1	1	1	1	1	1	1	1
끝	0	0	0	0	0	0	0	1

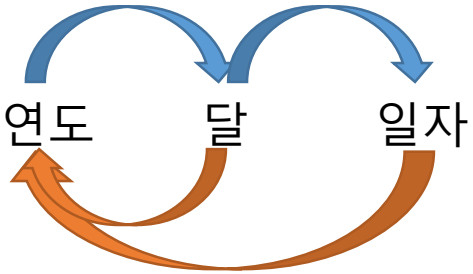
Start probability 초기값

한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
1	1	1	1	1	1	3	0

- Emission probability
  - ✓ 이름자의 경우, 통계 정보 활용하여 확률 값 고정했다. 성씨는 2000년 통계청 성씨 목록을, 상명자와 하명자는 erumy.com의 이름통계를 바탕으로 확률값을 추정했다.
  - ✓ 영문 이름은 네이버 한글 이름 로마자 표기 언어변환기를 이용했다.
  - ✓ 그 외 state emission probability는 초기값으로 모든 형태소에 동일한 확률 부여하고, 이후에 업데이트한다.

웹 아이디 구조 분석 모형

- 숫자 부분 HMM



20151113, 151113
112015
11132015

transition matrix 초기값

	연도	달	일자	그 외	끝
연도	0	1	0	0	1
달	1	0	1	0	1
일자	1	0	0	0	1
그 외	0	0	0	1	1
끝	0	0	0	0	1

Start probability 초기값

연도	달	일자	그 외	끝
1	1	0	3	0

- Emission probability
  - ✓ 연도 – 1960, 1961, …, 2015, 60, …, 99, 00, 01, …, 15 에 동일 확률 부여 후 업데이트
  - ✓ 달 – 01, 02, …, 12 에 동일 확률 부여하고 업데이트 없음
  - ✓ 일자 – 01, 02, …, 30 에 동일 확률 부여, 31에 반값 확률 부여, 업데이트 없음
  - ✓ 그 외 state emission probability는 초기값으로 모든 숫자 형태소에 동일한 확률 부여하고, 이후에 업데이트

구조 분석 실험

- 국내 유명 포털 사이트의 특정 카페에서 연령대 또는 성별을 확인할 수 있는 아이디 39,941개를 수집했다.
  - 이 카페에서는 연령대, 성별마다 게시판이 구분되어 있고, 자신의 신분과 맞지 않는 게시판에서 활동할 경우 엄중한 제재를 가할 것임을 명시하고 있다.
  - 제안하는 구조 분석 방법 자체는 비교사 학습으로 레이블 정보를 필요로 하지 않으나, 분석 이후 사용자 정보를 추론할 때 정확성 검증에 사용했다.
  - 게시판 별로 중복되는 아이디를 제거하여 29,544개를 최종적으로 사용했다.
  - 성별 분류기에는 교사 방법이 부분적으로 사용되어, 성별 레이블이 존재하는 27,408개를 훈련용 5,408개, 테스트용 2,000개로 사용했다.

구조 분석용

	10대	20대	30대	
남자	7661	8051	2136	
여자	6618	3078		

테스트용

	10대	20대
남자	500	500
여자	500	500

- 성별 분류기 테스트용 2,000개를 제외한 나머지 아이디로 앞 장에서 논의한 구조적 분석 수행했다.
  - Morfessor 알고리즘으로 형태소를 분해했다.

구조 분석 실험

- 문자 HMM로 학습된 전이 행렬과 초기 확률

	한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
한글 성	0	0.4863	0.0574	0	0	0	0.3636	0.0927
한글 상명	0	0	0.6780	0	0	0	0.3016	0.0203
한글 하명	0.0413	0	0	0	0	0	0.1389	0.8199
영문 성	0	0	0	0	0.2166	0.0808	0.3786	0.3240
영문 상명	0	0	0	0	0	0.4715	0.4949	0.0336
영문 하명	0	0	0	0.0293	0	0	0.2318	0.7389
그 외	0.0025	0.0090	0.0173	0.0158	0.0316	0.0724	0.4182	0.4332
끝	0	0	0	0	0	0	0	1

한글 성	4.96%
한글 상명	5.97%
한글 하명	0.06%
영문 성	5.67%
영문 상명	16.28%
영문 하명	0.89%
그 외	66.17%
끝	0

- 숫자 HMM로 학습된 전이 행렬과 초기 확률

	연도	달	일자	그 외	끝
연도	0	0.214416	0	0	0.785584
달	0.088762	0	0.644677	0	0.266562
일자	0.013316	0	0	0	0.986684
그 외	0	0	0	0.346859	0.653141
끝	0	0	0	0	1

연도	0.180333
달	0.147644
일자	0
그 외	0.672023
끝	0

구조 분석 실험

- 학습된 모형으로 저자를 포함한 7명의 아이디를 분석하고, 모형이 태깅한 의미 요소가 실제 아이디를 생성할 때 생각했던 의미와 일치하는지 비교했다.
  - 첫 번째 사용자는 hee가 상명자인데 하명자로 태깅이 되었다.
  - 마지막 사용자는 하명자가 형태소 두 개로 분리가 되어 이름 요소로 잡히지 않았다. 이는 구조 모형 학습 과정이 형태소 분리 알고리즘에는 영향을 주지 못하기 때문이다.
  - 나머지 사용자들은 대부분 의미 요소를 잘 잡아낼 수 있었다.

아이디	문자 토큰	문자 의미 태깅	숫자 토큰	숫자 의미 태깅
hee188	[hee]	[eng_back]	[188]	[number]
hank	[han, k]	[eng_front, word]		
hoseong	[ho, seong]	[eng_front, eng_back]		
jinwon	[jin, won]	[eng_front, eng_back]		
misuke88	[mi, suk, e]	[eng_front, eng_back, word]	[88]	[year]
wpgur0107	[wp, gur]	[kor_front, kor_back]	[01, 07]	[month, day]
zoon	[zoo, n]	[word, word]		

구조 분석 실험

- 구조 분석을 이용한 성별 분류
  - 남녀 레이블을 갖는 아이디를 각각의 셋으로 분리하여 2개의 HMM 모형을 학습했다.
  - 테스트 아이디가 주어지면 남녀 HMM 모형으로 각각 구조를 분석하고 모형을 적용했을 때 얻어지는 점수 값을 비교하여 점수가 높은 쪽으로 분류했다.

- 분류 결과를 아이디의 구조적인 분석 없이 형태소별 출현 확률만 고려한 나이브 베이즈 분류기(Jaech and Ostendorf, 2015)와 사람이 직접 레이블을 달아본 결과를 함께 비교했다.

구조 분석 실험

- 성별 분류 결과 방법 간에 성능 차이가 미미했다.
  - HMM 모형을 이용한 분류에서 남자 이름으로 보이는 아이디의 수집된 레이블은 여자인 경우가 종종 있어, 아이디만으로 70%이상의 성별 분류 정확도를 얻어내기는 어려워 보인다.

나이브 베이즈

Learning stage	Error rate
supervised	37.5%
semi sup iter 1	37.9%
semi sup iter 2	38.2%
semi sup iter 3	38.3%

HMM 각각

Error rate: 38.9%

Confusion matrix

	여	남
여	619	381
남	397	603

사람이 직접 레이블링

HMM 예측 score 차이 상위 20개 중 오분류된 아이디

segmented	states	predicted	real
[chan, chan, ace]	[eng_name, eng_name, word]	M	F
[tls, gh, cjf]	[kor_sur, kor_name, kor_name]	M	F
[kim, chul, ho]	[eng_sur, eng_name, eng_name]	M	F
[da, da, da]	[eng_name, eng_name, eng_name]	F	M
[qus, ckd, tjr]	[kor_sur, kor_name, kor_name]	M	F
[rh, tks, dnd]	[kor_sur, kor_name, kor_name]	M	F
[gwon, il, kwang]	[eng_sur, eng_name, eng_name]	M	F

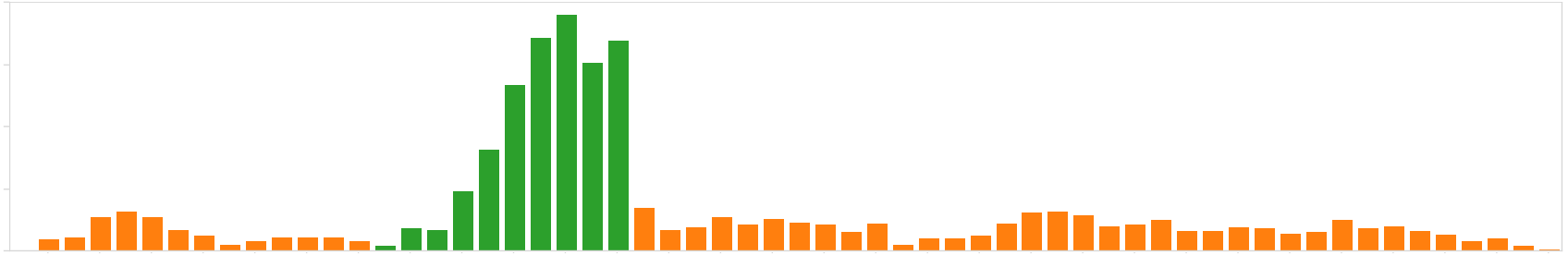
구조 분석 실험

- 구조 분석을 이용한 연령 예측
  - 숫자 부분 HMM을 이용해 연도 요소를 포함하고 있는 아이디에 대해, 해당 연도가 출생연도일 것이라고 추측하여 사용자의 연령을 예측한다
  - 연령을 추측할 수 있는 연도 요소를 포함하고 있는 아이디의 비율과 이때 추측한 연령대와 수집된 실제 연령대가 일치하는 비율을 평가 지표로 삼는다.
  - 웹 사용자의 연령을 예측하는 대부분의 모형이 어림 잡은 연령대로만 추측하는 반면, 아이디를 이용할 경우 정확한 출생년도를 추정이 가능할 수 있다.

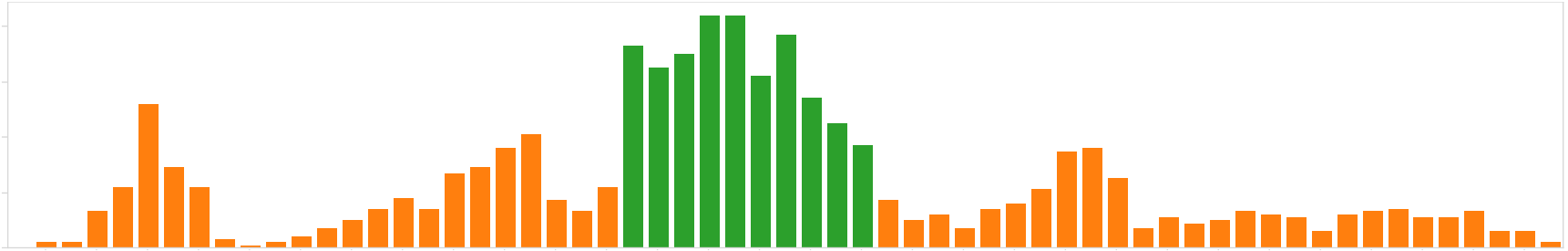
구조 분석 실험

- 구조 분석을 이용한 연령 예측

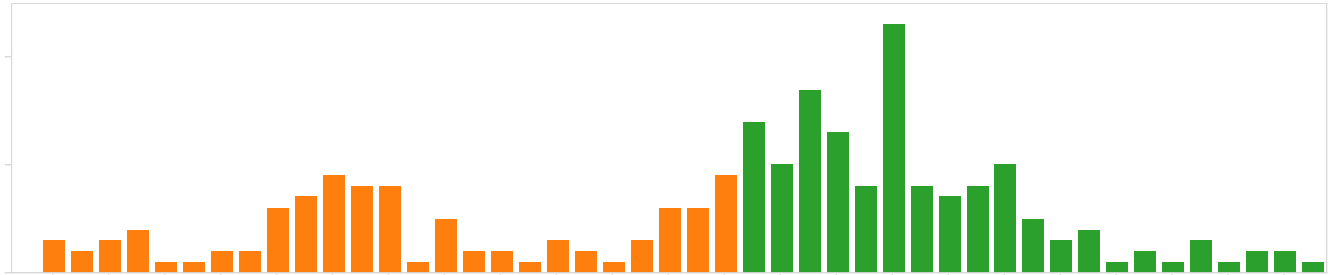
Coverage 12.8%, 연령 비율 53.8%



Coverage 12.7%, 연령 비율 46.1%



Coverage 14.7%, 연령 비율 59.6%



결론

Reference

- Jaech, A., & Ostendorf, M. (2015). What Your Username Says About You. *arXiv:1507.02045 [cs]*. Retrieved from http://arxiv.org/abs/1507.02045
- Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating Gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301–1309). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Creutz, M., & Lagus, K. (2006). Morfessor in the morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.