http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.21, No.1, February 2016, pp.65-77
http://dx.doi.org/10.7838/jsebs.2016.21.1.065

모바일 사용자의 성별 예측을 위한 식별 및 인기 단어 집합 기반 2단계 기기 내 분석

A Two-Phase On-Device Analysis for Gender Prediction of Mobile Users Using Discriminative and Popular Wordsets

최예림(Yerim Choi)*, 박규연(Kyuyon Park)**,
김소이(Solee Kim)***, 박종헌(Jonghun Park)****

초 록

모바일 기기 데이터를 활용한 분석에서 사용자의 프라이버시를 보호하는 것이 주요한 이슈로 대두됨에 따라 데이터를 외부로 전송하지 않고 모바일 기기 안에서 분석을 수행하는 기기 내 분석이 주목 받고 있다. 기기 내 분석을 활용하면 문자 메시지, 검색 단어, 북마크, 연락처 등 매우 개인적이지만 성별 구분에 효과적이라고 알려진 모바일 텍스트를 이용한 성별 예측이 가능하며, 사전에 선정된 단어들의 집합을 모바일 기기로 전송하여 이 단어들과 모바일 텍스트를 비교를 통해 성별을 예측하는 단어 비교 방식을 이용하면 모바일 기기의 제한된 자원 문제를 극복할 수 있다. 특히, 확실한 근거를 이용하여 필터링 한 후 예측을 수행하면 정확도를 극대화하고 복잡도를 낮출 수 있다. 따라서 본 논문에서는 단어의 식별력과 인기도를 순차적으로 고려하는 2단계의 기기 내 성별 예측 방법을 제안한다. 구체적으로, 제안하는 방법론은 소수의 높은 식별력을 가지는 단어를 이용하여 전체 사용자의 성별을 예측하고 이어서 인기도가 높은 단어를 활용하여 앞서 예측이 되지 않은 사용자의 성별을 예측한다. 실제 데이터를 이용한 실험에서 제안하는 방법론은 비교 방법론보다 우수한 성능을 나타내었다.

ABSTRACT

As respecting one's privacy becomes an important issue in mobile device data analysis, on-device analysis is getting attention, in which the data analysis is conducted inside a mobile device without sending data from the device to outside. One possible application of the on-device analysis is gender prediction using text data in mobile devices, such as text messages, search keyword, website bookmarks, and contact, which are highly private, and the limited computing power of mobile devices can be addressed by utilizing the word comparison method, where words are selected beforehand and delivered to a mobile device

본 연구는 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2013R1A2A2A03013947).

* Corresponding Author, Department of Industrial Engineering, Seoul National University (iangoozh@gmail.com)
** Department of Industrial Engineering, Seoul National University(mysnuky91@snu.ac.kr)
*** Department of Industrial Engineering, Seoul National University(kpsinw@gmail.com)
**** Department of Industrial Engineering, Seoul National University(jonghun@snu.ac.kr)
Received: 2016-01-12, Review completed: 2016-02-12, Accepted: 2016-02-15

of a user to determine the user's gender by matching mobile text data and the selected words. Moreover, it is known that performing prediction after filtering instances using definite evidences increases accuracy and reduces computational complexity. In this regard, we propose a two-phase approach to on-device gender prediction, where both discriminability and popularity of a word are sequentially considered. The proposed method performs predictions using a few highly discriminative words for all instances and popular words for unclassified instances from the previous prediction. From the experiments conducted on real-world dataset, the proposed method outperformed the compared methods.

키워드 : 기기 내 분석, 성별 예측, 모바일 텍스트, 2단계 기법, 식별 단어 집합, 인기 단어 집합
On-Device Analysis, Gender Prediction, Mobile Text, Two Phase Approach, Discriminative Wordset, Popular Wordset

1. 서 론

모바일 기기에서 발생하는 데이터의 수집이 폭발적으로 증가함에 따라 데이터에 담긴 개인 정보 보호가 중요 이슈로 대두되고 있다. 이에 따라 데이터를 이용하는 과정에서 개인 정보 보호를 목적으로 하는 기기 내 분석이 주목을 받고 있다. 기기 내 분석은 모바일 기기에서 생성된 데이터를 외부로 전송하지 않고 기기 내부에서 데이터 분석을 수행하는 방법론을 의미한다[5].

텍스트 데이터를 이용하는 성별 예측은 기기 내 분석의 주요 적용 분야이다. 텍스트 데이터는 성별 예측에 효과적이라고 알려져 있어[7] SNS나 블로그에 작성된 글[3, 12]이나 영화 리뷰[11] 등 웹 문서를 이용한 성별 예측 연구가 활발히 이루어지고 있다. 하지만 문자 메시지, 검색 단어, 주소록과 같은 모바일 텍스트의 수집은 개인 정보 유출의 소지가 있어[1, 9] 성별 예측에 활용이 제한되어 왔다. 따라서 기기 내 분석을 활용하면 개인 정보 유출 문제 없이 모바일 텍스트를 이용한 기기 사용자의 성별 예측이 가능하다.

하지만 모바일 기기의 경우 일반적인 컴퓨팅 기기에 비해 자원이 제한적이므로 단어 추출 및 단어 벡터 생성과 같이 높은 연산 능력이 요구되는 기존의 텍스트 분석 방법[10]을 도입할 수 없다. 예를 들어 대표적인 통계학습 기법인 Naïve Bayes 분류기를 사용하여 성별을 예측하려면 학습 데이터를 이용하여 분류기를 학습시키고 학습된 모델을 기기 내로 전송하여 성별 예측이 수행된다. 하지만 기기 내에서 학습된 모델을 통해 예측을 수행하기 위해서는 모바일 텍스트로부터 단어 벡터를 생성하는 복잡한 과정이 필요하다.

본 연구진의 기존 연구[5]에서는 제한된 모바일 기기의 자원 문제를 해결하기 위해 비교적 계산 복잡도가 낮은 단어 비교 방식을 이용한 성별 예측 기법을 제안하였다. 우선, 충분한 자원이 제공되는 상황에서 평가 지표에 따라 각 성별을 대표하는 단어를 선정하여 성별에 대한 단어 집합을 구성한다. 이후, 단어 집합을 모바일 기기로 전송하여 기기 내에서 단어 집합과 모바일 텍스트의 유사도 비교를 통해 기기 사용자의 성별을 예측한다. 이 방식은 단일 단계로 구성되며 단어를 하나의 지표로 평가하여 하나의 특성 밖에는 고려하지 못한다는

문제점이 있다.

기기 내 단어 비교를 이용한 성별 예측의 성능을 극대화시키기 위해서는 단어의 다양한 특징이 고려되어야 한다. 웹 문서와 비교하여 모바일 텍스트에는 등장하는 단어의 수가 적으므로 단어의 성별에 대한 식별력만큼이나 인기도도 중요하다[5]. 예를 들어, 어머니가 딸을 지칭할 때 사용하는 인터넷 용어인 '딸램'과 같이 성별 식별력이 매우 높은 단어를 사용하여 기기 내 성별 예측을 수행한다면 높은 예측 정확도를 얻을 수 있지만, 단어가 모바일 텍스트에 등장하지 않는 경우가 많아 높은 미분류율을 보일 것이다. 하지만 식별력과 인기도를 단순히 동시에 고려한다면 '액션'이나 '게임'과 같이 인기도가 높지만 식별력이 떨어지는 단어를 선정하게 되어 예측 정확도가 떨어지게 될 것이다.

이처럼 데이터의 다양한 특성을 서로 상쇄시키지 않으면서 동시에 고려하여 분석에 반영하기 위한 방법론들이 연구되어 왔다. 앙상블 학습 방식을 이용하면 서로 다른 특성을 가지는 데이터를 이용하여 별도로 예측을 수행한 후 그 결과를 조합하여 예측 성능을 극대화할 수 있다[6]. 또한, 확실한 근거가 주어진다면 예측을 순차적으로 수행하여 비교적 확실한 인스턴스를 먼저 필터링 할 수 있다. Han 등[4]은 문서 분류를 위해 소수의 식별력 높은 단어를 선정하여 분류가 확실한 문서를 일차적으로 필터링 하는 방식을 이용하여 좋은 성능을 얻을 수 있었다.

따라서 본 연구에서는 식별력과 인기도를 반영하는 단어를 순차적으로 고려하여 기기 내 예측을 수행하는 2단계 방법론을 제안한다. 웹 문서로부터 추출된 단어들의 식별력 및 인기도 점수를 계산한 후 미리 정해진 개수만큼 단어를 선정하여 식별 및 인기 단어 집합을 구축한다. 구축된 단어 집합과 모바일 텍스트를 이용하여 순차적으로 기기 사용자의 성별을 예측한다. 첫 번째 단계에서는 식별 단어 집합 내 소수의 식별력이 높은 단어를 이용하여 성별을 예측하고, 만일 사용자의 성별이 첫 단계에서 미분류 된 경우 인기 단어 집합을 이용하여 다시 예측을 수행한다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 제안하는 방법론을 구체적으로 설명하고, 제 3장에서는 성별 예측 성능의 평가를 위해 실제 데이터를 이용한 비교 실험을 수행한다. 마지막으로 제 4장에서 결론을 제시하여 논문을 마무리 짓는다.

2. 기기 내 성별 예측을 위한 2단계 기법

2.1 개요

본 연구에서는 성별에 따른 단어 집합과 모바일 텍스트를 활용하여 모바일 사용자의 성별을 기기 내 단어 비교 방식으로 예측하고자 한다. 특히, 예측의 정확도를 극대화하고 소요시간을 줄이기 위해 성별에 대한 식별력이 높은 소수의 단어들을 이용해 첫 단계 예측을 수행한다. 또한, 미분류율을 낮추기 위해 첫 단계 예측에서 성별이 분류되지 않은 사용자에 대해서만 다수의 인기 단어들을 이용해 두 번째 예측을 수행한다. 이때, 성별은 g로 표기하며 {female, male} 중 하나의 값을 가진다. 하나의 단어는 w로 표기하고, 각 g에 대해 식별력이

높은 w의 집합을 D_g로 표기하고 식별 단어 집합이라고 부르며, 인기가 많은 w의 집합을 P_g로 표기하고 인기 단어 집합이라고 부른다. 특정한 사용자의 모바일 텍스트는 t로 표기하며, 계산 복잡도를 낮추기 위해 t는 문서로부터 추출된 단어들의 집합이 아닌 추출 과정 없이 문자들이 이어진 시퀀스 형태로 사용한다.

제안 방법론은 <Figure 1>과 같이 크게 단어 선정과 성별 예측의 두 과정으로 나누어진다. 가장 먼저 웹에서 저자의 성별이 알려진 웹 문서들을 수집하고, 문서들로부터 단어를 추출한다. 그 후 w의 $g \in$ {female, male}에 대한 식별력과 인기도를 평가하여 그 결과를 기준으로 D_g와 P_g를 선정한다.

마지막으로, D_g와 P_g를 모바일 기기로 전송하여 성별 예측을 수행한다. 선정된 단어들의 t에서의 발생 빈도와 앞선 단어 평가 결과를 토대로 유사도를 계산하여 성별 예측을 수행한다. 이때, 사용하는 단어 집합에 따라 예측을 순차적으로 수행한다. 첫 번째 단계에서는 D_g를 이용하여 예측을 수행하고, 그 결과 분류되지 않은 경우 두 번째 단계에서 P_g를 이용하여 성별 예측을 수행한다. 단어 선정의 경우 충분한 컴퓨팅 능력이 요구되므로 모바일 기기 외부에서 미리 수행되고 그 결과물을 모바일 기기 내로 이전하여 성별 예측이 수행된다.

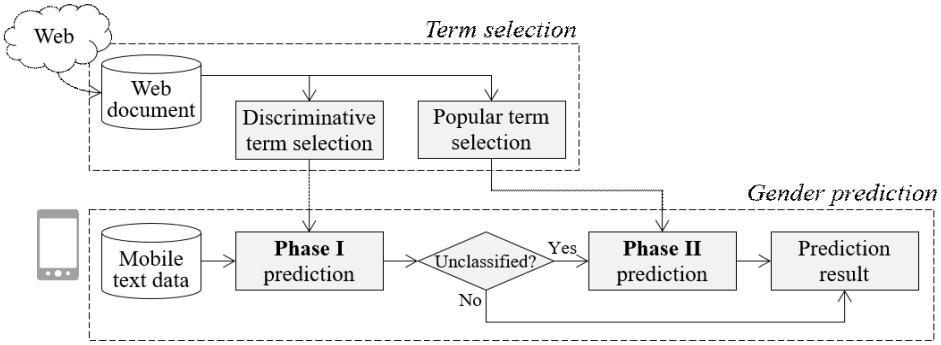
2.2 단어 선정

단어 선정 단계에서는 웹에서 수집된 문서로부터 추출한 w 중 g에 대한 식별력이 높거나 인기가 많은 단어를 선정하여 D_g와 P_g를 구성한다. 이를 위해 식별력 점수(DS: discriminative score)와 인기도 점수(PS: popularity score)를 정의한다.

w의 g에 대한 DS 값은 w의 두 성별을 구별하는 능력을 나타내며 ds_{gw}로 표기한다. 구체적으로, 식 (1)을 이용하여 w가 출현한 문서가 성별이 g인 사용자가 쓴 문서일 가능성을 측정한다.

$$ds_{gw} = \frac{p(g \mid w)p(g \mid \overline{w})}{p(g \mid w)p(g \mid \overline{w})} = \frac{df_{gw}df_{\bar{g}\overline{w}}}{df_{\bar{g}w}df_{g\overline{w}}}. \quad (1)$$

이때, df_{gw}는 w의 g에 대한 문서 빈도(document frequency) 값으로 기존 연구에서 널리 사용되며[13], <Table 1>에 나타난 바와 같이



〈Figure 1〉 Overview of the Proposed two Phase On-Device Gender Prediction Method

w가 출현하고 저자의 성별이 g인 문서의 수를 의미한다. $\bar{g}$는 g의 반대 성별을 나타내며, w와 $\bar{w}$는 해당 단어의 출현과 부재를 각각 나타낸다. 각 성별의 전체 문서 수 차이로 인한 불균형 문제를 해결하기 위해 정규화된 값을 사용하였다.

〈Table 1〉 Confusion Matrix of the Document Frequency According to g of the Author of a Document and the Existence of w in the Document

		Gender	
		g	$\bar{g}$
Word existence	w	df_{gw}	$df_{\bar{g}w}$
	$\bar{w}$	$df_{g\bar{w}}$	$df_{\bar{g}\bar{w}}$

w의 g에 대한 PS 값은 w의 성별이 g인 사용자 사이의 인기도를 나타내며 ds_{gw}로 표기한다. 구체적으로 식 (2)를 이용하여 성별이 g인 사용자가 저자인 문서에 w가 나타날 가능성을 측정한다.

$$ps_{gw} = P(w \mid g) = df_{gw}. \quad (2)$$

이때, 식 (2)에서 우변에 분모인 저자의 성별이 g인 문서 수는 성별에 대해 일정하기 때문에 생략하였다.

DS와 PS 값에 따라 각 성별에 대해 미리 정해진 수의 단어들이 기기 내 성별 예측을 위해 선정된다. DS를 기준으로는 n_d개의 단어를 선정하고 PS를 기준으로는 n_p개의 단어를 선정하며, 이는 모든 g에 대해 동일하다. 선정된 단어들은 기준 점수와 g에 따라 D_g와 P_g를 구성한다.

2.3 기기 내 성별 예측

각 g에 대한 D_g와 P_g를 사용자의 모바일 기기 내에서 t와 단어 비교 방식을 통해 사용자의 성별을 예측한다. 사용하는 단어 집합에 따라 두 단계로 구분되며, 첫 번째 단계에서 성별이 분류되지 않은 사용자에 대해서만 두 번째 단계에서 예측을 수행한다. 구체적으로, 첫 단계에서는 D_g가 두 번째 단계에서는 P_g가 사용된다. 따라서 첫 번째 단계에서는 식별력이 높은 단어들이 사용되어 높은 정확도가 기대되지만 인기도가 고려되지 않아 높은 미분류율을 보인다. 반면 두 번째 단계에서는 첫 번째 단계 대비 정확도는 떨어지나 낮은 미분류율을 보인다.

단어 비교를 이용한 성별 예측 방식은 두 단계에서 동일하게 구성된다. 우선, 성별에 따른 단어 집합과 t의 비교를 수행하여 빈도수 벡터(frequency vector)를 생성한다. 빈도수 벡터는 f_{gt}로 표기되며, 벡터의 원소는 대응하는 단어 집합의 원소의 t에서의 발생 빈도를 나타낸다. 예를 들어, g = female 일 때, D_g의 원소 '딸램'이 t에서 두 번 발견되었다면 f_{gt}에서 '딸램'에 해당하는 원소의 값은 2로 주어진다.

다음으로, f_{gt}와 이전 단계에서 얻어진 각 g의 단어 집합에 대한 평가 점수 벡터를 이용하여 t의 저자의 성별이 g일 가능성을 계산한다. 평가 점수 벡터는 s_g로 표기하며 s_g의 각 원소는 대응하는 단어의 DS 또는 PS 값을 의미한다. 예를 들어, g = female일 때, P_g의 원소 '사진'이 웹 문서에서 천 번 발견되었다면 s_g에서 '사진'에 해당하는 원소의 값은 1,000으로 주어진다. t의 저자의 성별이 g일 가능성은 f_{gt}와 s_g의 유사도로 정의되며, 본 연구에서는 단어 벡터 비교에 널리 사용되는[8] 코사인 유사도

(cosine similarity)를 도입하여 식 (3)과 같이 정의한다.

$$\mathrm{sim}(g,\ t) = \frac{f_{gt} \cdot s_g}{|f_{gt}\|s_g|}. \qquad (3)$$

이때, $f_{gt} \cdot s_g$은 두 벡터의 내적을 의미하고 f_{gt}와 $|s_g|$는 각 벡터의 크기를 나타낸다.

마지막으로 모든 $g \in$ {female, male}에 대한 $\mathrm{sim}(g, t)$ 값을 비교하여 t의 저자의 성별을 예측한다. 즉, t가 주어졌을 때, 저자의 성별의 예측값 $\hat{g}$는 식 (4)와 같이 주어진다.

$$\hat{g} = \mathrm{argmax}_g\ \mathrm{sim}\ (g, t) \qquad (4)$$

3. 모바일 기기 사용자 성별 예측 실험

3.1 수집 데이터

제안하는 모델의 성별 예측 성능을 알아보기 위해 실제 데이터를 이용한 성능 평가 실험을 수행하였다. 저자의 성별이 알려진 문서를 수집하기 위해 성별을 공개한 블로거의 블로그 문서를 수집하였으며, <Table 2>는 수집한 데이터에 대한 정보를 나타낸다. 수집된 문서의 총 수는 189,127건으로 여자 블로거의 문서는 53,382건, 남자 블로거의 문서는 135,745건이었다. 남자 문서의 평균 길이는 301단어로 359단어의 여자 문서에 비해 짧은 것을 알 수 있다. 구체적으로, 한글 블로그 서비스 중 가장 대표적인 네이버 블로그(http://blog.naver.com)에서 문서를 수집하였으며, 주제나 연령대의 편중으로 인한 단어의 편향을 막기 위해 서비스에서 제공하는 파워블로거 목록에서 무작위로 블로그를 선정하였다. 블로그 문서로부터 정교한 단어 추출을 위해 꼬꼬마 한글 분석기[14]와 Lucene 한글 분석기에서 동시에 추출되는 단어만을 실험의 대상으로 하였다.

제안하는 기기 내 분석 기법의 성능을 평가하기 위해서 피실험자로부터 모바일 텍스트를 수집하였다. 구체적으로, 자체 개발한 안드로이드 어플리케이션을 이용하여 모바일 기기로부터 문자 메시지, 검색 키워드, 북마크, 주소록으로 구성된 모바일 텍스트를 수집하였다. <Table 2>에 나타난 바와 같이, 남자 16명, 여자 16명으로 구성된 총 32명의 피실험자가 실험에 참여하였다. 성별에 따른 모바일 텍스트의 평균 길이는 남자 53,432단어, 여자 124,945단어로 여자의 모바일 텍스트의 경우 남자보다 두 배 가량 긴 것을 확인 할 수 있었다.

〈Table 2〉 Summary of the Collected Datasets in Terms of the Total Numbers and Average Lengths

Data type	Category	Total	Female	Male
Web document	The number of documents	189,127	53,382	135,745
	The average length of documents(in word)	317	359	301
Mobile text	The number of users	32	16	16
	The average length of texts(in character)	89,188	124,945	53,432

3.2 실험 환경

제안하는 방법론의 성능을 평가하기 위해 단일 단계 예측 방법과 충분한 자원이 제공된다고 가정하였을 때 사용할 수 있는 통상적인 방법을 추가적으로 구현하였다. 단일 단계 방법에서는 단어 선정을 위해 앞서 정의한 DS 및 PS와 Chi-square(CHI)를 단어 평가 지표로 사용하였다. CHI는 w와 g가 얼마나 의존적인지를 나타내는 지표로 웹 문서를 대상으로 한 성별 예측 연구에서 가장 우수한 성능을 나타내는 것으로 알려져 있다[16]. CHI는 식 (5)를 이용하여 계산된다.

$$chi_{gw} = \frac{N \times (df_{gw}df_{\bar{g}\bar{w}} - df_{\bar{g}w}df_{g\bar{w}})^2}{(df_{gw}+df_{\bar{g}w})(df_{g\bar{w}}+df_{\bar{g}\bar{w}})(df_{gw}+df_{g\bar{w}})(df_{\bar{g}w}+df_{\bar{g}\bar{w}})} \quad (5)$$

이때, N은 문서의 총 개수를 나타낸다.

통상적인 방법으로는 대부분의 분류 문제에서 좋은 성능을 보인다고 알려져 있는 지지기반벡터(support vector machine)[15]을 도입하였다. 성별이 알려진 블로그 문서를 단어 벡터(word vector) 형태로 표현하여 지지기반벡터를 학습하는 데에 사용하고, 학습된 지지기반벡터를 이용하여 단어 벡터 형태로 표현된 모바일 텍스트의 성별을 예측하였다. 단어 벡터 구성을 위한 단어는 CHI 지표를 기준으로 선정된 남/녀 단어의 합집합을 이용하였다. 지지기반벡터의 커널은 선형 커널(linear kernel)을 사용하였으며, 그 외 파라미터는 libsvm[2]에서 제공하는 기본값을 사용하였다.

성능 평가 지표로는 거시 F(macro F) 점수, 미분류율(uncalssification rate), 예측시간을 사용하였다. 거시 F 점수는 분류 문제에서 널리 사용되는 지표로 거시 정밀도(precision)와 재현율(recall)의 조화평균으로 정의된다. 거시 정밀도와 재현율은 각 성별의 정밀도와 재현율의 가중 평균이며, 각 g에 대한 정밀도 pr_g는 g로 예측한 인스턴스 중 실제로도 g인 인스턴스의 비율, 재현율 rc_g은 실제로 g인 인스턴스 중 g로 예측한 인스턴스의 비율로 각각 식 (6)과 식 (7)을 이용하여 계산된다.

$$pr_g = \frac{C_g}{C_g + I_g}. \quad (6)$$

$$rc_g = \frac{C_g}{C_g + I_{\bar{g}}}. \quad (7)$$

이때, C_g와 I_g는 <Table 3>에 나타난 것과 같이 각각 사용자의 성별을 g로 예측하였을 때 옳은(correct) 경우의 수와 옳지 않은(incorrect) 경우의 수를 나타낸다.

〈Table 3〉 Confusion Matrix of the Number of the Classified Instances According to the Actual and Predicted Genders with the Number of the Unclassified Instances

		Predicted gender		Unclassified instance
		g	$\bar{g}$	
Actual gender	g	C_g	$I_{\bar{g}}$	U_g
	$\bar{g}$	I_g	$C_{\bar{g}}$	$U_{\bar{g}}$

전체 인스턴스 중 분류되지 않은 인스턴스의 비율을 나타내는 미분류율도 평가지표로 사용되었다. 미분류율 un는 식 (8)과 같이 계산된다.

$$un = \frac{U_g + U_{\bar{g}}}{C_g + I_g + I_{\bar{g}} + C_{\bar{g}}}. \quad (8)$$

〈Table 4〉 List of the Top Five Selected Terms for two Genders According to the Term Evaluation Measures, DS, PS, and CHI with the Term Description

Measure	Female	Male
	Term	Term
DS	Cebu Pacific, a Philippines airline(세부퍼시픽)	Destroyer, a battle ship(구축함)
	Rose geranium(로즈제라늄)	US Navy(미해군)
	Overly expressing a fuss(난리난리)	Bull pen(불펜)
	A cute words for calling a daughter(딸램)	Starting(스타팅)
	Child-rearing mother(육아맘)	Comic Con(코믹콘)
PS	Thinking(생각)	Thinking(생각)
	Person(사람)	Person(사람)
	We(우리)	Degree(정도)
	Photo(사진)	Because(때문)
	Mind(마음)	We(우리)
CHI	Mom(엄마)	Director(감독)
	Women's old sister(언니)	Movie(영화)
	Completely(완전)	Action(액션)
	Husband(신랑)	Game(게임)
	Child(아이)	Battle(전투)

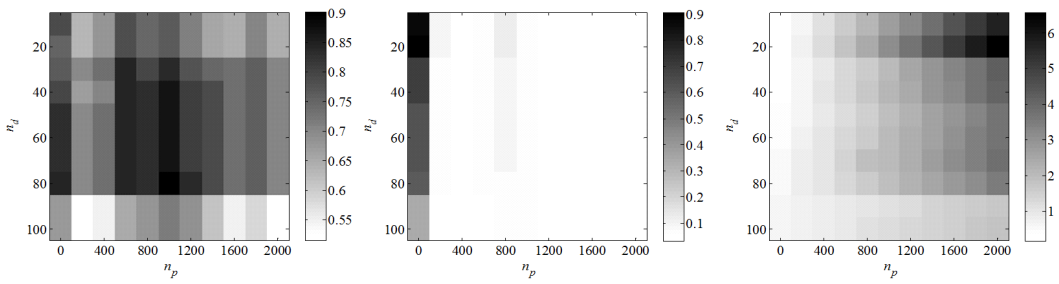
이때, U_g와 $U_{\bar{g}}$는 실제로 g와 $\bar{g}$인 인스턴스 중 분류가 이루어지지 않은 인스턴스의 수를 나타낸다. 또한, 테스트 인스턴스들이 주어졌을 때, 예측을 수행하는데 걸리는 시간인 예측 시간이 평가지표로 사용되었다. 예측 시간의 단위는 초를 사용하였다.

3.3 실험 결과

세 종류의 실험 결과를 통해 제안하는 모델의 성능을 평가하고 타당성을 검증하였다. 첫째, 단어 평가 지표인 DS, PS, CHI를 이용하여 수집된 단어 들 중 가장 높은 점수를 받은 단어들을 <Table 4>에 나타내었다. DS를 기준으로 선정된 단어들은 주로 'Cebu Pacific'이나 'Comic Con'과 같은 고유명사가 많았으며, '딸램', '육아맘'과 같은 신조어도 눈에 띄었다. 반면, PS는 매우 일반적인 단어를 골라 총 다섯 단어 중 세 단어가 동시에 남자와 여자 문서에서 선정되었다. CHI는 PS보다는 상대적으로 식별력이 높고 DS보다는 인기도가 높은 단어들이 선정되었다.

다음으로, 본격적인 성능 비교에 앞서 제안 방법론의 최적의 파라미터를 선택하기 위해 다양한 n_p와 n_d 값에 따른 제안 방법론의 성능을 확인하였다. <Figure 2>는 n_p와 n_d 값에 따른 제안 방법론의 성능을 나타낸다. 이때, n_p 값이 0인 경우는 두 번째 단계가 생략된 것으로, D_g만을 이용한 첫 단계의 성능을 의미한다.

n_p 값이 0인 경우 n_d가 80일 때 가장 높은 거시 F 값을 얻을 수 있었으며, n_d가 감소함에 따라 미분류율이 증가하는 것을 확인 할 수 있었다. 이는 n_d 값이 80보다 큰 경우 단어들의 성별 식별력이 낮은 단어들이 포함되어 구분 정확도가 떨어지는 것이라고 설명할 수 있다. 또한, n_d가 감소하면 그만큼 모바일 텍스트에서 발견될 확률이 낮아지므로 미분류율은 높아진다.



(a) Macro F (b) Unclassification rate (c) Prediction time

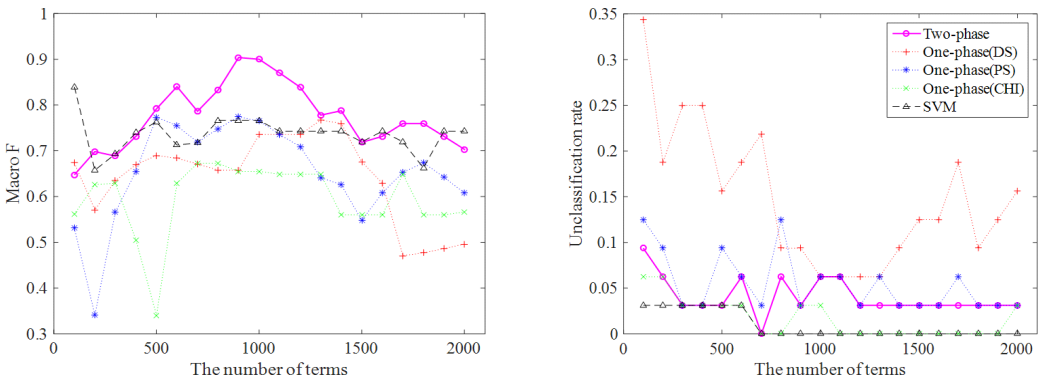
〈Figure 2〉 Prediction Performances of the Proposed Method According to n_d and n_p in Terms of Macro F, Unclassification Rate, and Prediction Time

제안 방법론은 n_d가 80이고 n_p가 1,000일 때, 0.9를 넘는 거시 F 값으로 가장 좋은 성능을 나타내었다. p_g를 사용하는 두 번째 단계까지 수행한 경우, 대부분의 경우에서 제안 방법론의 미분류율은 0.1 이하로 나타났다. 예측 시간은 n_d가 감소할수록, n_p가 증가할수록 증가한다. 이는 첫 번째 단계에서 예측이 이루어지지 않은 사용자에 대한 예측이 두 번째 단계에서 상대적으로 많은 수의 인기 단어를 이용하여 이루어지므로 n_d가 감소할수록 예측시간이 증가하게 된다.

마지막으로 제안 방법론과 DS, PS, CHI를 단어 평가 지표로 사용한 단일 단계 방법론, 지지기반벡터의 성능 비교 실험을 수행하였다. <Figure 3>은 예측을 위해 사용된 단어의 수에 따른 각 방법론의 성능을 나타낸다. 2단계 모델의 경우 단어의 수는 n_p를 의미하고 n_d는 앞선 실험에서 가장 높은 거시 F 값을 나타낸 80으로 고정하였다.

단어의 수가 100개와 2,000개인 양극단의 경우를 제외하고는 제안한 2단계 방법론이 다른 비교 방법론 대비 가장 높은 거시 F 값을 나타내었다. 지지기반벡터가 두 번째로 우수한 성능을 나타내었지만, 단어의 수가 1,000개 일 때 제안 방법론과 약 0.15의 거시 F 값 차이를 보였다. 또한, 미분류율 측면에서, 제안 방법론은 PS와 DS를 이용한 단일 단계 방법론 보다는 낮은 미분류율을 나타내었으며, CHI를 이용한 단일 단계 방법론이나 SVM과는 견줄만한 값을 보였다. 전반적으로 제안 방법론은 약 0.05 이하의 매우 낮은 미분류율을 나타내었다.

특히, 2단계 방법론과 PS를 이용한 단일 단계 방법론의 성능이 유사한 패턴을 나타내는 것을 확인 할 수 있다. <Figure 3>(a)에서 단어의 수가 600개 이상인 경우 제안 방법론과 PS를 이용한 단일 단계 방법론이 일정한 거시 F 값의 차이를 갖는다. 이는 전체 인스턴스 중에서 차이에 해당하는 인스턴스들이 DS를 이용한 첫 번째 단계에서 매우 높은 정확도로 선분류가 되어 전체적으로 평가하였을 때 PS를 이용한 단일 단계 방법론에서 일정 수준의 정확도를 높이는 효과를 가져온다는 것을 의미한다. 또한, <Figure 3>(b)에서는 단어의 수가 900개 이상인 경우부터는 제안 방법론과 PS를 이용한 단일 단계 방법론의 미분류율이 거의 동일한 값을 나타내는 것을 확인할 수 있다. 이는 n_p가 일정 수준을 넘어가면 전체 미분류율이 PS를 이용한 단일 단계 방법론에 의존한다는 것을 의미한다.



(a) Macro F (b) Unclassification rate

〈Figure 3〉 Performance Comparison Results of the Proposed Method, One-Phase Methods(DS, PS, CHI), and SVM in Terms of Macro F and Unclassification Rate

〈Table 5〉 Performance Comparison Results of the Proposed Method, One-Phase Method, and SVM in Terms of Prediction Time in Second

Method	Prediction time(s)
Two-phase	2.01
One-phase	4.79
SVM	844.88

<Table 5>는 제안 방법론, 단일 단계 방법론, 지지기반벡터의 예측 시간을 나타낸다. 제안 방법론이 단일 단계 방법론 대비 절반 정도의 시간이 소요되었다. 이는 제안 방법론의 경우 첫 번째 단계에서 소수의 식별 단어를 이용하여 확실한 사용자의 성별을 빠르게 예측하고 첫 번째 단계에서 미분류 된 사용자만을 대상으로 다수의 인기 단어를 이용한 예측을 수행하기 때문이다. 특히, 지지기반벡터는 단어 비교 방식을 이용한 기법들 대비 매우 오랜 시간이 소요되었는데, 학습된 지지기반벡터를 이용한 예측을 위해서는 모바일 텍스트로부터 단어를 추출하고 단어 벡터를 구성하는 단계가 선행되어야 하기 때문이다. 따라서 지지기반벡터와 같이 컴퓨팅 능력을 요하는 방법론을 이용하여 기기 내 예측을 수행하는 데에는 한계가 있음을 다시 한 번 확인할 수 있었다.

4. 결 론

본 논문에서는 모바일 텍스트를 이용한 모바일 기기 내 사용자의 성별 예측을 위해 D_g와 P_g를 순차적으로 활용하는 2단계 기법을 제안하였다. 성별이 알려진 웹 문서로부터 단어들이 추출한 후 DS와 PS를 이용하여 각 단어의 성별에 대한 식별력과 인기도를 평가한다. 이후 평가 결과에 따라 각 성별마다 정해진 수의 식별 단어들과 인기 단어를 선정하고 이들은 순차적으로 성별 예측에 사용한다. 첫 번째 단계에서는 식별 단어들과 모바일 텍스트를 비교해 사용자의 성별을 예측하고, 이 단계에서 분류되지 않은 사용자는 두 번째 단계에서 인기 단어들을 이용해 사용자의 성별을 예측한다.

실제 웹 문서와 피실험자로부터 수집한 모바일 텍스트를 이용하여 제안 방법론의 성능을 평가하였다. 제안 방법론은 DS와 PS를 이용한 단일 단계 방법론보다 정확도와 미분류율 측면 모두에서 우수한 성능을 나타내었으며, 충분한 컴퓨팅 능력을 요구하는 지지기반 벡터보다도 우수한 성능을 보였다. 제안하는 방법론은 기기 내 데이터 분석 기법을 도입함으로써 개인 정보를 담고 있는 모바일 텍스트를 기기 외부로 전송하지 않아도 된다는 장점을 가지며, 단어 비교를 이용하여 성별을 예측한다는 점에서 모바일 기기의 자원 한정 문제를 해결하였다. 또한, 단어의 성별에 대한 식별력과 인기도를 순차적으로 고려하여 예측 시간은 최소화하면서 예측 정확도를 최대화 하였다.

제안하는 방법론의 성능이 예측하는데 사용된 식별 단어와 인기 단어의 수에 따라 큰 변동성을 가지는 것을 확인 할 수 있었다. 따라서 추후 연구에서는 주어진 데이터 집합에 대해 빠르게 최적 단어 수를 찾는 방법을 연구할 예정이다. 또한, 제안하는 방법론은 보다 다양한 단어의 특성을 반영하기 위해 다 단계 방법으로 확장할 수 있을 것이다.

References

[1] Baek, S. and Choi, D., “Exploring User Attitude to Information Privacy,” The Journal of Society for e-Business Studies, Vol. 20, No. 1, pp. 45-59, 2015.

[2] Chang, C. C. and Lin, C. J., “LIBSVM: A Library for Support Vector Machines,” ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1-27, 2011.

[3] Goswami, S., Sarkar, S., and Rustagi, M., “Stylometric Analysis of Bloggers’ Age and Gender,” Proceedings of the International AAAI Conference on Weblogs and Social Media, pp. 214-217, 2009.

[4] Han, J., Park, M., and Kim, J., “Improving the Performance of Automatic Text Categorization by Using Phrasal Patterns and Keyword Sets,” Proceedings of the Korea Computer Congress, pp. 70-73, 1998.

[5] Kim, S., Choi, Y., Kim, Y., Park, K., and Park, J., “On-Device Gender Prediction Framework Based on the Development of Discriminative Word and Emoticon Sets,” KIISE Transactions on Computing Practices, Vol. 21, No. 11, pp. 733-738, 2015.

[6] Kim, Y., Choi, Y., Kim, S., Park, K., and Park, J., “An Ensemble Model for Gender Classification of Mobile Users,” Proceedings of the International Conference on Computer Technology and Development, 2015.

[7] Lakoff, R., “Language and Woman’s Place,” Language in Society, Vol. 2, No. 1, pp. 45-80, 1973.

[8] Lee, D. and Shim, J., “Survey on Vector Similarity Measures: Focusing on Algebraic Characteristics,” The Journal of Society for e-Business Studies, Vol. 17, No. 4,

pp. 209-219, 2012.

[9] Lee, J., Choi, H., and Choi, S., "Study on How Service Usefulness and Privacy Concern Influence on Service Acceptance," The Journal of Society for e-Business Studies, Vol. 12, No. 4, pp. 37-51, 2007.

[10] Lee, K., Kim, K., Lee, M., Kim, W., and Hong, J., "Post Clustering Method using Tag Hierarchy for Blog Search," The Journal of Society for e-Business Studies, Vol. 16, No. 4, pp. 301-319, 2011.

[11] Otterbacher, J., "Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata," Proceedings of the ACM International Conference on Information and Knowledge Management, pp. 369-378, 2010.

[12] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M., "Classifying Latent User Attributes in Twitter," Proceedings of the International Workshop on Search and Mining User-Generated Contents, pp. 37-44, 2010.

[13] Roh, J., Kim, H., and Jang, J., "Improving Hypertext Classification Systems through WordNet-based Feature Abstraction," The Journal of Society for e-Business Studies, Vol. 18, No. 2, pp. 95-110, 2013.

[14] Shim, K., "MADE: Morphological Analyzer Development Environment," Journal of Internet Computing and Services, Vol. 8, No. 4, pp. 159-171, 2007.

[15] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[16] Yang, Y. and Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the International Conference on Machine Learning, pp. 412-420, 1997.

저 자 소 개

최예림 (E-mail: iangoozh@gmail.com)
2010년 서울대학교 산업공학과 (학사)
2010년～현재 서울대학교 산업공학과 (석박사 통합과정)
관심분야 사물인터넷 및 빅데이터 기반의 인간 모델링

박규연 (E-mail: mysnuky91@snu.ac.kr)
2015년 서울대학교 산업공학과 (학사)
2015년～현재 서울대학교 산업공학과 (석사과정)
관심분야 모바일 서비스

김소이 (E-mail: kpsinw@gmail.com)
2014년 포항공과대학교 산업경영공학과 (학사)
2014년～현재 서울대학교 산업공학과 (석사과정)
관심분야 데이터마이닝, 추천 시스템

박종헌 (E-mail: jonghun@snu.ac.kr)
1990년 서울대학교 산업공학과 (학사)
1992년 서울대학교 산업공학과 (석사)
2000년 Georgia Institute of Technology 산업시스템공학과 (박사)
2004년～현재 서울대학교 산업공학과 교수
관심분야 모바일 인텔리전스, 산업 데이터 애널리틱스