공학석사 학위논문

스마트 기기의 다종 데이터를 이용한 사용자 성별 예측 기법

A Method for User Gender Prediction
Using Multi-type Smart Device Log Data

2015 년 8 월

서울대학교 대학원
산업공학과
김 윤 정

스마트 기기의 다종 데이터를 이용한 사용자 성별 예측 기법

지도 교수 박 종 헌

이 논문을 공학석사 학위논문으로 제출함

2015 년 7 월

서울대학교 대학원

산업공학과

김 윤 정

김윤정의 공학석사 학위논문을 인준함

2015 년 7 월

위 원 장 조 성 준 (인)

부위원장 박 종 헌 (인)

위 원 이 재 욱 (인)

초 록

스마트 기기의 사용이 보편화 됨에 따라 스마트 기기 사용자의 특성에 맞는 서비스를 제공할 필요성이 증가하고 있다. 사용자의 성별 정보는 개인화된 서비스를 위한 기초적이고 중요한 정보라 할 수 있다. 따라서 본 연구에서는 스마트 기기로부터 발생한 로그 데이터를 이용하여 사용자의 성별을 예측하는 앙상블 기법을 제안한다. 세 종류의 데이터에 기반한 사용자 성별 분류기의 예측 결과를 다수결 방식으로 앙상블하여 최종 성별을 예측한다.

텍스트 데이터를 이용한 분류기는 텍스트 데이터에 의한 사생활 침해 문제를 최소화하기 위해 사용자의 기기 내에서 성별 분류를 수행한다. 사전에 남녀를 대표하는 단어집합을 결정하고, 기기 내에서 텍스트 데이터와 비교하여 사용자의 성별을 분류한다. 단어집합 결정을 위해서 웹에서 문서를 수집하고 카이 제곱 통계량을 기준으로 남녀를 대표할 수 있는 단어를 추출한다.

어플리케이션 데이터에 기반한 분류기는 사용자가 실행한 어플리케이션들에 성별을 부여하고 높은 비율을 차지하는 성별로 사용자의 성별을 예측한다. 어플리케이션에 성별을 부여하기 위해 웹에 게시된 어플리케이션 설명글을 사용한다. 앞서 수집된 웹 문서로 문서 작성자의 성별을 분류하는 지지 벡터 기계를 학습하고 이를 통해 어플리케이션 설명글에 성별을 부여한다.

가속도 기반 분류기는 성별에 따른 사용자의 가속도 데이터 패턴을 학습한 지지 벡터 기계를 사용하여 주어진 사용자의 성별을 분류한다. 한 사용자에 대해서 단위 시간 동안 측정된 여러 개의 가속도 데이터를 남녀로 분류하고 이를 종합하여 사용자의 성별을 예측

한다.

자체 제작한 안드로이드 어플리케이션을 통해 수집된 실제 스마트 기기 로그 데이터를 사용하여 제안하는 기법을 평가하였다. 제안하는 방법론을 통해 0.95의 정확도를 얻을 수 있었다. 각 분류기 중 텍스트 기반 분류기와 어플리케이션 기반 분류기는 가속도 기반 분류기와 비교하여 좋은 성능을 나타내었다.

주요어 : 성별 예측, 스마트 기기 로그 데이터, 사생활 보호, 앙상블 기법, 통계학습, 기기 내 데이터 분석
학 번 : 2013 - 23204

목 차

표 목차

그림 목차

1. 서론

1.1. 연구 배경

스마트 기기의 사용이 대중화됨에 따라 스마트 기기를 이용한 개인화 서비스가 증가하고 있다. 마케팅 분야에서는 단순한 고객 관계 관리에서 나아가 모바일 커머스 시장에 적합한 일대일 개인화 마케팅이 화제로 떠오르고 있으며 한 대형 포털 사이트에서는 사용자에 따라 다른 검색어를 추천하는 추천 검색을 도입하였다. 웨어러블 스마트 디바이스의 발전에 힘입어 개인에게 최적화된 헬스케어 서비스도 확대될 전망이다.

개인화된 서비스를 제공하는 데에 있어서 사용자의 성별은 기초적이지만 중요한 정보가 된다. 마케팅 분야를 예로 들면, 남녀가 웹 상에서의 광고에 반응하는 성향이 매우 다르며 실제로 구매하는 구매하는 품목 또한 상이하다[1]. 성별 정보의 중요성에 기인하여 스마트 기기 사용자의 성별을 예측하는 연구가 활발하게 진행되어 왔다. 가속도 데이터와 설치된 어플리케이션의 정보를 통해서 사용자의 성별을 예측한 사례가 있었고[2, 3], 통화 기록과 위치 정보 등이 이용되기도 하였다[4].

한편 스마트 기기 로그 데이터는 매우 다양한 종류의 데이터를 포함하며 종류에 따라 형태와 특성이 매우 상이하다. 가속도, 자기장 등의 센서 데이터와 설치된 어플리케이션 정보, 통화 기록, 하드웨어 정보, 네트워크 정보 등이 모두 스마트 기기 로그 데이터에 해당한다. 이러한 데이터들은 모두 다른 특성을 가진다. 예를 들어 가속도 데이터는 센서에 의해 연속적으로 발생하며, 어느 한 순간의 가속도보다는 데이터가 모였을 때 사용자의 움직임에 대한 정보를

더 잘 나타낸다. 반면 기기에 설치되었거나 실행된 어플리케이션 목록은 사용자의 주요관심사를 반영하고 웹으로부터 얻을 수 있는 어플리케이션의 추가적인 정보와 결합될 수 있다. 또한 스마트 기기에서 발생하는 데이터에는 문자메시지, 검색 기록 등의 텍스트 데이터도 존재한다. 문자메시지는 성별에 따른 차이를 보이는 것으로 알려져 있으며[5], 인터넷 브라우저를 사용한 기록은 남녀의 서로 다른 관심사를 반영한다[6]. 더불어 텍스트 데이터는 사용자의 개인정보를 포함하기 때문에 기기 외부로 유출되면 사생활 침해 문제를 야기할 수 있다. 사용자들은 사생활 보호를 이유로 텍스트 데이터 공유를 꺼리는 경향을 보이기도 한다[7].

스마트 기기 사용자의 성별을 예측하는 데에 로그 데이터의 다양성을 반영하기 위한 방법으로 앙상블 기법이 적용될 수 있다. 앙상블 기법이란 여러 개의 분류기를 사용하여 분류한 다음 그 결과를 종합하여 최종적인 분류 결과를 얻는 방법이다[8]. 앙상블을 통해 얻어지는 분류 결과는 각각의 분류기보다 더 높은 성능을 보임이 수리적으로 증명되었다[9].

스마트 기기 로그 데이터 종류의 다양성은 분류기의 다양성을 요구하는 앙상블 기법에 부합한다. [8, 10]에 의하면 분류기들의 모델과 학습 방법, 입력 데이터가 다르면 분류기의 다양성이 보장되고, 각 분류기는 서로 연관성이 낮은 결과를 도출하게 된다. 이렇게 생성된 앙상블 분류기는 낮은 일반화 오류(generalization error)를 갖는다. 이를 스마트 기기 로그 데이터에도 적용하여 각 데이터의 특성에 부합하는 사용자의 성별 분류기를 여러 개 수립하고 앙상블하여 전체 예측 결과의 일반화 오류를 낮출 수 있다.

또한, 앙상블 기법을 적용하면 사용자의 사생활을 침해할 여지가 있는 텍스트 데이터를 사용할 수 있다. 앙상블 기법은 개인의 사

생활을 침해할 가능성이 있는 데이터를 이용해서 분류 문제를 해결하는 데에도 사용되어 왔다[11]. 분산된 저장소에 존재하는 데이터를 다른 저장소로 이동하는 것이 제한되면 분산된 데이터에 기반하여 분류 문제를 각각의 저장소에서 해결하고 결과값만을 중앙 시스템에 전송하여 앙상블을 통해 최종 결과를 얻는다. 스마트 기기 로그 데이터의 경우 스마트 기기 내에서 텍스트 데이터를 이용하여 성별을 예측하고 그 결과를 서버에서의 최종 앙상블에 사용함으로써 동일한 원리를 적용할 수 있다.

그러나 스마트 기기 내에서의 데이터 처리는 한정된 메모리 용량과 계산 자원이라는 제한조건을 갖는다. PC나 서버로 사용되는 컴퓨터에 비해 스마트 기기는 매우 낮은 계산 성능과 메모리 용량을 탑재하고 있는 것이 일반적이다. 기존 연구[12-15]에서도 스마트 기기 내에서 데이터를 처리하려는 시도가 존재했으며 한정된 자원 안에서 계산이 가능하도록 나이브 베이즈, 연관분석 등 계산 복잡도가 낮은 기법들이 사용되었다. 스마트 기기 내에서의 사용자 성별 예측 또한 한정된 자원 조건 하에서 계산이 가능한 방법론이 적용되어야 한다.

1.2. 연구 목적

본 연구에서는 스마트 기기에서 수집된 로그 데이터를 바탕으로 사용자의 성별을 예측하는 기법을 제안한다. 매우 다양한 데이터를 포함하는 스마트 기기 로그 데이터의 특성을 반영하기 위해 앙상블 기법을 도입한다. 서로 다른 데이터를 입력으로 가지는 분류기를 수립하고 그 결과를 결합하여 최종적인 성별을 예측하고자 한다.

제시하는 사용자 성별 예측 기법은 로그 데이터의 종류에 따라 데이터의 특성을 고려한 독립적인 분류기를 사용하여 데이터의 손실을 최소화할 수 있다. 더불어 앙상블에 사용되는 분류기의 다양성을 보장하여 예측 결과의 일반화 오류를 최소화한다. 또한 남녀의 차이가 존재한다고 알려져 있지만 사생활 침해 가능성으로 인해 적극적으로 활용되지 못했던 텍스트 데이터를 스마트 기기 내에서 처리함으로써 사생활 침해 문제를 최소화하고 성별 예측의 정확도를 높인다.

1.3. 연구 내용

본 연구에서는 최소한의 데이터 수집을 위하여 기존 연구에서 성별 예측에 효과적인 요인이라고 알려진 텍스트 데이터[5]와 어플리케이션 데이터[2], 가속도 데이터[3]를 사용한다. 입력 데이터에 따라 도출될 세 가지 분류기의 분류 결과를 앙상블하여 스마트 기기 사용자의 성별을 예측하는 기법을 제안한다.

텍스트 기반 분류기는 남녀를 구분할 수 있는 단어를 추출하는 단계와 추출한 단어를 이용하여 실제 스마트 기기 내에서 사용자의 성별을 예측하는 단계로 이루어진다. 남녀 단어 추출을 위해서 웹 문서에서 획득한 단어를 성별 구분력 척도인 카이 제곱 통계량(chi-square statistic)으로 분석하고 각 성별에 대해 구분력이 높은 단어들을 선택한다. 사용자 성별 예측 단계에서는 선택된 단어들과 스마트 기기 텍스트 데이터를 비교하여 사용자의 성별을 예측한다.

어플리케이션 기반 분류기는 웹에 게재된 어플리케이션 설명글을 활용하여 사용자의 성별을 예측한다. 작성자의 성별이 명확히 드러나 있는 웹 문서로 어플리케이션에 성별 레이블을 부여하는 지지 벡터 기계 모델을 학습한다. 사용자가 실행한 어플리케이션에 부여된 성별 레이블의 비율을 비교하여 사용자의 성별을 예측한다.

가속도 기반 분류기는 지지 벡터 기계를 학습하여 사용자의 성별을 예측한다. 지지 벡터 기계의 입력으로 사용되는 가속도 데이터의 요인으로는 [3]에서 사용한 가속도 데이터 요인에 실제 로그 데이터의 특성을 반영한 요인을 추가한다.

제 2장에서는 연구와 밀접한 관련이 있는 배경 이론과 기존 연

구를 정리하고 제 3장에서는 본 연구에서 제안하는 스마트 기기 사용자의 성별 예측 기법을 서술한다. 제 4장에서는 실험 설계와 실험 결과를 논의한다. 마지막으로 제 5장에서 결론을 도출한다.

2. 배경 이론 및 관련 연구

2.1. 배경 이론

2.1.1. 지지 벡터 기계

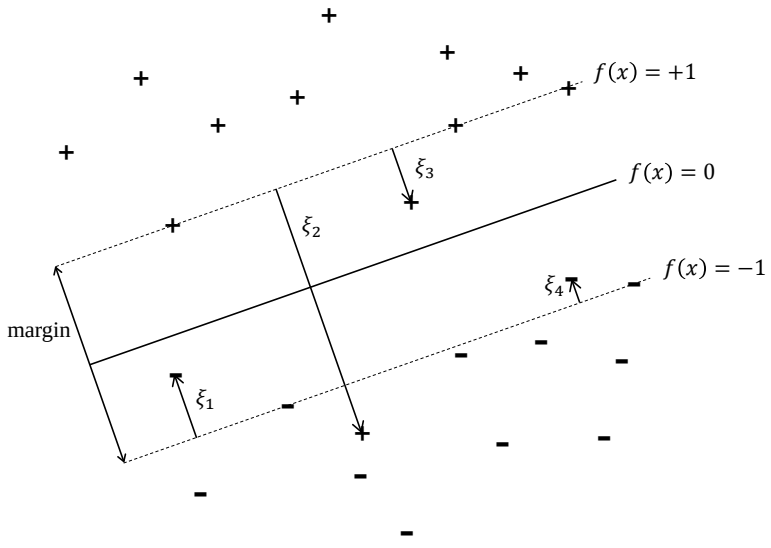
지지 벡터 기계는 V. Vapnik에 의해 처음 제안된 방법론으로 이진 분류 문제에서 두 클래스 간의 거리를 최대화 하는 초평면을 찾아 이를 기준으로 새로운 데이터를 분류한다[16]. 특히 약 마진(soft margin) 지지 벡터 기계는 오류를 허용하여 선형 분리가 불가능한 학습 데이터에 대해서도 학습이 가능하다.

이진 분류 문제에서 n개의 점으로 이루어진 학습데이터 집합 $\mathcal{D}$는 수식 (1)로 정의할 수 있다. $\mathbb{R}$은 실수의 집합, i번째 학습데이터 (x_i, y_i)의 x_i는 p차원의 실수 벡터, y_i는 x_i가 속한 클래스이다.

$$\mathcal{D} = \left\{ (x_i, y_i) \,\middle|\, x_i \in \mathbb{R}^p, y_i \in \{-1,1\} \right\}_{i=1}^{n} \tag{1}$$

$\beta \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$에 대해 함수 $f(x)$를 수식 (2)와 같이 정의하면 $p = 2$일 때 초평면 $\{x|f(x) = +1\}$와 $\{x|f(x) = 0\}$, $\{x|f(x) = -1\}$을 [그림 1]로 나타낼 수 있다.

$$f(x) = \beta^T x + \beta_0 \tag{2}$$



[그림 1] 이차원 이진 분류 문제

이때 약 마진 지지 벡터 기계는 학습데이터의 오분류를 일부 허용하되, 수식 (3)으로 표현되는 마진(margin)을 최대화하는 β와 β_0를 구한다.

$$\text{margin} = \frac{2}{\|\beta\|} \tag{3}$$

최적화 문제를 해결하여 얻어지는 (β^*, β_0^*)으로 추정된 분류 함수 수식 (4)를 사용하면 임의의 데이터 포인트에 대해 수식 (5)로 클래스를 예측할 수 있다.

$$\hat{f}(x) = \beta^{*T} x + \beta_0^* \tag{4}$$

$$\hat{y} = \text{sign}(\hat{f}(x)) \tag{5}$$

이 최적화 문제의 제약 조건은 수식 (6)과 (7), (8)로 표현된다. 여유변수(slack variable) ξ_i $(i = 1, .., n)$는 오분류를 허용하기 위해

도입한다. ξ_i는 0 또는 양의 실수 값을 가지며 학습 데이터 포인트 x_i가 오분류 되었거나 [그림 1]에서와 같이 초평면 $\{x|f(x)=+1\}$와 $\{x|f(x)=-1\}$ 사이에 점이 존재할 경우에만 양의 실수 값을 가진다. 단, 오분류의 수준을 제한하기 위해 수식 (7)에 양의 실수인 상수 C를 사용한다.

$$y_i f(x_i) \geq 1-\xi_i, \text{for } i=1,\dots,n \tag{6}$$

$$\sum_{i=1}^{n}\xi_i \leq C \tag{7}$$

$$\xi_i \geq 0, \text{for } i=1,\dots,n \tag{8}$$

계산의 편리성을 위해 마진 최대화 문제를 이차 계획법 최소화 문제로 변경하고 제약식 수식 (7)을 목적함수에 반영한다. 최종적으로 목적함수 수식 (9)와 제약조건 수식 (6), (8)을 갖는 최적화 문제의 쌍대문제를 라그랑지안 승수법으로 도출하여 최적해를 구한다. 학습데이터를 사용하여 (β^*,β_0^*)를 구하는 것을 지지 벡터 기계의 학습이라 한다. 자세한 설명은 [17]에서 찾을 수 있다.

$$\max_{\beta,\beta_0,\xi}\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \tag{9}$$

2.1.2. 카이 제곱 통계량

카이 제곱 통계량은 일반적으로 카이 제곱 분포에 기반하여 변수간의 독립성 검정에 사용된다. 텍스트 마이닝을 이용한 문서 분류에서는 단어로 표현되는 요인의 차원을 축소하기 위해서 사용되어왔다[18−21]. 문서의 이진 분류 문제에서 임의의 단어 w에 대한 문

서 출현 빈도수 분할표를 나타내면 [표 1]과 같다. 임의의 클래스 g에 대해 $\bar{g}$는 g가 아닌 클래스를 의미한다.

[표 1] 단어 w의 문서 빈도수 분할표

		문서의 클래스	
		g	$\bar{g}$
단어 w	존재	a	b
	부재	c	d

$N = a + b + c + d$ 으로 정의하면 하나의 단어 w와 클래스 g에 대해서 수식 (10)으로 카이 제곱 통계량 $\chi^2(w, g)$를 계산할 수 있다 [18].

$$\chi^2(w,g) = \frac{N \times (ad - bc)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)} \qquad (10)$$

$\chi^2(w, g)$의 값이 작을수록 단어 w가 클래스 g와 독립적이라고 말할 수 있으며 반대로 값이 클수록 단어 w는 클래스 g에 종속적이라고 할 수 있다. 또한 $\chi^2(w, g)$의 값이 크면 단어 w는 클래스 g와 $\bar{g}$에 속하는 문서를 구분할 수 있는 구분력을 가지는 것으로 해석된다. 그러나 단어의 전체 문서 출현 빈도수 $(a + b)$가 매우 작으면 카이 제곱 분포를 가정할 수 없으므로 계산되는 통계량의 의미를 신뢰할 수 없다.

본 연구에서는 위의 식을 다음의 수식 (11)과 (12)로 수정하여 임의의 단어 w에 대해 남성 카이 제곱 통계량과 여성 카이 제곱 통계량을 각각 계산한다. 남성 카이 제곱 통계량이 클수록 남성의 문서와 연관성이 높은 단어임을 의미하고, 여성 카이 제곱 통계량이

클수록 여성의 문서와 연관성이 높은 단어임을 의미한다.

$$\chi_m^2(w) = \begin{cases} \frac{N\times(ad-bc)^2}{(a+c)\times(b+d)\times(a+b)\times(c+d)}, & \text{if } (ad-bc) > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

$$\chi_f^2(w) = \begin{cases} \frac{N\times(bc-ad)^2}{(b+d)\times(a+c)\times(c+d)\times(a+b)}, & \text{if } (bc-ad) > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

2.1.3. 코사인 유사도

코사인 유사도는 임의의 두 벡터 $u, v \in \mathbb{R}^p$의 유사도를 벡터 사이의 각도를 이용하여 정의하는 방법이다. 수식 (13)으로 얻어지는 코사인 유사도를 계산한다. 벡터 사이의 각도가 클수록 작은 코사인 값을 가지므로 코사인 유사도가 클수록 사이각이 작음을 의미한다.

$$\text{sim}(u, v) = \frac{u^T v}{\|u\|\|v\|} \quad (13)$$

2.2. 관련 연구

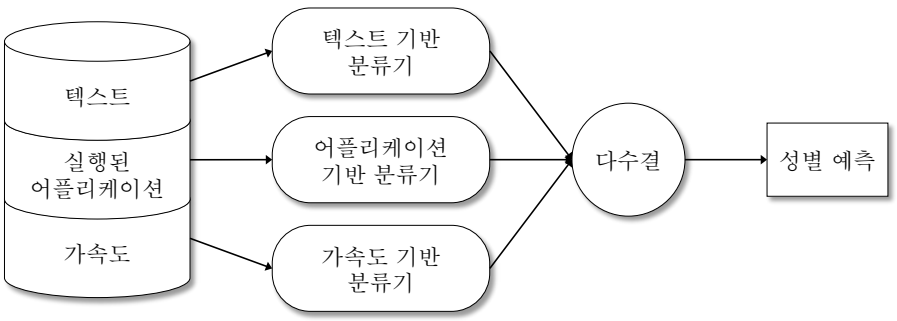
데이터 학습을 이용한 성별 예측에 관한 연구는 스마트 기기가 대중화되기 이전부터 진행되어 왔다. 웹 서비스 사용자의 성별 예측을 위해 웹페이지 방문 기록이나 소셜 네트워크 프로필의 색깔을 이용하는 방법들이 연구되었다[22, 23]. 특히 웹에서 발생하는 텍스트 데이터를 이용 하는 연구가 활발히 이루어졌다. [24, 25]는 이메일로부터, [26]은 채팅 기록으로부터 사용자의 성별을 예측하였다. 트위터를 비롯한 소셜 네트워크 사용자의 성별을 예측하고자 하는 시도도 있어왔다. [27-30]은 길이가 짧고 약어와 비표준어가 빈번하게 등장하는 소셜 네트워크 글의 특성을 반영하도록 n-gram 기법이나 글쓰기 스타일에 기반한 요인들을 사용하였으며 단어를 요인으로 설정하는 경우에는 카이 제곱 통계량과 같은 측정치를 이용하여 요인을 축소하였다. 트위터에서 사용자가 작성한 글이 아닌 프로필에 게시된 이름을 n-gram으로 요인화하여 사용자의 성별을 예측하는 방법론도 제안되었다[31]. [21]은 블로그 글로 작성자의 성별을 분류하고자 했다. 일반적인 단어 요인 이외에도 단어들의 의미적 상위 집합과 형태소의 상위 집합을 요인으로 사용하여 좋은 분류 성능을 보였다.

스마트 기기가 보편적으로 사용되면서 그로부터 얻어지는 데이터를 이용한 연구들이 활발히 이루어져왔다. 사생활 침해 여지가 있는 텍스트 데이터를 사용한 연구보다 텍스트 이외의 데이터를 이용한 연구가 주를 이뤘다. [3]은 기기에 장착된 가속도계를 이용하여 주어진 실험 환경에서 사용자의 가속도를 측정하고 이를 바탕으로 남녀를 구분하는 방법론을 제안하였다. 스마트 기기에 설치된 어플리케이션 목록과 웹에서 얻을 수 있는 관련 정보를 이용하여 사용자

의 성별을 예측하는 실험도 수행되었다[2]. 노키아에서 주관한 Mobile Data Challenge[4]에서는 다수의 스마트 기기에서 발생한 방대한 데이터를 참가자들에게 공개하고 사용자의 성별과 연령대, 직업, 가족의 수, 결혼 상태 등의 인구통계학적 정보를 예측하는 문제가 제시되었다. 많은 연구들이 해당 문제를 해결하기 위하여 다양한 통계학습 방법론을 제시하였다[32-35].

3. 제안 기법

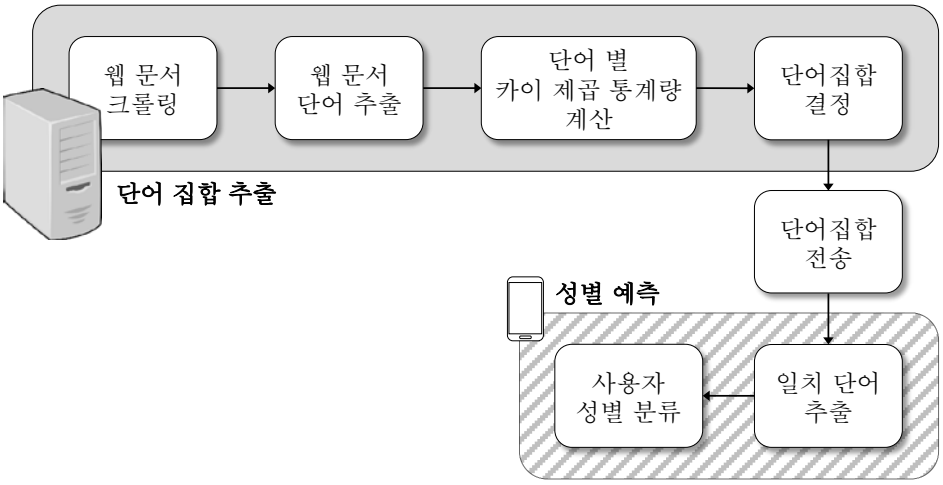
본 연구에서는 스마트 기기 사용자의 성별을 예측하기 위하여 서로 다른 데이터를 입력으로 가지는 분류기를 설정하고 각 분류기의 결과를 토대로 다수결(majority voting)에 의해 최종적으로 성별을 예측한다. [그림 2]에 나타난 바와 같이 텍스트 데이터, 어플리케이션 데이터, 가속도 데이터를 각각의 입력으로 갖는 분류기들을 이용한다.



[그림 2] 제안 기법의 사용자 성별 분류 과정

3.1. 텍스트 기반 분류기

텍스트 기반 분류기는 [그림 3]과 같이 단어집합 추출 단계와 성별 예측 단계로 나뉜다. 단어집합 추출의 모든 과정은 서버에서 수행된다. 웹 문서의 크롤링과 단어 추출, 카이 제곱 통계량 계산은 큰 저장소 용량을 요구하고 높은 계산 복잡도를 보이기 때문에 서버에서 수행하는 것이 적합하다. 서버에서 결정된 단어집합은 스마트 기기로 전송되어 기기 내에서 사용자의 성별을 예측한다. 사생활 침해 문제를 최소화하기 위해 텍스트 데이터는 사용자의 기기 내에서만 처리되며 스마트 기기의 한정된 저장 공간과 계산 능력을 고려하여 계산 복잡도가 낮은 방법을 적용한다.



[그림 3] 텍스트 데이터를 이용한 사용자 성별 분류 과정

.

3.1.1. 단어집합 추출

단어집합 추출은 웹 문서 크롤링에서 시작된다. 수집되는 웹 문서는 작성자의 성별이 명시적으로 드러나 있는 것으로 한정한다. 사

용될 수 있는 웹 문서의 예로는 블로그나 트위터, 페이스북 등의 소셜 네트워크에 게시된 글이 있다. 수집한 웹 문서로부터 스태밍과 불용어처리 과정 등을 거쳐 단어를 추출한다[36].

남성과 여성 단어집합을 결정하기 위해 단어의 카이제곱 통계량을 계산하고 주어진 단어집합 크기 N_r에 따라 단어를 선택한다. 남성과 여성의 글을 구분할 수 있는 단어들을 추출하기 위해 웹 문서에서 추출한 단어의 남성 카이 제곱 통계량 χ_m^2과 여성 카이 제곱 통계량 χ_f^2를 수식 (11)과 수식 (12)로 계산한다. 그러나 모든 문서에 대해 단어의 출현 빈도가 매우 낮으면 카이 제곱 분포를 가정할 수 없으므로 카이 제곱 통계량 계산에서 제외한다.

단어집합을 결정하기 위해 남성과 여성에 대한 카이 제곱 통계량에 대해 각각 내림차순으로 단어를 정렬한다. 카이 제곱 통계량의 크기가 큰 단어부터 단어집합으로 선택되고 단어집합의 크기가 N_r에 도달하면 단어집합 선택이 종료된다. 남녀 단어집합에 대해 동일한 과정이 이루어져서 최종적으로는 성별 $g \in \{$남성,여성$\}$에 대해 단어집합 R_g가 생성된다.

3.1.2. 텍스트 기반 성별 예측

서버에서 스마트 기기로 단어집합이 전송되면 사용자의 성별을 예측하기 위해 스마트 기기 텍스트 데이터로부터 일치 단어를 추출한다. 일치 단어 추출 과정을 통해서 성별 $g \in \{$남성,여성$\}$에 대한 일치 벡터가 생성된다. 먼저, 남녀 단어집합과의 일치 단어를 추출하기 위해 스마트 기기 내에서 텍스트 데이터를 수집하고 하나의 문자열로 연결한다. 계산 복잡도를 줄이기 위해서 텍스트 데이터에 대해 웹 문서와 같이 단어 추출 방법을 사용하지 않고, 확인하고자 하

는 단어의 길이를 갖는 연속적인 문자 단위로 문자열을 조사한다. 단어 r_g가 $r_g \in R_g$라 할 때 텍스트 데이터에 r_g가 존재하는지 여부를 나타내는 이진 일치 함수 $\mathrm{M}(r_g)$는 수식 (14)와 같이 표현된다.

$$\mathrm{M}(r_g) = \begin{cases} 1, & \text{if } r_g \text{ exists in the text data} \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

길이가 N_r인 이진 일치 벡터 v_g는 이진 일치 함수를 사용하여 다음의 수식 (15)로 정의한다.

$$v_g = < \mathrm{M}(r_g) > \quad (15)$$

한편 $\mathrm{M}'(r_g)$는 빈도 일치 함수로, 텍스트 데이터의 문자열에서 단어 r_g가 출현한 빈도수를 함수값으로 정의한다. 이를 사용하여 길이가 N_r인 빈도 일치 벡터 v'_g를 수식 (16)으로 생성한다. 빈도 일치 벡터 생성은 단 한번의 일치만 발생하면 해당 단어의 일치 여부 확인이 종료되는 이진 일치 벡터 생성과는 달리 문자열을 끝까지 확인하므로 생성 시간이 더 길다.

$$v'_g = < \mathrm{M}'(r_g) > \quad (16)$$

사용자의 성별을 예측하기 위해 각 성별에 대한 일치 벡터와 단어집합 R_g의 단어들을 벡터화한 기준 벡터의 코사인 유사도를 비교한다. 기준벡터는 벡터의 요소 값에 따라 두 가지를 생성할 수 있다. 첫째는 모든 단어에 대해 동일한 값을 부여하는 이진 기준 벡터 u_g이고 둘째는 단어의 카이 제곱 통계량을 값으로 갖는 점수 기준 벡터 u'_g이다. u_g는 길이가 N_r이고 모든 벡터 요소의 값을 1로 갖는다. 단어 r_g의 카이 제곱 통계량을 e_g라 하면 u'_g는 수식 (17)로 나타낼

수 있다.

$$u'_g = < e_g > \quad (17)$$

점수 기준 벡터 u'_g를 사용하기 위해서는 서버로부터 단어집합이 전송될 때 카이 제곱 통계량도 동시에 전송되어야 한다. 유사도 계산에 통계량을 반영하면 남녀를 대표하는 단어의 구분력에 대한 가중치로 작용한다.

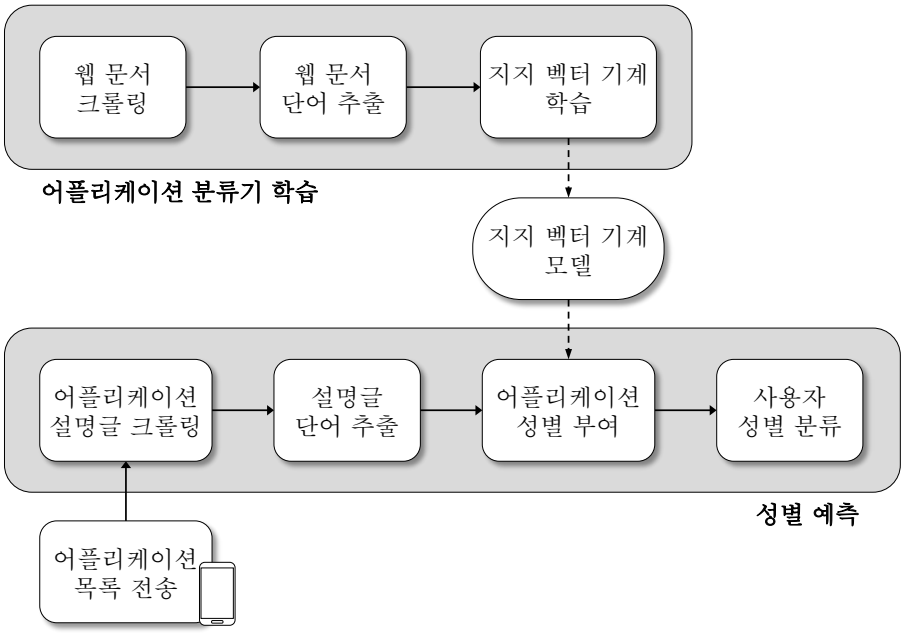
기준 벡터와 일치 벡터의 조합에 따라 네 가지 코사인 유사도 값을 계산할 수 있다. 이를 기준 벡터-일치 벡터로 형식으로 나타내면 이진-이진, 점수-이진, 이진-빈도, 점수-빈도로 표현된다.

$$\hat{g} = \arg\max_g \mathrm{sim}(u_g, v_g), g \in \{\text{남성}, \text{여성}\} \quad (18)$$

이진-이진 조합을 사용하여 코사인 유사도를 계산할 경우 수식 (18)과 같이 사용자의 성별을 예측한다. u_g를 u'_g로, v_g를 v'_g로 치환하여 다른 코사인 유사도 계산법에도 적용할 수 있다. 두 코사인 유사도의 값이 동일하면 해당 사용자에 대해 성별을 분류하지 않는다. 따라서 텍스트 기반 분류기를 거치면 사용자의 성별이 남성 혹은 여성으로 분류되거나 미분류 상태로 남을 수 있다.

3.2. 어플리케이션 기반 분류기

어플리케이션 기반 분류기는 스마트 기기에서 실제로 실행된 어플리케이션의 정보를 바탕으로 사용자의 성별을 분류한다. 어플리케이션 개발자가 게시한 어플리케이션 설명글은 사용자 성별을 예측하는데 도움을 주는 것으로 알려져 있다[2]. 본 연구에서는 설명글을 토대로 어플리케이션에 남성 혹은 여성의 레이블을 부여하고, 사용자가 실행한 어플리케이션들의 성별 레이블을 비교하여 사용자의 성별을 분류한다. 어플리케이션 데이터를 이용한 사용자의 성별 분류 과정을 [그림 4]로 도식화하였다. 스마트 기기로부터 사용 어플리케이션 목록 전송을 제외한 모든 과정이 서버에서 수행된다.



[그림 4] 어플리케이션 데이터를 이용한 사용자 성별 분류 과정

3.2.1. 어플리케이션 분류기 학습

어플리케이션에 남녀 레이블을 부여하기 위한 분류기를 학습하기 위해서 텍스트 기반 분류기와 같이 웹에서 작성자의 성별이 명시된 문서를 크롤링한다. 크롤링한 문서들로부터 단어를 추출하고 모든 문서에서 발생한 단어들로 단어 주머니(bag of words)를 생성한다. 이 집합에 속하는 단어들이 하나의 문서에서 발생하는 빈도수를 요인으로 하는 지지 벡터 기계를 학습한다. 웹 문서에서 발생한 서로 다른 단어의 총 수가 N_w라 하면, 학습 데이터는 $p = N_w$인 수식 (1)이 된다. 학습데이터 포인트 x_i는 i번째 웹 문서에서 각각의 단어가 발생한 빈도수를 요소로 갖는 벡터이며 문서 작성자가 여성이면 $y_i = +1$, 남성이면 $y_i = -1$ 값을 갖는다.

3.2.2. 실행 어플리케이션 기반 성별 예측

사용자의 스마트 기기에서 실행된 어플리케이션의 목록을 서버로 전송하여 성별 예측 단계를 시작한다. 어플리케이션에 대한 구분자로 패키지명을 사용이며, 이는 어플리케이션에 고유하게 부여된 영문 이름이다. 웹에서 어플리케이션 게시자가 작성한 어플리케이션 설명글을 크롤링 하고 단어와 단어의 발생 빈도수를 추출한다.

하나의 설명글을 하나의 문서로 가정하고 학습데이터와 동일한 요인을 갖는 벡터로 표현하여 학습한 지지 벡터 기계의 분류 함수 수식 (4)를 이용하여 설명글에 성별 레이블 $\hat{y}$을 부여한다. 여성 레이블은 $\hat{y} = +1$로, 남성 레이블은 $\hat{y} = -1$로 나타낸다.

이와 같은 방법으로 스마트 기기에서 실행된 모든 어플리케이션에 대해 성별을 부여한다. 사용자가 사용한 어플리케이션이 n_{app}개

라 하면 어플리케이션 $app_j, j = 1, \dots, n_{app}$에 대해 성별 클래스 $\hat{y}_j$가 부여된다.

최종적인 사용자의 성별은 부여된 어플리케이션 성별을 다수결하여 예측한다. 어플리케이션에 부여된 성비가 동일하면 미분류 처리한다. 다수결 함수 $g(\hat{y}_1, \dots, \hat{y}_{n_{test}})$를 수식 (19)로 정의하고 $n_{test} = n_{app}$이라 하면 수식 (20)으로 예측된 성별 $\hat{g}$을 나타낼 수 있다.

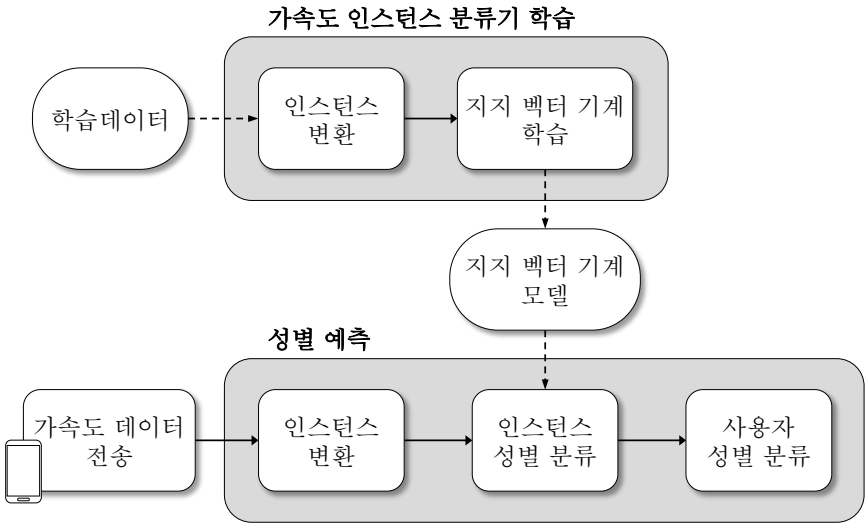
$$g(\hat{y}_1, \dots, \hat{y}_{n_{test}}) = \sum_{j=1}^{n_{test}} \hat{y}_j \tag{19}$$

$$\hat{g} = \begin{cases} \text{남성}, & \text{if } g(\hat{y}_1, \dots, \hat{y}_{n_{test}}) < 0 \\ \text{미분류}, & \text{if } g(\hat{y}_1, \dots, \hat{y}_{n_{test}}) = 0 \\ \text{여성}, & \text{if } g(\hat{y}_1, \dots, \hat{y}_{n_{test}}) > 0 \end{cases} \tag{20}$$

예를 들어 사용자가 사용한 5개의 어플리케이션 중 1개의 어플리케이션이 남성으로, 4개의 어플리케이션이 여성으로 분류되었다면 이 사용자는 여성으로 예측한다.

3.3. 가속도 기반 분류기

가속도 기반 분류기는 Weiss와 Lockhart의 연구[3]를 기반으로 설계되었다. [3]은 실험실 환경에서 얻어진 스마트폰 가속도 데이터를 이용하여 사용자의 성별을 예측하였다. 정해진 크기의 짧은 타임 윈도우(Example Duration, ED)에서 연속적으로 발생한 가속도 데이터를 사전에 정의된 요인을 갖는 하나의 데이터 인스턴스로 변환하여 분류기의 입력으로 사용하였다. 한 명의 사용자에게는 여러 인스턴스가 발생하고, 이 인스턴스들을 각각 남녀로 분류한다. 최종적으로는 인스턴스들의 성비를 계산하여 사용자의 성별을 예측한다. 이러한 예측 기법을 본 연구에 도입하면 여러 시간대에서 발생하는 사용자의 움직임을 반영하여 성별을 예측할 수 있다. [그림 5]로 가속도 데이터를 이용하여 사용자의 성별을 분류하는 과정을 나타내었다.



[그림 5] 가속도 데이터를 이용한 사용자 성별 분류 과정

3.3.1. 가속도 인스턴스 분류기 학습

스마트 기기에서 측정된 가속도 데이터는 서버로 전송되어 지정된 요인을 갖는 인스턴스로 변환된다. 기존의 연구에서 모든 데이터 측정 시 반복적인 동작을 취한 것과는 달리 본 연구에서는 실제 생활 속에서 스마트 기기를 사용할 때 발생하는 로그 데이터를 사용한다. 그러므로 본 연구에서는 기존에 사용된 요인들 중 반복 동작과 관련된 요인을 제외하고, 일주일 중 주중인지 여부와 시간대, 데이터 수집 플랫폼의 특성을 반영하는 요인을 추가한다. 데이터 수집 플랫폼의 특성을 반영한 요인은 데이터를 수집하기 위해 사용하는 안드로이드 운영체제에서 제공하는 API 중 센서 매니저 클래스의 특성을 반영한다. 안드로이드에서는 센서데이터 수집 시 데이터 수집 빈도를 네 가지 수준으로 선택할 수 있다. 정해진 수준은 절대적인 수집 빈도수를 나타내는 것이 아니기 때문에 실제로 일정한 시간 동안 수집되는 데이터의 수는 매번 상이하다. 수집되는 데이터의 수는 가속도의 변화와 비례하는 것으로 알려져 있다. 이외에도 가속도의 크기와 평균, 분산, 최소와 최대값 등을 요인으로 사용하였다. [표 2]에서 지지 벡터 기계의 입력으로 사용되는 16개의 요인을 정리하였다.

[표 2] 가속도 기반 분류기 지지 벡터 기계의 요인

요인 이름	종류	설명
weekday	진리값	주중 여부
hour	자연수	24시간 시간대
cnt	자연수	ED 동안 수집된 데이터의 수
norm	실수	ED 동안 수집된 가속도의 X, Y, Z 축 평균 값의 L2−norm
{X,Y,Z}Min	실수	ED 동안 수집된 가속도의 {X,Y,Z}축 최솟값
{X,Y,Z}Max	실수	ED 동안 수집된 가속도의 {X,Y,Z}축 최댓값
{X,Y,Z}Avg	실수	ED 동안 수집된 가속도의 {X,Y,Z}축 평균
{X,Y,Z}Var	실수	ED 동안 수집된 가속도의 {X,Y,Z}축 분산

분류를 위한 통계적 학습 기법으로는 지지 벡터 기계를 사용한다. 지지 벡터 기계의 학습데이터는 $p = 16$인 수식 (1)로 나타낼 수 있다. i번째 학습데이터 인스턴스 x_i는 [표 2]의 요인을 갖는 벡터이다. 데이터가 발생한 사용자가 여성이면 $y_i = +1$, 남성이면 $y_i = -1$ 값을 갖는다.

3.3.2. 가속도 기반 성별 예측

새로운 사용자의 가속도 데이터가 서버로 전송되면 성별 예측을 하기 위해 학습데이터와 동일한 요인을 가지는 인스턴스로 변환된다. 각각의 인스턴스는 학습된 지지 벡터 기계 모델에 의해 남성 혹은 여성으로 분류된다. 사용자의 모든 인스턴스가 분류되면 결과를 다수결하여 최종적으로 사용자의 성별을 예측한다.

한 명의 사용자의 가속도 데이터 인스턴스가 n_{acc}개라 하면 인스턴스 $acc_j, j = 1, \dots, n_{acc}$에 대해 성별이 분류된다. 이 결과를 다수결하여 사용자의 성별을 최종적으로 예측한다. 인스턴스의 성비가 동일하면 미분류 처리한다. 수식 (19)에서 $n_{test} = n_{acc}$라 하면 수식

(20)으로 예측된 성별 $\hat{g}$을 나타낼 수 있다. 단, 예측 대상이 되는 사용자의 가속도 데이터 인스턴스가 매우 적으면 다수결 시 편향이 발생할 수 있다. 따라서 일정 개수 이하의 인스턴스를 가진 사용자는 가속도 데이터로 성별을 예측하지 않는다.

3.4. 다수결에 근거한 앙상블

본 연구에서 제안하는 기법은 텍스트 기반 분류기와 어플리케이션 기반 분류기, 가속도 기반 분류기로부터 얻어진 결과를 앙상블하여 사용자의 성별을 예측한다. 분류기들로부터 얻어진 결과로 다수결에 의해 예측하되 미분류된 결과는 다수결에 포함하지 않는다. 따라서 동률이 발생할 수 있고, 이 경우에는 결과의 신뢰도가 가장 높은 분류기의 예측 성별을 최종 예측 성별로 선택한다.

각 분류기의 신뢰도는 0과 1사이의 값으로 정의한다. 텍스트 기반 분류기는 가장 높은 코사인 유사도 값을 나타낸 성별의 기준 벡터와 일치 벡터 사이의 각도를 정규화하여 각도 신뢰도로 정의하고 이를 신뢰도로 사용한다. 예측된 성별 $\hat{g}$의 기준 벡터를 $u_{\hat{g}}$, 일치 벡터를 $v_{\hat{g}}$라 하면, 각도 신뢰도는 수식 (21)과 같다.

$$\text{각도 신뢰도} = 1 - \frac{2}{\pi}\cos^{-1}\left(\text{sim}\left(u_{\hat{g}}, v_{\hat{g}}\right)\right) \qquad (21)$$

다수결을 적용한 어플리케이션 기반 분류기와 가속도 기반 분류기의 경우 다수를 차지한 성의 비율이 소수인 성의 비율보다 클수록 신뢰할 수 있는 결과라 할 수 있다. 따라서 두 분류기에 대해서는 두 성별의 비율 차이를 신뢰도로 정의하고 이를 차이 신뢰도라 한다. 수식 (22)에서 수식 (19)를 사용하여 차이 신뢰도를 수식으로 나타내었다. 어플리케이션 기반 분류기는 $n_{test} = n_{app}$, 가속도 기반 분류기는 $n_{test} = n_{acc}$를 적용한다.

$$\text{차이 신뢰도} = \left|\frac{g\left(\hat{y}_1, \ldots, \hat{y}_{n_{test}}\right)}{n_{test}}\right| \qquad (22)$$

4. 실험 및 결과

4.1. 실험 데이터

4.1.1. 스마트 기기 로그 데이터

본 연구에서 제안한 스마트 기기 사용자의 성별 예측 기법의 성능을 평가하기 위해 안드로이드 어플리케이션 imlabOHP를 개발하여 데이터를 수집하였다. 안드로이드는 구글에서 개발한 스마트 디바이스용 운영체제로 안드로이드 환경에서 구동되는 어플리케이션은 개발과 배포가 자유로운 것이 특징이다. [그림 6]은 imlabOHP의 첫 화면이다.

[그림 6] imlabOHP의 첫 화면

어플리케이션을 통해 텍스트 데이터인 주소록에 등록된 이름, 문자메시지, 브라우저 북마크, 검색 기록과 사용자가 실제로 실행한 어플리케이션, 기기에 설치된 어플리케이션, 주기적인 가속도 데이터를 수집하였다. 사용자의 실제 성별은 사용자가 직접 입력하였다. 텍스트 데이터의 경우 사용자의 신원을 확인할 수 있는 단어가 포함될 가능성이 있으므로 이름 등 본인식별이 가능한 문자열에 대해서는 수집 이전에 삭제되도록 하였다. 사용자가 실제로 실행한 어플리케이션을 조사하기 위해서 [37]의 연구 결과를 바탕으로 시스템의 가장 상위에서 구동되고 있는 프로그램을 1분 간격으로 수집하였다. 가속도 데이터는 15분마다 30초씩 수집되어 30초간의 데이터가 하나의 인스턴스를 이루도록 하였다.

20명의 스마트 기기 사용자로부터 데이터를 수집하였으며 실험에 참여한 사용자는 11명의 남성과 9명의 여성으로 구성되었다. [표 3]에 수집한 데이터의 특성을 요약하였다.

[표 3] 수집된 스마트 기기 로그 데이터 요약

성별	피실험자 수 (명)	평균 텍스트 데이터 길이 (글자)	설치된 어플리케이션 평균 개수	가속도 데이터 인스턴스 평균 개수
남성	11	45,633	325	251
여성	9	140,668	359	472
전체	20	88,399	340	347

평균 텍스트 데이터의 길이는 여성 사용자가 남성 사용자의 약 3배에 달하는 것으로 나타났다. 설치된 어플리케이션의 수는 여성이 다소 많았으며 가속도 데이터 인스턴스도 여성 사용자로부터 더 많이 수집되었다. 가속도 데이터 인스턴스의 경우 일부 여성 사용자의

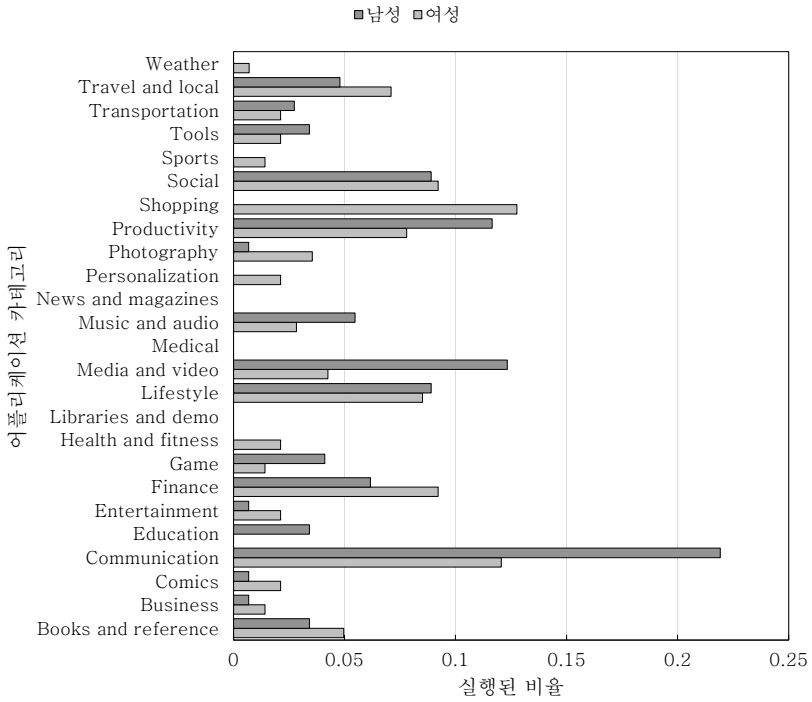
데이터 수집 기간이 다른 사용자와 비교하여 상대적으로 길었던 것으로 나타났다.

아래의 [표 4]는 수집된 텍스트 데이터에서 추출한 단어들 중 출현 빈도가 가장 높은 10개의 단어를 남녀에 대해 나타내었다. 남성의 텍스트 데이터에서 발생한 단어들로는 사용자의 성별을 직관적으로 파악할 수 없었지만 여성의 경우 사용자의 성별을 예상할 수 있는 “오빠”, “언니” 등의 단어가 빈번히 발생하였다.

[표 4] 수집된 텍스트 데이터의 출현 빈도수 상위 10개의 단어

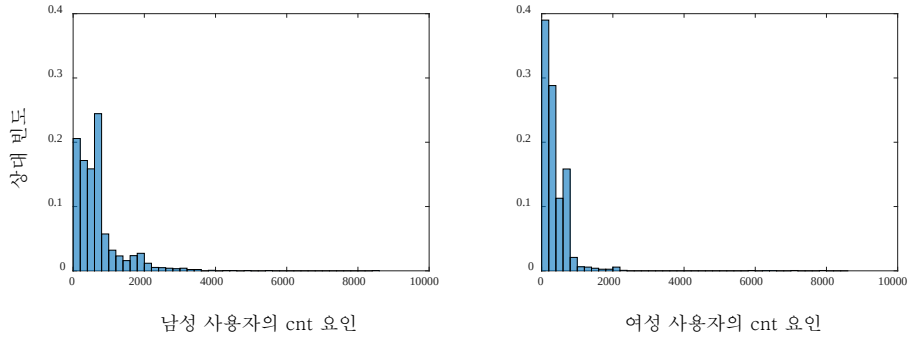
성별	단어
남성	네이버 웹툰 확인 생활 데이터 택시 이용 통화 기본 완료
여성	네이트 뉴스 오빠 오늘 통화 언니 쟈기야 내일 지금 감사

사용자가 사용한 어플리케이션도 성별에 따라 차이를 나타냈다. [그림 7]은 수집된 데이터 중 실행된 어플리케이션의 카테고리 분포를 보여준다. 그래프에서 x축은 데이터 수집 기간 동안 사용한 모든 어플리케이션 중에서 각 카테고리의 어플리케이션이 차지하는 비율을, y축은 구글에서 지정한 어플리케이션 카테고리의 명칭을 나타낸다. 남성 사용자의 데이터를 살펴보면 Productivity와 Media and video, Game, Communication 등의 카테고리 사용비율이 여성보다 높은 것으로 나타났다. 반면 여성 사용자는 남성 사용자는 실행하지 않은 Shopping 카테고리에 속하는 어플리케이션을 높은 비율로 이용하였다.



[그림 7] 실행된 어플리케이션의 카테고리 분포

가속도 데이터 수집 결과 cnt 요인의 분포가 성별에 따른 차이를 나타내었다. [그림 8]은 cnt 요인 분포를 보여준다. [그림 8]에서 알 수 있듯이 여성 사용자의 데이터에서 낮은 cnt값을 가지는 인스턴스가 남성에 비해 높은 빈도로 관측되었다.



[그림 8] 가속도 데이터의 cnt 요인 히스토그램

4.1.2. 웹 문서

텍스트 기반 분류기와 어플리케이션 기반 분류기는 작성자의 성별이 알려져 있는 웹 문서를 필요로 한다. 본 실험에서는 이를 위해 성별에 따른 서로 다른 관심사와 사용 단어를 포착할 수 있는 블로그 문서를 크롤링하였다. 다양한 주제를 갖는 블로그들 중 블로그 작성자인 블로거의 성별이 명시되어 있는 블로그에 한하여 문서를 수집하였다. 수집한 블로그 데이터를 [표 5]로 정리하였다.

[표 5] 수집된 웹 문서 요약

	블로그 수	문서의 수	추출된 단어의 수	평균 문서의 길이 (글자)
남성	97	135,745	137,743	1,327
여성	65	53,382	117,941	1,520
전체	162	88,399	141,509	1,381

[표 5]의 전체 웹 문서에 대한 추출된 단어의 수는 어플리케이션 기반 분류기의 N_w 에 해당한다. 텍스트 기반 분류기에서는 141,509개의 단어 중 전체 문서의 출현 빈도수가 5개 미만인 단어는 카이 제곱 통계량 계산에서 제외되었다.

4.2. 실험 환경 및 평가 지표

4.2.1. 실험 환경

본 실험은 Java 프로그래밍 언어를 사용하여 수행되었다. 블로그 문서와 어플리케이션 설명글을 크롤링하기 위하여 jsoup 라이브러리를 사용하였으며, 크롤링한 문서로부터 단어를 추출하기 위해 루신 아리랑 분석기와 꼬꼬마 형태소 분석기를 이용했다. 어플리케이션 기반 분류기의 지지 벡터 기계는 학습데이터의 희소 행렬 특성을 고려하여 SVM Light 라이브러리를 사용하였고 가속도 기반 분류기의 지지 벡터 기계는 LibSVM 라이브러리로 구현되었다. 모든 지지 벡터 기계 학습을 위해 데이터를 정규화하였고, 비복원 랜덤 샘플링하여 동일한 수의 남녀 학습데이터를 사용하였다.

어플리케이션 기반 분류기에서 학습된 지지 벡터 기계의 편향을 보정하기 위해 정규화된 비중으로 사용자의 성별을 다수결 하였다. 정규화를 위해서 imlabOHP로 수집된 모든 사용자의 설치 어플리케이션의 분류 결과를 사용하였다. 이때 지지 벡터 기계에 의해 성별이 부여되는 어플리케이션은 구글 안드로이드 플레이 스토어에 등록되어 설명글이 존재하는 것으로 한정하였다.

가속도 데이터의 경우 각 사용자의 데이터 인스턴스를 분류하기 위해 leave-one-out 교차타당화 방법을 적용하였다. 즉, 해당 사용자를 제외한 다른 사용자들의 인스턴스로 학습한 지지 벡터 기계를 사용하여 데이터 인스턴스를 분류하였다. 데이터 수집이 정상적으로 이루어졌을 경우에 하루 동안 수집되는 인스턴스의 개수 96개이다. 한 사용자에 대해 총 인스턴스의 수가 이보다 적으면 가속도 데이터로 성별을 분류하지 않았다.

4.2.2. 평가 지표

제안된 기법의 성능은 정확도와 미분류율로 평가되었다. [표 6]은 성능 평가를 위한 혼동 행렬이다. 정확도는 전체 사용자에 대한 정확도와 남녀 사용자에 대한 정확도로 나뉘며 각각 수식 (23), (24), (25)로 정의된다. 미분류율은 사용자의 성별이 분류되지 못한 경우를 성능으로 평가하기 위하여 도입하였다. 수식 (26)으로 미분류율을 계산한다.

[표 6] 성능 평가를 위한 혼동행렬

		예측		미분류
		남성	여성	
실제	남성	TM	FF	UM
	여성	FM	TF	UF

$$\text{전체 정확도} = \frac{TM + TF}{TM + FF + FM + TF} \tag{23}$$

$$\text{남성 정확도} = \frac{TM}{TM + FF} \tag{24}$$

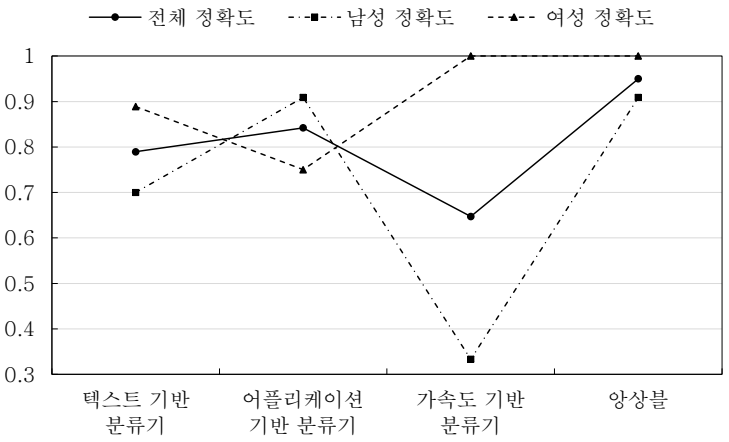
$$\text{여성 정확도} = \frac{TF}{FM + TF} \tag{25}$$

$$\text{미분류율} = \frac{UM + UF}{TM + FF + UM + FM + TF + UF} \tag{26}$$

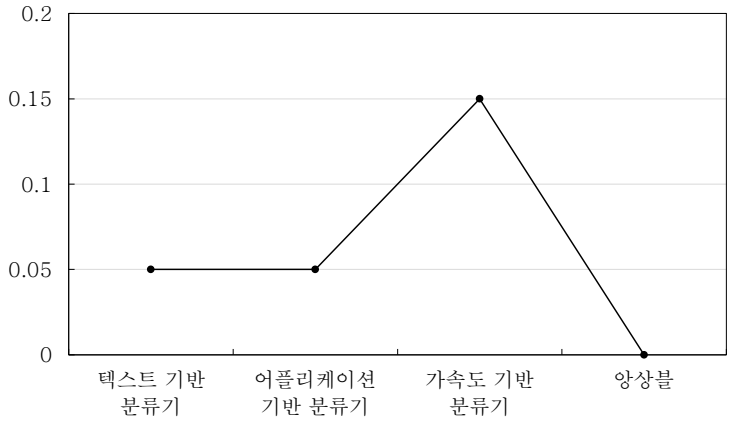
텍스트 기반 분류기는 계산 시간을 추가적인 평가 지표로 사용한다. 이 분류기는 스마트 기기 내에서 분류 과정이 수행되어야 한다. 따라서 코사인 유사도 계산 방법에 따라 상이한 계산 시간을 계산 복잡도에 대한 평가 지표로 고려한다.

4.3. 실험 결과

실험 결과 제안된 앙상블의 결과가 정확도 0.95, 미분류율 0으로 가장 높은 성능을 보였다. 남성 정확도와 여성 정확도 또한 앙상블한 결과가 세 분류기들보다 높았다. 모든 분류기와 앙상블한 결과의 전체, 남성, 여성 사용자에 대한 정확도를 그래프로 나타내면 [그림 9]와 같고, 미분류율은 [그림 10]에 나타내었다. 이를 바탕으로 분류기들과 다수결에 근거한 앙상블의 실험결과를 논의한다.



[그림 9] 제안 기법의 분류 정확도



[그림 10] 제안 기법의 미분류율

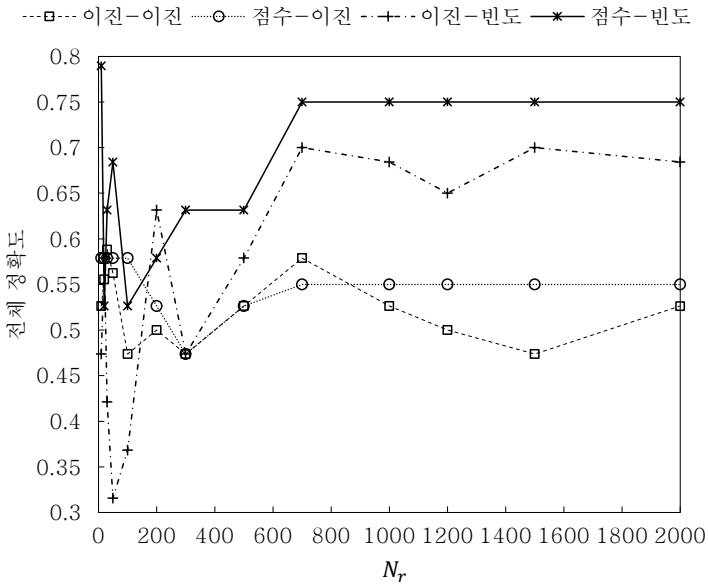
4.3.1. 텍스트 기반 분류기

텍스트 기반 분류기의 단어 추출 단계에서 얻어지는 카이 제곱 통계량을 기준으로 상위 20위까지의 단어들을 다음 [표 7]에 나타내었다. 여성 카이 제곱 통계량 1위~8위의 단어가 남성 단어 1위보다 더 큰 카이 제곱 통계량을 보였다. 또한 남성 단어는 주제 중심의 단어인 반면에 여성 단어는 여성들이 일상에서 사용하는 단어가 상위권에 자리하였다. 이는 남녀가 사용하는 단어가 매우 다르고 이를 통해 글 작성자 혹은 단어 사용자의 성별을 구분할 수 있음을 의미한다.

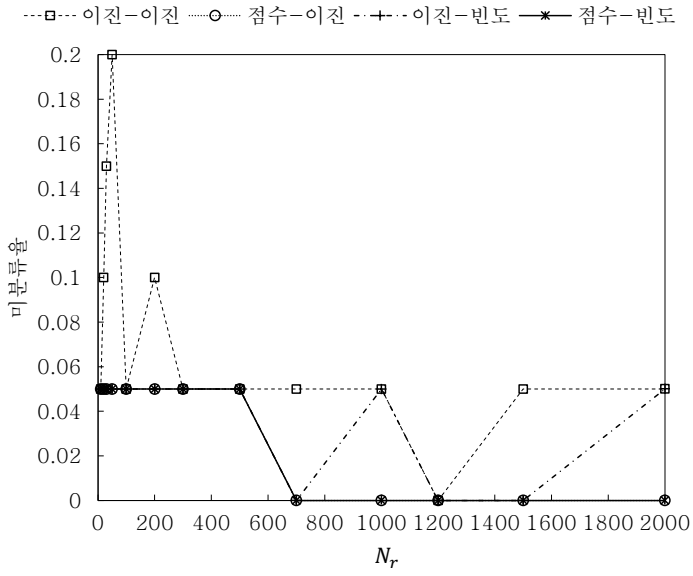
[표 7] 카이 제곱 통계량 기준 상위 20개의 단어

순위	남성		여성	
	단어	카이 제곱 통계량	단어	카이 제곱 통계량
1	감독	2961.14	엄마	8128.60
2	영화	2710.36	언니	6119.08
3	정답	2143.45	완전	4782.67
4	문제	2134.95	신랑	4638.55
5	개봉	1697.98	아이	4158.45
6	등장	1400.11	저희	4014.51
7	제작	1307.87	제가	3308.37
8	출연	1303.22	우리	3160.75
9	액션	1244.42	요즘	2955.54
10	원작	1230.39	요번	2936.11
11	게임	1219.24	요리	2911.67
12	시리즈	1157.04	열매	2893.51
13	전투	1155.52	허브	2844.01
14	사활	1151.47	이거	2843.81
15	출처	1130.51	오빠	2562.83
16	오브	1086.26	스크랩시	2502.30
17	버전	1071.22	꼬리말	2458.01
18	북미	1043.63	뮤비	2430.07
19	예고편	1001.82	남편	2396.67
20	흥행	963.29	퀘럼이	2359.19

텍스트 기반 분류기는 코사인 유사도 계산 방법과 단어집합 크기 N_r에 따라 분류 성능을 비교하였다. [그림 11]은 N_r에 따른 전체 정확도를 보여주며, [그림 12]는 N_r에 따른 미분류율을 보여준다.



[그림 11] 단어집합 크기에 따른 텍스트 기반 분류기의 전체 정확도

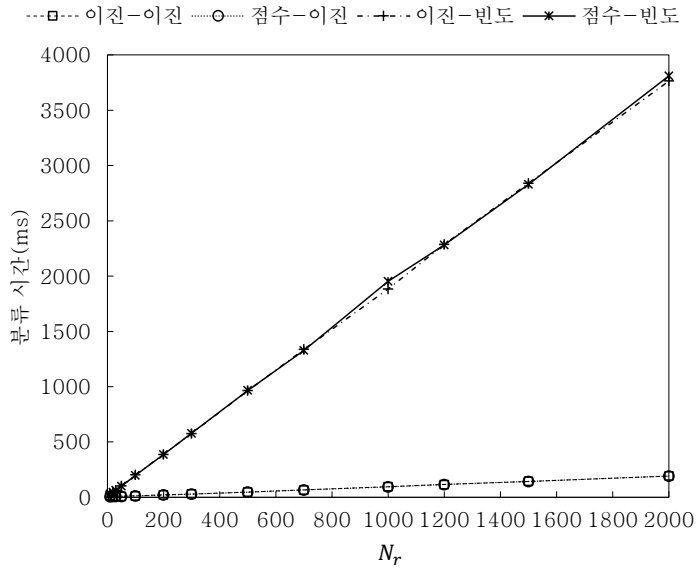


[그림 12] 단어집합 크기에 따른 텍스트 기반 분류기의 미분류율

네 가지 유사도 계산 방법들은 대체적으로 $300 \leq N_r \leq 700$일 때 N_r이 증가함에 따라 정확도도 증가하는 경향을 보였다. 점수-이진 방법과 점수-빈도 방법의 경우 단어가 많아져도 일정한 성능을 유지했지만 이진-이진 방법과 이진-빈도 방법은 단어가 많아지면 오히려 성능이 저하되기도 하였다. 이 경우 단어의 카이 제곱 통계량이 정확한 분류를 저해하는 것으로 해석된다.

미분류율은 가중치를 부여하지 않는 이진-이진 방법을 사용할 때 가장 높게 나타났다. 단어집합과 일치하는 텍스트 데이터 단어 수가 남성 단어집합과 여성 단어집합에서 동일하면 이 방법에 의해 계산되는 남녀 코사인 유사도 또한 같기 때문이다. 이진-이진 방법을 제외한 세 가지 유사도 계산 방법은 모든 단어집합의 크기에 대해 0.05 이하의 미분류율을 보였으며 미분류율이 0인 경우도 존재했다.

N_r에 따라 성별을 예측하는데 소요된 시간을 측정하여 [그림 13]에 나타내었다. 집합에 속한 단어의 수가 증가하면 모든 유사도 계산 방법에 대해 분류 시간 또한 증가하였다. 그러나 유사도 계산 방법에 따라 분류 시간의 증가율이 상이함을 확인할 수 있었다. 이진 일치 벡터를 사용하는 두 가지 유사도 계산법은 빈도 일치 벡터를 사용하는 방법과 비교하여 낮은 증가율을 보였다. 이러한 차이는 텍스트 데이터에서 단어를 확인할 때 일치 벡터를 생성하기 위한 계산 수준의 차이에 기인하다. 단어의 존재 여부만을 확인하는 것이 출현 빈도수를 모두 확인하는 것보다 계산 시간이 짧아지기 때문이다.



[그림 13] 단어집합 크기에 따른 텍스트 기반 분류기의 분류 시간

평균적으로 가장 좋은 성능을 보인 유사도 계산 방법은 단어의 카이 제곱 통계량과 사용자의 텍스트 데이터에서의 단어 출현 빈도 수를 모두 사용한 점수-빈도 방법이었다. $N_r \geq 700$이면 단어가 추가되어도 동일한 성능을 보였으며 다른 가중치 부여 방법들과 비교하여서도 가장 높은 정확도를 유지하였다. 특히 $N_r = 10$인 가장 작은 단어집합으로 모든 분류 결과들 중 가장 높은 전체 정확도 0.7894를 보였다. [표 8]은 가장 좋은 전체 정확도를 보인 $N_r = 10$이고 점수-빈도 방법으로 유사도를 계산하는 텍스트 기반 분류기로 사용자의 성별을 분류한 결과 얻어진 혼동행렬이다.

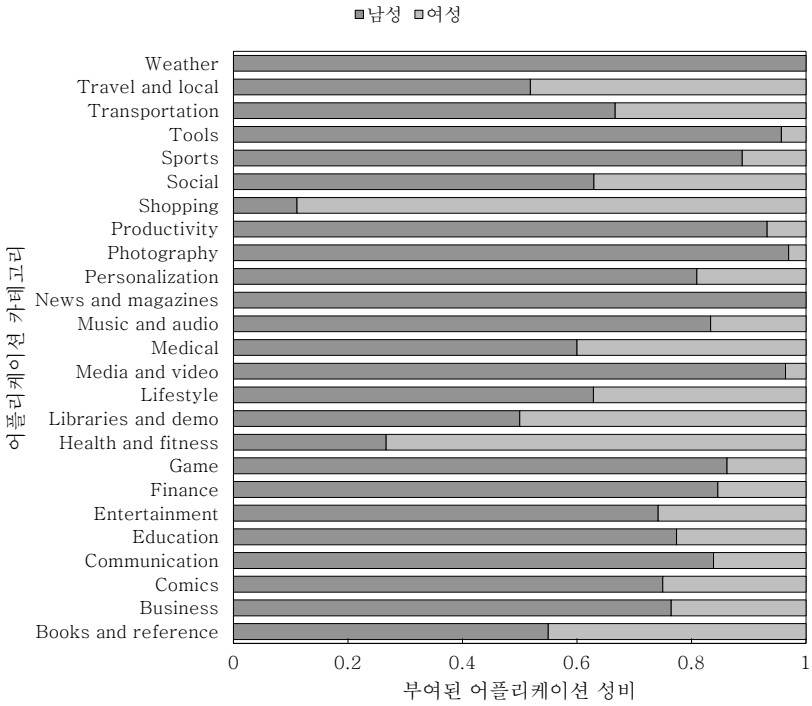
[표 8] 단어집합 크기가 10이고 점수-빈도 방법으로 유사도를 계산하는 텍스트 기반 분류기 분류 결과의 혼동행렬

		예측		미분류
		남성	여성	
실제	남성	7	3	1
	여성	1	8	0

특히 이 경우는 단어집합이 가장 작음에도 불구하고 미분류율은 0.05로, 한 명을 제외한 모든 사용자의 성별이 분류되었다. 미분류된 사용자의 경우 텍스트 데이터에 남녀 단어집합과 일치하는 단어가 전혀 존재하지 않았다.

4.3.2. 어플리케이션 기반 분류기

어플리케이션 기반 분류기를 수립하기 위하여 사용자의 스마트기기에 설치된 모든 어플리케이션을 사전에 조사하였다. 조사된 모든 어플리케이션들 중 구글 플레이에 등록된 어플리케이션은 총 663개였다. 이 663개의 어플리케이션 설명글을 웹에서 크롤링하고, 단어를 추출하여 웹 문서로 학습된 지지 벡터 기계 모델에 입력으로 넣어주었다. 전체 중 77%의 어플리케이션이 남성으로 분류되었고 23%는 여성으로 분류되었다. 각 어플리케이션 카테고리 별로 부여된 성별 레이블의 비율을 [그림 14]로 나타내었다.



[그림 14] 어플리케이션에 부여된 성별의 카테고리 별 비율

[그림 14]를 보면 남성 어플리케이션의 비율이 더 높은 것을 확인할 수 있다. 특히 Weather 카테고리의 어플리케이션은 모두 남성 어플리케이션으로 분류되었다. 그럼에도 불구하고 Shopping과 Health and fitness 카테고리에 속하는 어플리케이션은 50% 이상이 여성 어플리케이션으로 분류되었다. 이는 [그림 7]에서 여성 사용자가 이 두 카테고리에 속하는 어플리케이션을 남성보다 많이 사용한 것과 일치하는 결과이다.

[표 9] 어플리케이션 기반 분류기 분류 결과의 혼동행렬

		예측		미분류
		남성	여성	
실제	남성	10	1	0
	여성	2	6	1

[표 9]는 어플리케이션 기반 분류기로 사용자의 성별을 분류한 결과의 혼동행렬이다. 여성 사용자 한 명이 어플리케이션 데이터로 분류되지 않았다. 이 사용자의 기록을 조사한 결과, 실행한 어플리케이션이 모두 구글 플레이에 등록되지 않은 것으로 나타났다. 사용한 어플리케이션이 기기 구매 시부터 설치되어 있었던 어플리케이션이거나 개발자가 공식적인 어플리케이션 시장에서 배포하지 않은 어플리케이션인 것으로 추측된다.

[그림 9]의 어플리케이션 기반 분류기의 정확도를 살펴보면, 전체 정확도는 세 가지 분류기 중 가장 높은 0.8421을 기록하였다. 또한 남성 사용자에 대한 정확도가 여성 사용자에 대한 정확도보다 높다. 이는 위에서 언급한 바와 같이 웹 문서로 학습된 지지 벡터 기계 모델이 80%에 가까운 어플리케이션을 남성 어플리케이션으로 분류하였기 때문인 것으로 분석된다. 비록 지지 벡터 기계에 의해서 남녀로 레이블이 부여된 설치 어플리케이션들의 비율을 고려하여 다수결 시 정규화하였지만 모든 안드로이드 어플리케이션이 고려된 것은 아니기 때문에 편향이 발생한 것으로 보인다.

4.3.3. 가속도 기반 분류기

가속도 기반 분류기로부터 얻어지는 사용자 성별 예측 결과를 [표 10]의 혼동행렬로 나타냈다. [표 10]와 [그림 9]에서 볼 수 있듯이 가속도 기반 분류기로 사용자의 성별을 분류한 결과 앞에서 살펴본 다른 분류기들보다 분류 정확도가 낮다. 전체 정확도는 0.6470을 나타냈고 특히 남성 정확도는 0.4 이하로 [그림 9]의 모든 정확도 결과값 중 가장 낮았다. 여성 정확도가 남성보다는 높았지만 다른 분류기들과 비교할 때 여전히 가장 낮은 수준을 보였다.

[표 10] 가속도 기반 분류기 분류 결과의 혼동행렬

		예측		미분류
		남성	여성	
실제	남성	3	6	2
	여성	0	8	1

낮은 성능의 원인은 가속도 기반 분류기가 다른 분류기들과는 달리 수집된 스마트 기기 로그 데이터에 영향을 받는 통계적 학습 기법을 사용하는 데에 있다. 어플리케이션 기반 분류기는 웹에서 얻어진, 스마트 기기 로그 데이터와는 별개인 데이터를 사용하여 지지 벡터 기계를 학습한다. 그러나 가속도 기반 분류기의 경우 스마트 기기 사용자로부터 수집된 데이터를 바탕으로 지지 벡터 기계 모델을 학습하고 성별을 예측한다. 본 실험에서 수집된 가속도 데이터의 부족으로 지지 벡터 기계가 최적의 분류 함수를 찾지 못하고 그 결과 가속도 데이터 인스턴스를 올바르게 분류하지 못하였다.

또한 세 명의 사용자 데이터에 대해 성별을 분류하지 못하였다. 이 사용자들로부터 수집된 가속도 데이터 인스턴스의 수가 최소 분류 기준보다 적어서 분류되지 않았다. 이들의 가속도 인스턴스 수가 매우 적은 이유는 데이터 수집 어플리케이션의 오류이거나, 사용자들이 수집 어플리케이션을 너무 일찍 종료하였기 때문인 것으로 추측된다.

4.3.4. 다수결에 근거한 앙상블

본 논문에서 제시된 텍스트 기반 분류기와 어플리케이션 기반 분류기, 가속도 기반 분류기의 분류 결과를 모두 종합하여 다수결한 앙상블 결과를 [표 11]의 혼동행렬로 나타내었다. 이때 텍스트 기반

분류기는 가장 좋은 성능을 보인 $N_r = 10$, 점수−빈도 유사도 계산 방법을 사용하였다. 미분류된 경우는 없었으며 [그림 9]에 나타난 바와 같이 0.95의 전체 정확도를 보였다. 또한 남성 정확도는 0.9090, 여성 정확도는 1로 분류기들의 분류 결과와 비교하여 가장 높은 수준을 기록하였다.

[표 11] 앙상블 기법을 적용한 분류 결과의 혼동행렬

		예측		미분류
		남성	여성	
실제	남성	10	1	0
	여성	0	9	0

[표 12]는 모든 분류기와 제안된 앙상블 기법을 적용한 분류 결과와 신뢰도를 요약한 표이다. 사용자 11을 제외한 모든 사용자에 대하여 올바른 예측이 이루어졌다. 제안된 기법의 정성적 평가를 위해 올바르게 성별을 예측한 사례 중 만장일치와 신뢰도에 의한 예측 사례를 [표 12]의 사용자 2와 사용자 14의 분류 결과로 분석한다.

[표 12] 분류 결과와 신뢰도

사용자	실제 성별	분류 결과				분류 신뢰도		
		텍스트 기반 분류기	어플리케이션 기반 분류기	가속도 기반 분류기	앙상블	텍스트 기반 분류기	어플리케이션 기반 분류기	가속도 기반 분류기
1	여성	여성	남성	미분류	여성	0.4117	0.2934	–
2	여성	여성	여성	여성	여성	0.5543	0.6905	0.9218
3	여성	여성	여성	여성	여성	0.3765	0.3257	0.8949
4	여성	남성	여성	여성	여성	0.4145	0.1596	0.4248
5	여성	여성	여성	여성	여성	0.4600	0.2093	0.8755
6	여성	여성	남성	여성	여성	0.3829	1.0000	0.0340
7	여성	여성	여성	여성	여성	0.5157	0.0119	0.0183
8	여성	여성	미분류	여성	여성	0.4614	–	0.8205
9	여성	여성	여성	여성	여성	0.4865	0.0119	0.9066
10	남성	여성	남성	남성	남성	0.3546	1.0000	0.0635
11	남성	남성	여성	여성	여성	0.3606	0.2909	0.2307
12	남성	여성	남성	남성	남성	0.2725	0.6599	0.1314
13	남성	남성	남성	남성	남성	0.3773	0.1748	0.1378
14	남성	여성	남성	미분류	남성	0.3897	1.0000	–
15	남성	미분류	남성	여성	남성	–	1.0000	0.7623
16	남성	남성	남성	여성	남성	0.4840	1.0000	0.0476
17	남성	남성	남성	여성	남성	0.2958	0.6413	0.9060
18	남성	남성	남성	여성	남성	0.5417	0.6599	0.8046
19	남성	남성	남성	여성	남성	0.3773	0.1748	0.7508
20	남성	남성	남성	미분류	남성	0.3452	0.2387	–

4.3.4.1. 예측 사례: 만장일치

사용자 2는 앙상블에 의해서 만장일치로 여성으로 예측되었다. 사용자 2의 텍스트 기반 분류기의 분류 신뢰도는 다른 사용자들과 비교하여 가장 높았다. [표 13]은 카이 제곱 통계량을 기준으로 선택된 10개의 단어와 정규화된 점수, 사용자의 텍스트 데이터에서 실제로 발생한 횟수를 보여준다. 남성 단어집합과 일치하는 단어는 87회 출현하는데 그쳤지만 여성 단어집합과 일치하는 단어는 1,673회 출현하였다. 특히 점수의 비중이 높은 "엄마"와 "언니"가 가장 많이 출현하여 여성 단어집합과의 비교로 계산된 코사인 유사도가 높았다.

[표 13] 사용자 2의 일치 단어

남성 단어 집합	단어	감독	영화	정답	문제	개봉
	출현 빈도	2	34	0	47	1
	단어	등장	제작	출연	액션	원작
	출현 빈도	1	1	1	0	0
여성 단어 집합	단어	엄마	언니	완전	신랑	아이
	출현 빈도	332	691	175	0	84
	단어	저희	제가	우리	요즘	요번
	출현 빈도	46	81	227	35	2

사용자 2가 실행 어플리케이션 중 구글 플레이 스토어에 등록되어 있는 어플리케이션은 [표 14]와 같다. 이 사용자에게 여성 레이블이 부여된 어플리케이션이 남성 레이블의 어플리케이션보다 많음을 알 수 있다. 실제 여성이 많이 사용하는 소셜 커머스의 모바일 어플리케이션, 사진으로 대화하는 메신저 등이 여성으로 분류되었다.

[표 14] 사용자 2의 실행 어플리케이션

이름	카테고리	설명	부여 성별
카카오톡	Communication	메신저 서비스	남성
NH 스마트 뱅킹	Finance	모바일 뱅킹	남성
쿠팡	Shopping	쇼셜 커머스	여성
네이버 뮤직	Music and audio	음악 재생	여성
포토 원더	Photography	사진 편집	남성
네이버 캘린더	Productivity	일정 관리	여성
스냅챗	Social	사진 전송 메신저	여성
카카오홈	Personalization	런처	여성

이 사용자의 가속도 데이터는 여성으로 분류된 인스턴스가 246개, 남성으로 분류된 인스턴스가 10개로 올바른 분류 결과를 보였다. 세 가지 분류기가 모두 여성이라고 분류한 만장일치 사례이다.

4.3.4.2. 예측 사례: 분류 신뢰도에 의한 동률 문제 해결

사용자 14는 가속도 데이터 인스턴스가 44개에 불과하여 가속도 기반 분류기에 의해 미분류 처리되었다. 가속도 기반 분류기를 제외한 두 분류기는 서로 다른 분류 결과를 보였다. 앙상블에서 미분류 결과는 다수결에 포함하지 않기 때문에 동률이 발생하였다. 따라서 예측이 이루어진 두 분류기의 결과 중 분류 신뢰도가 큰 결과를 따라 최종 성별이 예측되었다.

텍스트 기반 분류기의 분류 결과 이 사용자 14는 여성으로 분류되었다. [표 15]의 자세한 일치 결과를 살펴보면 스마트 기기에서 수집된 텍스트 데이터 중 남녀 단어집합과 일치하는 단어가 매우 적다. 남성 단어집합과 일치하는 단어는 전혀 없었고 여성 단어집합과 일치하는 단어의 총 출현 빈도도 5회에 그쳤다. 남성 기준 벡터와

남성 일치 벡터의 코사인 유사도는 0, 여성 기준 벡터와 여성 일치 벡터의 코사인 유사도는 0.4472였다. 텍스트 데이터를 사용하여 예측한 성별의 신뢰도인 각도 신뢰도는 0.5747을 나타냈다. 사용자 14로부터 수집된 텍스트 데이터는 길이가 6,179 글자로 전체 사용자의 평균 길이인 88,398 글자에 크게 못 미쳤으며 전체 평균보다 더 짧은 남성 사용자 텍스트 데이터의 평균 길이의 13.5%에 불과했다.

[표 15] 사용자 14의 일치 단어

남성 단어 집합	단어	감독	영화	정답	문제	개봉
	출현 빈도	0	0	0	0	0
	단어	등장	제작	출연	액션	원작
	출현 빈도	0	0	0	0	0
여성 단어 집합	단어	엄마	언니	완전	신랑	아이
	출현 빈도	3	0	0	0	0
	단어	저희	제가	우리	요즘	요번
	출현 빈도	0	0	2	0	0

어플리케이션 데이터로 예측된 사용자 14의 성별은 남성으로, 올바르게 예측되었다. 이 사용자가 사용한 어플리케이션 목록 [표 16]을 보면 모든 어플리케이션이 남성으로 레이블이 부여된 Communication 카테고리에 속함을 알 수 있다. 여성 레이블을 가진 어플리케이션이 사용 기록에 전혀 존재하지 않으므로 신뢰도 1로 남성으로 분류되었다.

[표 16] 사용자 14의 사용 어플리케이션

이름	카테고리	설명	부여 성별
카카오톡	Communication	메신저 서비스	남성
NH 스마트 뱅킹	Finance	모바일 뱅킹	남성
크롬	Communication	인터넷 브라우저	남성

일치 단어가 매우 적게 존재하였던 텍스트 기반 분류기의 신뢰도는 0.3897인 반면 모든 사용 어플리케이션이 남성 레이블이었던 어플리케이션 기반 분류기의 신뢰도는 1이었다. 그러므로 신뢰도가 더 높은 어플리케이션 기반 분류기의 분류 결과를 따라 남성으로 최종 예측이 이루어졌다.

5. 결론

5.1. 요약 및 연구 의의

본 연구에서는 스마트 기기 로그 데이터를 이용하여 기기 사용자의 성별을 예측하는 앙상블 기법을 제안하였다. 제안된 앙상블 기법은 로그 데이터 중 텍스트와 어플리케이션, 가속도 데이터를 입력으로 갖는 세 가지 분류기를 바탕으로 다수결에 의해 최종적으로 사용자의 성별을 예측하였다. 제안된 기법의 성능을 평가하기 위하여 스마트 기기 로그 데이터 수집 어플리케이션을 개발하고 이를 이용하여 데이터를 수집하였다.

제안된 텍스트 기반 분류기는 사용자의 성별을 잘 구분할 수 있을 것으로 예상되지만 타인에게 노출될 시에는 사생활 침해 문제가 발생하는 문자메시지, 검색 기록 등을 사용하기 위하여 기기 내에서 데이터 처리가 가능하도록 하였다. 웹에서 수집한 문서를 이용하여 사용자의 성별을 구분할 수 있는 단어의 집합을 도출하고 스마트 기기 사용자의 텍스트 데이터의 일치 정도를 측정하여 사용자의 성별을 예측하였다.

어플리케이션 기반 분류기는 웹에서 수집한 어플리케이션의 설명글을 사용하여 어플리케이션에 성별 레이블을 부여하고 이를 바탕으로 사용자의 성별을 예측하였다. 사용자가 데이터 수집기간 동안 실행한 어플리케이션의 성별을 분류한 결과를 이용하였으며 사용자의 실제 어플리케이션 사용 성향을 반영하여 제안된 세 가지 분류기 중 가장 높은 분류 정확도를 얻었다. 특히, 남성 사용자의 성별을 예측하는데 높은 정확도를 보였다.

가속도 데이터를 사용하여 사용자의 성별을 예측하기 위하여 지지 벡터 기계를 이용한 통계 학습기반의 분류기를 제안하였다. 스마트 기기 로그 데이터와 데이터 수집 플랫폼의 특성을 반영하는 요인을 입력으로 갖는 지지 벡터 기계를 학습하여 일정 시간 간격 동안 발생한 가속도 데이터를 벡터화한 인스턴스의 성별을 분류하였다. 인스턴스 분류 결과를 바탕으로 사용자의 성별을 예측하였다.

제안된 앙상블 기법은 위의 세 분류기의 결과를 모두 사용하여 사용자의 성별을 예측하였다. 수집된 데이터를 통해 성능을 확인한 결과 각 분류기들과 비교하여 높은 분류 정확도를 보였으며 남성과 여성 사용자에 대해 모두 높은 분류 성능을 나타내었다. 더불어, 제안된 기법은 텍스트와 어플리케이션, 가속도 데이터 중 일부가 수집될 수 없는 상황에서도 사용자의 성별을 분류할 수 있다. 높은 분류 성능과 이러한 기법의 특성은 실제 스마트 기기 로그 데이터를 이용한 사용자 성별 예측에 제안된 기법이 실제로 적용 가능함을 의미한다.

5.2. 향후 발전 방향

본 연구에서는 기존의 연구에서 사용자의 성별 예측에 효과적이라고 알려진 텍스트, 어플리케이션, 가속도 데이터를 이용하였다. 그러나 스마트 기기 로그 데이터의 종류는 이외에도 매우 많으며 각각의 데이터의 특성과 활용 방안이 아직까지 심도 있게 연구되지 않았다. 따라서 향후 스마트 기기 사용자의 성별을 더욱더 정확하게 예측하기 위하여 본 연구에서 사용된 데이터 이외의 데이터를 사용하는 분류 방법을 연구할 예정이다.

연구의 배경이 된 스마트 기기를 이용한 개인화 서비스에 필요한 정보는 사용자의 성별 외에도 많은 인구통계학적 정보를 포함한다. 나이와 직업, 종교, 가족구성원의 수 등은 성별과 더불어 매우 중요하고 기초적인 사용자 정보라 할 수 있다. 그러므로 이 속성들에 대해서도 사용자를 분류할 수 있는 기법을 제안하고자 한다. 더 나아가서는 보다 정확하고 개인화된 서비스를 제공하기 위해 인구통계학적 정보 이외에도 스마트 기기 사용자의 관심 분야나 현재 상황을 예측하는 기법도 연구되어야 할 것이다.

참고 문헌

[1] L. D. Wolin and P. Korgaonkar, "Web advertising: gender differences in beliefs, attitudes and behavior," *Internet Research,* vol. 13, pp. 375–385, 2003.

[2] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, "Your installed apps reveal your gender and more!," *SIGMOBILE Mobile Computing and Communications Review,* vol. 18, pp. 55–61, 2015.

[3] G. M. Weiss and J. W. Lockhart, "Identifying user traits by mining smart phone accelerometer data," in *Proceedings of the International Workshop on Knowledge Discovery from Sensor Data*, 2011, pp. 61–69.

[4] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do*, et al.*, "From big smartphone data to worldwide research: The Mobile Data Challenge," *Pervasive and Mobile Computing,* vol. 9, pp. 752–771, 2013.

[5] T. Igarashi, J. Takai, and T. Yoshida, "Gender differences in social network development via mobile phone text messages: A longitudinal study," *Journal of Social and Personal Relationships,* vol. 22, pp. 691–713, 2005.

[6] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *Proceedings of the International Conference on World Wide Web*, 2007, pp. 151–160.

[7] S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: Methods and data," *Expert Systems with Applications,* vol. 39, pp. 9899–9908, 2012.

[8] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion,* vol. 16, pp. 3–17, 2014.

[9] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*: John Wiley & Sons, 2004.

[10] G. Zenobi and P. Cunningham, "Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error," in *Proceedings of the European Conference on Machine Learning*, 2001, pp. 576–587.

[11] K. Walkowiak, S. Sztajer, and M. Woźniak, "Decentralized

distributed computing system for privacy-preserving combined classifiers–modeling and optimization," in *Proceedings of the International Conference on Computational Science and Its Applications*, 2011, pp. 512-525.

[12] S. Jeong, S. Kalasapur, D. Cheng, H. Song, and H. Cho, "Clustering and naïve bayesian approaches for situation-aware recommendation on mobile devices," in *Proceedings of the International Conference on Machine Learning and Applications,* 2009, pp. 353-358.

[13] S. Jeong, D. Cheng, H. Song, and S. Kalasapur, "Non-collaborative interest mining for personal devices," in *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*, 2009, pp. 179-186.

[14] A. Mukherji, V. Srinivasan, and E. Welbourne, "Adding intelligence to your mobile device via on-device sequential pattern mining," in *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 1005–1014.

[15] V. Srinivasan, S. Moghaddam, A. Mukherji, K. K. Rachuri, C. Xu, and E. M. Tapia, "MobileMiner: Mining your frequent patterns on your phone," in *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 389–400.

[16] V. N. Vapnik, *The nature of statistical learning theory*: Springer-Verlag New York, Inc., 1995.

[17] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning* vol. 2: Springer, 2009.

[18] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the International Conference on Machine Learning*, 1997, pp. 412–420.

[19] M. Oakes, R. Gaaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu, "A method based on the chi-square test for document classification," in *Proceedings of the Annual International Conference on Research and Development in Information Retrieval*, 2001, pp. 440-441.

[20] P. Meesad, P. Boonrawd, and V. Nuipian, "A chi-square-test for word importance differentiation in text

classification," in *Proceedings of the International Conference on Information and Electronics Engineering*, 2011, pp. 110–114.
[21] P. H. Shahana and B. Omman, "Feature selection techniques for gender prediction from blogs," in *Proceedings of the International Conference on Networks Soft Computing*, 2014, pp. 355–359.
[22] M. Kakkar and D. Upadhyay, "Web browsing behaviors based age detection," *International Journal of Soft Computing and Engineering,* vol. 3, pp. 99–101, 2013.
[23] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 739–743.
[24] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, "Author gender prediction in an email stream using neural networks," *Journal of Intelligent Learning Systems and Applications,* vol. 4, pp. 169–175, 2012.
[25] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digital Investigation,* vol. 8, pp. 78–88, 2011.
[26] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," in *Proceedings of the International Conference on Advances in Information Systems*, 2006, pp. 274–283.
[27] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the International Workshop on Search and Mining User-generated Contents*, 2010, pp. 37–44.
[28] Z. Miller, B. Dickinson, and W. Hu, "Gender prediction on twitter using stream algorithms with N-gram character features," *International Journal of Intelligence Science,* vol. 2, pp. 143–148, 2012.
[29] F. Huang, C. Li, and L. Lin, "Identifying gender of microblog users based on message mining," in *Proceedings of the International Conference on Web-Age Information Management*, 2014, pp. 488–493.
[30] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proceedings of the International Workshop on Search and*

Mining User-generated Contents, 2011, pp. 37-44.
[31] J. S. Alowibdi, U. A. Buy, and P. Yu, "Empirical evaluation of profile characteristics for gender classification on twitter," in *Proceedings of the International Conference on Machine Learning and Applications*, 2013, pp. 365–369.
[32] S. Brdar, D. Ćulibrk, and V. Crnojević, "Demographic attributes prediction on the real-world mobile data," in *Proceedings of Mobile Data Challenge by Nokia Workshop*, 2012.
[33] S. Mohrehkesh, S. Ji, T. Nadeem, and M. C. Weigle, "Demographic prediction of mobile user from phone usage," in *Proceedings of Mobile Data Challenge by Nokia Workshop*, 2012.
[34] J. J.-C. Ying, Y.-J. Chang, C.-M. Huang, and V. S. Tseng, "Demographic prediction based on users mobile behaviors," in *Proceedings of Mobile Data Challenge by Nokia Workshop*, 2012.
[35] E. Zhong, B. Tan, K. Mo, and Q. Yang, "User demographics prediction based on mobile data," *Pervasive and Mobile Computing,* vol. 9, pp. 823-837, 2013.
[36] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*: Addison-Wesley Reading, 2010.
[37] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer, "Falling asleep with Angry Birds, Facebook and Kindle: A large scale study on mobile application usage," in *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011, pp. 47–56.

Abstract

A Method for User Gender Prediction Using Multi-type Smart Device Log Data

Yoonjung Kim
Industrial Engineering
The Graduate School
Seoul National University

As smart devices are propagating rapidly, needs for personalized services targeting smart device users are increasing. Meanwhile, the gender of a user is essential information to provide personalized services. Thus it is important to predict the gender of a smart device user accurately. In this study, an ensemble method for predicting the gender of a smart device user by using smart device log data is proposed. Since smart device log data consists of diverse types of data, the ensemble method utilizes majority voting scheme across the results of three classifiers based on different types of data.

First, a privacy preserving text based classifier is introduced. To alleviate privacy issues that occur when text data generated in a smart device are sent outside, a classification method which scans smart device text data only on the device and classifies the

gender of the user by matching text data with predefined sets of word.

Second, an application based classifier is presented. It first trains a support vector machine classifier with web documents and then assigns gender labels to applications which a smart device user actually used. By comparing the assigned gender label ratio, the majority label becomes the user's predicted gender.

Last, an acceleration based classifier is proposed. A support vector machine classifier is trained by smart device acceleration data instances with sixteen features including average acceleration of three axes and the hour of day at which data was collected. A new user's acceleration data instances are classified by the support vector machine and the majority class of instances is finally assigned as the user's gender.

The proposed method was evaluated with actual smart device log data. The results of an experiment showed that the method performed with high overall classification accuracy, 0.95. The text based classifier and the application based classifier performed better than the acceleration based classifier.

Keywords: gender prediction, smart device log data, privacy preserving, ensemble method, statistical learning, on-device analytics
Student Number: 2013 - 23204