http://dx.doi.org/10.3745/KTSDE.2013.2.9.603

Estimating Personal and Social Information for Mobile User

Jeong-Woo Son[†] · Yong-Jin Han[††] · Hyun-Je Song[††] · Seong-Bae Park[†††] · Sang-Jo Lee[††††]

ABSTRACT

The popularity of mobile devices provides their users with a circumstance that services and information can be accessed wherever and whenever users need. Accordingly, various studies have been proposed personalized methods to improve accessibility of mobile users to information. However, since these personalized methods require users' private information, they gives rise to problems on security. An efficient way to resolve security problems is to estimate user information by using their online and offline behavior. In this paper, for this purpose, it is proposed a novel user information identification system that identifies users' personal and social information by using both his/her behavior on social network services and proximity patterns obtained from GPS data. In the proposed system, personal information of a user like age, gender, and so on is estimated by analyzing SNS texts and POI (Point of Interest) patterns, while social information between a pair of users like family and friend is predicted with proximity patterns between the users. Each identification module is efficiently designed to handle the characteristics of user data like much noise in SNS texts and missing signals in GPS data. In experiments to evaluate the proposed system, our system shows its superiority against ordinary identification methods. This result means that the proposed system can efficiently reflect the characteristics of user data.

Keywords : User Identification, SNS, Proximity, SVM, Gaussian Mixture

모바일 사용자의 개인 및 소셜 정보 추정

손 정 우[†] · 한 용 진[††] · 송 현 제[††] · 박 성 배[†††] · 이 상 조[††††]

요 약

모바일 디바이스의 발달은 사용자가 언제 어디서나 원하는 서비스에 접근하고, 정보를 소비할 수 있는 환경을 마련했다. 이에 맞춰 다양한 연구들이 모바일 사용자의 정보 접근성을 향상 시키기 위한 개인화 방법을 제안해 왔다. 하지만, 이와 같은 개인화는 사용자 개인과 관련된 정보를 요구하기에, 사용자 정보에 대한 보안과 관련된 우려를 낳고 있다. 이를 해결할 수 있는 효과적인 방법 중 하나로 사용자 정보를 사용자의 온라인 혹은 오프라인 상의 행동 패턴으로부터 추정하는 것을 들 수 있다. 본 논문에서는 SNS(Social Network Service) 상에서의 사용자 패턴과 사용자 간 물리적인 근접성 패턴을 분석하여 사용자 개인의 정보와 타 사용자와의 사회 관계정보를 식별하는 사용자 정보 식별 시스템을 제안하고자 한다. 제안한 시스템은 SNS 텍스트와 GPS 데이터에 기반한 POI(Point of Interest) 패턴으로부터 사용자의 나이, 성별 등 개인 정보를 식별하고, 사용자 GPS 데이터를 이용하여 얻어진 사용자 간 근접성 패턴을 이용하여 두 사용자 간의 가족, 동료 등 관계 정보를 추정한다. 각각의 사용자 식별 모듈은 해당 데이터의 특성을 고려하여 SNS 데이터의 노이즈와 사용자 GPS 데이터의 손실을 감안함으로써 더 정확한 사용자 식별 성능을 보이도록 설계되었다. 이를 검증하기 위한 실험에서 제안한 시스템은 기존의 방법에 비해 더 나은 성능을 보였으며, 이는 본 논문에서 제안하는 방법이 사용자 데이터의 특성을 효과적으로 반영하고 있음을 의미한다.

키워드 : 사용자 식별, SNS, 근접성, SVM, Gaussian Mixture

1. 서 론

스마트폰, 태블릿 PC 등 모바일 디바이스의 고성능화와 대중화는 사용자가 언제, 어디서나 쉽게 정보나 컨텐츠에 접근할 수 있도록 하였다. 이에 따라, 모바일 디바이스를 위한 다양한 개인화 서비스들이 선보이고 있다. 개인화 서비스의 핵심 기술 중 하나는 사용자를 표현하는 기술이다. 사용자는 나이, 성별과 같은 개인 정보와 과거 컨텐츠 접근 정보, 그리고 최근 많이 활용되는 사용자 간 관계를 의미하

※ 본 연구는 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발)의 지원으로 수행되었음.
† 정 회 원 : 경북대학교 전자전기컴퓨터학부 박사
†† 비 회 원 : 경북대학교 전자전기컴퓨터학부 박사과정
††† 비 회 원 : 경북대학교 IT대학 컴퓨터학부 교수
†††† 정 회 원 : 경북대학교 IT대학 컴퓨터학부 교수
논문접수 : 2013년 7월 16일
수 정 일 : 1차 2013년 8월 6일
심사완료 : 2013년 8월 6일
* Corresponding Author : Sang-Jo Lee(sjlee@knu.ac.kr)

는 소셜 정보 등을 이용하여 표현될 수 있다[1]. 따라서 다양한 개인화 서비스에서는 이들 정보를 사용자에게 요구하고 있다. 하지만, 사용자의 개인 혹은 소셜 정보를 사용자에게 직접 받는 것은 최근 높아진 보안에 대한 관심이 반영하듯 쉽지 않은 일이다. 사용자의 승인을 받아 사용하는 방식을 따르더라도, 모든 개인 정보를 입력하는 번거로움을 감수하는 사용자는 많지 않다.

직접 사용자로부터 정보를 얻을 수 없을 때, 가장 효과적인 방법은 사용자 데이터로부터 사용자 정보를 식별하는 것이다. 이에 따라, 다양한 연구들이 사용자의 나이, 성별 혹은 사용자 간의 관계를 추정하기 위해 제안되고 있다. 본 논문에서는 사용자의 개인 정보와 소셜 정보를 자동으로 식별할 수 있는 시스템을 제안한다. 본 논문에서 제안하는 식별 시스템은 개인 및 소셜 정보를 모두 식별할 뿐 아니라, 식별된 정보 간의 공유를 통해 오류를 줄일 수 있다는 장점이 있다. 제안한 시스템은 크게 개인 정보 식별과 소셜 정보 식별 모듈로 이루어져 있다.

- 개인 정보 식별

개인 정보 식별은 사용자의 SNS(Social Network Service) 활동 내역과 POI(Point of Interest) 패턴을 토대로 사용자의 전기적 속성 (biographic attribute)을 식별한다. Facebook, Twitter, me2day 등 다양한 SNS들이 등장하면서 사용자들이 위 서비스들을 이용하여 다양한 주제의 텍스트들을 작성한다. 작성된 사용자의 텍스트에는 작성한 사용자의 속성 즉, 성별, 나이, 관심 분야 등의 정보가 명시적으로 또는 묵시적으로 포함되어 있다. 최근 이러한 텍스트들을 분석하여 사용자 맞춤 서비스에 사용하고자 하는 연구들이 많이 진행되고 있다.

사용자의 개인 정보 식별은 SNS 외 블로그, 전화 대화문서, 영화 리뷰 등을 이용한 연구가 이미 많이 진행되어 왔다[2, 3, 4, 5]. 이러한 연구들은 사용자 식별을 위한 자질들을 정의하는 데에 중점을 두었다. 즉, 기존 연구들은 주어진 도메인에 따라 사용자가 작성한 텍스트의 특성을 잘 반영할 수 있는 자질을 정의하려 했다. 정의된 자질을 이용하여 표현된 텍스트는 분류기를 이용하여 사용자 속성을 식별하는 데 사용되었다. 최근 제안된 SNS를 이용한 연구 역시 기존의 자질을 이용한 연구에서 벗어나지 못하고 있다.

하지만, 기존의 다른 문서들과 달리 SNS 텍스트들은 짧은 길이와 더불어 다양한 주제를 다루고 있다. 따라서, 사용자들이 작성한 텍스트들 중에는 사용자를 식별하는 데에 영향을 끼치지 않거나 잘못 식별하게 만드는 텍스트들이 존재한다. 하지만, 기존 연구들은 이들 모두를 사용하거나 간단한 방법으로 필터링하여 사용자 식별 모델을 구축하였다[2, 4, 5]. 이러한 방법들은 다음과 같은 문제점들이 있다. 첫 번째, 주어진 텍스트를 모두 학습에 사용할 경우, 주어진 사용자 속성과는 관련 없거나 잘못 태깅된 학습 데이터로 인해 사용자 식별 성능이 떨어질 수 있다. 두 번째, 특정 사용자 속성과 관련없는 데이터를 필터링할 경우, 필터링 오류가 다음 단계로 전파될 수 있다. 또한, 필터링을 위한 학습 데이터를 구축하는 것은 많은 비용이 든다.

제안한 시스템의 개인 정보 식별 모듈은 앞선 문제점들을 해결하기 위해 다중 인스턴스 학습(Multi-Instance Learning)을 사용한다. 다중 인스턴스 학습이란 인스턴스 하나를 학습 단위로 간주하는 개별 인스턴스 학습(Single-Instance Learning)과는 달리 인스턴스 집합을 하나의 학습 단위로 수행하는 학습 방법이다[6]. 학습시 인스턴스 집합에서 “도움이 되는” 인스턴스 만을 찾아 학습을 수행하므로 관련 없는 또는 잘못 태깅된 인스턴스들로부터의 성능 하락을 최소화할 수 있다. 위 학습 방법은 사용자들이 작성한 텍스트들을 개별로 태깅하는 것이 아닌 텍스트 집합에 태깅을 수행하기 때문에 개별 결과를 합치는 모듈이 필요하지 않으며 학습 데이터를 쉽게 구축할 수 있는 장점을 가진다.

POI 패턴의 경우, 특정 속성을 가지는 사용자들이 유사한 장소를 선호할 것이라는 가정 하에 사용자 정보 식별을 위해 사용되었다. 예로 “남자”의 경우, “여자”에 비해 미용과 관련된 POI의 방문 빈도나 시간에서 차이를 보일 수 있다. 뿐만 아니라, “20대”와 “40대”의 주요 POI도 차이를 보일 수 있다. 이러한 차이점을 이용하여 제안한 시스템의 개인 정보 식별 모듈은 사용자의 주요 POI와 이들의 방문 빈도를 기반으로 사용자 식별을 수행한다.

- 소셜 정보 식별

소셜 정보 식별은 두 사용자 간의 근접성 패턴(proximity pattern)을 이용하여 가족, 동료, 친구 등의 소셜 정보를 식별하기 위한 모듈이다. 사람들이 소통하는 시공간적 상황은 높은 자유도(degree of freedom)와 변형을 보이지만, 생활공간과 사회적 관계의 제약을 받는다. 이러한 제약은 사용자의 일상, 즉, 평일에 직장과 집을 오고가거나, 일과 중에는 동료와 직장에서 업무를 진행하고, 일과 이후 친구를 만나거나 집에서 가족과 휴식을 취하는 것 등, 평범한 일상에서 쉽게 찾을 수 있다[1. 7]. 따라서 소셜 정보 식별 모듈은 집이나 직장과 같은 공간적 상황 정보를 인지함으로써 사용자들 간의 가족, 친구, 동료와 같은 관계 정보를 식별한다.

위치 식별 센서가 탑재된 모바일 폰의 사용이 확대되면서 이를 이용한 사용자들 간의 친밀도 및 사회적 관계 분류 연구가 활발히 진행되어 왔다 [1, 8, 9, 10]. Eagle et al.은 사용자들 간의 시공간적인 근접성을 측정하고 이들 간의 친밀도를 분류하는 방법을 제안하였다 [8]. Li et al.은 사용자들의 공통된 위치 변화를 이용하여 사회적 관계 분류 방법을 제안하였다 [1]. 이와 같은 기존연구들은 두가지 문제점을 가진다. 먼저, 특정 장소에 의존하여 사용자 간의 관계를 모델링했기에, 일상적인 사용자 패턴을 반영하기 힘들다. 다음으로 근접성 측정을 위해 사용한 GPS 데이터의 손실이나 오류를 감안하지 않고 있다.

본 시스템에서는 이러한 문제점을 해결하기 위해, 특정 장소와 관계없이 사용자 간의 전반적인 근접성 패턴을 이용하여 특정 관계를 모델링한다. 이때, GPS 데이터의 손실이나 오류를 감안하여 근접성 패턴을 확률 함수로 모델링한

다. 제안한 시스템에서는 먼저 안드로이드 기반 상황인지 플랫폼[11]을 이용하여 개별 사용자들의 일상적인 시공간적 상황을 추출한다. 추출된 개별 사용자들의 정보를 이용하여 사용자들이 상황을 공유하는 확률 분포를 추정한다. 이때, 사용자 간의 근접성 패턴을 정규 분포의 혼합으로 가정하고 Gaussian mixture model을 이용한다. 마지막으로 특정 관계에 속한 사용자 간의 근접성 분포를 이용하여, 가족, 친구 동료 등에 대한 일반화된 패턴을 학습함으로써 새로운 사용자 간의 데이터가 주어졌을 때, 이들의 관계를 식별한다.

본 논문에서는 제안하고자 하는 시스템의 성능을 실제 사용자 데이터를 이용하여 검증하였다. 실험 결과에서는, 본 논문에서 제안하는 시스템이 개인 정보 식별 및 소셜 정보 식별에서 기존 방법에 비해 유의미한 성능 향상을 이룰 수 있음을 보였다. 이러한 성능의 향상은 본 시스템의 두 식별 모듈이 사용자 데이터의 특성을 효과적으로 반영하고 있음을 증명한다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 정보 식별과 관련된 기존 연구들을 살펴본다. 3장에서는 제안한 시스템의 구조를 살펴보고, 개인 정보 식별 모듈과 소셜 정보 식별 모듈에 대한 자세한 설명은 4장과 5장에서 다룬다. 6장에서는 제안한 시스템의 성능 검증을 위한 실험 및 결과를 분석한다. 7장에서는 결론과 향후 연구를 다룬다.

2. 관련 연구

2.1 개인 정보 식별

블로그, 이메일, 방문한 웹 페이지 및 다른 사용자와의 대화문서 등 사용자가 생성 및 작성한 문서로부터 사용자들의 성별, 나이 등을 식별하고자 하는 연구는 오래 전부터 많이 진행되어 왔다. Boulis와 Ostendorf는 전화로 주고받은 대화 문서에서 사용자의 성별을 식별하고자 하였다[2]. 이들은 전화 대화 문서에서 성별 간의 단어 사용이 다름을 파악하고, 이를 반영하기 위해 n-gram모델로 대화를 표현하고 기계학습을 사용하여 성별을 식별하였다. Garera와 Yarowsky는 더 나아가 전화 대화문서와 이메일에서 사용자의 성별, 나이와 모국어로 말하는지(native speaker) 여부를 식별하였다[3]. 이들은 사회 언어학(sociolinguistic) 자질과 담화(discourse) 자질을 정의하고 이로부터 사용자의 속성을 식별하였다. 뿐만 아니라, 영화 리뷰 데이터를 이용한 성별 식별 연구도 있었다[4, 5]. 이들은 IMDb(www.imdb.com)에서 성별에 따라 언어 사용에 있어 다른 점을 반영하고자 하였다. 즉, 여성이 남성에 비해 사교적인 스타일(social style)로 리뷰를 쓰는 반면 남성은 여성에 비해 제 3자가 다른 사람에게 알리고자 하는 스타일(broadcast style)로 리뷰를 쓴다는 것을 발견하였다. 이를 위해 단어와 문장의 복잡도(complexity)[12]와 대명사 사용 빈도, 관용구(hedging phrase) 등을 자질로 사용하였다.

최근 SNS에서 사용자 정보를 식별하고자 하는 연구들이 많이 진행되고 있다. 이들 연구들은 기존의 사용자 식별을 위한 방법들에 SNS에서 얻을 수 있는 정보 및 특징을 추가로 반영하는데 초점을 두고 있다. Rao et al.은 Twitter에서 성별, 나이, 지역출신, 정치적 성향을 식별하고자 하였다[13]. 이들은 딸림벗(follower), 따름벗(following)을 통한 네트워크 관계 및 트윗(tweet), 리트윗(retweet) 등으로 얻어지는 사용자간의 커뮤니케이션 횟수에 기존 연구의 사회 언어학 자질을 합쳐 사용자 속성을 식별하였다. Pennacchiotti와 Popescu는 기계학습 기반의 방법을 이용하여 Twitter 사용자를 분류하였다[14]. 이 연구에서는 Bio 필드에서 추출하는 프로파일 자질과 트윗 작성 비율 및 리트윗 비율과 작성한 트윗이 URL을 포함하는지 여부를 사용하였다. 뿐만 아니라 토픽 모델을 사용하여 트윗을 분석한 자질과 소셜 네트워크 관계를 자질로 사용하여 사용자의 정치적 성향, 스타벅스 팬(starbucks aficionado)인지를 식별하였다. Burger et al.은 Twitter에서 성별을 식별하고자 하였다[15]. 이들은 다량의 Twitter 사용자와 트윗을 수집하였으며, 언어와 독립적인 모델을 구축하기 위해 Bio 필드 및 닉네임과 작성한 트윗에서 단어 레벨 n-gram 모델과 음절 레벨의 n-gram 모델을 사용하여 자질을 추출하고 이를 기반으로 Winnow 알고리즘을 사용하여 성별을 식별하였다.

앞선 연구들은 사용자가 작성한 정보들을 잘 반영할 수 있는 자질에 대해 주로 연구하였다. 하지만, 사용자들이 작성한 정보들 중에는 개인 정보 식별에 직접적으로 또는 간접적으로 영향을 끼치는 것이 있는 반면 그렇지 않는 것들도 존재한다. 이들 모두를 식별에 사용하면 영향을 끼치지 않는 다수의 텍스트들로 인해 성능 하락이 발생할 수 있다. 제안한 개인 정보 식별 모듈에서는 개별 인스턴스 학습 방법이 아닌 다중 인스턴스 학습 방법으로 사용자의 프로파일을 식별하여 이러한 문제들을 피하고자 한다.

2.2 소셜 정보 식별

제안한 시스템의 소셜 정보 식별 모듈은 사용자 간의 근접성 패턴을 분석하여, 이들의 관계를 식별한다. 사용자 간 근접성을 식별하기 위한 방법으로 크게 두 가지 접근이 연구되어 왔다. 하나는 블루투스를 이용해 모바일 폰 간의 근접성을 명시적으로 추출하는 것이다[8]. 이러한 접근은 사용자들과 인접한 블루투스 장치에 대한 지도를 미리 구성하여야 한다. 다른 하나는 GPS와 같은 위치 식별 센서를 이용해 개별 사용자들의 위치를 추출하고 추출된 위치 정보의 시간적 공간적 거리를 계산하여 사용자 간의 근접성을 추론하는 것이다[1, 9, 10]. 후자의 경우 전자에 비해 정확도는 떨어지지만 쉽게 확장할 수 있는 장점이 있다.

최근 사용자의 위치 데이터 접근이 용이해지면서 이를 이용한 사용자 분석 연구들이 소개되고 있다. Cho et al.은 사용자가 하루 중 집과 직장에 머무는 확률 분포의 차이를 보였다[9]. 이러한 분석을 바탕으로 Eagle et al.은 사용자 간의 관계에 따라 MIT 내와 밖에서 근접성의 확률 분포의 차이를 분석하고, 이를 이용해 사용자 간 관계를 식별하였다[8].

기존 소셜 정보 식별을 위한 연구는 특정 장소를 기준으로 사용자 간의 근접성 패턴을 분석하였다. 제안한 시스템에서는 사용자 간의 지리적 근접성을 식별하기 위해 안드로이드 기반 상황인지 플랫폼[11]을 이용한다. 상황인지 플랫폼은 GPS 및 가속도, 고도 센서 등을 이용해 사용자의 POI (point of interest)와 해당 POI에 대한 레이블, 그리고 사용자가 걷는지 정지해 있는지 등의 행동 유형을 저장한다. 따라서, 제안한 시스템은 특정 장소에 의존하지 않고 사용자 간의 근접성 패턴을 얻을 수 있다. 제안한 시스템에서는 이를 이용하여 관계별 패턴을 학습하고 새로운 사용자 간의 관계를 식별한다.

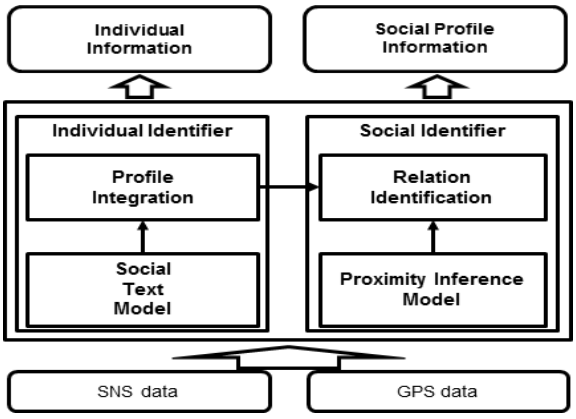


Fig. 1. Overall structure of the proposed system

3. 사용자 식별 시스템

제안한 시스템은 사용자의 SNS 데이터와 GPS 데이터를 이용하여 개인 정보와 소셜 정보를 식별한다. Fig. 1은 제안한 시스템의 구조를 보여준다. 그림에서 보듯이, 먼저 사용자의 SNS 데이터를 이용하여 개인 정보를 식별한다. 이때, 다중 인스턴스 기반의 분류 방법을 사용한다. 소셜 정보의 경우, GPS 데이터를 확률 분포로 모델링하여 근접성 패턴을 분석한다. 분석된 결과를 바탕으로 먼저 사용자 간 관계를 식별하고, 이를 이전에 추정된 각 사용자의 개인 정보를 이용하여 검증한다. 검증을 통해 “남자”와 “남자”간의 부부 관계나, “어머니”와 “아들” 간의 부부 관계, “남자”와 “여자” 간의 자매 관계 등 사회 통념상 쉽게 인지할 수 있는 오류를 줄인다.

4. 개인 정보 식별

본 논문에서 제안하는 시스템의 개인 정보 식별 모듈은 사용자의 SNS 데이터와 POI 패턴을 사용하여 개인 정보를 식별한다. 이때, 사용자가 작성한 SNS 데이터의 특성을 반영하기 위해 다중 인스턴스 학습 기반의 기계학습 방법 중 하나인 Multi-Instance Support Vector Machine(MI-SVM)를 사용한다.

4.1 다중 인스턴스 학습을 이용한 텍스트 기반 정보 식별

다중 인스턴스 학습이란, 일반적인 기계학습 방법에서 개개의 인스턴스를 하나의 데이터 단위로 보는 것과 달리, 인스턴스의 집합인 bag을 가정하고, bag 단위의 학습을 수행하는 것이다[6]. SNS 데이터를 예로 들어보면, 데이터 내에 나타나는 개개의 텍스트는 인스턴스라 볼 수 있으며, 단일 사용자가 작성한 모든 텍스트의 집합을 하나의 bag으로 간주한다. Bag 단위의 학습이 이 경우 의미 있는 이유는 다중 인스턴스 학습에서 일반적으로 하는 가정에 있다. 다중 인스턴스 학습에서는 bag에 레이블링을 하게 되는데, 이때, bag 내에 모든 인스턴스가 해당 클래스에 속할 필요가 없다. 즉, bag에 속한 인스턴스가 하나라도 해당 클래스에 속한다면, 그 클래스로 레이블링하게 된다. 이는 SNS 텍스트의 대부분이 특정 사용자 속성을 표현하지 못하는 상황에서 학습을 위한 데이터 구축을 쉽게 해준다. 학습 데이터로 여러 bag이 주어지면, 다중 인스턴스 학습은 이들 bag에 속한 인스턴스 중, 현재 결정할 클래스에 적합한 인스턴스를 찾고 이를 바탕으로 모델을 학습한다 [16, 17, 18, 19]. 따라서, 특정 사용자의 SNS 텍스트가 모두 사용자의 성별을 표현하지 못하더라도, 다중 인스턴스 학습을 통해 성별을 식별하는데 유의미한 데이터만을 이용하는 효과를 가져올 수 있다.

제안한 시스템에서는 다양한 다중 인스턴스 학습 방법 중, 분류 문제에 가장 좋은 성능을 보이는 Multi-Instance Support Vector Machine(MI-SVM)을 사용한다. MI-SVM는 기존의 SVM에서 다중 인스턴스 학습을 고려하도록 Andrews et al.이 제안한 알고리즘이다[20]. 기존의 SVM의 경우 모든 인스턴스를 고려하여 마진(Margin)이 최대화되는 결정 경계를 찾는 데 반해, MI-SVM는 Bag 간의 마진을 최대화하는 결정 경계를 찾는다.

이를 위해 MI-SVM는 긍정과 부정 클래스에 속한 bag 중, 긍정 bag에는 최소 하나의 유의미한 인스턴스가 존재함을 가정한다. 이를 Witness 인스턴스라 하는데, MI-SVM에서는 부정 bag에 속한 인스턴스로부터 가장 멀리 떨어진 인스턴스들을 Witness 인스턴스로 정의한다.

모든 인스턴스를 고려하여 bag 간의 마진을 최대화 하는 결정 경계를 찾는 soft-margin SVM은 아래 수식과 같이 표현될 수 있다.

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_I \xi_I \\ &\text{s.t.} \quad \forall I: \; Y_I \max_{i\in I}(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_I, \; \xi_I \geq 0 \end{aligned} \tag{1}$$

수식 (1)에서 X는 인스턴스를 나타내고 Y는 Bag을 나타낸다. Y_I는 I번째 Bag의 레이블을 나타내며, X_i는 I번째 Bag에 존재하는 인스턴스를 나타낸다. 수식 (1)의 최적화 문제에 다중 인스턴스 학습의 가정인, “긍정 bag에 최소 하나의 witness 인스턴스가 존재해야 한다.”를 반영하면 수식 (2)와 같이 표현된다.

Table 1. Features for SNS text model

Feature	Description/Example
Bag-of-Word	Noun, Unknown Word, Symbol Ex) 공채, 소개팅, LOL, 진보, 선거, ...
Emoticons	If each emoticon is contained a given text, the value of emoticon feature is set to 1. Otherwise, 0. Ex) ㅡㅡ, ㅠㅠ, ㅠ_ㅠ, ㅜㅜ, T_T, ㅇㅁㅇ, ㅇㅅㅇ, ^^, ^_^, :), :(, ♡, ㄱ-, ...
Alphabetic Character/Symbol Repetition	If the pattern of [ㅋ|ㅎ|ㄷ|ㅋㄷ|.|!|;|~]+ are matched a given text, the value of pattern feature is set to 1. Otherwise, 0. Ex) ㅋㅋㅋㅋ, ㅎㅎ, ㄷㄷ, ㅋㄷㅋㄷ, ...

$$\min_{s}\min_{\mathbf{w},b,\xi}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{I}\xi_I$$
$$\text{s.t.}\quad \forall I:\ Y_I=-1\ \wedge\ -\langle\mathbf{w},\mathbf{x}_i\rangle-b\geq 1-\xi_I,\ \forall i\in I,\quad (2)$$
$$\text{or}\ \ Y_I=1\ \ \wedge\ \langle\mathbf{w},\mathbf{x}_{s(I)}\rangle+b\geq 1-\xi_I,\ \ \text{and}\ \xi_I\geq 0$$

위의 수식에서 $s(.)$는 해당 인스턴스가 Witness 인스턴스 인지를 반환하는 선택 변수(selector variable)이다. 따라서 수식에서 알 수 있듯이, 마진에 대한 제약 사항은 긍정 bag의 Witness 인스턴스들에게만 적용된다. 위 수식은 혼합 정수 계획법(Mixed Integer Programming)으로 변환하여 최적화할 수 있다 [20].

1) SNS를 이용한 개인 정보 식별

제안한 시스템에서는 SNS 텍스트의 특성을 고려하여 MI-SVM를 이용한 개인 정보를 식별 모듈을 사용한다. 먼저, 사용자가 작성한 SNS 텍스트들은 아래와 같은 형식으로 주어진다.

$$D'=\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$$

X_i는 i번째 사용자를, $Y_i\in\{-1,+1\}$는 i번째 사용자의 클래스를 나타낸다. $X_i=\{x_{i1},x_{i2},\ldots,x_{in_i}\}\in R^m$는 i번째 사용자가 작성한 SNS 텍스트들 전체를 나타내며, $x_{ij}\in R^d$은 i번째 사용자가 작성한 j번째 텍스트, 즉 개별 텍스트를 의미한다. n_i는 i번째 사용자가 작성한 SNS 텍스트 전체 개수이다. 개인 정보 식별 모듈에서는 사용자가 작성한 모든 텍스트를 하나의 학습 단위로 사용하며, 위 단위에 대해 클래스를 추정한다.

개인 정보 식별 모델은 데이터 D'로부터 함수 $f:R^m\rightarrow Y$를 추정하는 것이며, 이는 수식 (2)를 사용하여 추정한다. MI-SVM은 부정 bag에서의 긍정과 "가장 가까운" 인스턴스와 긍정 bag에서의 인스턴스들과 거리가 멀도록, 단 긍정 Bag에 적어도 하나의 인스턴스가 포함되도록 결정 경계를 선택한다. 따라서 현재 학습하고자 하는 클래스와 관련 없는 텍스트 또는 잘못된 텍스트들은 결정 경계를 선택함에 있어 영향을 끼치지 않게 된다. 이로 인해 학습 시 자동으로 긍정 오류를 최소화할 수 있다.

본 시스템에서는 사용자가 SNS에서 작성한 텍스트를 Table 1에 나열된 자질들을 사용하여 표현하였다. 사용된 자질을 살펴보면, 기본적으로 문서 분류에서 널리 사용되는 Bag-of-Word가 사용되었다. 이때, 모든 단어를 사용하지 않고, 형태소 분석을 통해 얻어진 명사, 미등록어 및 기호만을 사용하였다. 추가로, SNS 텍스트가 이모티콘, 알파벳 반복 등이 자주 사용되는 특성[13]을 반영하기 위해서, 이모티콘, 알파벳 반복 유무 및 문자 반복 유무등의 자질을 사용하였다. 이모티콘을 반영하기 위해 SNS에서 널리 쓰이는 이모티콘을 수집하여 이모티콘 벡터를 미리 정의하였다. 마찬가지로 알파벳 반복 유무 및 문자 반복을 반영하기 위해 특정 패턴들을 미리 구축하여 사용했다.

다중 인스턴스 학습을 사용하여 개인 정보 식별 모델을 구축할 경우, 어떤 클래스를 긍정으로 선택하느냐에 따라 모델의 결정 경계가 달라진다. 예를 들어, 성별을 식별함에 있어, 긍정 클래스를 남성으로 선택하면, 긍정이 아닌 부정, 즉 여성에는 모두다 여성과 관련된 글만 존재해야 함에도 불구하고 실질적으로는 그렇지 않게 된다. 본 논문에서는 위 문제점을 해결하기 위해 클래스별 모델을 따로 구축한다. 즉, 앞선 예에서 긍정 클래스를 남성으로 한 후 남성 모델을 구축하고, 긍정 클래스를 여성으로 한 후 여성 모델을 구축한다. 새로운 사용자가 입력되면 클래스별로 마진을 계산한다. 마진이 가장 큰 클래스를 선택하여 최종적으로 개인 정보를 식별한다.

4.2 모빌리티 기반 프로파일 식별 모델

사용자 모빌리티 정보에는 사용자를 식별할 수 있는 단서들이 포함되어 있다. 예를 들어, 여성이 남성에 비해 미용실에 자주 가고 오래 머물러 있는 반면, 남성이 여성에 비해 오락실 및 축구장을 주로 방문한다. 또한, 10대, 20대들은 집에서 학교로, 학교에서 집으로 이동하는 반면, 40대, 50대는 집에서 회사, 회사에서 집으로의 이동 히스토리를 보인다. 한 사용자가 미용실에 자주 가고 오래 머물러 있으며, 집에서 학교, 학교에서 집으로 주로 이동한다면 앞서 사용자들의 이동 정보 중 여성과 10대, 20대와 유사함을 통해 이 사용자를 10대, 20대 여성으로 식별할 수 있다. 즉, 모빌리티 간 유사성을 이용하여 사용자의 속성을 식별할 수 있다.

Table 2. Features for mobility based model

Features	Example
Frequencies of the point[1)]	Nail salon: 2 times Coffee shop: 4 times ...
Time slot of the point	Coffee shop: 1 at afternoon, 3 times at evening Football statdium: 2 times at morning, 1 at evening ...
Stay time of the point	Nail salon: 200 min. Coffee shop: 360 min. ...
Point to Point	From Home to School: 3 times From School to Coffee shop: 3 times ...

사용자 프로파일을 식별하기 위해 먼저, 상황인지 플랫폼[11]을 이용하여 스마트 폰, 테블렛 PC(Tab.t PC)와 같이 GPS 센서가 부착된 기기로부터 매초마다 기록된 GPS 데이터를 수집한다. 수집된 사용자의 모빌리티 정보를 기반으로 POI에 대한 사용자의 패턴을 알 수 있다.

사용자의 정보를 식별하기 위해서는 모빌리티 정보를 d 차원의 벡터로 표현해야 한다. 개인 정보 식별 모듈에서는 Table 2에서 설명된 자질들을 사용하여 모빌리티 정보를 표현한다. 거점 방문 빈도(frequencies of the point)는 거점에 방문한 빈도수를 나타낸다. 이는 거점에 상대적으로 자주 방문하는 것을 반영하기 위함이다. 다음으로 거점 별 방문 시간대(time slot of the point)는 거점에 방문한 시간대를 오전(06~12시), 오후(12~18시), 저녁(18~24시), 밤(24~06시)으로 4등분하여 표현한 것으로 같은 거점을 방문하였을지라도 방문 시간대가 다르면 다른 의도로 방문함을 표현하기 위함이다. 거점 별 머문 시간(stay time of the point)는 방문 거점에 얼마나 머물렀는지를 나타내며, 이는 같은 거점을 방문하더라도 머문 시간이 특정 클래스에 따라 다른 경우를 반영하기 위함이다. 예를 들어, 앞선 예에서 여성과 남성이 미용실을 방문하였을지라도 여성이 오래 머무는 것을 표현할 수 있다. 마지막으로 거점에서 거점(point to point)은 거점에서 거점간의 이동 정보를 표현하며, 같은 거점을 방문하였을지라도 거점 사이의 방문 순서를 통해 나타나는 특징을 반영한다.

자질로 표현된 사용자 모빌리티 정보로부터 개인 정보를 식별하기 위해 본 논문에서는 k-Nearest Neighborhood (kNN) 알고리즘을 사용한다. kNN 알고리즘은 클래스를 추정함에 있어 주어진 인스턴스와 "가까운" k개의 인스턴스로부터 클래스를 추정하는 알고리즘이다. kNN 알고리즘을 통해 본 논문에서는 주어진 사용자의 모빌리티 패턴과 유사한 k명의 사용자 개인 정보들로부터 사용자 모빌리티 프로파일을 식별하게 된다. 본 연구에서는 자질별 중요도가 반영된 코사인 유사도(cosine similarity) 함수를 사용한다. 자질별 중요도는 학습 시 선형 모델(linear model)을 사용하여 미리 계산해둔다. 마지막으로 유사한 사용자들로부터 과반수 투표 방식(majority voting)을 사용하여 최종적으로 사용자 정보를 식별한다.

1) 사용자가 방문한 거점의 이름은 주어진다고 가정한다.

4.3 통합 사용자 개인 정보 식별 모듈

모빌리티 정보와 소셜 미디어 텍스트로부터 식별된 개인 정보들은 서로 다른 특징이 반영되어 있다. 예를 들면, 사용자의 정치적 성향은 모빌리티 정보로는 잘 들어나지 않는 반면, 텍스트에서는 사용자의 정치적 성향이 잘 들어난다. 성별의 경우, 거점 방문 횟수 및 머문 시간 등이 포함된 모빌리티들이 텍스트에 비해 더 많은 정보를 가진다. 각 식별된 사용자 정보들은 최종적으로 하나로 통합되어야 하며, 이를 위해 단순히 합치는 것이 아닌 속성에 따라 모델별로 서로 다른 중요도를 가지고 결합하여야 한다.

제안한 시스템에서는 사용자의 최종 개인 정보를 식별하기 위해 모빌리티 모델과 텍스트 모델을 선형 모델로 결합한다. 사용자 i가 주어질 때, 사용자의 최종 개인 정보 y_{iv}^{*}은 수식 (3)과 같이 정의될 수 있다.

$$y_i^{*} = \underset{v \in V}{\operatorname{argmax}}\, y_{iv}^{*} \tag{3}$$

V는 속성별 클래스의 집합을 나타낸다. 예를 들어, 속성이 성별인 경우, V = {남성, 여성}이 된다. $v \in V$는 속성이 가지는 개별 클래스를 나타낸다. 각 클래스별 사용자 i의 개인 정보 y_{iv}^{*}는 수식 (4)로 계산된다.

$$y_{iv}^{*} = w_{Tv} Text_v(x_i) + w_{Mv} Mobility_v(x_i)$$
$$w_{Tv} + w_{Mv} = 1 \tag{4}$$

$Text_v(\cdot)$는 클래스 v에 대해 텍스트 모델에서 나온 결과값을 나타내며, $Mobility_v(\cdot)$는 클래스 v에 대해 모빌리티 모델에서 나온 결과값을 나타낸다. w_{Tv}는 클래스 v에 대한 텍스트 모델의 중요도를 나타내며, w_{Mv} 클래스 v에 대한 모빌리티 모델의 중요도를 나타낸다. 각 모델의 중요도 w_{Tv}와 w_{Mv}는 주어진 학습 데이터의 에러를 최소화하도록 결정된다. 에러는 클래스 v에 대한 정답인 사용자 i의 클래스 y_{iv}와 모델에서 추정한 클래스 y_{iv}^{*}의 차로 정의할 수 있다.

$$E = (y_{iv} - y_{iv}^{*}) \tag{5}$$

텍스트 모델의 중요도 w_{Tv}는 다음과 같이 에러 E를 최소화하도록 중요도를 업데이트 한다.

$$w_{Tv} = w_{Tv} + \Delta w_{Tv}$$

$$\Delta w_{Tv} = \lambda (y_{iv} - y_{iv}^{*}) Text_v(x_i)$$

λ는 학습 비율로, 얼마만큼 업데이트를 할지에 대한 파라미터이다. 즉, 수식 (5)인 에러 E에 현재 텍스트 프로파일 모델의 결과값과 학습 비율을 곱하여 중요도를 업데이트 한다. 모빌리티 모델의 중요도 w_{Mv} 또한 텍스트 모델과 동일한 방식으로 업데이트를 하며, 다음과 같다.

$$w_{Mv} = w_{Mv} + \Delta w_{Mv}$$

$$\Delta w_{Mv} = \lambda(y_{iv} - y^*_{iv}) Mobility_v(x_i)$$

위 중요도 업데이트는 학습 사용자에 대해 반복적으로 수행하며, 더 이상 모델의 중요도에 변화가 없거나 기존의 중요도와 현재 업데이트된 중요도의 차이가 특정 값 이하가 될 때까지 반복한다.

5. 소셜 정보 식별

5.1 관계 별 근접성 패턴 학습

소셜 정보 식별 모듈은 먼저, 사용자와 폰북(phone book)의 인물 사이에 유의미한 사회적 관계가 있다고 가정한다. 이러한 가정을 바탕으로 소셜 정보 식별 모듈은 사용자 간 가족, 친구, 동료 이렇게 세 가지 사회적 관계를 식별한다.

사회적 관계 분류를 위해 먼저 두 사용자 간의 시공간적 상황 공유 여부를 인지한다. 상황인지 플랫폼[11]을 이용해 각 사용자들로부터 POI에 대한 레이블, 행동 유형 등의 상황 정보를 축적한다. 두 사람으로부터 이러한 상황 정보를 얻으면, 그들이 어떤 POI에 함께 머문 시간을 얻을 수 있다. 이 때 두 사람이 POI에 근접해 있는지 여부는 이들의 지리적 위치가 해당 POI에 특정 임계값 이하의 거리인지에 따라 결정한다.

소셜 정보 식별 모듈은 두 사람이 시공간적 상황을 공유할 확률 분포를 정규 분포 혼합 모델로 가정한다. 즉, 어떤 시간 t이후 1시간 동안, 특정 POI l에 두 사람이 함께 있을 확률 분포 f_l을 수식으로 정의하면 다음과 같다.

$$f_l(t;\theta) = \sum_{k=1}^{K} p_k g_l(t; m_k, \sigma_k),$$

$$where\ g_l(t; m_k, \sigma_k) = \frac{p_l}{\sqrt{2\pi}\sigma_k} e^{-\left(\frac{\pi}{12}\right)^2 \frac{(t-m_k)^2}{2\sigma_k{}^2}},$$

$$and\ \theta = ((p_1, m_1, \sigma_1), \ldots, (p_K, m_K, \sigma_K))$$

가정에 따라 두 사람이 함께 있을 확률 분포는 K개의 정규 분포들로 구성된다. 확률 p_k는 각 정규 분포에 대한 사전확률이다. 함수 g_l은 m_k, σ_k를 각각 평균과 분산으로 하는 k번째 정규 분포를 의미한다. 확률 p_l은 시간에 독립적으로 두 사람이 l에 함께 있을 확률을 의미한다. 정규 분포 확률을 정의할 때, 하루 24시간이 원형의 연속성을 갖는 특징을 고려하였다 [9]. 사용자 POI가 L개라고 할 때, 각 POI에 대해 동일한 방법으로 사용자 간 상황 공유 분포를 정의할 수 있다.

확률 분포 학습을 위해 두 사람이 시간대 t에 POI l에 함께 머문 초 단위 횟수만큼 시간대 t를 생성한다. 두 사람이 해당 POI에 N번 함께 머문 시간대 정보가 주어졌을 때, 각 시간대 t를 생성할 수 있는 확률 분포 파라미터 Θ는 다음의 최적화 문제를 통해 얻을 수 있다.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} \sum_{k=1}^{K} p_k g_l(t_n; m_k, \sigma_k)$$

본 논문은 확률 분포 f_l을 학습하기 위해 EM알고리즘을 이용하였다. E-step에서는 현재까지 추정된 파라미터 Θ를 이용해 어떤 시간대 t가 주어졌을 때, k번째 정규 분포로부터 시간대 t가 생성될 확률을 추정하고, M-step에서 앞서 추정된 확률을 이용해 위 수식을 만족하는 Θ를 새롭게 결정한다. 동일한 방법으로 각 POI에 대해 두 사람이 함께 있을 확률 분포를 추정한다.

관계에 대한 확률적 패턴은 관계가 알려진 여러 개의 사람 쌍들로부터 얻은 시간대 정보를 이용해 학습된 확률 분포로 정의한다. 이때, 지리적으로 서로 다른 POI들 중 사용자의 위치적 상황(context)이 같은 경우 동일한 것으로 보고, 확률 분포를 추정한다. 위치적 상황은 상황 인지 프레임워크로부터 얻은 POI 레이블을 이용하여 정의한다. 본 논문은 위치적 상황을 집, 직장, 그 외 장소(ETC) 이렇게 세 가지로 정의한다. 두 사용자 간의 상황 공유 분포 또한 이러한 위치적 상황에 따라 정의할 수 있다.

Fig. 2는 대학교에 재학 중인 학생과 대학원생 16명을 대상으로 한 달 간 수집한 데이터로부터 관계별 확률적인 centroid를 학습한 결과이다. 각 그래프는 시간에 따라 어떤 장소에서 두 사람이 만날 확률을 보이고 있다. K는 각 관계에 대해 추론된 정규 분포의 개수이다. Fig. 2에서 일과 중에 직장 동료 간에 만날 확률이 높게 나타나고 가족의 경우 이러한 확률이 전혀 없다. 반면 가족 관계는 주중 일과 이후, 주말 대부분의 시간을 집에서 함께 보내고, 친구나 동료가 집을 방문하는 일은 드물게 나타난다. 이와 같이 확률적 패턴을 통해 임의의 시공간적 상황에 대한 관계별 특징을 시각화하여 표현할 수 있다.

5.2 확률 패턴을 이용한 소셜 정보 식별 모델

소셜 정보 식별 모듈은 기본적으로 관계별 패턴와 두 사용자 간 상황 공유 패턴의 유사도를 이용하여 이들 간의 관계를 식별한다. 이를 위해, 두 사람이 어떤 관계를 가질 확률이 이러한 유사도에 선형적으로 비례한다고 가정한다. 이때, 관계 별 유사도의 중요도를 학습하기 위해 로지스틱 회귀 모델 (logistic regression model)을 이용한다. 두 사용자 간의 상황 별 확률 분포의 모음 F가 주어졌을 때, 두 사람의 관계 y가 $r \in R$ = {가족, 친구, 동료}일 확률은 다음과 같이 정의된다.

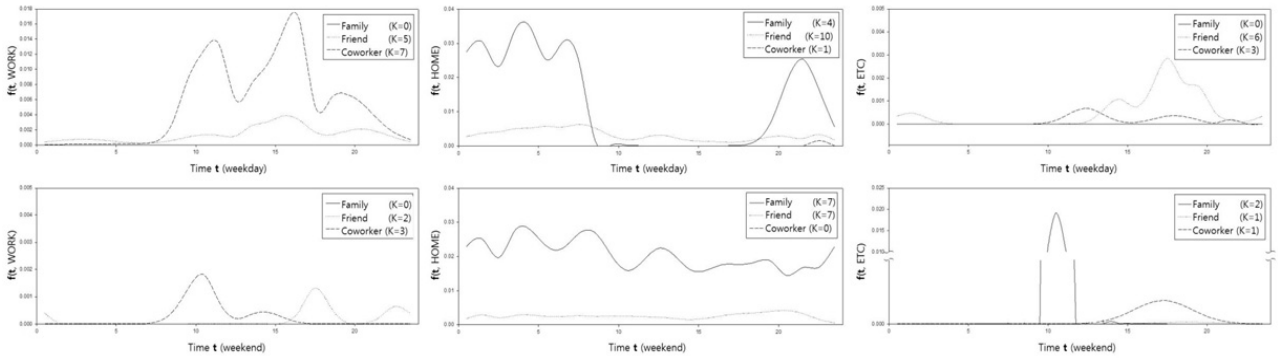


Fig. 2. Patterns for social relations estimated from university students and graduate school students

$$p(y=r|F)=\frac{e^{w_r{}^T\emptyset(F)}}{\sum_j^R e^{w_j{}^T\emptyset(F)}}, where\ F\equiv\{f_{home}, f_{work}, f_{etc}\}$$

함수 ϕ는 입력 F의 각 분포를 관계 별 동일 상황의 확률 패턴 과의 유사도로 맵핑하여 관계와 상황 별 유사도 벡터를 반환한다. 벡터 w_r은 관계 r에 대한 맵핑된 유사도의 가중치이다.

하루 동안 두 사람이 어떤 위치적 상황을 공유할 확률은 한 시간 단위의 이산 확률로 정의하여 벡터로 표현할 수 있다. 따라서 확률 패턴에 기반한 유사도는 벡터 유사도 측정 방법으로 얻을 수 있다.

5.3 개인 정보에 기반한 소셜 정보 검증

두 사용자의 패턴과 관계별 패턴으로부터 우리는 쉽게 두 사용자가 특정 관계를 가질 확률을 추정할 수 있다. 추정된 결과를 바탕으로 가장 확률이 높은 관계를 선택하여, 두 사용자 간의 관계를 식별할 수 있다. 하지만, 앞서 설명한 확률 모델을 이용한 이와 같은 방법의 관계 식별은 완벽할 수 없기에, 오류를 피할 수 없다. 여러 오류들 중에는 사람과 사람 사이의 관계를 고려했을 때, 쉽게 해결할 수 있는 것들이 있다. 제안한 시스템에서는 사용자의 소셜 정보뿐 만 아니라, 사용자의 개인 정보도 식별하고 있다. 따라서 식별된 관계에 대해, 사용자의 개인 정보를 바탕으로 검증이 가능하다. 이를 위해 사람과 사람 사이의 관계를 명세하는 온톨로지를 구축하였다. 구축된 온톨로지를 통해 식별된 관계가 개개의 사용자 속성 측면에서 타당한지 검증하고 사람이 봤을 때, 납득할 수 없는 오류들을 줄인다.

Fig. 3은 제안한 시스템에서 사용하는 사람 사이의 관계를 표현한 온톨로지이다. 그림에서 보듯이, 온톨로지는 “person”이라는 concept을 중심으로 사람 사이의 관계에 나타나는 제약들을 표현했다. 예로, 부부 관계는 2명 사이에 1:1로 나타나는 관계이며, 부모-자식 관계는 1:n으로 부모는 여러 자식을 가질 수 있다. 뿐만 아니라, “person” concept는 나이와 성별을 속성으로 가지며, 관계를 검증할 때, 사용한다. 예로 부모-자식 사이에는 어느 정도 나이 차이를 보여야 한다던가, 부부 사이의 두 사람은 성별이 달라야한다 등의 추론 규칙을 함께 사용하여 식별된 관계를 검증한다. 기존의 연구들은 개인 정보나 소셜 정보 중, 하나를 식별하고자 하였기에 이러한 검증을 사용할 수 없다. 하지만 제안한 시스템은 사람의 개인 및 소셜 정보를 함께 식별함으로서 서로 간의 정보 공유를 통해 식별 성능을 올릴 수 있다.

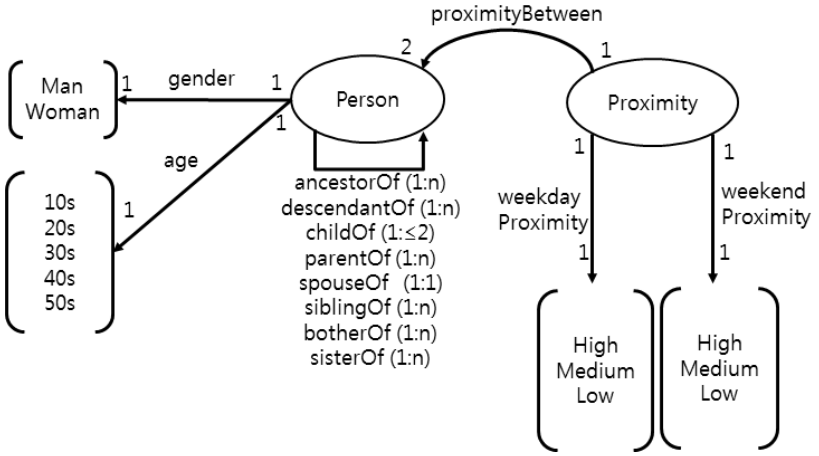


Fig. 3. Ontology for validation of identified relations

6. 실 험

6.1 개인 정보 식별 성능

개인 정보 식별 성능 검증을 위해 대학교 재학생 및 대학원생, 15명의 사용자가 2달 동안 이동한 경로를 사용하였다. 사용자의 이동 경로를 1주일 단위로 Table 2에서 정의한 자질을 사용하여 표현하였다. 텍스트 기반 식별 방법의 경우, Facebook, Twitter, me2day에서 사용자와 사용자의 정보를 수집하여 Table 1에서 정의한 자질을 사용하여 표현했다. Facebook을 통해 수집한 사용자의 정보는 명시적으로 나타나 있는 것을 사용하였으며, Twitter나 me2day의 경우 사용자가 작성한 텍스트와 약력 등으로부터 수작업으로 구축하였다. Table 3은 실험에 사용한 학습 데이터 및 실험 데이터의 통계 정보를 보여준다. 모빌리티 사용자 데이터 수집의 어려움으로 수집한 사용자는 총 15명이다. 실험에서 각 데이터 인스턴스는 1주일 단위의 모빌리티 데이터를 표현한다. 따라서, 모빌리티 데이터의 총 인스턴스 수는 120개이며, 이중 40개(5명)는 학습을 위해, 나머지 80개(10명)는 테스트를 위해 사용되었다.

본 논문에서는 사용자 속성 중 성별(gender), 나이(age), 연애 유무(marital status)를 식별하였다. 성별의 경우, 클래스는 남성, 여성으로, 나이는 10대-20대, 30대, 40대-50대, 연애 유무는 싱글, 연애 중, 기혼으로 정의하였다. 비교 모델(baseline)로는 모빌리티 모델의 경우 거점 방문 빈도만을 사용한 kNN을 사용하였으며 [21], 텍스트 모델의 경우 텍스트 개별을 인스턴스로 사용하는 Single-Instance SVM (SI-SVM)를 사용하였다 [13, 14]. SI-SVM에서는 사용자가 작성한 텍스트 각각에 대해 클래스를 추정하고, 추정된 클래스 중 가장 많이 분류된 클래스를 사용자 정보로 식별하였다. 비교 모델의 kNN과 본 논문의 모빌리티 기반 모델의 k는 3으로 고정하였다. SI-SVM의 분류기로는 LIBSVM [22]을 사용하였으며 제안한 방법과 LIBSVM에서 모두 선형 커널(linear kernel)을 사용하였다. 다중 클래스를 처리하기 위해 본 논문에서는 One-Versus-All [23] 방법을 사용하였다. 모든 실험은 임의 추출(random samling)을 통해 5회 반복 수행하여 유의미한 결과를 내고자 했다.

Table 3. Simple statistics of experimental data for personal information identification

	Information	Value
Training	No. of users for text model	260
	No. of users for mobility model	5
	No. of social texts	3,854
	Average No. of social texts per user	14.82
Test	No. of users	10
	No. of social texts	114
	Average No. of social texts per user	11.4
	No. of places to visit	48

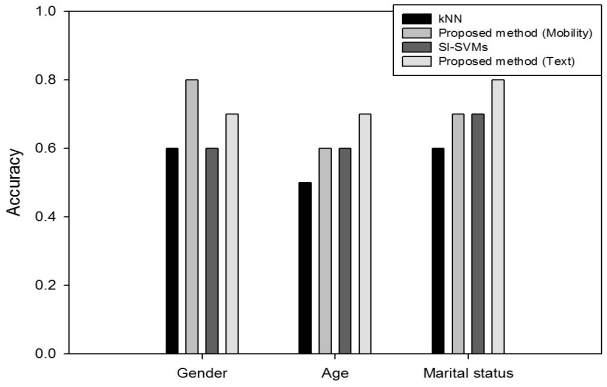


Fig. 4. Experimental results on personal information identification

Fig. 4는 성별, 나이, 연애 유무에 대해 모빌리티와 텍스트 관점에서 제안한 개별 모델과 비교 모델의 성능을 보여준다. Fig. 4에서 보듯이 제안한 방법이 모든 속성에 대해 비교 모델보다 좋은 성능을 보인다. 모빌리티에서는 거점별 방문 빈도 수 뿐만 아니라 본 논문에서 제안한 자질이 프로파일을 식별함에 있어 도움을 주는 것을 볼 수 있다. 텍스트에서는 싱글 인스턴스 기반 방법보다 다중 인스턴스 기반 방법이 프로파일 식별에 도움을 주는 것을 볼 수 있다. 이를 통해 제안한 방법들이 프로파일 식별에 도움이 되는 텍스트만이 고려되었음을 의미한다.

개별 분류 방법의 성능 검증에서 성별의 경우, 모빌리티를 이용한 방법이, 나이나 연애 유무의 경우 텍스트에 기반한 방법이 더 나은 성능을 보였다. 이는 두 데이터가 사용자의 서로 다른 측면을 고려하고 있음을 의미한다. 따라서 이들 두 방법을 결합하여 제안한 개인 정보 식별 모듈은 실험에서 더 나은 성능을 보였다. Table 4는 개인 정보 식별 모듈의 성능을 보여 준다. 표에서 보이듯이, 성별에 대해서는 0.9의 정확도를, 나이에 대해서는 0.8의 정확도, 연애 유무에 대해서는 0.9의 성능을 보였다. 위 성능은 Fig. 4에서 언급한 모빌리티 또는 텍스트 모델만을 사용한 성능보다 높으며, 이는 각각 모델들의 결과를 합쳐 그 사용자 개인 정보를 식별하는 모델이 유의미하다는 것을 입증한다.

Table 4. Performance of personal information identification module

Information	Accuracy
Gender	0.9
Age	0.8
Marital status	0.9

6.2 소셜 정보 식별 성능

소셜 정보 식별 모듈의 성능은 개인 정보 식별에 사용된 15명의 모빌리티 데이터를 이용하여 검증하였다. 이들 실험 참가자들 사이에 각 1주일 데이터로부터 추론된 상황 공유

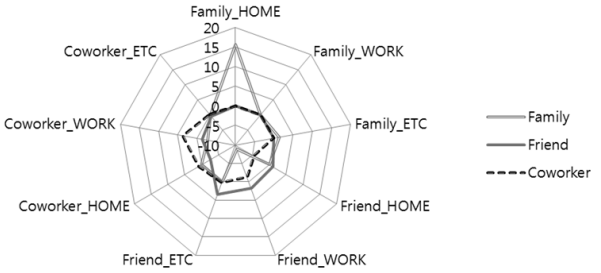


Fig. 5. Weights of patterns on POIs for classes

확률 분포를 하나의 인스턴스로 사용하였다. 학습에 사용된 총 인스턴스는 380개이다. 성능 평가를 위해 5회 교차 검증을 수행하였다.

Fig. 5는 제안한 소셜 정보 식별 모듈에서 학습한 각 유사도 자질 별 가중치를 보여준다. 9각형의 꼭지점은 관계와 상황 별 패턴과의 유사도에 해당한다. 유사도의 가중치가 높을수록 각 꼭지점에 가까운 좌표로 표현된다. 가족 관계 추론 시에는 가족이 집에 함께 있을 확률적 패턴이 중요한 요소로 표현되었다. 친구와 동료 관계의 경우, 각각 친구와 그 밖 장소(ETC)에 있을 패턴과 동료와 직장에 있을 패턴이 중요한 요소가 된다. 가족 관계 추론 시, 친구와 직장에 있을 패턴에 대해 음의 가중치를 부여함으로써 입력 분포와의 유사도가 작을수록 가족 관계로 분류할 가능성이 높다.

평가를 위해 제안한 소셜 정보 식별 모듈에서 사용한 로지스틱 회귀 모델(LR), 분류 문제에 우수한 SVM, 그리고 베이지안 네트워크(BN)를 비교하였다. 제안한 방법과 동일한 방법으로 각 모델에 맵핑 함수 ϕ를 적용하였다.

Table 5는 네 가지 모델에 대한 성능 평가 결과이다. 모든 모델이 가족 관계에 대해 정확한 분류 결과를 보였다. 가족 관계는 다른 관계들과 확연히 구분되는 특징이 있다. 예를 들어, 가족 관계는 주말 거의 모든 시간 집에 함께 있는 반면 주중 직장에 함께 있는 경우는 없다. 친구 관계에 비해 동료 관계에 대한 분류 성능이 높은 것은 비교적 규칙적인 패턴을 보이기 때문으로 이해된다.

Mandan et al.[24]은 본 논문과 비슷한 조건의 실험 데이터를 이용해 SVM과 베이지안네트워크 모델을 이용한 친밀도 분류 결과를 보고한 바 있다. 실험 설정 상 모델 성능에 대한 직접적인 비교는 어렵지만, 본 논문의 실험 설정에서는 제안한 방법이 이들 두 모델에 비해 높은 성능을 보였다.

친구와 동료 관계의 경우 제안한 식별 모듈은 다른 모델에 비해 매우 우수한 성능을 보였다. 친구 관계에 대해서는 10%이상 더 높은 F-measure를 동료 관계에서는 5% 높은 성능을 보였다. 이는 온톨로지를 활용한 검증 단계의 효과를 입증한다. 다만, 가족 관계의 경우, 기존의 방법들이 모두 1.00의 성능을 보이는데 반해 제안한 방법은 5% 낮은 성능을 얻었다. 이는 개인 정보 식별 모듈의 오류에서 오는 성능 저하로 보이며, 이와 관련한 추후 연구가 필요할 것으로 보인다. 그럼에도 불구하고, 소셜 정보 식별 모듈의 전체 정확도는 85%로 타 방법에 비해 제안한 시스템의 소셜 정보 식별 모듈이 매우 우수함으로 입증하였다.

7. 결 론

본 논문에서는 사용자의 다양한 정보를 식별하기 위한 시스템을 제안하였다. 작은 크기의 고성능 디바이스가 발달하고, 보급되는 현재, 다양한 서비스에서 사용자 정보에 대한 요구가 높아지고 있다. 제안한 시스템은 이와 같은 요구에 맞춰 사용자의 개인 정보 및 소셜 정보를 자동으로 식별한다. 제안한 시스템에서의 식별은 사용자의 온라인 행동 패턴인 SNS 데이터와 오프라인 행동 패턴인 GPS 데이터에 기반하여 사용자의 나이, 성별 등의 개인 정보 뿐만 아니라, 사용자 간의 친구, 동료, 가족 등 관계를 식별한다. 이는, 개인 정보 혹은 소셜 정보 등 하나의 정보만을 추정하는 기존 방법에 비해 진일보한 방식이다.

실험에서는 각 식별 모듈이 기존의 방법에 비해 뛰어남을 보였다. 이는 제안한 시스템에 탑재한 각 모듈이 데이터의 특성을 효과적으로 다루고 있음을 의미한다. 통합된 사용자 식별 시스템의 입장에서는, 두 종류의 정보를 함께 식별함으로써 더 나은 결과를 유도할 수 있음을 보였다. 실험에서

Table 5. Performances on social information identification

방법	acc.	class	precision	recall	F-measure
SVM	72.6%	Family	1.00	1.00	**1.00**
		Friend	0.74	0.49	0.52
			0.76	0.80	0.76
BN	75.0%	Family	1.00	1.00	**1.00**
		Friend	0.73	0.58	0.61
		동료	0.78	0.79	0.77
LR	79.7%	Family	1.00	1.00	**1.00**
		Friend	0.78	0.50	0.61
		동료	0.80	0.90	0.84
Proposed Method	**85.0%**	Family	1.00	0.90	0.95
		Friend	0.93	0.59	**072**
		동료	0.82	0.98	**0.89**

는 소셜 정보를 식별한 후, 이를 개인 정보에 기반하여 검증함으로써 기존 방법에 비해 제안한 시스템이 오류를 효과적으로 줄일 수 있음을 보였다. 결론적으로 이러한 결과는 모바일 디바이스 사용자 식별에 제안한 방법이 매우 효과적임을 의미한다.

참 고 문 헌

[1] Q. Li, Y. Zheng, X. Xie, and W.-Y. Ma, "Mining User Similarity based on Location History," In *Proceedings of GIS2008*, pp.298-307, 2008.

[2] C. Boulis and M. Ostendorf. "A Quantitative Analysis of Lexical Differences between Genders in Telephone Conversations," In *Proceedings of ACL*, pp.435-442, 2005.

[3] N. Garera and D. Yarosky, "Modeling Latent Biographic Attributes in Conversational Genres," In *Proceedings of ACL and IJCNLP*, pp.710-718, 2009.

[4] J. Otterbacher, "Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata," In *Proceedings of CIKM*, pp.369-378, 2010.

[5] L. Hemphill and J. Otterbacher, "Learning the Lingo? Gender, Prestige and Linguistic Adaptation in Review Communities," In *Proceedings of CSCW*, pp.305-314, 2012.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. "Solving the Multiple-Instance Problem with Axis-Parallet Rectangles," *Artificial Intelligence*, Vol.89, No.1-2, pp.31-71, 1997.

[7] J. Goldenberg and M. Levy. "Distance is Not Dead: Social Interaction and Geographical Distance in the Internet Era," CoRR, abs/0906.3202, 2009.

[8] N. Eagle, A. Pentland, and D. Lazer, "Inferring Friendship Network Structure by Using Mobile Phone Data," In *Proceedings of the National Academy of Sciences of the United States of America*, pp.15274-15278, 2009.

[9] E. Cho, S. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," In *Proceedings of SIGKDD*, pp.1082-1090. 2011.

[10] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. "Inferring Social Ties from Geographic Coincidences," *PNAS*, 2010.

[11] H.-J. Lee, J.-H. Choi, and Y.-T. Park, "Semantic Point of Interest Detection from Large-scale GPS Data of Mobile Users," *KISSE: Software and Application*, 39(3), pp.175-184, 2012.

[12] P. W. Foltz, D. Laham, and T.K Landauer, "Automated Essay Scoring: Applications to Educational Technology," In *Proceedings of EdMedia*, 1999.

[13] D. Rao, D. Yarosky, A. Shreevats, and M. Gupta, "Classifying Latent User Attributes in Twitter," In *Proceedings of SMUC*, pp.37-44, 2010.

[14] M. Pennacchiotti and A.-M. Popescu, "A Machine Learning Approach to Twitter User Classification," In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp.281-288, 2011.

[15] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating Gender on Twitter," In *Proceedings of the EMNLP,* pp.1301-1309, 2011.

[16] O. Maron and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," In *Proceedings of ICML*, pp.341-349, 1998.

[17] C. Yang and T. Lozano-Perez, "Image Database Retrieval with Multiple-Instance Learning Techniques," In *Proceedings of ICDE*, pp.233-243, 2000.

[18] J. Ramon and L. De Raedt, "Multi Instance Neural Networks," In *Proceedings of ICML-2000 Workshop on Attribute-Value and Relational Learning*, 2000.

[19] J. Wang and J.-D. Zucker, "Solving the Multiple-Instance Problem: a Lazy Learning Approach," In *Proceedings of ICML*, pp.1119-1125, 2002.

[20] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multi-Instance Learning", In *Advances in NIPS*, pp.577-584, 2002.

[21] Y. Matsuo, N. Okazaki, K. Izumi, Y. Nakamura, T. Nishmura, K. Hasida, and H. Nakashima, "Inferring Long-term User Properties based on Users' Location History," In *Proceedings of IJCAI*, pp.2159-2165, 2007.

[22] C. Chang and C. Lin, "LIBSVM : a Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, 2(27), pp.1-27, 2011.

[23] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *Journal of Machine Learning Research*, Vol.5, pp.101-141, 2004.

[24] A. Madan, and A. Pentland, "Modeling Social Diffusion Phenomena Using Reality Mining," In *Proceedings of AAAI Spring Symposium on Human Behavior Modeling*, 2009.

손 정 우

e-mail : jwson@sejong.knu.ac.kr

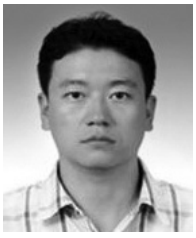
2005년 경북대학교 컴퓨터공학과(학사)

2007년 경북대학교 컴퓨터공학과(석사)

2012년 경북대학교 전자전기컴퓨터학부 (박사)

2013년~현 재 한국전자통신연구원 연구원

관심분야 : 기계학습, 자연어처리, 온톨로지



한 용 진

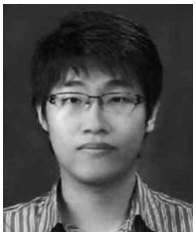
e-mail : yjhan@sejong.knu.ac.kr

2006년 경북대학교 컴퓨터공학과(학사)

2008년 경북대학교 전자전기컴퓨터학부 (석사)

2008년~현 재 경북대학교 전자전기 컴퓨터학부 박사과정

관심분야 : 시맨틱웹, 자연어처리, 기계학습



송 현 제

e-mail : hjsong@sejong.knu.ac.kr
2008년 경북대학교 컴퓨터공학과(학사)
2010년 경북대학교 전자전기컴퓨터학부 (석사)
2010년~현 재 경북대학교 전자전기 컴퓨터학부 박사과정
관심분야 : 기계학습, 자연언어처리

박 성 배

e-mail : seongbae@knu.ac.kr
1994년 한국과학기술원 컴퓨터과학과 (학사)
1996년 서울대학교 컴퓨터공학과(석사)
2002년 서울대학교 컴퓨터공학과(박사)
2004년~현 재 경북대학교 IT대학 컴퓨터학부 교수
관심분야 : 기계학습, 자연언어처리, 텍스트 마이닝, 정보추출, 생명정보학

이 상 조

e-mail : sjlee@knu.ac.kr
1974년 경북대학교 수학교육과(학사)
1976년 한국과학기술원 전산학과(석사)
1994년 서울대학교 컴퓨터공학과(박사)
1976년~현 재 경북대학교 IT대학 컴퓨터학부 교수
관심분야 : 자연어처리, 기계번역, 정보검색, 정보추출