Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.23, NO.6, Dec. 2013 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2013.23.6.1103

소셜 네트워크 서비스에 노출된 개인정보의 소유자 식별 방법

김 석 현,† 조 진 만, 진 승 헌, 최 대 선‡
한국전자통신연구원

A Method of Identifying Ownership of Personal Information exposed in Social Network Service

Seok-hyun Kim,† Jin-man Cho, Seung-hun Jin, Dae-seon Choi‡
Electronics and Telecommunications Research Institute

요 약

본 논문에서는 소셜 네트워크 서비스 상에 공개된 개인정보의 소유자 식별 방법을 제안한다. 구체적으로는 트위터 상에 언급된 지역 정보가 게시자의 거주지를 의미하는지를 자동으로 판단하는 방법이다. 개인정보 소유자 식별은 특정인의 개인정보가 온라인 상에 얼마나 노출되어 있는지 파악하여 그 위험도를 산정하기 위한 과정의 일부로서 필수적이다. 제안 방법은 트윗 문장의 어휘 및 구조적 특징 13개를 자질(feature set)로 활용한 소유자 식별 규칙들을 통해 지역정보가 게시자의 거주지를 의미하는지 판단한다. 실제 트위터 데이터를 이용한 실험에서 제안방법이 n-gram을 자질로 사용한 나이브베이지안 같은 전통적인 문서 분류 모델보다 더 높은 성능 (F1값 0.876)을 보였다.

ABSTRACT

This paper proposes a method of identifying ownership of personal information in Social Network Service. In detail, the proposed method automatically decides whether any location information mentioned in twitter indicates the publisher's residence area. Identifying ownership of personal information is necessary part of evaluating risk of opened personal information online. The proposed method uses a set of decision rules that considers 13 features that are lexicographic and syntactic characteristics of the tweet sentences. In an experiment using real twitter data, the proposed method shows better performance (f1-score: 0.876) than the conventional document classification models such as naive bayesian that uses n-gram as a feature set.

Keywords: Big Data, Pricacy, Personal Information, Social Network Service, Location Information

I. 서 론

2013년 2월 한글 트위터(twitter)이용자 수는 약 600만 명으로 추정되며, 2012년 9월부터 올해 3월까지 한 달 평균 트윗 개수는 약 1억 5천만 개[1], 페이스북(facebook) 이용자는 월 1천 100만 명 이상이 사용하고 있다[2]. 이런 현상은 소셜 네트워크 서비스(Social Network Service; SNS)가 이용자들 간에 새로운 정보 생성과 공유의 수단으로 널리 이용되고 있음을 의미한다. 하지만 이용자들에 의해서 공개적으로 게시되는 정보에는 본인 또는 타인의 이름, 거주지, 생일 등과 같은 신상정보가 노출되는 경우가 많이 발생되고 있다. 게다가 소셜 네트워크 서비스인 트위터는 오픈 서비스로 제공되고 있기 때문에 트위터에

접수일(2013년 7월 8일), 수정일(2013년 9월 9일), 게재확정일(2013년 9월 12일)
† 주저자, ksh4uu@etri.re.kr
‡ 교신저자, sunchoi@etri.re.kr(Corresponding author)

공개된 모든 정보는 타인에 의해 쉽게 수집되고 가공될 수 있다. 이런 문제는 소셜 네트워크 서비스를 통한 개인 신상털기, 보이스 피싱, 스토킹 등과 같은 불법적인 행위에 악용될 수 있는 문제가 존재한다. 따라서 SNS 상에 공개되는 정보에 본인 또는 타인의 신상정보가 존재하는지 분석하고 소유자를 식별하여, 이용자들의 개인정보 노출을 사전에 경고할 수 있는 수단이 필요하다.

본 논문의 실험은 개인정보 노출 위협이 가장 높은 소셜 네트워크 서비스를 선택했으며. 그 중에서도 개인들 간에 정보 공유가 가장 활발하게 이루어지고 있는 트위터를 통해서 실험 데이터를 수집했다. 수집된 트윗 문장은 본인의 지역정보를 노출한 문장과 그렇지 않은 문장으로 구분하고, 노출된 지역정보의 소유자를 식별하기 위한 자질(feature set) 추출 방법을 연구한다. 그리고 기존의 문서 분류 모델을 사용해서 트윗 문장에 노출된 지역정보의 소유자를 식별 가능한지 분석하고, 최적의 소유자 식별 모델의 구조를 도출한다. 마지막으로 더욱 향상된 성능을 나타낼 수 있는 소유자 식별 방법을 제안하고 추가 실험을 진행한다.

본 논문은 개인정보 분야에서 문장에 노출된 개인정보 소유자 식별 방법을 고찰한 점과 실제 트윗 문장의 구조적 자질을 분석하여 기존의 문서 분류 모델 보다 더 뛰어난 성능을 나타낼 수 있는 방법을 제안한 점에서 의의가 있다. 또한 소셜 네트워크 이용자가 서비스를 이용하면서 발생될 수 있는 개인정보 노출 문제를 사전에 탐지하고 경고할 수 있는 시스템 개발에 도움이 될 것으로 기대한다. 논문의 구성은 2장에서 SNS 환경에서 발생되는 프라이버시 문제와 그 연구 동향을 소개한다. 3장에서는 개인정보 소유자 식별 방법을 설명하고 4장에서는 실험 환경 및 실험 결과를 분석 한다. 마지막 5장에서는 논문의 결론과 향후 연구 방향을 기술한다.

II. 관련 연구

본 장에서는 현재 SNS에서 나타나고 있는 프라이버시 노출 문제와 관련 연구 동향을 설명 한다.

2.1 SNS 프라이버시 노출 문제

소셜 네트워크 서비스는 이용자들 스스로 정보를 생성하고 공개하는 시스템이다. 그래서 소셜 네트워크 서비스에서 발생하는 프라이버시 책임은 이용자 본인에게 있다고 간주 할 수 있다. 하지만 이용자는 본인이 작성한 정보를 공개하는 시점에서 개인정보의 위험도를 쉽게 알 수 없다. 또한 추후 그 정보가 누구에 의해 어떻게 쓰일 수 있는지 짐작할 수 없고, 익명이라도 생각하고 공개한 정보라도 다른 정보와 결합하여 또 다른 신상정보를 유추할 수 있는 문제가 존재한다[3]. 이러한 문제의 가능성은 이미 2011년 한국인터넷진흥원에서 트위터 사용자 200명의 ID만으로 파악할 수 있는 개인정보를 조사해서 발표했다. 보고서에 따르면 총 34개의 개인정보 항목을 분석하여 이름(88%), 인맥정보(86%), 사진 등 외모정보(84%), 위치정보(83%), 관심분야 등 취미정보(64%), 스케줄 정보(63%), 가족 정보(52%) 등 조사 대상 중 절반 이상의 항목을 쉽게 파악할 수 있었다. 또한 의료정보(29%), 정치성향 정보(19%) 등 민감 정보로 분류되는 항목까지도 상당히 높은 수치를 보인 것으로 파악되었다[4]. 이는 현재 소셜 네트워크 서비스 상에서 공유되고 있는 정보에 개인정보가 과다 노출되고 있으며 그러한 정보들이 타인에 의해서 쉽게 수집되고 가공될 수 있음을 확인시켜준 것이다.

2.2 SNS 프라이버시 연구 동향

최근 소셜 네트워크 서비스는 프라이버스 문제와 관련된 다양한 연구가 진행되고 있다. 대표적으로 시빌 공격(sybil attack). 신원 도난 공격(identity theft attack), 인가 받지 않은 데이터 수집, 프라이버시 블리칭(privacy bleaching), 평판표백 등이 있다[5]. 특히 가짜 ID(Fake IDentification) 탐지와 인가받지 않은 데이터 수집에 대한 연구가 가장 활발하게 진행되고 있다. 가짜 ID 탐지에 대한 연구는 SybilGuard [6]을 시작으로 최근에는 SNS 사용자 프로파일 유사도, 사용자 신뢰도 판단과 같은 다양한 연구[7,8] 내용이 발표되었으며, 인가 받지 않은 데이터 수집 방지 방법으로는 접근통제 기술을 사용해서 인맥 관계에 따라 이용자 프로파일의 접근을 제한하도록 하는 방법[9]이 발표되었다. 또한 소셜 네트워크 서비스에 노출된 개인정보보호 방안으로는 문서 분류 알고리즘을 사용해서 트위터 문장의 성격을 분류[10]한 연구, 트윗 문장에 개인정보 포함 여부를 분류[11]하는 모델 등이 연구되고 있다. Table 1.은 SNS에서 발생될 수 있는 공격 및 취약점을 정리한 것이다.

Table 1. SNS attacks and vulnerabilities

공격 및 취약점	설 명
시빌 공격	허위 신원을 통한 평판 조작
신원 도난 공격	유사 또는 복제 ID를 통한 개인정보 수집 및 평판 조작
인가 받지 않은 데이터 수집	소프트웨어를 통한 사용자 프로파일 자동 수집
프라이버시 블리칭	의도지 않게 발생하는 프라이버시 침해 (eg. 리트윗, 파도타기)
평판표백	기존의 악의적인 신원을 버리고 새로운 신원으로 재가입하는 방법

III. 개인정보 소유자 식별 방법

본 장은 개인정보 소유자 식별을 위한 방법으로써 트윗 문장에 노출된 지역정보의 소유자를 식별하기 위한 자질 추출 방법과 본 논문에서 제안한 소유자 식별 규칙을 소개한다.

3.1 자질 추출 방법

3.1.1 데이터 정제

본 논문은 트윗 문장에 노출된 지역정보의 소유자를 식별하기 위한 것으로써, 먼저 트윗 문장에 대한 특징을 분석할 필요가 있다. 트윗 문장은 한글과 영문을 1글자로 동일하게 취급하며, 140자 이내로 제한된 짧은 문장이다. 그리고 일반 문서와 비교했을 때 문법적 오류가 많고, 인터넷에서 사용되는 신조어나 이모티콘 등과 같은 특수문자를 많이 포함하고 있다. 그래서 트윗 문장은 일반적인 문장보다 충분한 단어가 발생되지 않고 노이즈(noise)가 많기 때문에 어절의 모든 형태를 의미의 최소단위로 분석하는 형태소 분석[12]이나 언어적 특성과 상관없이 글자 단위로 분석하는 n-gram[13]을 사용해서 의미 있는 자질을 추출하기에는 어려움 있다[14]. 그래서 본 논문에서는 알파벳, 특수문자는 모두 제거하고, 띄어쓰기(spacing) 단위로 트윗 문장을 재구성했다. 하지만 'RT', '?'는 노출된 지역정보의 소유자를 식별하는데 중요한 자질로 판단되어 제거하지 않는다. Table 2.는 데이터 정제 과정을 설명한 것이다.

3.1.2 자질 추출 및 선택

소셜 네트워크 이용자는 본인의 개인정보를 트윗

Table 2. Data purification methods

원본 문장	RT @iam???: [투표] 신화 서울가요 대상 본상 후보. 투표는 11월 27일부터 http://t.co/3AHXUkQG
정제된 문장	RT 투표 신화 서울가요 대상 본상 후보 투표는 11월 27일부터

문장을 통해서 공개할 때, 문장의 길이 제약 때문에 대부분 간단명료하게 작성하는 경우가 많다. 그래서 트윗 문장에 노출되는 개인정보는 그 속성(이름/지역/나이 등)에 따라 특정 단어와 문장 구조로 많이 이루어져 있음을 확인할 수 있다. 이러한 특징을 이용하여 본 논문에서는 형태소 분석이나 n-gram 방식을 사용하지 않고 개인정보 속성에 따른 문장의 구조적 특징을 자질로 선택했다. Table 3.은 본인의 지역정보를 트윗 문장에 공개할 때 주로 사용되는 문장 구조 패턴을 나타낸 것이다 [15,16,17].

Table 3. Pattern of the sentence structure of a regional information exposed in twitter

좌측 품사 정보	문장 구조 패턴 분석 기준	우측 품사 정보
1인칭 대명사, 부사	지역 정보	동사, 명사, 형용사, 조사

Table 4.는 트윗 문장에 노출된 지역정보의 소유자를 식별하는데 사용한 품사별 단어 유형을 나열한 것이다.

Table 4. Word types by the part of speech

품 사	품사별 단어 유형
1인칭 대명사	나, 난, 저, 전, 나는, 나두, 나도 등
동사	살다, 사는, 삽니다, 살았다, 오다, 오면, 가다, 가면 등
명사	사람, 출신, 거주지, 토박이, 눈, 비, 바람, 안개 등
형용사	덥다, 춥다, 쌀살 하다 등
조사	에, 는, 에도 등
부사	지금, 현재, 방금, 다시 등

Table 5.는 위에 언급한 문장 구조 패턴를 사용하여 트윗 문장에 노출된 지역정보의 소유자를 식별 할 수 있는 자질 13개를 나타낸 것이다. Table 5.에서 글씨가 진하고 기울어지게 표기한 부분은 해당 자질에 대한 주요 특징이고, '_'은 띄어쓰기를 의미한다. 자질

Table 5. Features for identifying the ownership of a regional information exposed in twitter.

자질명	자질 설명 및 문장 구조
S	1인칭 대명사+지역
	'***나도_서울***'...와 같이 2,3어절로 끝난 문장
R	지역+명사
	서울사람입니다...
RT	'RT'로 시작하는 문장
	RT @xxx: 〔투표〕 신화 서울가요 대상
FR	1인칭 대명사+지역+동사
	언니는_어디?_***나는_서울_살아***요...
TR	부사+지역
	또_***다시_서울***사람으로_살아요
RW	지역+형용사
	오빠_***강릉은추워***요_낼올때옷따숩게입고와
QN	지역+물음표
	서울_사세요?
FLI	지역+'살아' 단어 존재 여부
	서울_살아서_콘서트_갈_수_있어
FRE	1인칭 대명사+지역+명사+문장 끝
	나는_서울_거주.
FJBR	지역+조사+동사
	나는_***서울에_살아***요
NEGA	지역+왕래동사(가다)
	나는_***서울_갔다***...
COME	지역+왕래동사(오다)
	서울_오면_맛있는거_사줄게...
LIVE	지역+'살아요' 단어 존재 여부 + 문장 끝
	서울_살아요.

추출 방법은 탐지된 지역 정보를 기준으로 앞 뒤 첫 번째 단어에 대한 식별과 지역정보에 붙는 조사, 특수 문자(물음표)의 존재 유무 및 그 위치 등을 분석하여 자질을 추출한다.

3.2 제안하는 지역정보 소유자 식별 규칙

본 논문에서 제안한 지역정보의 소유자 식별 규칙은 Table 5.에서 설명한 자질 13개를 사용해서 12개의 소유자 식별 규칙 생성했고, 그 식별 규칙을 Fig.1.과 같은 트리(Tree) 구조로 표현했다. Fig.1.에 그려진 실선 화살표는 식별 규칙 조건의 만족을 의미이고, 점선 화살표는 식별 규칙 조건의 불만족을 의미한다. 소유자 식별 규칙은 대부분 사람의 직관적인 식별 규칙과 의사결정트리 알고리즘 결과에서 정확도가 100%인 규칙만을 사용했다.

IV. 실험 결과 및 분석

4.1 실험 환경 및 데이터

본 논문의 실험은 이미 여러 도메인에서 검증된 문서 분류 모델인 서포트 벡터 머신(Support Vector Machine)〔18〕, 문서가 가진 자질에 따라 특정 범주에 속할 확률을 계산하여 문서를 분류하는 나이브 베이즈안 분류기(Naive Bayesian Classifier) 〔19〕, 감독 학습 기반의 데이터 분류 모델인 의사결정트리(Decision Tree)〔20〕을 사용해서 기계학습을 통한 소유자 식별 가능성을 분석한다. 실험 데이터는 임의의 트위터 사용자 1000명에 대한 100일치 트윗 문장(801,234문장)을 사용했다. 그리고 ETRI의 개체명 분석기〔21〕를 사용해서 문장에서 지역 정보가 포함된 문장 27,056개를 추출했다. 추출된 27,056개 문장에서 노출된 지역정보의 소유자 구성비(작성자/제3자)를 조사하기 위해 임의의 문장 7,000개를 추출해서 사람이 직접 분석했다. 분석 결과 7,000개 문장에 약 100개 문장이 작성자 본인의 지역정보를 노출한 것으로 확인되었다. 이는 실제 트위터 환경에서도 사용자의 개인정보(지역정보)가 약 1/70의 비율 정도로 노출 될 수 있다고 고려하고, 본 논문의 실험에서도 1/70의 비율로 실험 데이터를 구성했다. Table 6.은 본 논문의 실험에 사용한 실험 데이터 구성표이다.

Table 6. Structure of the experiment data

	작성자의 지역	제 3자의 지역
학습 데이터	100	100
실험 데이터	60	3,500

실험 결과를 평가하기 위해서는 3배 교차 검증(3 fold cross validation)을 진행하였고, 재현률(recall), 정확률(precision), F1-척도(F1-measure)를 이용하여 성능을 비교했다. 그리고 오픈소스 툴킷인 SCIKIT〔22〕을 사용해서 소유자 식별 모델을 만들고 실험을 진행했다.

4.2 실험 결과 분석

본 논문의 실험 결과를 통해서 트윗 문장에 노출된 지역정보의 소유자를 식별할 수 있는 최적의 자질과 식별 모델의 설계 방식을 도출하고, 기존의 문서 분류 모델을 통해서 그 실효성을 검증한다.

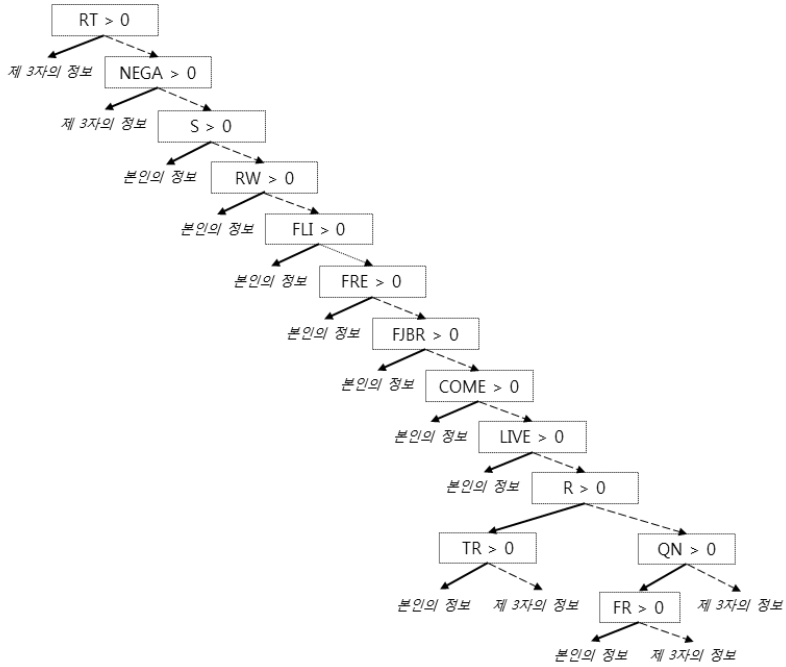

Fig.1. A decision tree for identifying the ownership of a regional information exposed in twitter

4.2.1 소유자 식별 모델 분석

Fig.2.는 자질 유형에 따른 소유자 식별 성능에 대한 결과로 n-gram 방식으로 추출한 8,684개 자질을 사용할 때와 문장의 구조적 특징으로 추출한 13개 자질을 사용했을 때의 성능을 비교한 것이다. 실험 결과, n-gram 방식보다 문장의 구조적 특징을 이용했을 때 모든 실험 모델에서 2배 이상 향상된 성능을 보였다. Fig.3.은 벡터 유형에 따라 소유자 식별 성능을 표시한 것으로, TF-IDF Vector 방식과 Counter Vector 방식을 각각 적용하여 비교 실험했다. 그 결과 Counter Vector 방식이 모든 실험 모델에서 약 2배 더 좋은 성능을 보였고, 그 중에서 의사결정트리가 SVM과 NBC보다 조금 더 높은 성능을 나타냈다. 이는 본 실험에서 문장에 1회만 출현한 문장구조 자질을 사용하므로, 여러 번 출현할 수 있는 토근의 횟수를 세는 TF-IDF 보다 자질의 출현 여부만을 반영한 Counter Vector가 더 적합함을 보여주는 것으로 해석된다.

결론적으로, 우리는 두 실험을 통해서 트윗 문장에 노출된 개인정보(지역정보)의 소유자 식별 모델은 문장의 구조적 특징을 자질로 사용하고, Counter

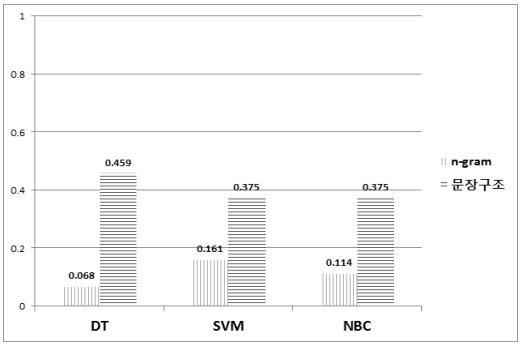

Fig.2. A result of the experiment for feature type

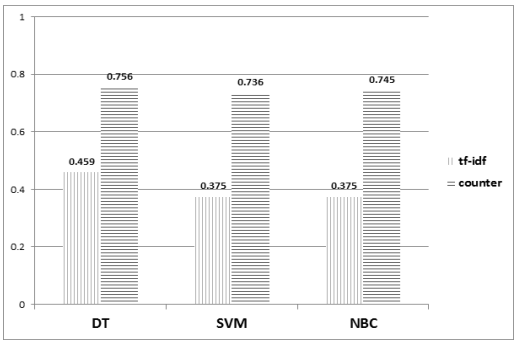

Fig.3. A result of the experiment for vector type

Vector 방식의 모델이 가장 최적의 소유자 식별 모델임을 확인할 수 있었다.

4.2.2 소유자 식별 규칙을 통한 성능 향상

본 논문에서는 사람의 직관적인 접근법을 사용해서 만든 Fig.1.과 같은 12개의 소유자 식별 규칙을 제안했고, 그 식별 규칙에 따라 소유자 식별 모델을 구현하여 추가 실험을 했다. Fig.4.는 본 논문에서 제안한 소유자 식별 규칙 모델과 기존 문서 분류 모델의 실험 결과를 비교한 것으로, 소유자 식별 규칙 모델이 다른 문서 분류 모델보다 최대 12% 더 향상된 성능을 보였다.

하지만 본 논문에서 제안한 소유자 식별 규칙 모델은 실험에 사용된 3,560개 문장 중에서 단 한 개의 식별 규칙에도 만족하지 못한 문장 3개가 존재했다. 식별 규칙에 포함되지 못한 3개 문장은 기존 문서 분류 모델을 통해서 추가적으로 실험한 결과 모두 올바르게 식별할 수 있었다.

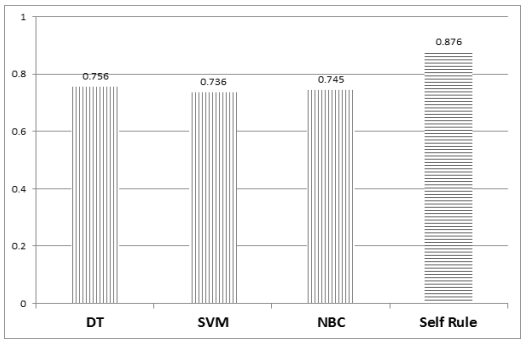

Fig.4. A result of the experiment for comparing performance by the models for distinguishing ownership of a regional information exposed in twitter.

V. 결 론

본 논문에서는 트윗 문장에 노출된 지역 정보의 소유자를 식별하기 위한 자질 생성 방법과 소유자 식별 규칙을 제안하고 실험했다. 그 결과 문장의 구조적 특징을 자질로 사용하고 Counter Vector 방식의 모델이 지역 정보의 소유자를 식별하는데 가장 최적의 모델임을 확인했다. 그리고 본 논문에서 제안한 소유자 식별 규칙 모델이 기존의 문서 분류 모델 보다 최대 12% 더 향상된 성능을 보였다. 이는 개인정보 소유자 식별 분야에서는 확률·통계적인 기계학습 모델보다 개인정보 속성에 따른 휴리스틱 규칙 모델이 더 효과적인 방법이 될 수 있음을 보인다.

하지만 본 논문에서 제안한 소유자 식별 규칙은 전체 실험 데이터를 모두 식별하지 못했고, 식별하지 못한 문장들은 기존의 문서 분류 모델을 통해서 분류할 수 있음을 확인했다. 비록 3개의 문장을 통해서 확인한 결과이지만, 이는 휴리스틱 모델과 문서 분류 모델을 결합한 하이브리드 모델이 실제 서비스 환경에서는 더욱 신뢰성 있는 결과를 보일 것으로 예상할 수 있다.

향후 연구로는 실험 데이터 확장을 통한 하이브리드 모델의 실효성 검증하고, 자연어 처리에 대한 다양한 문제를 보완해서 이름, 직업, 생일과 같은 다양한 개인정보들에 대한 소유자 식별 자질과 규칙을 추가적으로 연구할 계획이다.

References

[1] Seong-mi Sim, Korea people leave twitter because of it tired, The Korea Economic Daily. May. 2013.

[2] Ha-na Jun, Facebook user is 11 million people per month and enhance the marketing, ZDNET Korea, Feb. 2013.

[3] Dae-seon Choi, Seok-hyun Kim, Jin-man Cho and Seung-hun Jin, "Analysis technique for privacy in bigdata," Journal of The Korea Institute of Information Security & Cryptology, 23(3), pp. 56-60, June 2013.

[4] Press release, "How much my information expose in twitter?," Korea Communications Commission, 2011.

[5] Tae-kyeong Yoon, Do-won Hong, "Trend of technology of reliability reinforcement of social network service," Electronics and Telecommunications Trends, 26(4), pp. 134-145, Aug. 2011.

[6] Y, Haifeng and K. Michael, "Sybilguard: defending against sybil attacks via social networks," ACM SIGCOMM Computer Communication, vol. 36, no. 4, pp. 267-278, 2006

[7] Kolaczek and Grzegorz, "An approach to identity theft detection using social net-

work analysis," Intelligent Information and Database Systems ACIIDS First Asian Conference, pp. 78-81. Apr. 2009.
〔8〕 Kumar, Nitesh, and Ranabothu Nithin Reddy. "Automatic detection of fake profiles in online social networks," Bachelor Thesis, National Institute of Technology Rourkela, May. 2012.
〔9〕 Abhishek Kumar, Subham Kumar Gupta, Animesh Kumar Rai and Sapna Sinha, "Social networking sites and their security issues," International Journal of Scientific and Research Publications, vol. 3, no. 4, pp. 1-5, Apr. 2013.
〔10〕 Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu and Murat Demirbas, "Short text classification in twitter to improve information," Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841-842, 2010.
〔11〕 Nam-won Kim, Jin-su Park, "Personal information detection by using naive bayes methodology," Journal of The Korea Intelligent Information System Society, 18(1), pp. 91-107, Mar. 2012.
〔12〕 Morphological analysis 〔Online〕 : htto://ko.wikipedia.org/wiki/형태_분석
〔13〕 William B. Cavnar and John M. Trenkle, "N-gram-based text categorization," Ann Arbor MI, vol. 48113, no. 2, pp. 161-175, 1994.
〔14〕 Cho-hui Hong, Hak-su Kim, "Comparative study of various machine-learning features for tweets sentiment classification," Journal of The Korea Contents Association, 12(12), pp. 471-478, Dec. 2012.
〔15〕 Gyeong-ryeol Kim, Dong-hyeon Choi, Eun-gyeong Kim, Gi-seon Choi, "Feature selection for meeting location from non-itemized meeting email announcement in korean," Journal of The Korean Institute of Information Scientists and Engineers, 37(2), pp. 50-51, Nov. 2010.
〔16〕 Ui-gyu Park, Min-hui Cho, Seong-won Kim, Dong-ryeol Na, "A method for extracting dependency relations using chunking and segmentation," Journal of The Korean Institute of Information Scientists and Engineers, 16(1), pp. 131-137, Oct. 2004.
〔17〕 Jong-su Im, Tae-yeong Kim, Dong-ryeol Na, "Korean dependency parsing based on machine learning of feature weights," Journal of The Korean Institute of Information Scientists and Engineers, 38(4), pp. 214-223, Apr. 2011.
〔18〕 Chang, Chih-Chung, and Chih-jen Lin. "LIBSVN: A library support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, 2011.
〔19〕 Harry Zhang, "The optimality of naive bayes," Proceedings of the FLAIRS Conference, vol. 1, no. 2, pp. 3-9, 2004.
〔20〕 Decision Trees〔Online〕.Available : http://scikit-learn.org/stable/modules/tree.html
〔21〕 Chang-gi Lee, Myeong-gil Jang, "Named entity recognition with structural SVMs and pegasos algorithm," Journal of The Korean Society for Cognitive Science, 21(4), pp. 665-667, Dec. 2010.
〔22〕 SCIKIT 〔Online〕.Available: http://scikit-learn.org

〈저 자 소 개〉

김 석 현 (Seok-hyun Kim) 정회원
2008년 2월: 충주대학교 전자통신학과 졸업
2010년 3월: 전남대학교 정보보호협동과정 석사
2010년 7월~현재: 한국전자통신연구원 선임연구원
〈관심분야〉 정보보호, 비소셜 네트워크 보안/분석

조 진 만 (CHO, JIN-MAN) 종신회원
1989년: 충남대학교 계산통계학과 졸업
1991년: 충남대학교 전자계산학과 석사
1991년~현재: 한국전자통신연구원 책임연구원
〈관심분야〉 개인정보보호, 스마트카드

진 승 헌 (Seung-Hun Jin) 종신회원
1993년: 숭실대학교 전자계산학과 졸업
1995년: 숭실대학교 전자계산학과 석사
2004년: 충남대학교 컴퓨터과학과 박사
1994년~1996년: 대우통신
1996년~1999년: 삼성전자
1999년~현재: 한국전자통신연구원 인증기술연구실 실장/책임연구원
〈관심분야〉 컴퓨터/네트워크 보안, PKI, ID관리, 개인정보보호, 모바일 지불결제 보안

최 대 선 (Daeseon Choi) 정회원
1995년: 동국대학교 컴퓨터공학과 졸업
1997년: 포항공과대학교 컴퓨터공학과 석사
2009년: 한국과학기술원 전산학과 박사
1997년~1999년: 현대정보기술
1999년~현재: 한국전자통신연구원 책임연구원
〈관심분야〉 인증, 개인정보보호, 빅데이터 분석