Journal of The Korea Society of Computer and Information
Vol. 20, No. 1, January 2015

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.1.255

범죄발생 위험요소와 연관된 SNS 데이터의 효율적 추출 방법에 관한 연구

이 종 훈*, 송 기 성*, 강 진 아*, 황 정 래*

A study on the efficient extraction method of SNS data related to crime risk factor

Jong-Hoon Lee*, Ki-Sung Song*, Jin-A Kang*, Jung-Rae Hwang*

요 약

본 연구에서는 매년 증가하는 범죄에 대한 예방 측면에서 범죄발생 위험요소에 관한 정보를 사전에 파악하고 범죄발생을 예방하기 위해 SNS 데이터를 활용하는 방안을 제시한다. 최근에 SNS(Social Network Service) 데이터는 다양한 분야에서 선제적 예방 대응체계를 구축하는데 활용됨에 따라 그 중요성 또한 점점 증가하고 있다. 하지만 SNS 데이터를 단순 키워드로 수집하는 경우 관련되지 않은 데이터가 다수 포함되어 정확도 저하와 데이터 분석에 혼란을 초래할 우려가 있다. 이에, SNS 데이터의 텍스트 마이닝 분석을 통해 범죄발생 위험요소의 검색 정확도를 향상시켜 효율적으로 추출할 수 있는 방안을 제시한다.

▸ Keywords : 범죄발생 위험요소, SNS 데이터, 데이터 검색, 텍스트 마이닝

Abstract

In this paper, we suggest a plan to take advantage of the SNS data to proactively identify the information on crime risk factor and to prevent crime. Recently, SNS(Social Network Service) data have been used to build a proactive prevention system in a variety of fields. However, when users are collecting SNS data with simple keyword, the result is contain a large amount of unrelated data. It may possibly accuracy decreases and lead to confusion in the data analysis. So we present a method that can be efficiently extracted by improving the search accuracy through text mining analysis of SNS data.

▸ Keywords : Crime risk factor, SNS data, Data retrieval, Text mining

•제1저자 : 이종훈 •교신저자 : 황정래
•투고일 : 2015. 1. 3. 심사일 : 2015. 1. 14, 게재확정일 : 2015. 1. 21.
* 공간정보산업진흥원(Spatial Information Industry Promotion Institute)
※ 본 연구는 국토교통부 국토공간정보연구사업의 연구비지원(14NSIP-B081051-01)에 의해 수행됨.

I. 서 론

통계청의 자료에 의하면 살인, 강도, 방화, 강간 등의 강력 범죄가 매년 늘고 있으며 여성 피해자의 비율은 2012년 85.6%(2014년 자료 기준)로 지난 3년간 꾸준히 증가하는 추세다. 연령별로는 20대 여성이 가장 많았고 13세 이하 여아도 크게 늘어난 것으로 나타났다[1]. 이러한 범죄에 대한 대책으로 정부에서는 시민들이 주변의 범죄 위험에 관심을 가지고 위험으로부터 대처할 수 있도록 치안 및 범죄 정보를 제공하고 있으며 지자체에서는 안심귀가 서비스를 시행하는 등 범죄 예방을 위한 사회적 관심이 증가하고 있다. 범죄의 선제적 예방 차원에서 범죄발생 위험요소를 사전에 파악하고 대처하기 위해서는 담당 공무원의 노력과 시민들의 적극적인 민원 제기가 필요하나, 인적·물적 한계 등으로 정보를 수집하는데 한계가 있다.

해외의 연구에서는 SNS와 같은 비정형데이터의 방대한 정보를 분석하여 경향을 찾아내고 패턴을 찾아냄으로써 트위터의 데이터가 실세계를 반영하고 있다는 것을 증명하였다[2]. SNS상의 데이터와 위치정보를 수집하고 공간적 특성을 분석해 범죄발생 위험요소 및 이상 징후를 사전에 포착 및 대응한다면 범죄예방에 기여할 수 있을 것으로 판단한다.

하지만 범죄발생 위험요소와 관련된 다량의 정보를 수집하기 위해 SNS 데이터를 활용하는 경우, 단순 키워드만으로 검색하게 되면 범죄와 관련되지 않은 불필요한 SNS 데이터가 포함되어 데이터 분석에 혼란을 초래할 수 있다.

본 연구는 트위터에서 범죄발생 위험요소를 검색할 때 검색되는 정보의 정확도를 향상시켜 효율적으로 추출할 수 있는 방안을 제시하는 것을 목적으로 한다. 구체적으로는 트위터에서 데이터를 검색하기 위해 범죄발생 위험요소와 관련된 키워드를 선정하고 텍스트 마이닝 분석을 통해 검색되는 트위터 데이터의 정확도를 향상시킬 수 있는 방안을 연구하였다.

연구방법으로는 첫째, 범죄발생, SNS 데이터 활용, 비정형 데이터 분석과 관련된 선행연구를 고찰하였다. 둘째, 범죄발생 위험요소와 관련된 문헌조사를 통해 기본 키워드를 정리하였다. 셋째, 트위터 API를 활용하여 기본 키워드를 검색함으로써 국내 트위터 데이터를 수집하였다. 셋째, 범죄발생 위험요소와 관련된 트위터 데이터의 텍스트 마이닝 분석을 통해 검색 정확도를 향상시킬 수 있는 방안을 제시하였다. 넷째, 일반적인 기본 키워드로 검색한 결과와 본 연구에서 제시한 방안으로 검색한 결과를 비교하여 검색 정확도가 향상된 것을 검증하였다.

II. 관련 연구

범죄 발생을 사전에 예방 및 예측하기 위한 선행 연구로는 범죄예방활동으로써 범죄를 예측하기 위해 범죄자처우, 범죄가능성, 재범예측 관점에서 연구를 진행한 경우[3]가 있었고 범죄 위험도를 평가하기 위해 범죄 영향요인들의 관계분석을 하고 범죄 위험도 지표를 선정하여 평가방법을 제안한 사례[4]가 있었다. 또한 과거 범죄 통계자료를 바탕으로 범죄 위험 등급지수에 따라 범죄를 예측할 수 있는 지도 서비스를 제안하는 연구가 있었다[5].

SNS 데이터를 분석하여 타 분야에서 활용한 선행 연구로는 소셜 네트워크 서비스를 활용한 기존의 메쉬업 시스템들이 정보 확산을 목적으로만 활용하는 점을 개선하여, 소셜 네트워크 서비스의 데이터를 수집, 분석하여 재해 상황에서 활용할 수 있도록 하는 방법을 연구한 사례가 있었다[6]. 또한 트위터 데이터 중 주거환경에 대한 평가와 관련된 트위터 데이터를 추출하여 공간적 분석을 수행하고, 도시정책지표를 보완하는 데에 트위터 데이터의 분석결과가 활용될 수 있는지 가능성을 확인한 연구가 있었다[2].

비정형 데이터 분석 관련 연구사례는 자동문서분류 시스템의 성능을 향상시키기 위해 나이브 베이즈 분류기를 이용하여 문서분류 실험을 하고 문서분류 성능 향상을 위한 단어 가중치 기법을 제안한 연구[7]와 미래예측 시 전문가들의 정성적인 의견과 평가를 보조할 수 있는 정량적이고 객관적인 자료를 도출하기 위하여 인터넷과 네트워크기법 및 텍스트 마이닝 기법을 활용한 연구가 있었다[8].

기존 선행연구를 종합해보면 범죄예방 및 예측 관련 연구로는 이론 및 과거 통계 분석 측면에서 접근한 사례가 있으며 SNS 데이터는 재난·재해나 도시정책 보완 등의 연구에서 활용되었고 비정형 데이터 분석과 관련해서는 문서 및 정성적 자료 등을 분석하는 연구들이 있었다. 결과적으로 범죄예방 차원에서 범죄발생 위험요소를 파악하기 위해 SNS 데이터를 활용한 사례는 없는 것으로 나타났다.

III. 키워드 선정 및 트위터 데이터 추출

3.1 SNS의 개념 및 특성

소셜 네트워크 서비스(Social Network Service)는 사용자 간의 자유로운 의사소통과 정보 공유, 그리고 인맥 확대

등을 통해 사회적 관계를 생성하고 강화시켜주는 온라인 플랫폼을 의미한다. 대한민국 내 SNS 시장을 주도하고 있는 페이스북과 트위터 이용자 수는 2011년 1천만 명을 돌파한 이후 계속 증가 추세이다[9].

최근 이용률이 증가하고 있는 카카오스토리, 라인, 밴드 등은 폐쇄형 SNS로써 OpenAPI를 제공하지 않아 데이터 검색 및 수집이 불가하다. 이에 따라 본 연구에서는 OpenAPI를 활용하여 데이터 검색이 가능한 트위터를 대상으로 진행하였다.

여러 소셜 네트워크 서비스 중의 하나인 트위터는 140자 이내의 단문을 게재하거나 구독하면서 일생생활 이야기, 유용한 정보, 새로운 소식 등을 공유할 수 있다[2]. 트위터는 기존의 인위적인 실험 환경이나 구조화된 설문 방식 등과 차별화되어 자발적인 의견이 표현되고 실시간으로 방대한 데이터 수집이 가능한 특징 등이 나타나므로 연구 분석의 대상으로 각광받고 있다[10].

일반적으로 사람들은 트위터에서 형식에 구애받지 않는 구어체를 사용하는 특징이 있기 때문에 비정형 데이터 분석을 위해서는 형태소 분석을 해야 한다. 형식이 정해져 있는 텍스트 위주의 데이터와 달리 트위터와 같은 비정형 데이터는 유형이 불규칙하기 때문에 형태소 분석을 하여 불규칙적인 문법에서 단어를 구분하고 품사의 모호성을 해결할 필요가 있다[11].

3.2 CPTED의 원리에 따른 트위터 검색 키워드 선정

3.2.1 CPTED의 의의

CPTED(Crime Prevention Through Environmental Design)란 '환경설계를 통한 범죄예방'으로 표현하고 있으며 "장소와 범죄"사이의 관계를 중심으로 범죄예방을 위한 건축 및 도시 환경 설계를 연구하는 이론이다[12].

CPTED는 건축 및 도시 환경을 범죄에 대한 방어적인 디자인으로 설계함으로써 범죄를 예방하고 불안감을 감소시켜 삶의 질을 향상시킬 수 있다는 가정에서 출발하였다. 이는 결국 실패한 건축 및 도시 환경의 설계는 범죄 및 범죄 두려움을 증가 시키고 삶의 질을 저하시키는 결과를 초래할 수 있다는 것을 의미한다. 이처럼 CPTED는 공간적(장소) 특성을 고려하여 설계함으로써 범죄 불안감과 발생범위를 감소시키는 이론으로 CPTED 기본원리를 활용하여 키워드를 선정하는 것이 효과적일 것으로 판단하였다.

CPTED 기법이 적용되는 기본적인 원리로는 감시, 접근통제, 영역성 강화, 활용성 증대, 유지관리가 있으며 각각에 대한 개념은 표 1과 같다[12].

표 1. CPTED 원리와 개념
Table 1. CPTED Strategies and Definition

CPTED원리	개념
감시	가시권을 최대화시킬 수 있도록 건물이나 시설물을 배치하는 원리를 말하며 자연적/기계적 감시로 구분
접근 통제	도로, 보행로 등을 일정한 공간으로 유도함과 동시에 허가받지 않은 사람의 출입을 차단하는 원리로서 자연적/기계적 접근통제로 구분
영역성 강화	사적공간과 공적공간을 명확히 구별하여 주민들이 권리를 주장할 수 있는 점유 영역을 가상으로 통제하는 원리
활용성 증대	주민들의 활발한 사용을 유도함으로써 자연스런 감시를 강화할 수 있도록 지역 시설을 보강하거나 행사 등을 개최하는 원리
유지 관리	시설물이나 장소를 처음 설계된 대로 지속적으로 이용 가능하도록 관리하여 사용자의 일탈행동을 자제시키는 원리

(출처 : 경찰청, 환경설계를 통한 범죄예방(CPTED) 방안, 2005, 요약 재정리)

3.2.2 CPTED의 원리에 따른 키워드 정리

장소와 관련된 범죄 키워드를 추출하기 위해서 CPTED의 기본원리를 유형별로 분류하고 유형에 해당하는 관련 키워드를 선정해야 한다. 본 연구에서는 CPTED와 관련된 선행연구를 참고하여 관련 키워드를 정리하였다. 관련 연구로는 공동주택 주민의 만족도를 조사하기 위해 CPTED 원리 중 접근통제, 감시, 강화를 활용한 사례[13]가 있었고 활용성 증대 원리를 추가로 검토하여 공동주택 단지의 CPTED 도입현황과 문제점을 분석한 연구[14], 유지관리 원리를 추가하여 대학교 캠퍼스에 적용하기 위한 변수를 도출한 연구가 있었다

표 2. CPTED 원리에 따른 기본 키워드 선정
Table 2. Basic Keywords Selection by CPTED

CPTED 원리	기본 키워드	
	기존 키워드	추가된 키워드
감시	조명, 가로등, 순찰, CCTV, 출입구, 현관, 조경, 사각지대, 안전거울	씨씨티비, 감시카메라, 외진 곳, 외진 지역, 드문 곳, 드문 지역, 인적 드문, 어두운 곳
접근 통제	개폐기, 차단기, 외부인, 경비실, 경비원, 담장, 방범창, 보도, 조경	차단시설
영역성 강화	울타리, 표지판, 주차장	펜스, 안내판
활용성 증대	공원, 놀이터, 체육시설, 정자, 벤치, 노인정	근린공원, 녹지공원
유지 관리	청소, 공공시설, 시설/설비 관리	노후된, 낙후된, 청소/시설 불량

[15]. 또한 공동주거단지와 지구단위 계획에 적용 가능하도록 CPTED를 유형화하기 위해서 감시, 접근통제, 영역성 강화, 활용성 증대, 유지관리의 5가지 원리를 활용한 연구들이 있었다[16][17].

기존의 연구 사례에서 CPTED 유형에 따른 키워드를 정리한 결과는 표 2와 같으며, 본 연구에서는 트위터에서 포괄적인 검색이 가능하도록 관련 키워드를 추가하였다. 추가된 키워드는 인터넷상에서 키워드가 다양하게 표현되는 것을 고려하여 영문과 한글을 전환한 경우와 키워드를 구체화 할 수 있는 수식어 등을 포함하여 검색이 가능하도록 하였다.

3.3 선정된 키워드를 통한 트위터 검색

트위터의 오픈 API를 활용하여 키워드로 검색한 결과를 확인할 수 있도록 검색 사이트를 구현하였으며 구성은 그림 1과 같다.

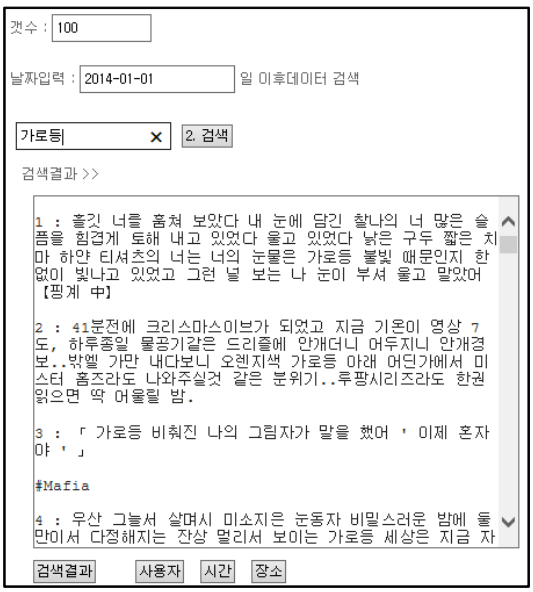


그림 1. 검색 사이트 구성
Fig. 1. Search Site Composition

본문 3.2.2의 선행연구를 통해 선정한 기본 키워드 48개를 토대로 트위터에서 데이터를 검색한 결과 총 3,649건이 검색되었다. 트위터의 특성상 리트윗 되어 내용이 중복 검색된 결과를 제외하고 2,235건을 바탕으로 분석한 결과 범죄발생 위험요소와 관련된 데이터는 그림 2와 같이 총 42건으로 나타났다. 결과적으로 범죄발생 위험요소 관련 트위터 데이터의 비율은 1.88%로 매우 낮은 것으로 나타났다.

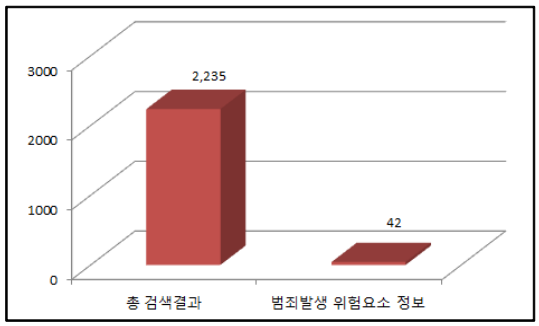


그림 2. 기본 키워드 검색결과
Fig. 2. Search Result of Basic Keywords

단순히 CPTED 원리에 사용된 키워드를 검색할 경우 표 3과 같이 범죄발생 위험요소와 관련되지 않은 트위터 데이터가 다수 포함되어 있어 검색 정확도가 현저하게 떨어지는 것으로 나타났다.

검색결과를 토대로 범죄발생 위험요소와 관련된 트위터 데이터가 검색되지 않은 기본 키워드는 제외하였고 관련 데이터가 검색된 키워드 10개만 선정하여 본 연구의 이후 단계에서 활용하였으며 선정된 키워드는 표 4와 같다.

표 3. 기본 키워드 트위터 검색 결과
Table 3. Twitter Search Results of Basic Keywords

기본 키워드	트위터 검색 결과 예시
가로등	반짝반짝 빛나는 작은 별들 그 보다는 가까운 가로등 불 어딘가에 여기 어디쯤인가 함께했던 그대와의 발걸음
	맑은 날만 반기는 먼 저 하늘 달빛보다는 늘 당신 가까이서 빛을 내는 가로등이었으면
	가로등이 필요 없네요. 보름달이 너무 밝아서.
순찰	코난군, 이거 봐. 순찰차야.
	여~ 파트너. 순찰 안돌고 뭐하냐?
	#순찰 #짤랑 #짤랑
외진	단점은 골목 외진 곳에 있어서 찾기 힘듬. 밖에서 간판도 잘 안 보임.
	당신도 참 별난 사람이네~ 이런 외진 곳 까지 보물을 찾으러 오다니.
	내가 진짜 공연보러 해외진출을 할 줄은.....

표 4. 검색결과에 따른 기본 키워드 선정
Table 4. Basic Keywords Selection by Search Results

CPTED원리	기본 키워드
감시	가로등, 순찰, 출입구, 현관, 씨씨티비, 외진, 인적
접근통제	방범창
유지관리	낙후된, 시설 불량

IV. 검색 정확도 향상 방안 도출

4.1 검색 정확도 향상을 위한 비정형 데이터 분석

본문 3.2.2의 선행연구를 통해 키워드를 선정하고 검색하였지만 범죄발생 위험요소와 관련된 다량의 트위터 데이터를 추출하지는 못하였다.

기존 연구 문헌에 따르면 어떤 문서에서 정보를 검색할 때, 단순히 키워드뿐만 아니라 그와 관련된 단어를 같이 검색하게 되면 검색하고자 하는 범위가 축소되기 때문에 검색의 정확성을 높일 수 있다고 언급하고 있다[18].

트위터와 같이 데이터의 크기와 내용이 달라 통일된 구조로 정리하기 어려운 데이터에서 특정 주제와 관련된 단어를 도출하려면 비정형 데이터 분석을 통해 문서에 포함된 단어들의 중요도를 추출하여야 한다[19].

이에, 본 연구에서는 검색된 트위터 데이터를 토대로 비정형 데이터 분석을 하여 최적 키워드를 추가로 선정하고 검색의 정확도를 높이는 방안을 제시한다.

트위터의 특성에서 언급하였듯이 트위터에 게재되는 글은 구어체이기 때문에 비정형 데이터 분석 이전에 형태소 분석을 우선시해야 한다. 형태소 분석을 통해 불규칙한 문장이나 어절을 형태소로 분석하고 품사의 모호성을 해소하였다. 또한 분석 결과의 정확성을 높이기 위해 맞춤법에 어긋나는 단어를 바른 표기법으로 변환하여 모호성을 해결하였다. 형태소 분석에는 한나눔 한국어 형태소 분석기를 사용하였으며 플러그인은 Informal Sentence Filter, Sentence Segmentor, Chart Morph Analyzer, Unknown Morph Processor를 일부 활용하였다.

4.1.1 비정형 데이터 분석 기법

비정형 데이터를 분석하는 세부 기법들은 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 군집 분석 등이 있다. 본 연구는 텍스트 내에서 다른 정보와의 연결성을 파악하고 의미 있는 정보를 추출하기 위해 텍스트 마이닝 기법을 활용하였다.

구체적으로는 트위터 데이터(문서) 내에 출현하는 모든 단어를 대상으로 각 단어들이 해당 문서에서 출현하는 빈도를 측정하기 위해 텍스트 마이닝 기법의 TF-IDF모델을 활용하였고 키워드별 중요도를 분석하기 위해 벡터공간모델을 활용하였다.

○ TF-IDF모델

정보추출방법은 텍스트마이닝에서 가장 중요한 부분이며 그중 간단하면서 가장 강력한 방법으로 TF-IDF(Term Frequency - Inverse Document Frequency) 방식을 많이 사용하고 있다[8]. TF-IDF는 Spark(1972)에 의해 여러 문서에 동시에 출현하는 단어의 빈도수를 계산하는 공식 (1)이 제시되었다.

$$w_{i,j} = TF \times IDF = tf_{i,j} \times \log(\frac{N}{n_i}) \quad (1)$$

$tf_{i,j}$: 문서내용어의총빈도
N : 전체문서의수
n : 용어가포함된문서의수

○ 벡터공간모델

벡터공간모델은 문헌과 용어에 대한 중요도를 분석하기 위하여 사용되며 SMART(1960)에 의해 최초로 적용되었다. 이 모델은 용어에 대한 가중치를 주는 방식으로 기존의 불리언 모델보다 우수한 검색결과를 나타내는 특징이 있으며 식 (2), (3)과 같다[8].

$$COS\theta = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}||\vec{q}|} = \frac{(w_{1j} \times w_{1q} + .. + w_{tj} \times w_{tq})}{\sqrt{w_{1q}^2 + ... + w_{tq}^2} + \sqrt{w_{1j}^2 + ... + w_{tj}^2}} \quad (2)$$

$$C(d,q) = \frac{(\sum_{i=1}^{t} w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{iq}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{ij}^2}} \quad (3)$$

k_i : 용어
d_i : 문서
w_{ij} : $term\ Weight$
$\vec{q} = (w_{1q}, w_{2q}, ..., w_{tq}), t = term$ 빈도수
$\vec{d_j} = (w_{1j}, w_{2j}, ..., w_{tj})$

4.1.2 텍스트 마이닝 분석 결과

TF-IDF모델과 벡터공간모델을 활용하여 10개의 기본 키워드에 따른 203개 트위터 데이터(문서) 내의 단어별 중요도 및 검색 키워드별 중요도를 분석하였으며 각각의 중요도는 표 5, 표 6과 같다.

표 5. 문서 내 단어별 중요도
Table 5. Importance of words in each document

구분	가로등
가로등	13
어둡다	6
너무	3
집	2.614394
무서워서	2.09691
걷다가	2
꺼졌다	2
밤	2
없다	2
여기	2

구분	순찰
순찰	11
민·경	4
합동	4
경찰서	3
예방	3
야간	2
School-비	1
강화	1
경·학	1
경찰관	1

구분	외진
곳	4.18303
외진	4
무서워서	2.09691
대학	2
배달	2
많은	1.39794
걱정	1
골목	1
공연	1
기숙사	1

구분	방범창
방범창	4
가스배관	2
강도	1
그나마	1
금품	1
도둑	1
두려움	1
듣고	1
School-비	0
가로등	0

구분	낙후된
낙후된	3
곳	0.522879
School-비	0
가로등	0
가스배관	0
강도	0
강화	0
거기가	0
걱정	0
걷다가	0

구분	인적
인적	2.09691
드문	1.39794
집	1.045757
바글바글	1
안심	1
안전한	1
옆집	1
위험	1
장기매매	1
주위	1

구분	씨씨티비
씨씨티비	2
통네	1
카페	1
화장실	1
School-비	0
가로등	0
가스배관	0
강도	0
강화	0
거기가	0

구분	현관
혼자	2
문	1
여자	1
유흥	1
학교마크	1
현관	1
School-비	0
가로등	0
가스배관	0
강도	0

구분	시설불량
불량	1
소방시설	1
요양시설	1
School-비	0
가로등	0
가스배관	0
강도	0
강화	0
거기가	0
걱정	0

구분	출입구
센서	1
출입구	1
집	0.522879
School-비	0
가로등	0
가스배관	0
강도	0
강화	0
거기가	0
걱정	0

표 6. 키워드별 중요도
Table 6. Importance of Keywords

구분	가로등	순찰	외진	방범창	낙후된
중요도	150.715	89.898	6.7449	4.4090	1.4811

구분	인적	씨씨티비	현관	시설불량	출입구
중요도	1.0139	0.5719	0.1621	0.0936	0.0814

4.1.3 분석결과를 토대로 최적 키워드 선정

키워드별 중요도가 낮아 검색 결과의 정확도 하락에 영향을 주는 키워드(중요도 5미만)와 문서 내의 단어(중요도 2미만)는 본 연구의 이후 단계에서 제외하였다.

결과적으로 키워드별 중요도가 높고 문서 내 중요도가 높은 단어만 선정하여 재검색에 사용할 수 있도록 최적 키워드로 도출하였으며 표 7과 같다.

표 7. 최적 키워드 선정
Table 7. Optimal Keywords Selection

기본 키워드	최적 키워드
가로등	어둡, 어두, 너무, 집, 무섭, 무서
순찰	민·경, 합동, 경찰서, 예방
외진	곳, 무서, 무섭

4.2 트위터 재검색 결과

텍스트 마이닝 분석을 통해 도출한 키워드를 추가하여 재검색한 결과, 범죄발생 위험요소와 무관한 트위터 데이터가 제외되고 표 8과 같이 관련 트위터 데이터가 다수 검색되는 것을 확인하였다.

표 8. 최적 키워드 트위터 검색 결과
Table 8. Twitter Search Results of Optimal Keywords

기본 키워드	최적 키워드	트위터 검색 결과 예시
가로등	무섭 or 무서 or 어둡 or 어두 or 너무 or 집	저 상대원에 사는데 집이 좀 높은데 잇고 가로등도 별로없어서 완전 어두운데 집올라가는길이 너무 어둡고 무서워서 발목아픈줄고 모르고 겁나 빠르게 올라왓는데 집도착하니까 다리가 후들후들 발목이 욱신욱신 심장이 덕쿵덕쿵
		사실 아까 실기실가는데 가는길이 너무 무서웠다. 이상하게 가로등불 몇개가 꺼져있어서 평소 보다 어두운길 누가 갑자기 나타나서 납치당하면 어카지하고 두려운 마음에 빠른걸음으로 달려감
		네네! 진짜 오는데 무섭기도 무서웠어요 가로등은 다 꺼져있고...사람은 별로 없고...으으..왜 가로등은 안켰는지 몰라요...
		밖에 나가서 뭐좀 사고 싶은데 진짜 진심으로 무섭다 여기 가로등도 없으 ㅁ;′
순찰	합동 or 민·경 or 예방 or 경찰서	화산지구대는 생활안전협의회 회원들과 관내 중화산동의 유흥가 주변을 청소년 탈선예방 및 4대 사회악 등 각종 범죄예방을 위해 합동순찰을 실시하였습니다.
		지난 4일 수원시 팔달산 등산로에서 토막시신이 발견되고 시민들이 불안에 떨자 수원시장까지 나서 야간합동순찰을 벌였다.
		아산경찰서, "민·경 합동 힐링폴 순찰 활동" 전개
외진	무서 or 무섭 or 곳	그니까 외진곳 다니면 안되고 사람많은곳도 무서움....아 지금 이렇게 트위터에서 얘기하는것도 무섭다
		밤에 집올때 길이 외진곳이라 무서워

4.3 검색 방식별 범죄발생 위험요소 데이터 추출 정확도 비교

검색 결과를 비교해 보면 그림 3과 같으며 일반적인 방법으로 기본 키워드 검색을 하였을 경우 트위터 데이터는 총 2,235건이 검색되었고 그 중 범죄발생 위험요소와 관련된 트위터 데이터는 42건으로써 정확도가 1.88%로 나타났다.

텍스트 마이닝 분석을 통해 중요도가 높은 최적 키워드를 추가하여 검색한 결과, 트위터 데이터가 총 203건 검색되었으며 그 중 범죄발생 위험요소와 관련된 트위터 데이터는 103건으로 정확도가 50.73%로 나타났다. 키워드별 검색 결과는 그림 4와 같다.

기본 키워드 검색 결과와 본 연구에서 추가한 키워드 검색 결과를 비교한 결과, 검색 정확도는 약 27배 증가한 것으로 나타났다. 이는 기본 키워드에 영향력이 높은 키워드를 추가

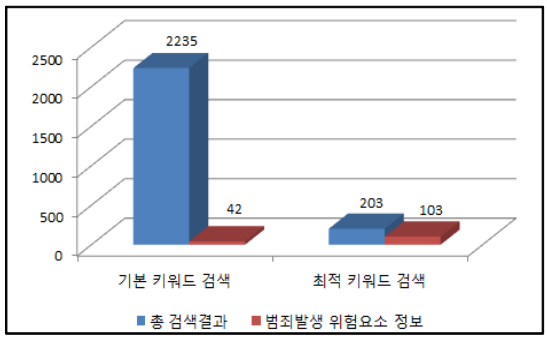


그림 3. 기본 및 최적 키워드 검색 결과
Fig. 3. Search Results of Basic and Optimal Keywords

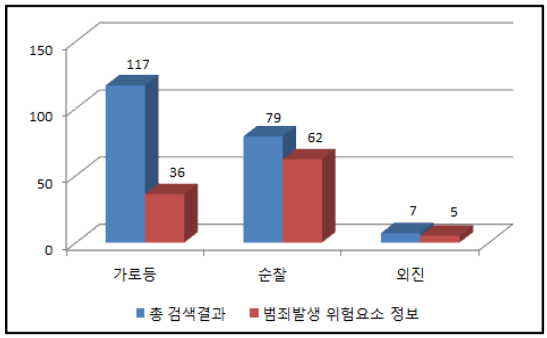


그림 4. 키워드별 검색 결과
Fig. 4. Search Results of Keywords

로 검색하면, 검색범위가 축소되고 관련 데이터를 다수 포함하게 되므로 검색 정확도가 향상되는 것으로 판단된다. 본 연구에서 제안하는 방안을 시스템 알고리즘으로 나타내면 그림 5와 같은 구성으로 나타낼 수 있다.

V. 결 론

본 연구는 기존의 범죄발생 위험요소와 관련된 데이터를 수집하는 방법에 한계점이 있음을 명시하고 SNS 데이터를 활용하여 해결하는 방안을 도출하였다.

하지만 단순 키워드 검색으로 SNS 데이터를 수집할 경우, 불필요한 정보가 포함되어 데이터 분석에 혼란을 초래할 수 있다. 이에, 본 연구에서는 범죄발생 위험요소와 관련된 SNS 데이터를 효율적으로 추출하는 방법을 제시하는 것을 목적으로 진행하였다.

우선적으로 문헌조사를 통해 범죄발생 위험요소를 검색하기 위한 키워드를 선정하였고 SNS 서비스 중 트위터 API를 활용하여 데이터를 수집하였다. 처음 선정한 48개의 키워드로 수집한 트위터 데이터에는 범죄발생 위험요소와 관련되지 않은 데이터가 다량 포함되어 있어 검색 정확도가 1.88%로 현저히 떨어지는 것으로 나타났다. 이후 범죄발생 위험요소와

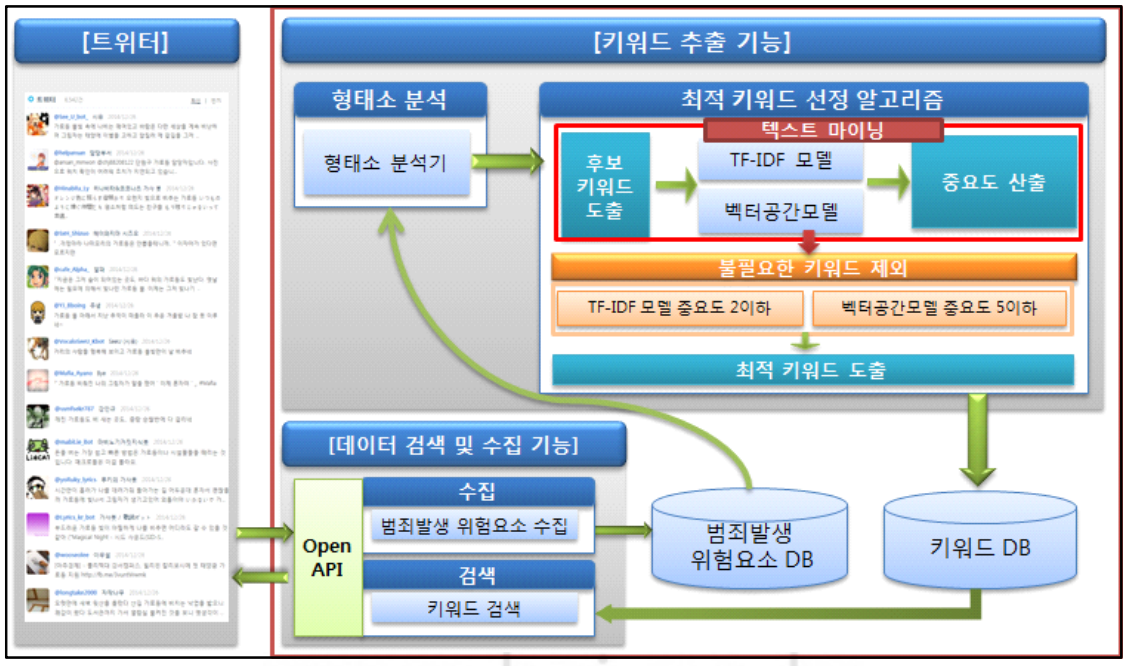


그림 5. '트위터 키워드 검색 기반 범죄발생 위험요소 추출 시스템' 구성도
Fig. 5. 'The Extraction System of the Crime Risk Factor based on Twitter Keyword Search' Block Diagram

관련된 트위터 데이터만을 대상으로 텍스트 마이닝 분석을 시행하였고 범죄발생 위험요소와 관련하여 주로 사용되는 키워드들의 중요도를 도출하였다. 도출된 키워드의 중요도에 따라 중요도가 낮은 키워드들을 제외하고 중요도가 높은 키워드들을 선정하여 처음 선정한 키워드에 추가하여 재검색하였다. 도출된 최적 키워드를 추가하여 트위터에서 재검색한 결과 범죄발생요소와 관련된 트위터 데이터의 비율이 50.73%로 크게 향상된 것을 확인할 수 있었다.

본 연구에서 제안한 SNS 데이터 추출 방법으로 범죄발생 위험요소와 관련된 데이터를 수집할 경우 검색범위가 축소되고 관련 데이터를 다수 포함하게 되므로 신속하게 정확도 높은 데이터를 추출해 낼 수 있을 것이다.

향후 연구에서 트위터의 데이터뿐만 아니라 좌표까지 수집하여 지도위에 매핑한다면 범죄발생 위험요소의 공간적 특성을 분석할 수 있을 것으로 기대한다.

참고문헌

〔1〕 Statistics Korea, "Women's lives to see the statistics", Report, June, 2014.

〔2〕 J. H. Park, "A study on the Applicability of SNS Data for the Urban Policy Indicator : Tweet data on the Spatial Characteristics of the Residential Environment Satisfaction", The Graduate School of Ewha Womans University, Master's Thesis, July, 2013.

〔3〕 S. J. Kang, K. H. Lee, "A study on the Assessment Variables and Method for the Crime Risk Assessment - Focused on the burglary composed of invasion and street crime", Journal of Safety and Crisis Management, Vol. 6, No. 3, pp. 144-171, Sep. 2010.

〔4〕 J. S. Lee, C. S. Hwang, T. H. Kim,, "A Study on the Crime Mapping and Monitoring System Development", Journal of Cadastre, Vol. 44, No. 1, June, 2014.

〔5〕 Y. C. Jo, "A Review on the Crime Prevention Through the Crime Prediction", Korea Association of Criminal psychology, Vol. 1, No. 1, pp. 273-294, Jan. 2005.

〔6〕 T. W. Seo, M. G. Park, C. S. Kim, "Design and Implementation of the Extraction Mashup for Reported Disaster Information on SNSs", Journal of Korea Multimedia Society, Vol. 16, No. 11, pp. 1297-1304, Nov. 2013.

〔7〕 M. H. Kim, Y. S. Kwon, "A study on Term Weighting Methods for performance improvements of text classification", Fall Conference - The Korean Institute of Industrial Engineers, Vol. 2011, No. 11, pp. 453-464, Nov. 2011.

〔8〕 G. H. Jung, "A Study of foresight method based on textmining and complexity network analysis", Korea Institute of S&T Evaluation and Planning, Report, Dec. 2010.

〔9〕 wikipedia, "http://ko.wikipedia.org", 2014

〔10〕 M. H. Kim, J. H. Jung, LG Business Insight Report, Feb. 2013.

〔11〕 S. A. Kim, "A study on SNS(Social Network Service) service Improvement plan through SNS users's interest grasp", DongGuk University, Master's Thesis, June 2010.

〔12〕 National Police Agency, "Crime Prevention Through Environmental Design", Report, Sep. 2005.

〔13〕 H. S. Choi, H. H. Park, "Satisfaction Realization of Apartment House Inhabitants for CPTED Design Element: To with Group by CPTED Application Level, Reciprocal Action Effect of Crime Prevention Effort", Korea Security, Science Association, Vol. 22, pp.231-258, Mar. 2010.

〔14〕 S. S. Kim, D. K. Kim, "A Study on Schemes to Activate CPTED in Housing Complex", Korea Association of Criminal Psychology, Vol. 7, No. 1, pp. 55-78, July, 2011.

〔15〕 S. J. Yun, S. J. Lee, S. J. Kang, "A Study on the Applicable Factors for the Crime-free Campus Focused on the CPTED", The Architectural Institute of Korea, Vol. 28, No. 3, pp. 119-126, Mar. 2012.

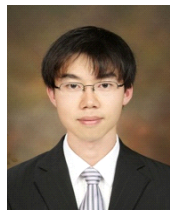
〔16〕 E. H. Lee, S. J. Kang, K. H. Lee, "A Study on the Application of Crime Prevention Through Environmental Design for the District Unit

Plan", The Architectural Institute of Korea, Vol. 24, No. 2, pp. 129-138, Feb. 2008.
[17] H. T. Shin, S. C. Bahn, "A Study on Analysis for the applications of CPTED in Urban Residential Neighborhood", The Architectural Institute of Korea Branch Association Studies, Vol. 2010, No. 1, pp. 109-118, Dec. 2010.
[18] J. H. Choi, D. S. Choi, S. Y. Park, H. K. Oh, "A Method for Improving Recall Precision on Information Retrieval Systems Using Multiple Terms", The Korean Institute of Information Scientists and Engineers, Vol. 25, No. 2, pp. 150-152, Oct. 1998.
[19] E. J. Kim, H. S. Lee, "A Study on Alternative Design Research Model using Unstructured Online Data -through Design Ethnography Methodology-", Society of Design Convergence, Vol. -, No. 42, pp. 205-223, Oct. 2013.

저 자 소 개

이 종 훈

2012: 경남대학교 건축공학과 공학사.
2014: 한양대학교
첨단건축도시환경공학과 공학석사.
현 재: 공간정보산업진흥원 연구원
관심분야: 공간정보, 빅데이터
Email : jh.lee@spacen.or.kr



송 기 성

2007: 인하대학교
지리정보공학과 공학사.
2012: 인하대학교
지리정보공학과 공학석사.
2007 ~ 2012: 지능형국토정보기술
혁신사업단 연구원
2013 ~ 현 재: 공간정보산업진흥원
선임연구원
관심분야: 공간정보기술 테스트베드,
3차원 공간정보, 공간정보
융합산업 등
Email : ks.song@spacen.or.kr

강 진 아

2006: 인하대학교
지리정보공학과 공학사.
2008: 인하대학교
지리정보공학과 공학석사
2008 ~ 2012: 한국건설기술연구원
ICT융합연구실 연구원
현 재: 공간정보산업진흥원 선임연구원
관심분야: 영상 처리, 3차원 공간정보,
사물통신, 공간정보 융복합
Email : ja.kang@spacen.or.kr

황 정 래

2007: 부산대학교
지형정보공학과 공학박사
2007 ~ 2008: 부산대학교 연구교수
2008 ~ 20013: 한국건설기술연구원
박사후연구원
2013 ~ 현 재: 공간정보산업진흥원
수석연구원
관심분야: 3차원 공간정보,
공간데이터모델링,
공간정보표준, BIM/GIS
Email : jr.hwang@spacen.or.kr