KIPS Tr. Comp. and Comm. Sys.
Vol.5, No.9 pp.229~236 pISSN: 2287-5891

http://dx.doi.org/10.3745/KTCCS.2016.5.9.229

An Analysis of Relationship Between Word Frequency in Social Network Service Data and Crime Occurences

Yong-Woo Kim† · Hang-Bong Kang††

ABSTRACT

In the past, crime prediction methods utilized previous records to accurately predict crime occurrences. Yet these crime prediction models had difficulty in updating immense data. To enhance the crime prediction methods, some approaches used social network service (SNS) data in crime prediction studies, but the relationship between SNS data and crime records has not been studied thoroughly. Hence, in this paper, we analyze the relationship between SNS data and criminal occurrences in the perspective of crime prediction. Using Latent Dirichlet Allocation (LDA), we extract tweets that included any words regarding criminal occurrences and analyze the changes in tweet frequency according to the crime records. We then calculate the number of tweets including crime related words and investigate accordingly depending on crime occurrences. Our experimental results demonstrate that there is a difference in crime related tweet occurrences when criminal activity occurs. Moreover, our results show that SNS data analysis will be helpful in crime prediction model as there are certain patterns in tweet occurrences before and after the crime.

Keywords : Social Network Service, Crime Record, Latent Dirichlet Allocation, Tweet Frequency

소셜 네트워크 서비스의 단어 빈도와 범죄 발생과의 관계 분석

김 용 우† · 강 행 봉††

요 약

기존의 범죄 예측 방법들은 범죄 발생을 예측하기 위해 기존 기록을 이용하였다. 그러나 이러한 범죄 예측 모델은 데이터를 갱신하는데 어려움이 있다. 범죄 예측을 향상시키기 위해서 소셜 네트워크 서비스(SNS)를 이용하여 범죄를 예측하는 연구들이 진행되었지만, SNS 데이터와 범죄 기록 사이의 관계에 대한 연구는 미흡하다. 따라서, 본 논문에서는 SNS 데이터와 범죄 발생 사이의 관계를 범죄 예측의 관점에서 분석하였다. 잠재 디리클레 할당(LDA)을 이용하여 범죄 발생과 관련된 단어를 포함하는 트윗을 추출하였고, 범죄 기록에 따른 트윗 빈도의 변화를 분석하였다. 범죄 관련 단어를 포함하는 트윗의 빈도를 계산하고, 범죄 발생에 따라서 트윗 빈도를 분석하였다. 범죄가 발생하였을 때, 범죄와 관련된 트윗의 빈도가 변화하였다. 게다가, 범죄 발생 전후에 트윗 빈도가 특정 패턴을 보이기 때문에 SNS 데이터가 범죄 예측 모델에 도움이 될 것이다.

키워드 : 소셜 네트워크 서비스, 범죄 기록, 잠재 디리클레 할당, 트윗 빈도

1. 서 론

범죄는 언제 어디서나 발생할 수 있다. 범죄 발생을 억제하고 예방하기 위해서 범죄가 언제 어디서 일어날지 예측하는 것이 중요하다. 기존의 범죄예측 방법은 범죄 발생기록과 지리학적 정보를 사용하였다[1-4]. 분석가들은 범죄 예측을 위해서 범죄 발생기록을 이용하여 범죄가 많이 발생하는 지역을 확인하고 시각적으로 표현하였는데 이를 “범죄 핫스팟(Crime hotspots)”이라 한다[1]. 특히, 새로운 범죄는 주로 기존에 범죄가 발생했던 지점 근처에서 발생하기 때문에, 이런 범죄 예측을 위해서 핫스팟과 커널 밀도 추정(Kernel Density Estimation, KDE) 방법을 이용하였다. 하지만, 이러한 방법들은 과거의 기록만을 이용하는데, 미래에 발생할 범죄의 정확한 예측을 위해서는 해당 지역 커뮤니티에서 현재 발생하고 있는 다양한 정보를 반영하는 연구가 필요하다.

해당지역 커뮤니티의 정보는 소셜 네트워크 서비스(Social Network Service, SNS) 데이터로부터 취합할 수 있다. ‘Twitter’와 ‘Facebook’과 같은 소셜 네트워크 서비스는 시간

※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2015R1A2A1A10056304).
† 준 회 원 : 가톨릭대학교 디지털미디어학과 석사과정
†† 종신회원 : 가톨릭대학교 디지털미디어학부 교수
Manuscript Received : August 9, 2016
Accepted : August 29, 2016
* Corresponding Author : Hang-Bong Kang(hbkang@catholic.ac.kr)

과 장소에 구애받지 않고 사용자들이 자신 및 해당지역에 관련된 메세지를 게시하기 때문에, 이런 SNS를 통해 해당 지역의 커뮤니티에 관련된 중요한 정보를 쉽게 획득할 수 있다. 이러한 소셜 네트워크 서비스 데이터로부터 범죄를 예측하는 연구는 SNS 메세지의 주제 분석을 통한 예측 연구[6], 사용자의 감성을 통한 예측 연구[9-11] 및 사람들이 범죄에 대해 가지는 범죄 공포심(fear of crime)을 측정하기 위해 주변 환경을 분석한 연구들이 있다[12-16]. 또한 도시에서의 안전에 대한 연구도 있다[17-19]. 하지만, 소셜 네트워크 서비스 데이터의 변화로부터 범죄 발생 예측을 위한 연구는 미흡한 편이다.

본 논문에서는 소셜 네트워크 서비스 데이터로부터 범죄 발생을 예측하기 위해, 세계적으로 사용되는 소셜 네트워크 서비스인 트위터로 부터 취득한 트윗을 기존의 범죄 기록과의 비교를 통해 소셜 네트워크 서비스 데이터와 범죄 발생 사이에 존재하는 관계를 분석하였다. 먼저, 트윗의 주제를 검출하기 위해 Latent Dirichlet Allocation(LDA)를 사용하였는데, LDA를 통해 검출된 트윗의 주제 중 범죄와 관련이 있는 단어의 빈도를 계산하여 범죄 발생에 따른 변화를 확인하였다. 그리고, 변화의 유의미한 차이를 확인하기 위해, SPSS를 이용하여 대응표본 T-검정을 실시하였다. 이런 통계적 분석을 통해 SNS의 단어 빈도수의 변화와 범죄 발생간의 관계를 분석하였다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 범죄 발생 예측에 관련된 기존 연구들을 기술하였다. 제 3장에서는 실험 방법을 기술하였고, 제 4장에서는 실험 결과 및 기존의 연구와의 비교를 통한 분석 결과를 기술하였다.

2. 관련 연구

소셜 네트워크 서비스를 범죄예측에 이용한 연구는 다음과 같다[1-3]. Sun and Ng [5]는 SNS 메세지의 인기를 보여주는 “Comment Arrival Model”을 개발하고, 마약 남용에 대한 분석을 통해 범죄 예측을 위한 모델을 제안하였다. Wang et al. [6]은 뉴스 트윗의 주제를 추출하고, 이 주제와 범죄기록을 이용하여 ‘Generalized Linear regression Model (GLM)’을 통해 뺑소니 범죄를 예측하였다. Gerber [7]는 소셜 네트워크 서비스 정보를 포함하면 KDE를 이용한 기존 연구들에 비하여 범죄예측 성능이 향상되는 것을 발견하였다. Chen et al. [8]은 트윗의 감정을 추출하고 온도, 이슬점, 습도 등의 날씨데이터와 범죄기록을 수집하여 KDE를 통해 범죄를 예측하였다.

소셜 네트워크 서비스와 범죄의 관계에 대한 연구로서는 SNS 메세지의 주제, 작성자의 감정 및 특정 메시지에 대한 관심 등의 정보와 범죄와의 관계를 분석한 연구가 있다[11-16]. Heverin and Zach [11]는 소셜 네트워크 서비스의 글을 감정과 의견으로 분류하였고, 소셜 네트워크 서비스가 위험과 관련된 정보를 전달하는데 좋은 방법이라는 것을 발견하였다. Willams [12]는 범죄 공포심(fear of crime)이 주변 환경에 따라서 변화하는 것을 발견하였다. Wang and Taylor [13]는 탈출구까지의 거리와 주변 사람들을 볼 수 있는 시야를 범죄 공포심에 영향을 미치는 주변 환경에 포함하였다. 범죄 공포심과 소셜 네트워크 서비스를 이용하여 범죄에 대해 분석한 연구도 있다. Kounadi et al. [14]는 살인사건에 대한 소셜 네트워크 서비스 메세지의 빈도를 시간과 공간으로 나누어 분석하였고, 살인에 대한 대중의 반응을 확인하였다. Furstenberg [15]는 범죄률이 높은 곳에 사는 사람보다 범죄률이 낮은 곳에 사는 사람이 범죄에 대한 고려를 더 많이 한다는 것을 발견했다. McCord et al. [16]는 범죄 발생지표와 범죄인지 사이의 비례관계를 발견했다.

3. 실험 방법

본 논문에서는 소셜 네트워크 서비스 데이터의 변화로부터 범죄 발생 예측을 위해 트윗을 수집하고 트윗의 주제를 검출하기 위해 Latent Dirichlet Allocation (LDA)를 사용한다. 범죄 발생과 트윗과의 관계를 분석하기 위해 LDA를 통해 검출된 트윗의 주제 중 범죄와 관련이 있는 단어의 빈도를 계산하여 범죄 발생에 따른 변화를 검출한다. 끝으로, 변화의 유의미한 차이를 확인하기 위해, 통계 분석을 통하여 SNS의 단어 빈도수의 변화와 범죄 발생간의 관계를 분석한다.

3.1 데이터 수집

본 논문에서는 소셜 네트워크 서비스의 단어빈도를 측정하기 위해 트위터 정보를 수집하였고, 범죄와의 관계를 확인하기 위해서 범죄기록을 수집하였다.

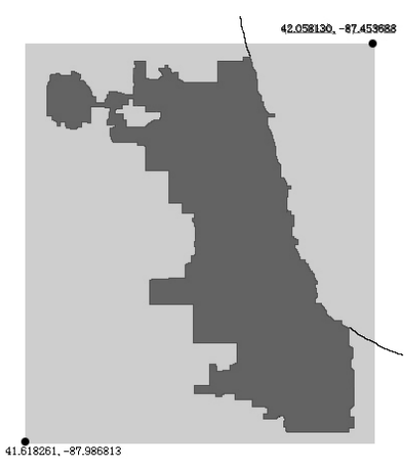

Fig. 1. Data acquisition area (Chicago, USA)

해당 지역은 미국의 시카고 도시로서 수집기간은 2016년 4월1일부터 4월30일까지이며, 이때 시카고에서 발생한 약 80만개의 트윗을 수집하였다. 트위터 정보는 트위터에서 제공하는 Python Streaming API인 ‘Tweepy’를 이용하여 수집하였다. 트위터 수집범위는 위도와 경도를 이용하였는데, 시

카고 데이터를 수집하기 위해 시카고가 포함되는 위도와 경도인 [41.618261, -87.985813]와 [42.058130, -87.453688] 사이에서 발생한 트윗을 수집하였다. 수집한 트윗 데이터는 발생 시간, 내용, 발생위치, 위도 및 경도를 기록하였다. Fig. 1은 수집한 지역을 나타낸 그림이다. 회색 부분이 트윗을 수집한 부분이고, 빨간색 선은 시카고 도시 지역을 나타낸다.

소셜 네트워크 서비스의 단어빈도와 범죄와의 관련성을 확인하기 위해 과거의 범죄기록을 수집하였다. 범죄기록은 'City of Chicago Data Portal [20]'를 통해 시카고 데이터를 수집하였으며, 트위터 데이터와 동일하게 2016년 4월 1일부터 4월 30일까지 발생한 범죄기록을 수집하였다. 범죄종류는 살인만을 수집하였는데, 살인은 범죄 발생 후 신고되지 않는 경우가 적기 때문에 살인에 대한 범죄기록을 수집하였다.

3.2 데이터 분석

수집한 데이터에는 시카고에서 발생하지 않은 트윗이 포함되어 있는데, 이를 제거하기 위해서 트윗 발생위치와 위도 및 경도를 기록하였다. 그러나 위도와 경도가 수집되지 않는 트윗이 많기 때문에 발생위치를 이용하여 시카고에서 발생한 트윗을 분류하였다.

트윗의 위치를 분류한 다음, 각각의 트윗의 주제를 확인하기 위해서 LDA를 사용하였다. LDA는 문서 [doc_n]에서 주제 [T_1, T_2, …, T_k]를 추출하는 것이다[21-22]. LDA는 단어 수 분포를 분석하여 주제를 추출하는 확률모형이며, 효율적으로 단어기반의 주제를 발견한다. LDA를 이용하여 각 트윗마다 3개의 주제 [T_1, T_2, T3]를 추출하였다. 4개의 주제를 추출하면 글의 주제와 관련 없는 단어가 나오는 트윗이 많기 때문에 3개의 주제를 추출하였다. 그리고 범죄와 관련된 단어를 선정하고 LDA를 통해 추출한 주제와 비교하여 범죄관련 주제의 빈도를 계산하였다. Table 1은 범죄와 관련된 트윗을 선별하기 위해 사용한 주제이다. 주제는 범죄에 사용될 수 있는 무기, 피해자, 가해자, 수상한 사람에 대한 단어를 선정하였고, 피해자가 느낄 수 있는 두려움과 가해자가 범행을 저지르기 전 느낄 것이라고 판단되는 감정을 선정하였다.

Table 1. The words related to crime

Crime	Homicide	Weapon	Knife
Gunfight	Victim	Criminal	Attack
Strange	Fear	Disgust	Loathing

범죄기록을 이용하여 범죄가 발생하지 않은 시점으로부터 한 시간 단위로 주제가 범죄와 관련이 있는지 확인하고, 범죄와 관련이 있는 주제의 빈도를 계산하였다. 주제의 빈도는 해당 시점의 6시간 전부터 6시간 후까지 계산하였다. 또한, 같은 방법으로 범죄가 발생한 시점에서의 주제 빈도를 계산하였다. Fig. 2는 데이터 분석 과정을 도식화한 것이다. 소셜 네트워크 서비스인 트위터의 데이터를 수집하고, LDA를 이용하여 주제를 추출하였다. 시카고의 과거 범죄기록을 수집하여, 범죄 발생 및 미발생 시점을 생성하였다. 범죄기록을 통해 구한 시점에서 LDA를 통해 추출한 주제와 범죄관련 단어로 선정한 것을 비교하여 범죄관련 트윗 빈도를 구하였다.

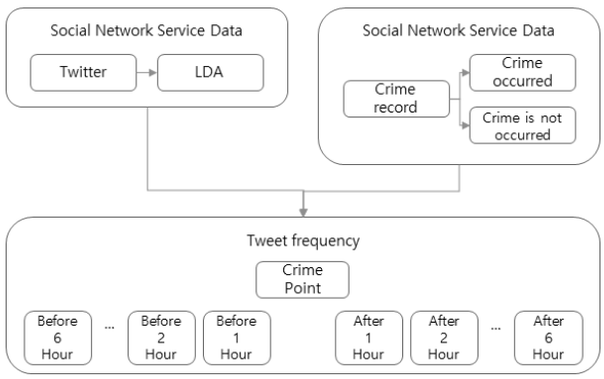

Fig. 2. Data analysis procedure

범죄가 발생하였을 때 시간의 흐름에 따른 차이를 확인하기 위해, 범죄기록을 이용하여 범죄가 발생한 시점과 발생하지 않은 시점에서의 전후 6시간의 트윗 빈도를 구하여 차이를 비교하였다. 그리고 범죄가 발생했을 때의 변화를 확인하기 위해, 범죄 발생 전 6시간부터 범죄 발생 후 6시간까지 시간의 흐름에 따른 트윗 빈도 변화를 확인하였다.

범죄와 관련된 주제 빈도의 유의미한 차이를 확인하기 위해, 주제 빈도에 대한 통계분석을 실시하였다. 동일한 범죄에 대해 발생한 트윗이고 트윗의 발생시간만 다르기 때문에 SPSS를 이용하여 대응표본 T-검정(Paired Samples T-Test)을 실시하였다. 통계분석을 진행한 다음, 그래프를 통해 범죄 발생에 따른 트윗 빈도의 일정한 경향을 확인하였다. 전체적인 경향과 다른 부분은 다르게 발생한 이유에 대해서도 분석하였으며, 전체적인 평균을 구하여 범죄 발생 전후의 경향을 분석하였다.

4. 실험 결과

4.1 범죄 미발생

범죄 발생의 유무에 따른 차이를 확인하기 위해, 범죄가 발생하지 않은 5시점에서의 빈도와 범죄가 발생한 17시점의 빈도를 확인하였다.

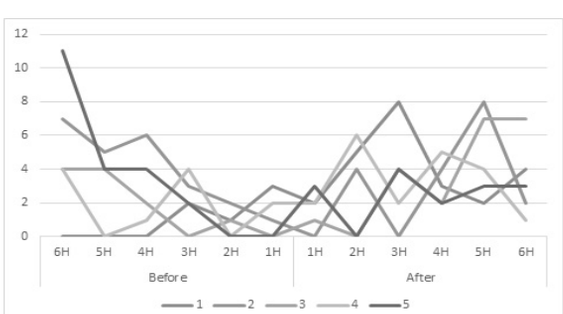

Fig. 3. Word frequency when crime is not occurred

Table 2. T-test on paired samples in topic frequency of tweets of 6 hours before a certain point (no crime case)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Before_1H - Before_2H	.40000	1.51658	.67823	-1.48308	2.28308	.590	4	.587
Pair 2	Before_2H - Before_3H	-1.40000	1.81659	.81240	-3.65559	.85559	-1.723	4	.160
Pair 3	Before_3H - Before_4H	-.40000	2.70185	1.20830	-3.75479	2.95479	-.331	4	.757
Pair 4	Before_4H - Before_5H	.00000	1.22474	.54772	-1.52072	1.52072	.000	4	1.000
Pair 5	Before_5H - Before_6H	-2.60000	2.96648	1.32665	-6.28337	1.08337	-1.960	4	.122

Table 3. T-test on paired samples in topic frequency of tweets of 6 hours after a certain point (no crime case)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	After_1H - After_2H	-1.40000	3.20936	1.43527	-5.38495	2.58495	-.975	4	.385
Pair 2	After_2H - After_3H	-.60000	4.21900	1.88680	-5.83859	4.63859	-.318	4	.766
Pair 3	After_3H - After_4H	.40000	3.78153	1.69115	-4.29539	5.09539	.237	4	.825
Pair 4	After_4H - After_5H	-1.60000	2.79285	1.24900	-5.06778	1.86778	-1.281	4	.269
Pair 5	After_5H - After_6H	1.40000	3.13050	1.40000	-2.48702	5.28702	1.000	4	.374

Fig. 3은 범죄가 발생하지 않은 시점에서의 주제 빈도를 도식화한 것이며, 5시점에서 모두 특정한 경향을 보이지 않는다. 선정한 시점에서 전후보다 비교적 적은 빈도를 보였으나, 통계적으로 유의미한 차이는 없었다.

그래프를 통해 특정 경향을 보이지 않는 것을 확인하였으나, 유의미한 차이를 확인하기 위해 대응표본 T-검정을 실시하였다. Table 2와 Table 3은 범죄가 일어나지 않은 시점의 데이터에 대한 대응표본 T-검정의 결과이다. 모든 시간대에서 모두 유의미한 차이를 보이지 않았다. 범죄가 발생했을 경우에는 유의미한 차이를 보였으며, 4.2장에서 자세히 기술한다.

4.2 범죄 발생

계산한 트윗 빈도가 유의미한 차이를 보이는 것을 확인하기 위해, 대응표본 T-검정을 실시하였다. Table 4는 범죄 발생 이전의 주제 빈도에 대하여 대응표본 T-검정을 실시한 결과이다. 범죄 발생 4시간 전까지는 계속 유의미한 차이를 보이지만, 5시간 이후부터는 유의미한 차이를 보이지 않는다. Table 5는 범죄 발생 이후의 주제 빈도에 대하여 대응표본 T-검정을 실시한 결과이다. 범죄 발생 이후 3시간과 4시간 사이를 제외하면 모두 유의미한 차이를 보였다. 소셜 네트워크 서비스가 위험을 전달하기에 좋은 방법이라는 연구 [11]처럼, 소셜 네트워크 서비스는 범죄와 관련된 정보를 전달하고 있다.

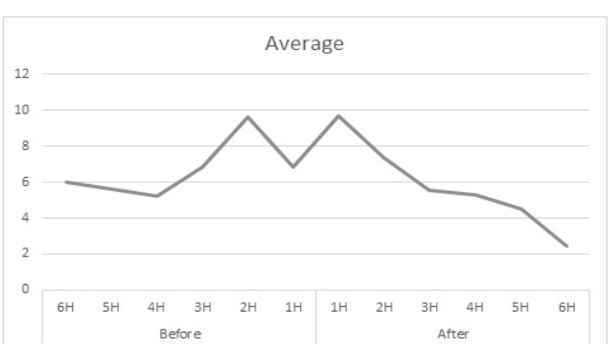

Fig. 4. Average graph of topic frequency when crime is occurred

Table 4. T-test on paired samples in topic frequency of tweets of 6 hours before crime

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Before_1H - Before_2H	-2.82353	5.10190	1.23739	-5.44669	-.20037	-2.282	16	.037
Pair 2	Before_2H - Before_3H	2.82353	4.47542	1.08545	.52248	5.12458	2.601	16	.019
Pair 3	Before_3H - Before_4H	1.58824	2.78520	.67551	.15622	3.02025	2.351	16	.032
Pair 4	Before_4H - Before_5H	-.35294	2.71434	.65832	-1.74853	1.04264	-.536	16	.599
Pair 5	Before_5H - Before_6H	-.41176	3.62386	.87891	-2.27498	1.45145	-.468	16	.646

Table 5. T-test on paired samples in topic frequency of tweets of 6 hours after crime

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	After_1H - After_2H	2.35294	4.40087	1.06737	.09022	4.61566	2.204	16	.042
Pair 2	After_2H - After_3H	1.82353	3.50420	.84989	.02184	3.62522	2.146	16	.048
Pair 3	After_3H - After_4H	.23529	2.53795	.61554	-1.06960	1.54019	.382	16	.707
Pair 4	After_4H - After_5H	.76471	1.43742	.34863	.02565	1.50376	2.193	16	.043
Pair 5	After_5H - After_6H	2.05882	2.56102	.62114	.74207	3.37558	3.315	16	.004

Fig. 4는 주제 빈도의 평균을 도식화한 것이다. 범죄 발생 4시간 전부터 2시간 전까지 범죄와 관련된 주제를 가진 트윗의 빈도가 증가하다가 1시간 이전에 감소한다. 범죄 발생 5시간 전과 6시간 전 사이에는 유의미한 차이가 없는데, 그래프 또한 변화가 없다. 범죄 발생 이후에는 범죄와 관련된 주제를 가진 트윗의 빈도가 지속적으로 감소하는 것을 확인할 수 있다. 범죄 발생 이후 3시간부터 4시간 사이에는 변화가 없었으며, 범죄와 관련된 트윗의 빈도가 비슷하게 유지된다. 범죄 발생 이후 시간이 지남에 따라서 트윗의 빈도가 감소하는 것은 기존 연구[14]와 비슷한 결과를 보인다.

Fig. 5는 각 범죄에 대한 주제 빈도를 도식화한 것이다. 대부분 범죄 발생 2시간 전까지 증가하다가 범죄 발생 1시간 전에 감소하고, 범죄 발생 이후 다시 증가하였다가 지속적으로 감소한다. 7번과 9번, 10번, 12번, 13번, 15번, 16번은 범죄 발생 5시간부터 6시간 전의 빈도가 4시간 전의 빈도보다 높게 나타나는데, 이는 이전 범죄와 관련된 트윗이 함께 수집되었기 때문이다. 또한, 2번, 3번, 17번은 범죄 발생 이후 트윗 빈도가 감소하다가 다시 증가하는데, 이후에 일어난 범죄와 관련된 트윗이 함께 수집되었기 때문이다.

4.3 분석

범죄가 발생하지 않았을 때는 트윗 빈도의 유의미한 차이가 없었고, 범죄가 발생했을 때는 트윗 빈도의 유의미한 차이가 있었다. 범죄가 발생하면, 범죄관련 트윗의 빈도가 특정 경향에 따라서 변화한다는 것을 알 수 있다. 범죄 발생에 따른 변화가 존재하기 때문에 범죄관련 트윗의 빈도를 범죄 예측 요소로써 사용할 수 있다.

소셜 네트워크 서비스가 위험과 관련된 정보를 전달하는데 좋은 방법이라는 것을 발견한 연구[11]처럼, 범죄가 발생하지 않은 시점에는 낮은 트윗 빈도를 보였고 일정한 경향이 나타나지 않았지만, 범죄가 발생한 시점에는 높은 트윗 빈도를 보였고 범죄 발생 전후에 일정한 경향을 나타냈다. 특히, 본 연구에서는 범죄가 발생한 경우, 범죄 발생 2시간 전까지 범죄와 관련된 트윗이 증가하는 것은 범죄에 대한 사람들의 걱정과 범죄 모의 때문이고, 1시간 전에 감소하는 것은 범죄를 일으키는 사람들이 트윗 사용을 자제하기 때문이라는 사실을 발견하였다.

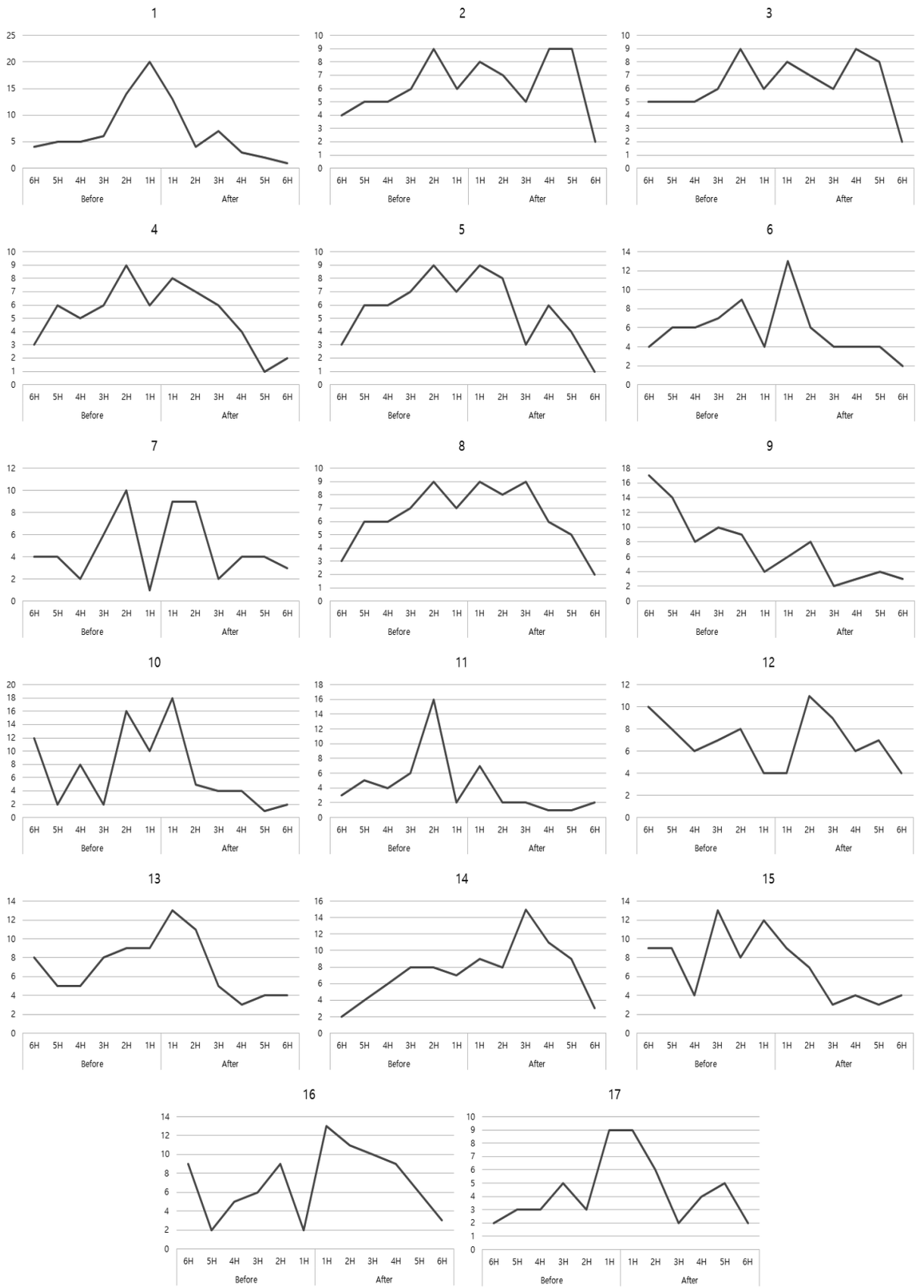

Fig. 5. Topic frequency for each crime when crime is occurred

범죄 발생 이후에는 범죄와 관련된 트윗이 지속적으로 감소한다. 범죄관련 트윗의 70%가 한달 이내에 발생하고 그중 25%가 하루 이내에 발생한다는 기존 연구[14]와 유사하게, 본 논문에서는 범죄 발생 6시간 이후 범죄관련 트윗 빈도가 낮은 것을 알 수 있었다. 특히, 해당 범죄와 관련된 트윗 중 대다수의 트윗이 범죄 발생 후 6시간 이내에 발생하였다. 범죄소식을 접했을 때 사람들이 범죄에 대한 두려움, 걱정 및 관심 때문에 범죄관련 트윗의 빈도가 증가하지만, 시간이 지남에 따라 범죄가 해결되거나 범죄에 대해 잊어버리는 등의 이유로 해당 범죄와 관련된 트윗의 빈도가 줄어든다.

유사한 시간에 여러 범죄가 발생하는 경우, 다른 범죄의 트윗 빈도에도 영향을 미친다. 위치정보를 이용하여, 이를 해결할 수 있을 것으로 예상된다. 본 논문에서는 위치정보를 가지는 트윗이 적기 때문에 위치정보를 활용하지 못하였다. 그러나 본 논문에서 측정한 전후 시간보다 시간을 짧게 지정하여 다른 범죄의 영향을 줄임으로써 특정 범죄에 대한 트윗을 구별할 수 있다.

5. 결 론

본 논문에서는 범죄와 SNS 데이터 사이의 관계를 확인하기 위하여, 트위터와 범죄기록 데이터를 수집하고 범죄 발생 예측 관점에서 분석하였다. LDA를 이용하여 발생 트윗 주제를 추출하였다. 추출한 주제와 범죄 관련 단어가 일치하는 트윗의 개수를 계산하였고, 범죄가 발생한 경우와 발생하지 않은 경우에 따라 통계 분석을 진행하였다. 범죄가 발생하지 않은 경우, 대응표본 T-검정에서 유의미한 차이를 보이지 않았고, 그래프 상으로 특정 경향을 보이지 않았다. 범죄가 발생한 경우, 대응표본 T-검정에서 유의미한 차이를 보였다. 범죄에 대한 걱정과 범죄 모의 때문에 범죄 관련 트윗 빈도는 범죄 발생 2시간 전까지 증가하다가 1시간 전에 줄어든다. 아울러, 해당 범죄에 대한 사람들의 관심 때문에 범죄 발생 이후에 다시 증가하였다가 점차적으로 감소한다. 따라서, 범죄 관련 트윗 빈도는 범죄가 발생하면 유의미한 차이를 보이며, 발생 전후에 특정 경향을 보이기 때문에 SNS 정보가 범죄 예측 요소로써 중요하다는 사실을 발견하였다.

향후 연구로는 살인뿐만 아니라 다른 다양한 범죄 종류에 따른 트윗 빈도 변화를 확인하고, 범죄 종류에 따른 트윗의 증감 차이를 확인하는 것이다. 아울러, 본 논문에서는 위치정보(GPS)를 가지는 데이터의 양이 너무 적기 때문에 분석에 사용하지 못하였지만, 정확한 범죄 발생 예측을 위해서는 범죄 발생 위치에 따른 SNS 메세지의 빈도 변화를 확인하는 것도 필요하다.

References

[1] J. Eck, S. Chainey, J. Cameron, M. Leitner, and R. Wilson, "Mapping Crime: Understanding Hot Spots," U.S. Department of Justice, 2005.

[2] S. Chainey, L. Tompson, and S. Uhlig, "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security Journal*, Vol.24, No.1-2, pp.4-28, 2008.

[3] I. Van Patten, J. McKeldin-Conor, and D. Cox, "A Microspatial Analysis of Robbery: Prospective Hot Spotting in a Small City," *Crime Mapping: A Journal of Research and Practive*, Vol.1, No.1, pp.7-32, 2009.

[4] J. Ratcliffe, "Crime Mapping: Spatial and Temporal Challenges," *Handbook of Quantitative Criminology*, Springer New York, pp.5-24, 2009.

[5] B. Sun and V. Ng, "Lifespan and popularity measurement of online content on social networks," *Intelligence and Security Informatics(ISI)*, pp.379-383, 2011.

[6] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic Crime Prediction Using Events Extracted from Twitter Posts," *Social Computing, Behavioral - Cultural Modeling and Prediction*, pp.231-238, 2012.

[7] M. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, Vol.61, pp.115-125, 2014.

[8] X. Chen, Y. Cho, and S. Jang, "Crime Prediction Using Twitter Sentiment and Weather," *Systems and Information Engineering Design Symposium*, pp.63-68, 2015.

[9] B. Back, I. Ha, B. Ahn, "An Extraction Method of Sentiment Infromation from Unstructed Big Data on SNS," *Korea Multimedia Society*, Vol.17, No.6, pp.671-680, 2014.

[10] M. Nam, E. Lee, and H. Shin, "A Method for User Sentiment Classification using Instagram Hashtags," *Korea Multimedia Society*, Vol.18, No.11, pp.1391-1399, 2015.

[11] T. Heverin, and L. Zach, "Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in the Seattle-Tacoma, Washington Area," *ISCRAM*, 2010.

[12] C. Williams, "Mapping the fear of crime - a micro-approach in Merton, London," *Crime Mapping Case Studies: Practice and Research*, 2008.

[13] K. Wang, R. Taylor, "Simulated walks through dangerous alleys: Impacts of features and progress on fear," *Journal of Environmental Psychology*, Vol.26, No.4, pp.269-283, 2006.

[14] O. Kounadi, T. Lampoltshammer, E. Groff, I. Sitko, and M. Leitner, "Exploring Twitter to Analyze the Public's Reaction Patterns to Recently Reported Homicides in London," *PLOS ONE*, Vol.10, No.3, pp.1-15, 2015.

[15] F. Furstenberg, "Public reaction to crime in the streets," *The American Scholar*, Vol.40, No.4, pp.601-610. 1971.

[16] S. McCord, H. Ratcliffe, M. Garcia, and B. Taylor, "Nonresidential crime attractors and generators elevate perceived neighborhood crime and incivilities," *Journal of Research in crime and delinquency*, Vol.44, No.3, pp.295-320, 2007.

[17] H. Kang and H. Kang, “A Safety Score Prediction Model in Urban Environment Using Convolutional Neural Network,” *Korea Information Processing Society*, Vol.5, No.8, pp.393-400, 2016.

[18] H. Kang and H. Kang, “A New Context-Aware Computing Method for Urban Safety.” *International Conference on Image Analysis and Processing*, pp.298-305, 2015.

[19] A. Crawford and J. Flint, “Urban safety, anti-social behaviour and the night-time economy,” *Criminology and Criminal Justice*, Vol.9, No.4, pp.403-413, 2009.

[20] City of Chicago Data Portal [Internet], https://data.cityofchicago.org/.

[21] D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” *Signal Processing Magazine*, Vol.27, No.6, pp.55-65, 2010.

[22] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, pp.993-1022, 2003.

김 용 우

e-mail : kyw93@catholic.ac.kr

2016년 가톨릭대학교 디지털미디어학부 (학사)

2016년～현 재 가톨릭대학교 디지털미디어학과 석사과정

관심분야 : Computer Vision, Artificial Intelligence, Machine Learning, Big Data

강 행 봉

e-mail : hbkang@catholic.ac.kr

1980년 한양대학교 전자공학과(학사)

1986년 한양대학교 전자공학과(석사)

1989년 Ohio State Univ. 컴퓨터공학(석사)

1993년 Rensselaer Polytechnic Institute 컴퓨터공학(박사)

1993년～1997년 삼성종합기술원 수석연구원

1997년～현 재 가톨릭대학교 디지털미디어학부 교수

관심분야 : Computer Vision, Machine Learning, HCI, Artificial Intelligence, Computer Graphics, Big Data