Journal of the Korea Convergence Society
Vol. 9. No. 5, pp. 15-20, 2018

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2018.9.5.015

범죄 발생 빈도수와 웹 검색 빈도수의 관계 분석 연구

박정민, 박구락*, 정영석
공주대학교 컴퓨터공학과

Analysis of relationship between frequency of crime occurrence and frequency of web search

Jung-Min Park, Koo-Rack Park*, Young-Suk Chung
Dept of Computer Science, Kongju National University

요 약 현대사회에서 범죄는 큰 사회문제 중의 하나이다. 범죄는 피해자뿐만 아니라 피해자 주변인들에게도 큰 영향을 미친다. 범죄는 발생하기 전에 예측하여 범죄 발생을 막는 것이 중요하다. 범죄를 예측하기 위한 다양한 연구가 진행되었다. 범죄 예측에 중요한 요소 중에 하나가 범죄 발생 빈도수 이다. 범죄 발생 빈도수는 범죄를 예측하는 분야의 기본 데이터로 많이 사용되고 있다. 그러나 범죄 발생 빈도수는 통계처리기간을 거쳐 약 2년 뒤에 발표된다.
본 논문은 범죄 발생 빈도수를 간접적으로 파악할 수 있는 방법으로 웹에서 검색되는 범죄 관련 키워드의 빈도수 분석을 제안한다. 범죄 발생 빈도수의 키워드와 실제 범죄 발생빈도수의 관계를 상관 계수로 분석하여 관련이 있음을 확인하였다.

주제어 : 범죄, 상관계수, 빅 데이터, 트렌드

Abstract In modern society, crime is one of the major social problems. Crime has a great impact not only on victims but also on those around them. It is important to predict crimes before they occur and to prevent crime. Various studies have been conducted to predict crime. One of the most important factors in predicting crime is frequency of crime occurrence. The frequency of crime is widely used as basic data for predicting crime. However, the frequency of crime occurrence is announced about 2 years after the statistical processing period. In this paper, we propose a frequency analysis of crime - related key words retrieved from the web as a way to indirectly grasp the frequency of crime occurrence. The relationship between the number of frequency of crime occurrence and frequency of actual crime occurrence was analyzed by correlation coefficient.

Key Words : Crime, Correlation coefficient, Big Data, Trend

1. 서론

현대 사회의 큰 문제 중의 하나가 범죄이다. 범죄는 범죄 피해자뿐만 아니라 피해자 가족에게도 큰 상처로 남는다. 범죄는 발생하기 전에 예측하여 예방활동을 하는 것이 중요하다. 그래서 범죄 발생을 예측하는 기본 데이터로 범죄발생빈도수가 필요하다. 그러나 범죄발생 빈도수는 대검찰청 또는 나라지표에서 확인 할 수 있다. 나라지표는 공공데이터를 활용하는데 큰 역할을 하고 있다[1,2]. 그러나 범죄 발생 빈도수의 기초자료는 각 경찰서에서 모아져서 대검찰청에서 최종적으로 처리하므로, 범죄 통계 자료 처리에 시간이 걸려 현재를 기준으로 약 2년 전의 자료만 알 수가 있다[3]. 그러나 범죄는 계속 발생하고 있다. 범죄 발생 빈도수를 간접적으로 확인 할 수 있다면, 범죄 예측 및 예방 정책 수립에 도움을 줄 수 있다. 범죄 발생을 간접적으로 알 수 있는 방법 중의 하나가 범죄에 대한 관심도이다. 특정 범죄에 대한 관심도가 높다는 것은 그 범죄가 증가 하고 있다는 것을 의미 할 수 있다. 예를 들어 인터넷에서 관심도에 의해 발생하는 키워드를 분석해 실제 사건과 비교하는 구글 트렌드를

*Corresponding Author : Koo-Rack Park (ecgrpark@kongju.ac.k)
Received March 6, 2017 Revised April 23, 2018
Accepted May 20, 2018 Published May 28, 2018

활용한 독감예측이 있다[4]. 그러나 독감예측시스템은 특정 키워드와의 연관성을 입증하거나 분석하는 알고리즘이 누락되어, 독감예측이론이 유행하자 실제 독감 현상과 관련 없는 독감 키워드 검색 횟수가 갑자기 늘어 독감 예측이 어려워졌다.

본 논문에서는 검색 키워드와의 연관성을 입증하기 위해 범죄에 대한 키워드와 실제 범죄 발생 빈도수와의 관계를 상관계수를 적용하여 분석하였다. 상관계수는 두 변수 사이의 관계를 분석하는데 사용된다. 실제 발생한 범죄의 발생 빈도수와 검색 빈도수의 관계에 대해 분석하고 논의하였다. 본 논문의 구성은 다음과 같다. 2장에서 관련연구인 상관계수와 웹 검색트래픽에 대해 논의한다. 3장에서 웹 검색 키워드의 빈도수를 얻기 위한 알고리즘에 대해 논의하고 4장에서 실제 범죄수와 웹 검색 키워드 사이의 관계를 상관계수를 적용하여 분석하고 5장에서 결론 및 향후 연구관제에 대해 논의한다.

2. 관련연구

2.1 상관계수

두 개의 변수 x 와 y가 있을 때, x값의 데이터 값에 따라 y값의 데이터 값도 변화하는 관계를 상관관계라고 한다. x 와 y 두 변수 간 상관관계 유무를 수치로 판단하는 지표로 상관계수를 적용한다. 상관계수는 r이라는 기호로 표기되고, -1에서 1까지의 값을 취한다.

$$-1 \leq r \leq 1$$

상관계수 수식은 다음과 같다[3,4].

$$S(xx)=\sum_{i=1}^{n}(x_i-\overline{x_i})^2=\sum_{i=1}^{n}x^{2_i}-(\sum_{i=1}^{n}x_i)^2/n \quad (1)$$

$$S(yy)=\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}y^{2_i}-(\sum_{i=1}^{n}y_i)^2/n \quad (2)$$

$$S(xy)=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)/n \quad (3)$$

$$r=\frac{S(xy)}{\sqrt{S(xx)S(yy)}} \quad (4)$$

단, $S(xx)=x$의 편차제곱의 합,
$S(yy)=y$ 의 편차제곱의 합,
$S(xy)=x$ 와 y의 편차 곱의 합,
$r=$ 상관계수

상관성의 밀접도를 정의하는 상관계수는 주장하는 논문마다 차이가 있으나, 보편적인 조건은 다음과 같다[5,6].

$0.8 \leq |r| \rightarrow$ 강한 상관관계
$0.6 \leq |r| \leq 0.8 \rightarrow$ 상관관계 있음
$0.4 \leq |r| \leq 0.6 \rightarrow$ 약한 상관관계
$|r| \leq 0.4 \rightarrow$ 상관관계 거의 없음

2.2 빅 데이터

현재 IT 뉴스의 핵심 키워드로 빅데이터(Big-data)가 있다[7,8]. 빅 데이터란 대용량 데이터 활용하고 분석하여 가치 있는 정보를 추출하고, 이를 바탕으로 대응 방안 도출 또는 변화를 예측하기 위한 정보화 기술을 말한다[9-11]. 세계 각국의 정부와 주요 민간 기업들은 빅데이터가 새로운 경제적 가치의 원천이 될 것으로 예상하고, 빅데이터를 활용한 시장 동향, 신산업 발굴 등 경제적 가치 사례 및 효과를 제시하고 있다[12,13]. 빅데이터를 활용하기 위해 IT기업들은 다양한 서비스를 제공하고 있는데 그 예로 구글 트렌드와 네이버 트렌드가 있다. 인터넷에서 발생하는 빈도수 예측에 활용한 연구로 구글 트렌드를 활용한 독감 예측이 있다. 실제 독감 발생 빈도수와 구글 검색 빈도수 사이에 밀접한 관계가 있다는 것이 증명되었다[4]. 소비자의 웹 검색 트래픽을 정보를 활용하여 네트워크 모델링의 방법을 적용하여 소비자가 원하는 태블릿 PC 제품의 알 수 있는 지능형 브랜드 포지셔닝 시스템에 관련된 연구가 있었다[14]. 한국여행을 준비하는 중국인 관광객들의 웹 검색 수치를 활용하여 여행 수요예측에 활용한 연구가 있었다[15,16].

3. 범죄 관련 키워드 빈도수 수집 및 범죄 발생 빈도수 연관성 분석 모델

본 논문은 웹에서 생성되는 범죄 관련 단어의 빈도수와 실제 범죄발생 빈도수의 관련성에 대해 연구 하였다.

범죄 관련 키워드 수집 및 범죄 발생 빈도수 연관성 분석 모델의 처리 절차는 다음의 Fig. 1과 같다.

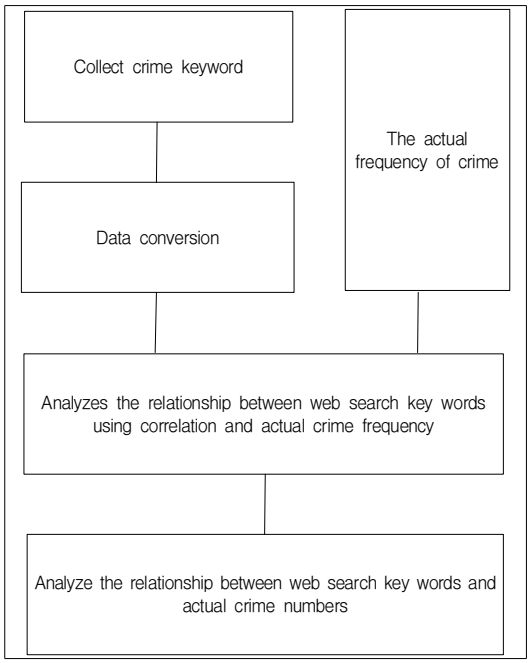

Fig. 1. Procedures for collecting crime-related key words and analyzing the frequency of crime occurrence associations

첫 번째, 범죄 키워드 수집을 한다.

웹에 발생하는 범죄와 관련된 키워드의 빈도수를 수집한다. 빈도수를 수집하기 위해서 한글 검색 량이 많은 네이버 트렌드 서비스를 활용하였다. 네이버 트렌드 서비스를 활용하면 특정 단어의 빈도수를 시간 진행 순서에 따른 빈도수를 확인 할 수 있다.

두 번째, 데이터 변환을 한다.

네이버 트렌드 서비스는 기간별 키워드의 조회 빈도수를 수치로 제공하고 있다. 수치는 일정 기간의 조회 빈도수 중 가장 큰 값을 100으로 하여 나누어준 값이다. 일별로 정리된 키워드의 조회 빈도수를 실제 범죄 발생 빈도수와 비교하기 위해 월별 빈도수로 변환한다.

세 번째, 실제 범죄 발생 빈도수를 정리한다.

대 검찰청에서는 매년 각종 범죄의 발생 빈도 및 처리 현황을 종합한 범죄분석 자료를 공개하고 있다. 공개된 자료는 통계처리를 거치므로 2년 전의 자료이다. 예를 들면 2018년 현재 얻을 수 있는 범죄 분석 자료는 2016년의 자료를 공개하고 있다.

네 번째, 웹 검색 키워드와 실제범죄 발생 빈도수의 관계를 분석한다.

두 변수간의 관계의 정도를 알아보기 위해 상관 계수가 활용된다. 본 연구에서도 상관관계를 분석하여 웹 검색 키워드의 빈도수와 실제 범죄 발생 빈도수를 비교 분석한다.

마지막 단계로 웹 검색 키워드와 실제 범죄 발생 빈도수와의 관계를 확인하다.

웹에서 검색된 키워드의 횟수와 실재 범재 발생 빈도수의 관계를 분석하여 어떤 범죄 키워드가 실제 범죄 발생 빈도수와 관계가 있는지를 확인한다.

4. 범죄 관련 키워드 빈도수 수집 및 범죄 발생 빈도수 연관성 분석 모델에 적용

다양한 범죄 중 5대 범죄에 속하는 강도, 절도. 폭력을 본 연구에서 제안하는 범죄 관련 키워드 수집 및 범죄 발생 빈도수 연관 모델에 적용하였다. 범죄 관련 키워드의 수집기간은 대 검찰청 범죄 발생 빈도수의 자료와 같아야 하므로 2016년 1월1일부터 2016년 12월31일까지의 데이터를 수집하였다.

4.1 강도

5대 범죄 중 강도의 네이버 트렌드 서비스를 이용한 기간별 키워드 검색 조회 그래프는 다음 Fig. 2와 같다.

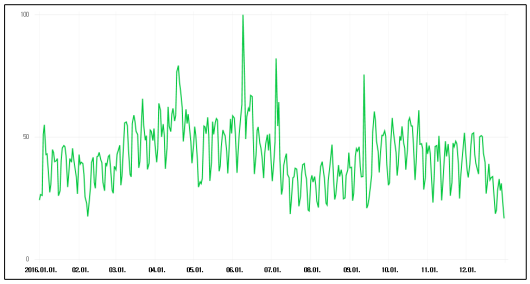

Fig. 2. Robber Naver Trend Graph

네이버 트렌드의 검색횟수를 월별로 정리한 것과 실제 범죄 발생 빈도수는 다음의 Table 1과 같다.

강도 범죄 발생 빈도수와 트렌드 검색회수를 상관관계 식(4)에 넣어 분석한 결과는 -0.06으로 음의 상관관계를 나타내었다.

Table 1. Robber Frequency of Crime and Naver Trend

Month of occurrence	Frequency of crime	Trend search count
January	130	1174
February	96	1014
March	99	1470
April	97	1674
May	78	1440
June	81	1594
July	95	1128
August	83	1013
September	87	1250
October	69	1412
November	68	1200
December	72	1106

4.2 절도

5대 범죄 중 절도의 네이버 트렌드 서비스를 이용한 기간별 키워드 검색 조회 그래프는 다음 Fig. 3과 같다.

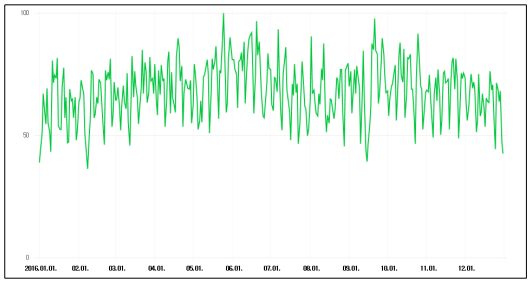

Fig. 3. Theft Naver Trend Graph

절도 범죄를 네이버 트렌드에서 검색된 검색횟수를 월별로 정리한 것과 실제 범죄 발생 빈도수는 다음의 Table 2와 같다.

Table 2. Theft Frequency of Crime and Naver Trend

Month of occurrence	Frequency of crime	Trend search count
January	15,502	1860
February	14,689	1841
March	16,644	2085
April	16,901	2111
May	18,315	2271
June	18,189	2309
July	18,782	2073
August	18,449	2071
September	17,817	2104
October	17,623	2186
November	15,294	2046
December	15,368	1956

절도 범죄 발생 빈도수와 트렌드 검색회수를 상관관계 식(4)에 넣은 결과는 0.76 으로 강한 상관관계를 나타내었다.

4.3 폭력

5대 범죄 중 폭력의 네이버 트렌드 서비스를 이용한 기간별 키워드 검색 조회 그래프는 다음 Fig. 4와 같다.

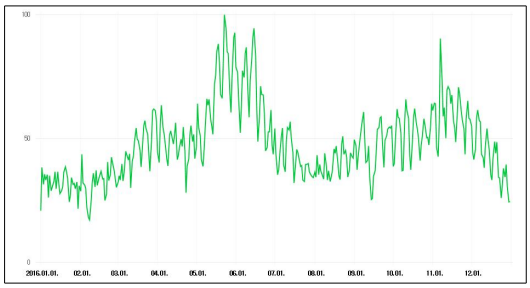

Fig. 4 Violence Naver Trend Graph

폭력 범죄의 네이버 트렌드의 검색된 횟수를 월별로 정리한 것과 실제 범죄 발생 빈도수는 다음의 Table 3 과 같다.

Table 3. violence Frequency of Crime and Naver Trend

Month of occurrence	Frequency of crime	Trend search count
January	18288	977
February	17786	932
March	20578	1416
April	22710	1431
May	22415	2123
June	22676	1995
July	22788	1307
August	22153	1226
September	22982	1402
October	21512	1605
November	18781	1828
December	19200	1308

폭력 범죄 발생 빈도수와 트렌드 검색회수를 상관관계 식(4)에 넣어 분석한 결과는 0.46으로 양의 상관관계를 나타내었다.

5. 결론 및 향후 연구과제

범죄는 피해자가 발생하기 전에 예측하는 것이 필요하다. 범죄를 예측하기 위해서는 실시간으로 발생하는 범죄 발생 데이터가 필요하다. 그러나 현재 발표되는 범죄 발생 데이터는 대 검찰청에서 1년에 한번 씩 발표하는 백서인 범죄 분석이 있으나 2년 전의 범죄통계자료이다. 그래서 현재 발생하는 범죄의 빈도수를 파악하기는 부족하다.

본 논문은 빅데이터 분석기법의 하나인 트렌드를 활용하여 5대 범죄 중 강도, 절도, 폭력에 대한 실제 범죄 발생빈도와 인터넷에서의 검색 횟수를 나타내는 트렌드 검색 횟수를 비교하였고 그 결과를 강도 범죄는 -0.06으로 음의 상관관계, 절도는 0.76으로 강한 상관관계, 폭력은 0.46으로 양의 상관관계를 나타내었다. 강도의 경우의 여러 의미로 사용되어 음의 상관계수를 나타낸 것으로 추정할 수 있다. 예를 들면 범죄의 강도도 있으나 물건의 단단함을 나타낼 때도 강도를 사용한다. 절도의 경우 강한 상관관계를 보이고 있으며 절도의 경우에도 양의 상관관계를 가지고 있으므로 실제 범죄의 빈도수를 추정하는데 도움을 줄 수 있다. 향후 연구 과제로 범죄 중 다른 단어와 중복되는 것을 구별하기 위해 구문분석을 통해 범죄의 의미만을 추출한 범죄 발생 빈도수 측정 모델을 연구할 예정이다.

REFERENCES

[1] M. Yun Kim & D. J. Seo. (2014) An Analysis of the Public Data for Making the Ambient Intelligent Service. *Journal of Digital Convergence,12(12).* 313-321.

[2] Y. I. Cha, S. K. Choi & K. S. Han. (2017). An Empirical Study on the Influence on Public Data Usage in Private Business Sectors. *Journal of Digital Convergence,15(6),* 9-17.

[3] Supreme (Public) Prosecutors' Office(2017) *Crime analysis. http://www.spo.go.kr/spo/info/stats/stats02.jsp*

[4] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski & L. Brilliant. (2009). Detecting influenza epidemics using search engine query data, *Nature*, 457, 1012-101.

[5] J. M. Park, Y. S. Chung, D. H. Kim, K. R. Park & H. R. Kim, (2011). A study of Using Correlation coefficient Analysis of Crime Association"*ksii,* 163-164.

[6] J. M. Park. (2011). *A crime prediction modeling using Markov chains,* Master's degree request paper. kongju National University. konngju.

[7] Wikipedia. *Big data* https://en.wikipedia.org/wiki/Big_data

[8] Y. K. Jung, M. G Suk & C. J. Kim. (2014). A study on the success factors of Big Data through an analysis of introduction effect of Big Data. *Journal of Digital Convergence,12(11),* 241-248.

[9] J. H. Kim & J. M. Lee.. (2017). Analysis of Waterpark Status and Recognition Using Big Data Analysis. *Journal of Digital Convergence,15(10),* 525-535

[10] C. S. Kim. (2012) Big data Utilization and related Technique and Technology Analysis. *The Korea Contents Association Review,* 10(1), 34-40

[11] M. M. Kang, S. R. Kim & S. M. Park. (2012). Analysis and Utilization of Big Data, *Communications of the Korean Institute of Information Scientists and Engineers, 30(6),* 25-32

[12] D. Mun, Namchul Do,(2016) Big Data, *CDE REVIEW18(2),* 26-29

[13] J. S. Kim. (2016). Subway Congestion Prediction and Recommendation System using Big Data Analysis. *Journal of Digital Convergence,14(11)* 289-295

[14] S. P. Jun & D. H. Park. (2013) Intelligent Brand Positioning Visualization System Based on Web Search Traffic Information : Focusing on Tablet PC, *Journal of Intelligence and Information Systems 19(3),* 93-11.

[15] Y. J. Choi & D. H. Park. (2017) Development of Yóukè Mining System with Yóukè's Travel Demand and Insight Based on Web Search Traffic Information, *Journal of Intelligence and Information Systems 23(3),* 155-17.

[16] S. J. Hwang & D. I. Kim. (2017). Analysis of Duty-Free Shopping Attributes and Shopping Satisfaction of Chinese Tourists : Focusing on duty free shops in Busan. *Journal of Intelligence and Information Systems, 15(12),* 137-145.

박 정 민(Jung-Min Park) [정회원]

- 2007년 2월 : 공주대학교 정보과학과 (공학사)
- 2011년 2월 : 공주대학교 멀티미디어공학과 (공학석사)
- 2014년 2월 : 공주대학교 컴퓨터공학과(박사수료)
- 2014년 3월 ~ 현재 : 공주대학교 시간강사
- 관심분야 : 사물인터넷, 빅데이타, 클라우드컴퓨팅,소프트웨어엔지니어링, 정보통신
- E-Mail : parkjm711@gmail.com

박 구 락(Koo Rack Park) [정회원]

- 1986년 2월 : 중앙대학교 전기공학과 (공학사)
- 1988년 2월 : 숭실대학교 전자계산학과 (공학석사)
- 2000년 2월 : 경기대학교 전자계산학과 (이학박사)
- 1992년 2월 ~ 현재 : 공주대학교 컴퓨터공학부 교수
- 관심분야 : 사물인터넷, 분산처리, 소프트웨어엔지니어링, 정보경영, 정보통신, 전자상거래
- E-Mail : ecgrpark@kongju.ac.kr

정 영 석(Young suk Chung) [정회원]

- 2009년 2월 : 공주대학교 멀티미디어공학과 (공학석사)
- 2013년 2월 : 공주대학교 컴퓨터공학과 (공학박사)
- 2009년 3월 ~ 현재 : 공주대학교 시간강사
- 관심분야 : 사물인터넷, 영상처리, 분산처리, 클라우드컴퓨팅, 시뮬레이션
- E-Mail : merope@kongju.ac.kr