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.2015.20.1.093

웹 검색 트래픽 정보를 이용한 범죄 예측 모델링에 관한 연구

박 정 민*, 정 영 석*, 박 구 락*

A study to Predictive modeling of crime using Web traffic information

Jung-Min Park *, Young-Suk Chung*, koo-Rack Park*

요 약

현대 사회는 다양한 범죄가 발생하고 있다. 범죄를 예방하기 위해서는 범죄를 예측 하는 것이 필요하고, 범죄 예측에 관한 다양한 연구가 진행 중에 있다. 범죄 관련 데이터는 검찰청에서 1년에 한번 통계처리를 하여 발표하고 있다. 그러나 통계처리 된 자료는 현재 시점을 기준으로 약 2년 전의 자료로 현재 발생하는 범죄에 대한 데이터로 적합하지 않다.

본 논문은 범죄를 예측하는 데이터로 네이버 트랜드를 적용했다. 네이버 트랜드의 웹 검색 트래픽을 이용하면, 현재 발생하는 범죄에 대한 관심도 데이터를 얻을 수 있다. 네이버 웹 검색 트래픽 데이터를 이용하여 범죄를 예측할 수 있는 모델링을 구성하였고, 예측 이론으로 마코프 체인을 적용하였다. 다양한 범죄 중 살인, 방화, 강간을 대상으로 예측 모델링에 적용하였고, 결과 값을 분석하였다. 그 결과 실제 발생한 범죄 발생 빈도수를 기준으로 20% 이내의 유사한 결과를 얻었다. 향후에는 계절의 특성을 고려한 범죄 예측 모델링에 대한 연구를 진행할 예정이다.

▸ Keywords : 범죄 예측, 검색 트래픽, 마코프 체인, 트랜드 예측

Abstract

In modern society, various crimes is occurred. It is necessary to predict the criminal in order to prevent crimes, various studies on the prediction of crime is in progress. Crime-related data, is announced to the statistical processing of once a year from the Public Prosecutor's Office. However, relative to the current point in time, data that has been statistical processing is a data of about two years ago. It does not fit to the data of the crime currently being generated.

In This paper, crime prediction data was apply with Naver trend data. By using the Web traffic Naver trend, it is possible to obtain the data of interest level for crime currently being generated. It was

•제1저자 : 박정민 •교신저자 : 박구락
•투고일 : 2014. 9. 27, 심사일 : 2014. 10. 15, 게재확정일 : 2014. 11. 10.
* 공주대학교 컴퓨터공학과(Dept. of Computer Science & Engineering, Kongju national University)

constructed a modeling that can predict the crime by using traffic data of the Naver web search. There have been applied to Markov chains prediction theory. Among various crimes, murder, arson, rape, predictive modeling was applied to target. And the result of predictive modeling value was analyzed. As a result, it got the same results within 20%, based on the value of crime that actually occurred. In the future, it plan to advance research for the predictive modeling of crime that takes into the characteristics of the season.

▸ Keywords : Crime prediction, search traffic, Markov chains, trend prediction

I. 서 론

범죄로 인한 피해는 피해자뿐만 아니라 피해자 주변인에게도 큰 피해를 주고 범죄 피해로 인한 사회적 손실도 크다. 범죄로 인한 피해자를 돕기 위해 정부에서는 범죄 피해 구조금 제도를 시행 중에 있다. 나라 지표에서 공개하고 있는 범죄 피해 구조금 지급현황 그래프는 다음의 그림 1과 같이 계속증가하고 있다. 범죄 피해 구조금의 지급 금액으로 2006년 1,063,000,000원에서 2013년 7,912,273,000원이 지급되었고, 이것은 2006년보다 약 7배 증가했다[1].

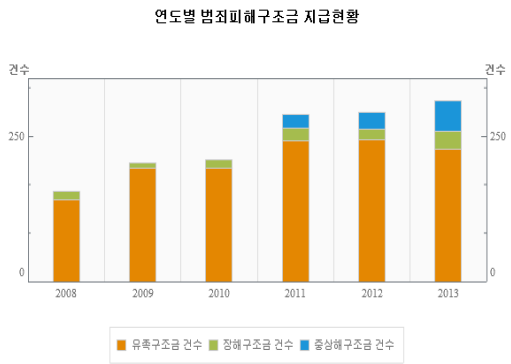


그림 1. 연도별 범죄 피해구조금 지급현황
Fig. 1. Crime Victim relief fund Status By Year

범죄로 발생하는 피해자와 사회적 간접비용을 절감하기 위해서는 범죄를 발생하기 전에 예측하는 것이 필요하다. 그래서 범죄를 예측하기 위한 다양한 연구는 진행되었다. 살인, 강도, 강간 등 주요 범죄의 단기적 발생 전망을 할 수 있는 시계열 예측 모형을 구현하여, 범죄 발생을 예측한 연구가 진행되었다[2]. 마코프 체인을 기반으로 범죄 발생 위험도를 확률지도로 생성하는 연구도 수행되었다.[3]. 그러나 위의 연구들은 경찰청, 검찰의 범죄 발생 데이터를 적용하여 예측을 수행하였는데, 범죄 발생 데이터는 통계처리 과정을 거치므로 매년 1회씩 발표되고 있다. 그런데 범죄 발생 통계로 발표되는 데이터는 현재를 기준으로 2년 전에 발생한 데이터가 발표된다. 예를 들어 2014년 7월을 기준으로 보면 2012년도에 발생한 범죄 발생 데이터를 얻을 수 있다. 결국 예측하고 싶은 달을 기준으로 2년 전의 데이터를 적용하므로, 현재 발생하는 범죄들의 예측에 사용되는 데이터로 부족하다. 그래서 본 논문에서는 빅 데이터의 하나인 웹 검색 트래픽을 적용하여 범죄를 예측하였다. 검색 트래픽은 구글에서 빅 데이터 분석 방법으로 시작한 것으로 구글을 검색하는 사용자들의 검색어를 통해 예측을 하는 것을 의미한다. 실제로 웹 검색어 중 독감의 검색 빈도수를 이용하여 독감의 유행을 예측하는 연구를 수행하였고, 현재는 구글 트랜드라는 이름으로 서비스를 하고 있다[4]. 국내 포털 사이트 중 네이버도 비슷한 서비스를 진행 중이다. 본 논문은 구글 트랜드와 네이버 트랜드 중 범죄에 대한 한글 검색 트래픽이 많은 네이버 트랜드를 선택하였다. 네이버 트랜드의 웹 검색 트래픽을 이용하여 범죄를 예측할 수 있는 모델링을 구성하였고, 실제 웹 검색 트래픽 데이터를 적용하여 범죄를 예측하였고, 실제 발생한 범죄와 비교하였다. 예측 모델링 이론으로는 마코프 체인을 적용하였다. 마코프 체인은 과거의 동적인 특성을 분석하여 미래를 예측하는 수학적 기법을 의미한다[5]. 본 논문의 구성은 다음과 같다. 2장에서 관련연구인 웹 검색 트랜드와 마코프 체인에 대해 논의한다. 3장에서 웹 검색 트래픽을 적용한 범죄 예측 모델링을 제안하고, 실제 데이터를 적용하여 범죄를 예측하고 실제 발생 빈도수와 비교한다. 4장은 각 범죄 별로 범죄 예측

빈도수와 실제 범죄 발생 빈도수의 비율을 분석하여 범죄 발생 예측 모델링의 정확도를 분석한다. 마지막으로 5장에서 결론 및 향후 연구 과제에 대해 논의한다.

II. 관련 연구

2.1 웹 검색 트래픽

2012년 세계 경제포럼에서 2012년 떠오르는 신기술 10중1위로 선정된 빅 데이터는 기존 데이터베이스의 데이터 수집,관리, 분석의 역할을 넘어서는 대량의 정형 비정형 데이터의 집합에서 가치를 찾아 분석하는 기술을 의미한다[6,7]. 빅 데이터를 기반으로 하는 웹 검색 트래픽을 분석하여, 미래에 발생할 사건을 예측하려는 다양한 연구가 진행되었다.

구글은 매일 수백만 명이 이용하는 인터넷 검색 동향을 모니터링 하여 독감의 발생 시기를 예측하는데 적용했다[8].

사회적 변화를 예측하고 설명하기 위해 사용되는 기술수명주기(hype cycle)의 이론적인 구조와 관계 및 실증에 대한 연구를 진행하기 위해, 웹의 검색 트래픽을 활용하여 국내와 해외의 기대 주기를 비교한 연구가 진행되었다[9].

소비자의 웹 검색 트래픽 정보를 통해 네트워크 모델링의 방법을 적용한 시스템을 적용하여, 소비자가 선호하는 제품을 가시화 할 수 있는 방법을 제안한 연구가 진행되었다[10].

2.2 마코프 체인

마코프 프로세스는 과거의 동적 특성을 분석하여 미래에 있을 변화를 예측하기 위한 수학적 기법이다. 마코프 프로세스는 상태간의 전이가 이전 n개의 상태에 의존하는 것을 의미한다. 마코프 체인은 전체 사건에서 가능한 상태들을 집합으로 구성한 상태집합, 초기화 확률 벡터인 초기확률과 각 상태간의 전이를 확률로 나타낸 전이 확률로 구성되어 있다[11]. 마코프 체인은 예측이 필요한 다양한 분야에 적용되고 있다. 웜, 바이러스 등 사이버 공격에 대한 그 피해 정도를 예측하는 모델에 적용되었다[12]. 네트워크의 잠재적 위협을 예측할 수 있는 확률적 통계모형의 모델링에 적용되었다[13].

III. 웹 검색 트래픽 정보를 적용한 범죄 예측 모델링

본 논문은 웹 검색 트래픽 정보를 적용한 범죄 예측 모델링을 제안했고, 실제 범죄 데이터를 적용하여 예측 하였고, 예측 값과 실제값을 비교하였다. 다양한 범죄 중 살인, 방화, 강간 범죄의 예측을 진행하였다.

3.1 범죄 예측 모델링 구성

본 논문은 웹 검색 트래픽 데이터를 적용하여 범죄를 예측하기 위한 모델링을 구현 하였다.

첫 번째, 웹 검색 트래픽 정보를 분류한다. 웹 검색 트래픽을 제공하는 네이버 트랜드는 일정 기간 동안 검색어의 검색 횟수를 그래프와 수치 데이터 형태로 제공한다. 제공된 데이터를 월별로 정리하여 범죄예측 모델링에 적용될 데이터의 형태로 변환 한다.

두 번째, 범죄 예측 확률 생성단계이다. 이전 단계에서 얻어진 웹 검색 트래픽 데이터를 바탕으로 범죄 예측 확률을 생성 한다. 예측 이론은 마코프 체인을 적용하였다.

마지막으로 범죄 예측 확률을 바탕으로 범죄가 발생할 빈도수를 예측 한다. 이전 단계의 예측 확률을 일정기간 월별 실제 발생한 범죄 빈도수의 최대값에 적용하여 범죄 발생 빈도수를 예측한다.

본 논문은 여러 범죄 중 살인, 방화, 강간을 대상으로 연구하였다. 현재 대 검찰청에서 공개하는 범죄 발생 통계 자료는 2013년 백서를 공개하고 있다. 2013년 백서에는 2012년도에 발생한 범죄데이터가 포함되어 있다. 그래서 범죄 발생 예측값과 실제값을 비교하기 위해 2009년~2011년도의 데이터를 추출하여 비교하였다[14].

3.1.1 웹 검색 트랙픽 정보 분류

본 논문은 범죄에 대한 웹 검색 트래픽 정보를 얻기 위해 네이버 트랜드를 이용하여 범죄에 대한 검색 추이 정보를 얻었다.네이버 트랜드는 특정 키워드가 통합 검색에서 가장 많이 사용된 지점(단위: 주)을 기준(100)으로 나머지 기간의 검색 횟수의 상대값을 환산하여 보여주는 것이다. 예를 들어 날씨라는 키워드의 가장 많이 입력된 횟수가 1000회라고 가정하면, 이것을 기준인 100으로 한다. 만약 날씨라는 키워드가 500회,350회가 검색된 지점은 상대값인 50, 35로 환산되어 나타낸다[15].

네이버 트래픽에서 범죄에 대한 검색 데이터를 처리하는 방법은 다음의 그림 2와 같다.

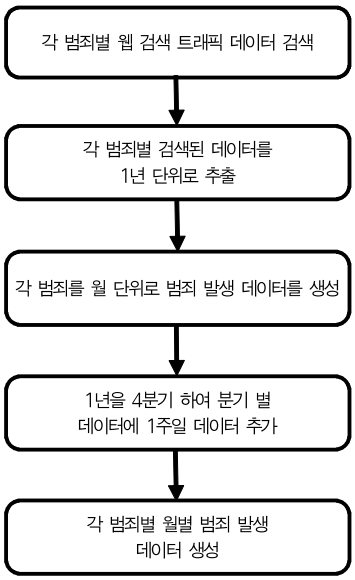


그림 2. 각 범죄별 웹 검색 트래픽 데이터 생성 단계
Fig. 2. Each crime traffic data generated by Web search step

첫 번째, 네이버 트랜드를 적용하여 각 범죄별 검색어로 검색 횟수를 구한다. 본 논문은 강력 사건 중 키워드로 살인, 방화, 강간을 선택하였고 기간은 2009년~2011년에 발생한 데이터를 추출하였다.

두 번째, 각 범죄별 검색된 데이터를 1년 단위로 나눈다.

세 번째, 각 범죄별 검색된 데이터를 월 단위로 나눈다. 그런데 각 범죄 별 네이버 트랜드 검색을 하면 2009년 1월 5일부터 시작하여 2012년1월 1-2일 까지 웹 검색 트래픽 데이터를 얻을 수 있다.

네 번째, 단계로 웹 트래픽 데이터를 월별로 구분하기 위해, 3개월(3월,6월,9월,12월)마다 5주로 하여 데이터를 분류한다.

마지막으로 각 범죄별 월별 범죄 발생 데이터를 생성한다.

표 1은 살인을 검색어로 사용하여 얻은 데이터를 월별로 분류한 것이다.

표 1. 웹 검색 트래픽 데이터 분류 (살인)
Table 1. Web traffic data categories (murder)

구분	2009년	2010년	2011년
1월	163	129	107
2월	194	141	134
3월	198	187	134
4월	176	124	108
5월	190	136	147
6월	208	188	168
7월	147	141	100
8월	157	124	104
9월	192	166	124
10월	154	142	101
11월	168	124	107
12월	181	140	113

표 2는 방화를 검색어로 사용하여 얻은 데이터를 월별로 분류한 것이다.

표 2. 웹 검색 트래픽 분류 (방화)
Table 2. Web traffic data categories (Arson)

구분	2009년	2010년	2011년
1월	68	57	55
2월	71	56	51
3월	72	71	76
4월	64	61	60
5월	60	57	56
6월	71	65	62
7월	48	54	44
8월	47	53	44
9월	63	68	55
10월	65	149	53
11월	80	82	62
12월	70	97	65

표 3은 강간을 검색어로 사용하여 얻은 데이터를 월별로 분류한 것이다.

표 3. 웹 검색 트래픽 분류 (강간)
Table 3. Web traffic data categories (Rape)

구분	2009년	2010년	2011년
1월	205	126	99
2월	178	125	106
3월	146	149	116
4월	124	105	103
5월	146	116	105
6월	175	161	117
7월	142	130	104
8월	147	136	94
9월	175	147	108
10월	122	117	87
11월	107	115	73
12월	130	115	81

3.1.2 웹 검색 트래픽 정보를 적용한 범죄 예측 확률 생성

범죄를 예측하기 위한 이론으로 마코프 체인을 적용했다. 마코프 체인을 이용하여 미래 발생할 범죄의 발생 확률을 구하였다. 마코프 체인으로 예측 확률 값을 구하기 위해서는 상태집합, 초기확률, 전이행렬이 필요하다[11,12].

- 상태집합: 네이버 트랜드 데이터를 이용한 각 범죄 검색어의 상태들의 집합이다. 본 논문에서는 웹 검색 트래픽 데이터를 바탕으로 임계값을 설정하고, 상태들을 정의하였다. 상태는 각 범죄마다 웹 검색 트래픽의 전체 평균값을 계산하여, 전체 평균 값 보다, 낮은 상태(S_1)와 평균 보다 높은 상태(S_2)인 두 가지 상태로 구성하였다. 각 범죄 별로 웹 검색 트래픽 자료가 다르므로 임계값과 상태 값이 다르다.

- 초기확률: 초기상태에 발생할 범죄 발생 확률이다. 본 논문에서는 웹 검색 트래픽 정보의 최근 상태를 적용하였고 식 (1)로 정의 한다.

$$P(S_1, S_2, \cdots, S_n) = P(\frac{\alpha}{F}, \frac{\beta}{F}, \cdots, \frac{\delta}{F}) \quad (1)$$

단, α, β, δ는 각 상태$(S_1, S_2, \cdots, S_n)$ 가 가지는 각 범죄의 웹 검색 트랙픽의 횟수이고, F는 α, β, δ의 합이다. 단, 확률이므로 총 합은 1이다.

- 전이 행렬: 상태 집합에서 정의한 상태간의 전이 상태를 확률로 계산 후, 행렬로 나타낸 것이다. 각 범죄발생의 웹 검색 트래픽 데이터를 월별로 분류한다. 그리고 월 별로 분류된 자료를 열거한 후 상태 집합과 매칭한다. 그리고 열거된 하나의 상태에서 다른 상태로의 전이 횟수를 구한 후 이를 전이행렬로 나타낸다. 각 열은 하나의 상태에서 다른 상태로 전이한 확률로서, 각 행의 합은 1이다. 식(2)의 P는 전이행렬이다.

$$P = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \cdots & \cdots & P_{ij} & \cdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{pmatrix} \quad (2)$$

식(3)은 마코프 체인 확률값으로 미래에 발생할 사건의 확률 값이다. 마코프 체인은 초기확률과 전이행렬로 구성되어 있다. 마코프 체인의 결과 값을 범죄 발생 예측 확률로 정의 하였다.

$$P(S_k) = \sum_{i=1}^{n} P(S_i) P_{ik} \quad (3)$$

여기서, $P(S_i)$: 초기 확률 P_{ik}:전이행렬

$P(S_k)$: 마코프 체인 확률 값

3.1.3 범죄 예측 확률을 적용한 범죄 발생빈도수 예측

범죄 발생 빈도수는 년 도별, 계절별, 월별로 발생하는 빈도수가 다르다. 본 논문은 월 별 범죄 발생 빈도수를 예측하기 위해 본 논문이 대상으로 하는 기간 중 각 월에 발생한 최대 범죄 발생 빈도수를 예측 확률에 적용하였다. 대상기간 중 각 월에 발생한 최대 범죄수를 발생빈도수 예측에 적용함으로써 그 다음 달에 발생한 범죄 뿐 만 아니라, 5개월 정도 지난 기간도 예측 가능했다. 본 논문은 이전 단계에서 구한 범죄 발생 확률에 실제 발생된 범죄 발생 빈도수 중 2009년~2011년의 각 월에 발생한 범죄발생 빈도수 중 최대값을 적용하여 범죄 발생 빈도수 예측에 적용하였다. 식(4)은 범죄 발생 예측 빈도수를 구하는 식이다.

$$\text{범죄 발생 예측 빈도수} = \sum_{i=1}^{n} P(S_i)\, MVIC(S_i) \quad (4)$$

n: 상태집합의 상태 수

$P(S_i)$: 범죄 발생 예측 확률

$MVIC(S_i)$: 매월 발생한 각 범죄 발생의 최대값

3.2 범죄 예측 모델링 적용

본 논문에서 제시한 범죄 예측 모델링을 적용하기 위해 다음의 조건하에 데이터를 적용하였다.

첫째, 다양한 범죄 중 살인, 방화, 강간 범죄에 적용하였다.
둘째, 예측 기간은 2012년 1월~5월의 범죄 발생 빈도수를 예측하였고, 실제 범죄 발생 빈도수와 비교하였다.

3.2.1 범죄예측(살인)

살인 범죄의 웹 검색 트래픽인 표1의 데이터를 이용하여, 범죄 발생 상태를 정의 하였다. 각 상태별 임계값의 범위는 S_1 은 0~148, S_2 는 149~296 로 하였다. 표 1에 있는 웹 검색 트래픽의 최근 3개월 (2011년 10월~12월)데이터를 이용하여 초기 확률을 계산하면, P (1, 0)으로 계산 할 수 있다. 웹 검색 트래픽 데이터와 임계값의 범위를 매핑하여 상태들을 나열한 후 전이횟수를 계산한 후, 식(2)에 적용하여 전이행렬을 구한다. 그리고 초기확률과 전이행렬의 결과 값을 식(3)에 적용하면, 살인에 대한 범죄 발생 예측 확률을 구할 수 있다.

$$(1 \quad 0)\begin{pmatrix} 0.75 & 0.25 \\ 0.4 & 0.6 \end{pmatrix} \tag{5}$$

$$= (0.75 \quad 0.25)$$

범죄발생확률은 식(5)에 따라 상태가 S_1일 때 0.75로 예측 되었다. 즉, 다음 달에 발생할 웹 검색 트래픽은 0~148 사이에 발생될 것으로 예측 할 수 있다. 범죄 발생 예측 빈도수를 구하기 위해서는 각 범죄 발생 빈도수의 당월 최대값을 구해야 한다. 본 논문에서는 2009년~2011년에 발생한 범죄 발생 데이터를 적용하였다. 각 월에 발생한 범죄 발생 빈도수의 최대값($MVIC$)을 구하였다. 표 4는 범죄 발생 데이터 및 범죄 발생 데이터의 최대값이다.

표 4. 범죄 발생 빈도수 및 당월 범죄 발생 빈도수의 최대값 (살인)
Table 4. Incidence of crime and The maximum number of the current month incidence of crime(Murder)

구분	범죄 발생 빈도수(살인)			범죄 발생 빈도수의 최대값($MVIC$)
	2009년	2010년	2011년	
1월	79	82	92	92
2월	76	84	90	90
3월	103	109	137	137
4월	111	112	101	112
5월	130	116	133	133

범죄 발생 빈도수를 예측하기 위해 범죄 발생 빈도수 및 당월 최대값($MVIC$)을 식(4)에 적용하여 얻은 범죄 발생 예측 빈도수와 실제 발생한 범죄 발생 데이터의 결과 값을 표 5로 정리하였다. 단, 범죄 발생 빈도수의 예측값의 소수점 이하는 반올림하였다.

표 5. 범죄발생 예측 빈도수 및 실제 빈도수(살인)
Table 5. The number of predicted and actual incidence of crime(Murder)

구분	범죄 발생 빈도수 예측값	범죄 발생 빈도수 실제값
1월	69	71
2월	68	79
3월	103	96
4월	84	92
5월	100	97

범죄 발생 빈도수 예측값과 실제값을 비교하면 그림3과 같다.

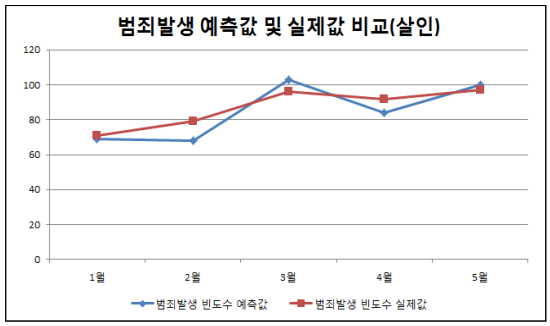


그림 3. 범죄발생 예측 빈도수 및 실제 범죄 발생빈도수 비교(살인)
Fig 3. Comparison of incidence of the predicted and actual value of crime (Murder)

3.2.2 범죄예측(방화)

방화 범죄의 웹 검색 트래픽인 표2의 데이터를 이용하여, 범죄 발생 상태를 정의 하였다. 각 상태의 임계값의 범위는 S_1 0~65, S_2는 66~150로 하였다. 초기확률은 식(1)를 이용하여, 웹 검색 트래픽의 최근 3개월 (2011년 10월~12월)데이터를 이용하여, P (1, 0)로 구하였다. 웹 검색 트래픽 데이터와 임계값의 범위를 매핑하여 상태들을 나열한 후 전이횟수를 계산한 후, 식(2)에 적용하여 전이행렬을 구한다. 그리고 초기확률과 전이행렬의 결과 값을 식(3)에 적용하면, 방화에 대한 범죄 발생 예측 확률을 구할 수 있다.

$$(1 \quad 0)\begin{pmatrix} 0.78 & 0.22 \\ 0.5 & 0.5 \end{pmatrix} \tag{6}$$

$$= (0.78 \quad 0.22)$$

범죄발생확률은 식(6)에 따라 상태가 S_1일 때 0.78로 예측 되었다. 즉, 다음 달에 발생할 웹 검색 트래픽은 0~65 사이에 발생될 것으로 예측 할 수 있다. 범죄 발생 예측 빈도수를 구하기 위해서는 각 범죄 발생 빈도수의 당월 최대값을 구해야 한다. 본 논문에서는 2009년~2011년에 발생한 범죄

표 6. 범죄 발생 빈도수 및 당월 범죄 발생 빈도수의 최대값 (방화)
Table 6. Incidence of crime and The maximum number of the current month incidence of crime(Arson)

구분	범죄 발생 빈도수(방화)			범죄 발생 빈도수의 최대값($MVIC$)
	2009년	2010년	2011년	
1월	143	96	128	143
2월	131	105	189	189
3월	169	149	210	210
4월	162	163	166	166
5월	203	162	175	203

발생 데이터를 적용하였다. 각 월에 발생한 범죄 발생 빈도수의 최대 값($MVIC$)을 구하였다. 표 6은 범죄 발생 데이터 및 범죄 발생 데이터의 최대값이다.

범죄 발생 빈도수를 예측하기 위해 범죄 발생 빈도수 및 당월 최대값($MVIC$)을 식(4)에 적용하여 얻은 범죄 발생 예측 빈도수와 실제 발생한 범죄 발생 데이터의 결과 값은 표 7로 정리하였다. 단, 범죄 발생 빈도수의 예측값의 소수점 이하는 반올림하였다.

표 7. 범죄발생 예측 빈도수 및 실제 빈도수(방화)
Table 7. The number of predicted and actual incidence of crime(Arson)

구분	범죄 발생 빈도수 예측값	범죄 발생 빈도수 실제값
1월	112	133
2월	147	130
3월	164	170
4월	129	145
5월	158	177

범죄 발생 빈도수 예측값과 실제값을 비교하면 그림4와 같다.

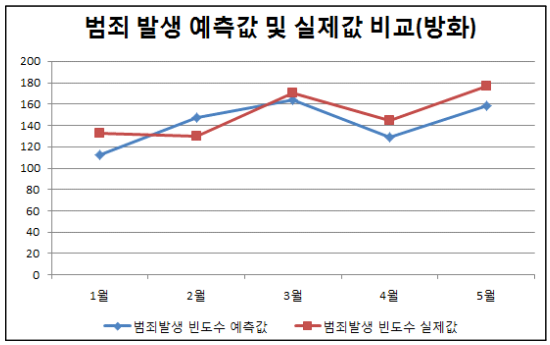


그림 4. 범죄 발생 예측 빈도수와 실제 발생 빈도수 비교 (방화)
Fig 4. Comparison of incidence of the predicted and actual value of crime (Arson)

3.2.3 범죄예측(강간)

강간 범죄의 웹 검색 트래픽인 표3의 데이터를 이용하여, 범죄 발생 상태를 정의 하였다. 각 상태의 임계값의 범위는 S_1은 0~126, S_2는 127~252로 하였다. 초기확률은 식(1)를 이용하여, 웹 검색 트래픽의 최근 3개월 (2011년 10월~12월)데이터를 이용하여, P (1, 0)로 구하였다. 웹 검색 트래픽 데이터와 임계값의 범위를 매핑하여 상태들을 나열한 후 전이횟수를 계산한 후, 식(2)에 적용하여 전이행렬을 구한다. 그리고 초기확률과 전이행렬의 결과 값을 식(3)에 적용하면, 강간에 대한 범죄 발생 예측 확률을 구할 수 있다.

$$(1 \quad 0)\begin{pmatrix} 0.81 & 0.19 \\ 0.36 & 0.64 \end{pmatrix} \tag{7}$$

$$= (0.81 \quad 0.19)$$

범죄발생확률은 식(7)에 따라 상태가 S_1일 때 0.81로 예측 되었다. 즉, 다음 달에 발생할 웹 검색 트래픽은 0~126 사이에 발생될 것으로 예측 할 수 있다. 범죄 발생 예측 빈도수를 구하기 위해서는 각 범죄 발생 빈도수의 당월 최대값을 구해야 한다. 본 논문에서는 2009년~2011년에 발생한 범죄 발생 데이터를 적용하였다. 각 월에 발생한 범죄 발생 빈도수의 최대값($MVIC$)을 구하였다. 표 8은 범죄 발생 데이터 및 범죄 발생 데이터의 최대값이다.

표 8. 범죄 발생 빈도수 및 당월 범죄 발생 빈도수의 최대값 (강간)
Table 8. Incidence of crime and The maximum number of the current month incidence of crime(Rape)

구분	범죄 발생 빈도수(방화)			범죄 발생 빈도수의 최대값($MVIC$)
	2009년	2010년	2011년	
1월	876	963	1,310	1,310
2월	938	974	1,548	1,548
3월	1,145	1,478	1,843	1,843
4월	1,141	1,562	1,528	1,562
5월	1,261	1,863	1,980	1,980

범죄 발생 빈도수를 예측하기 위해 범죄 발생 빈도수 및 당월 최대값($MVIC$)을 식(4)에 적용하여 얻은 범죄 발생 예측 빈도수와 실제 발생한 범죄 발생 데이터의 결과 값은 표 9로 정리하였다. 단, 범죄 발생 빈도수의 예측값의 소수점 이하는 반올림하였다.

표 9. 범죄발생 예측 빈도수 및 실제 빈도수(강간)
Table 9. The number of predicted and actual incidence of crime(Rape)

구분	범죄 발생 빈도수 예측값	범죄 발생 빈도수 실제값
1월	1,061	1,082
2월	1,254	1,081
3월	1,493	1,362
4월	1,265	1,472
5월	1,604	1,889

범죄 발생 빈도수 예측값과 실제값을 비교하면 그림5와 같다.

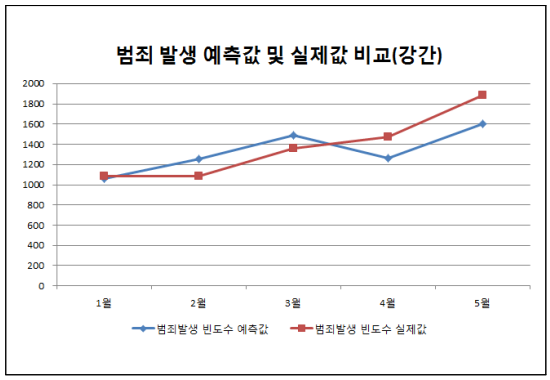


그림 5. 범죄 발생 예측 빈도수와 실제 발생 빈도수 비교 (강간)
Fig 5. Comparison of incidence of the predicted and actual value of crime (Rape)

Ⅳ.범죄 예측 결과 분석

본 논문은 살인, 방화, 강간을 대상으로 범죄 예측 모델링에 적용하였다. 범죄 예측 빈도수값과 실제 발생 빈도수를 비교하기 위해 실제 발생한 범죄 발생 빈도수와 예측빈도수의 비율을 계산하면, 다음의 표 10과 같다.

표 10. 범죄발생 예측 빈도수와 실제 범재 발생 빈도수의 비율
Table 10. Rate of the number of predicted and actual incidence of crime

구분	비율		
	살인	방화	강간
1월	0.97	0.84	0.98
2월	0.86	1.13	1.16
3월	1.07	0.96	1.10
4월	0.91	0.89	0.86
5월	1.03	0.89	0.85

각 범죄별 월 별 유사 비율을 보면 살인은 1-5월 범죄발생 예측 비율 중 2월을 제외하고는 10% 이내로 범죄 발생 빈도수를 예측하였다. 방화는 1-5월 범죄발생 예측 비율이 20% 이내로 범죄 발생 빈도수를 예측하였다. 강간은 1-5월 범죄 발생 예측 비율이 20% 이내로 범죄 발생 빈도수를 예측하였다.

본 논문에서 제시한 범죄 발생 예측 모델링을 적용한 결과 실제 발생한 범죄 발생 빈도수와 비교하여 살인의 경우는 10%, 방화와 강간은 20% 이내로 예측되었다.

Ⅴ. 결 론

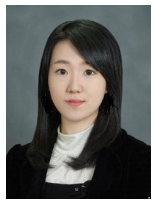
범죄는 피해자 뿐 만 아니라 피해자 주변에도 큰 영향을 미친다. 범죄로 인한 피해를 막기 위해서는 범죄가 발생하기 전에 예측을 하는 것이 필요하다. 범죄를 예측하기 위한 다양한 모델이 제시되었다. 그러나 범죄 예측에 사용된 데이터는 대검찰청 또는 경찰청에서 발표하는 자료는 현재를 기준으로 2년 전의 자료를 이용할 수 밖 에 없어서 현 시점에서 발생하는 범죄발생 예측 자료로는 부족함이 있다. 그래서 본 논문에서는 현 시점에서의 범죄를 예측하기 위해 네이버 트랜드의 웹 검색 트래픽을 데이터로 적용하였고, 예측을 위한 이론으로는 마코프 체인을 적용하였다. 그리고 예측에 적용될 수식에 사용될 데이터로, 예측하려는 각 범죄의 데이터로 사용된 기간 중 각 달에 발생한 범죄 발생 빈도수 중 가장 큰 값을 적용하여 발생할 범죄 빈도수의 정확성을 높였다. 본 논문은 살인, 방화, 강간 범죄를 대상으로 2012년 1월~5월에 발생될 범죄 발생 빈도수를 예측 하였다. 제시한 범죄예측 모델링을 적용한 결과 각 범죄 발생 빈도수를 기준으로 살인의 경우는 2월을 제외하고는 10%, 방화와 강간은 20% 이내에서 범죄 발생 빈도수를 예측 할 수 있었다. 결론적으로 본 논문에서 대상으로 적용한 범죄인 살인, 방화, 강간의 경우 20%이내에서 예측이 가능하였다. 향후에는 계절적인 요인을 적용한 범죄 예측 모델링에 대해 연구할 예정이다.

참고문헌

[1] Statistics Korea, *http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2809*
[2] Il-Yeob Joo "A Case Study on Crime Prediction using Time Series Models" Korean security science review, No.30 pp.139-169, 2012.
[3] Chan-Sook Noe, Dong-Hyun Kim, "A Crime Occurrence Risk Probability Map Generation Model based on the Markov Chain", The Journal of Korean Institute of Information Technology, Vol. 10, No.10, pp.89-98, 10, 2012.
[4] Google trends, *http://www.google.co.kr/trends/*
[5] Charles M. Grinstead, "Introduction to Probability: Second Revised Edition", American

Mathematical Society, pp405-406, 1997.
〔6〕 The top 10 emerging technologies for 2012 *http://forumblog.org/2012/02/the-2012-top-10-emerging-technologies/*
〔7〕 John Gantz, David Reinsel, " Extracting Value from Chaos ", IDC IVIEW June, 2011.
〔8〕 Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, Larry Brilliant "Detecting influenza epidemics using search engine query data", Nature, 457, pp.1012-1014, February 2009.
〔9〕 Seung-Pyo Jun, You Eil Kim, Hyoung Sun Yoo, "A Comparative Study of Consumer's Hype Cycles Using Web Search Traffic of Naver and Google", Journal of Korea Technology Innovation Society, pp.1109~1133, December 2013.
〔10〕 Seung-Pyo Jun Do-Hyung Park," Intelligent Brand Positioning Visualization System Based on Web Search Traffic Information : Focusing on Tablet PC", Journal of Intelligence and Information Systems , Vol.19, No.3, 93-111, 9. 2013.
〔11〕 Young-Gab Kim, Young-kyo Baek, Hoh Peter In, Doo-Kwon Baik, "A Probabilistic Model of Damage Propagation based on the Markov Process", Journal of KIISE, Vol33, No8, pp.524-535, 8. 2006.
〔12〕 Won-Hyung Park, Young-Jin Kim, Dong-Hwi Lee, Kui-Nam J Kim, "A Study on Prediction of Mass SQL Injection Worm Propagation Using The Markov Chain", Journal of the Korea Institute of Information Security and Cryptology, Vol 8, No4, pp.174-181, 12. 2008.
〔13〕 Kim Hyun-Woo, Shin Seong-Jun, Lee Seung-Min, Jeong Seok- Bong,"Network-based Intrusion Detection Scheme using Markov Chain Model",Journal of Decision Science, Vol.20, No.1, pp.75-88, 2012.
〔14〕 Crime statistics, SUPREME PROSECUTORS' OFFICE, http://www.spo.go.kr/spo/info/stats/stats02.jsp
〔15〕 naver trand help, http://help.naver.com/ops/step2/faq.nhn?fcatid=13953

저 자 소 개



박 정 민

2007: 공주대학교 정보과학과 공학사.
2011: 공주대학교 멀티미디어공학과 공학석사.
현 재: 공주대학교 컴퓨터공학과 박사수료
관심분야: 멀티미디어, 시뮬레이션, 클라우드 컴퓨팅, 모바일컴퓨팅,
Email : sweetmin71@kongju.ac.kr

정 영 석

2000: 배재대하교 물리학과 이학석사.
2009: 공주대학교 멀티미디어공학과 공학석사.
2013: 공주대학교 컴퓨터공학과 공학박사.
현 재: 대전보건대학교 겸임 교수
관심분야: 시뮬레이션, 클라우드 컴퓨팅, 정보보안, 모바일컴퓨팅
Email : merope@kongju.ac.kr

박 구 락

1986: 중앙대학교 전기공학과 공학사.
1988: 숭실대학교 전자계산학과 공학석사.
2000: 경기대학교 전자계산학과 이학박사.
현 재: 공주대학교 컴퓨터공학부 교수
관심분야: 정보경영, 정보통신, 전자상거래 클라우드 컴퓨팅, 정보보안, 모바일 컴퓨팅
Email : ecgrpark@kongju.ac.kr