군산 범죄예방 시스템에 관한 연구

한동엽*, 온병원**
군산대학교 통계컴퓨터과학과
e-mail: dongyup@kunsan.ac.kr*, bwon@kunsan.ac.kr**

A study on Gunsan crime mapping system

Dong-Yup Han*, Byung-Won On**
Department of Statistics and Computer Science, Kunsan National University

요 약

최근 다양한 공공 데이터가 속속 개방되고 있지만, 지역 내의 범죄 데이터는 통계 데이터 이외에는 공개되지 않고 있다. 이처럼 공공 데이터의 확보가 어려운 경우에는 해당 지역 내의 범죄 관련 모든 온라인 뉴스 기사를 주기적으로 수집하고 범죄 현황에 관한 정보를 자동으로 추출하여 맵(map)에 시각화 하여 보여주는 프레임워크의 개발이 필요하다. 본 논문에서는 프레임워크 개발에 필요한 주요 알고리즘들과 효과적인 시각화 방안을 제안한다. 또한 이미 공개된 군산시의 범죄 발생 통계 자료를 비교하여 제안 시스템의 효용성을 평가한다.

1. 서론

최근 정부3.0 이니셔티브를 통해 다양한 종류의 공공 데이터가 개방되고 있으며, 이러한 데이터를 활용하여 새로운 어플리케이션과 서비스가 개발됨으로써 국민 생활에 편익을 제공하고 관련 산업의 매출 증대를 꾀하고 있다. 현재 다양한 종류의 공공 데이터가 하루가 멀다 하고 속속 개방되고 있지만 데이터의 질이 떨어지고 쓸 만한 데이터는 많지 않은 실정이다. 또한 특정 데이터는 개인 프라이버시와 지적 재산권 문제로 인해 개방이 요원하다. 대표적인 데이터가 범죄 데이터라고 할 수 있다. 범죄는 실생활에 미치는 영향이 매우 크지만 우리 동네에 어떤 범죄가 발생했는지, 성범죄자들이 거주하는지, 또는 해당 지역의 CCTV 영상이 존재하지만 이러한 데이터에 접근하기는 원천적으로 불가능하다. 단지 5대 범죄에 대한 통계 데이터를 신청하여 얻을 수 있을 뿐이다. 이와 같이 데이터의 확보가 어렵고 1차 가공되어 공개된 데이터를 사용하여 서비스할 어플리케이션은 많지 않다. 결국 내가 살고 있는 지역에서 발생하는 범죄에 무방비로 노출될 수 밖에 없다.

이러한 데이터 확보의 어려움과 함께 더욱 큰 문제는 수도권과 지방간의 데이터 확보와 이를 활용할 수 있는 환경에서 정보 격차가 발생하고 있으며, 이는 시간이 갈수록 심화될 것으로 예상된다. 요즘 빅데이터가 국내외에서 큰 이슈가 되고 있지만, 공공 데이터와 같은 빅데이터를 확보하고, 이러한 데이터를 분석하고 시각화하여 공공 서비스에 활용하는 사례는 전무한 실정이다.

본 연구에서는 범죄 데이터와 같이 지역 내 파급 효과가 매우 크지만 데이터 확보가 쉽지 않고, 공공 데이터 활용이 미미한 지방 소도시에서 누구라도 공공 데이터를 쉽게 확보하고 이를 활용할 수 있는 프레임워크를 개발하는 것을 목표로 한다. 경찰서에서 작성하는 범죄 파일과 같은 데이터는 확보할 수 없지만 지역 내의 신문과 언론에서 기사화된 범죄 데이터를 온라인에서 정기적으로 수집하고, 범죄 데이터를 자동으로 분류하고 필요한 정보를 추출하는 알고리즘을 개발하고 구현하여, 군산 지역의 각 동읍면에서 발생하는 범죄 현황을 네이버 맵에 시각화 하여 보여주는 웹 기반의 프로토타입 시스템을 개발한다. 또한 최근 2년간 군산시에서 발생한 범죄 통계 데이터를 군산경찰서로부터 확보한 후에 개발된 시스템의 효용성을 평가하기 위해 통계적인 분석을 수행하여 제안 시스템의 우수성을 입증한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안 시스템을 구성하는 각 소프트웨어 컴포넌트와 알고리즘에 대해 자세히 설명한다. 3장은 제안방안과 관련 있는 주요 기존 연구에 대해 논의하고, 제안방안과의 차이점을 설명한다. 마지막으로 4장에서 결론을 맺고 향후 연구 방향에 대해 토의한다.

2. 제안방안

그림 1은 제안방안의 시스템 구성도를 나타낸다. 트위터와 네이버 검색엔진으로부터 군산 지역의 범죄 사건을 다루는 뉴스기사를 수집하기 위해서는 먼저 쿼리(query)를 생성하는 것이 필요하다. 쿼리는 군산의 동읍면과 살인/강도/절도/폭력/성범죄 등의 단어를 조합하여 생성된다. 검색결과는 검색 키워드를 포함하는 뉴스기사의 URL 주소의 집합이다. 이러한 검색결과에서 중복된 URL 주소

와 중복된 내용을 포함하는 뉴스기사를 필터링한 후에 각 기사의 메타 정보들이 데이터베이스에 저장된다. 메타 정보는 각 뉴스기사의 ID, 날짜, 언론사명, 뉴스제목, 동읍면, 살인/절도/강도/폭력/성범죄의 유무, 뉴스기사 파일의 저장 위치 등의 정보를 의미한다. 이러한 데이터의 수집과 저장은 배치(batch) 방식에 의해 주기적으로 업데이트 된다. 그리고 데이터베이스에 저장되어 있는 데이터는 클라이언트-서버 프로그램 방식으로 웹 브라우저 또는 스마트폰에서 시각화 되어 서비스 된다.

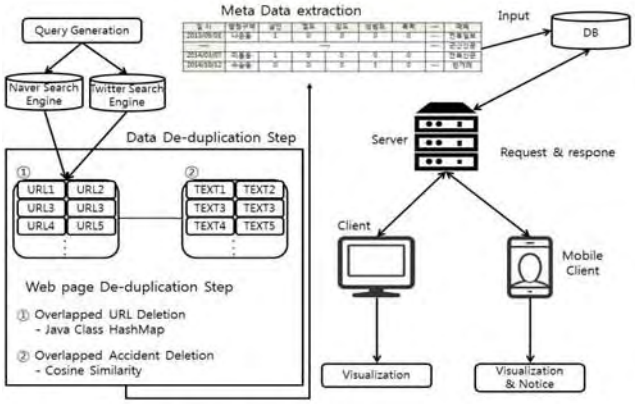


그림 1. 제안방안의 시스템 구성도

2.1 데이터 수집

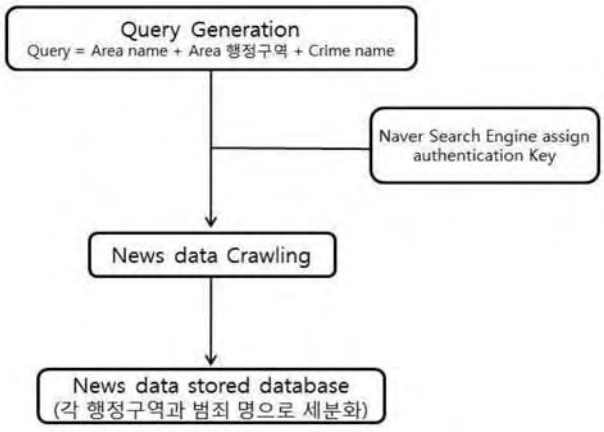


그림 2. 뉴스기사 수집과정

그림 2에서 보는 것처럼 뉴스기사의 수집과정의 첫 번째 작업은 쿼리를 생성하는 것이다. 쿼리 생성은 '군산', 동읍면, 살인/강도/절도/폭력/성범죄 등의 단어들을 조합하여 만든다. 예를 들면, 트위터와 네이버 검색엔진에 질의하기 위하여 <군산+미룡동+성폭행> 또는 <군산+나운동+강도>과 같은 쿼리를 생성할 수 있다. 군산시는 1읍 10면 16동으로 구성되어 있어 27개의 행정구역으로 나눌 수 있다. 따라서 생성되는 쿼리의 총 개수는 $1\times27\times5=135$ 개의 쿼리들이 생성되고 관련 뉴스기사들을 수집하게 된다. 다음, 트위터와 네이버 검색엔진의 인증키를 받은 후에 워크벤치(workbench)의 자바 라이브러리 파일인 제이숩(Jsoup)에 있는 매치(matches) 정규식을 이용하여 쿼리를 포함하는 뉴스기사들의 URL 주소들을 수집한다.

2.2 중복 데이터 제거

앞절에서 이미 전술한 바와 같이 다양한 쿼리를 생성하여 검색을 수행하기 때문에 어떤 뉴스기사 웹페이지는 여러 쿼리들에 의해서 검색되게 된다. 이와 같이 중복된 URL 주소들을 제거하는 것이 필요하다. 이것은 그림 3의 웹 주소 중복성에 해당한다. 더욱이 비록 두 개의 뉴스기사들의 URL 주소가 다르다고 할지라도 그 두 뉴스기사의 내용이 상당히 유사할 수 있다. 이러한 일은 흔히 발생한다. 사건 사고를 담당하는 뉴스기사는 사건이 발생한 관할 경찰서의 브리핑을 통해 기사를 작성하기 때문에 여러 언론사의 뉴스기사들의 내용은 상당히 유사하게 된다. 이와 같이 내용이 유사한 뉴스기사들을 찾아 필터링하는 것이 필요하다. 이것은 그림 3의 사건 중복성에 해당한다.

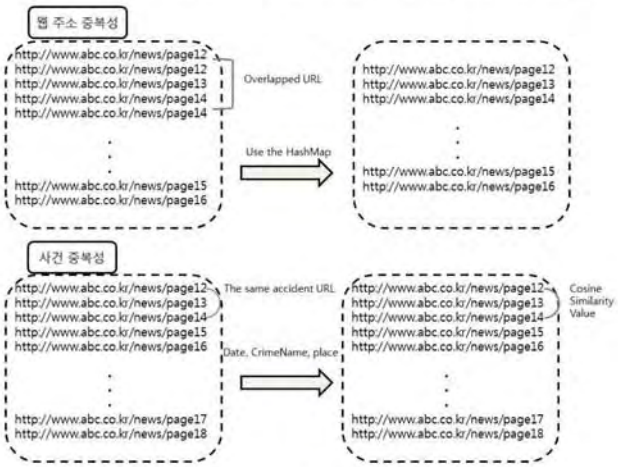


그림 3. 중복 데이터 필터링

제안 시스템에서는 본문 내용이 유사한 뉴스기사를 찾기 위해 코사인 유사도(cosine similarity)를 이용한다. 예를 들면, 두 개의 뉴스기사 X와 Y는 다음과 같이 표현할 수 있다. X=<a, b, c, d>와 Y=<a, b, c, m>. X는 a, b, c, d 단어들과 Y는 a, b, c, m 단어들을 포함한다. X와 Y는 대부분의 단어들을 서로 포함하고 있기 때문에, 두 뉴스기사의 내용이 유사함을 유추할 수 있다. 이러한 유사도에 대한 정확한 수치를 계산하기 위해 다음과 같은 공식을 사용한다.

$$\text{사인 유사도}(a,a_j)=\frac{\sum_{k=1}^{n} w_{ik}\cdot w_{jk}}{\sum_{k=1}^{n} w_{ik}^2\cdot\sum_{k=1}^{n} w_{jk}^2} \quad (1)$$

a_i와 a_j는 뉴스기사를 의미하며, w_{ik}와 w_{jk}는 a_i와 a_j에 있는 k번째 단어를 나타낸다. 만일 두 문서의 코사인 유사도가 1에 가까우면 그 두 문서의 내용의 거의 동일하고, 반대로 0에 가까우면 두 문서의 내용은 완전히 다르다고 판단할 수 있다. 본 연구의 실험에 따르면, 두 문서의 유사도가 0.5보다 높으면 그 두 문서는 사건 중복성에 해당한다고

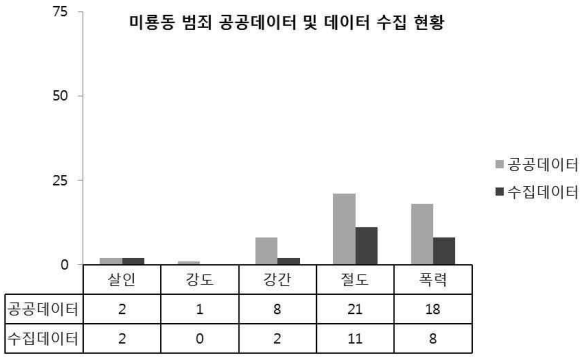
판단하여 필터링하였다. 웹 주소 중복성에 대해서는 해시셋(HashSet)이라는 자바 클래스를 사용하여 동일한 URL 주소를 제거하였다.

이러한 방식으로 2003년부터 2014년 동안 총 21,552건의 뉴스기사들을 수집하였다. 그중에서 살인 477건(2.2%), 절도 3,366건(15.6%), 성범죄 5,991건(27.8%), 강도 6,634건(30.7%), 폭력 5,114건(23.7%) 등이 발생하였다.

2.3 수집된 데이터의 품질 평가

이 절에서는 제안방안을 통해 수집된 데이터와 실제 경찰청에서 집계한 범죄 통계 자료와의 상관관계(correlation)가 있는지를 검사하여 본 연구에서 제안한 방안이 효과적인지를 알아본다. 즉, 경찰청의 공공 데이터와 제안방안을 통해 수집된 데이터의 양을 비교함으로써 제안 시스템의 데이터 수집 방안이 효과적인지를 평가한다. 그리고 SPSS를 이용하여 두 데이터의 독립표본 t-검정을 수행하여 정제된 데이터의 차이를 측정하도록 한다.

그림 4는 군산시 행정구역 중 하나인 미룡동을 예시로 공공 데이터와 수집된 데이터의 개수를 측정하여 실제 발생한 범죄의 수(공공 데이터의 개수)와 수집된 데이터의 수(제안방안을 통해 얻어진 데이터의 개수)를 비교한 것이다.



	살인	강도	강간	절도	폭력
공공데이터	2	1	8	21	18
수집데이터	2	0	2	11	8

그림 4. 공공 데이터와 수집된 데이터의 현황

살인/강도/절도/폭력/성범죄 등 5대 범죄를 대상으로 현황을 조사한 결과, 살인 사건의 경우에는 공공 데이터와 수집 데이터의 수가 일치하였다. 살인 사건은 사회적으로 비중이 큰 사건이기 때문에 뉴스기사화가 쉽게 됨을 간접적으로 알 수 있다. 반면에 살인 사건을 제외한 나머지 경우에는 공공 데이터의 개수에 미치지 못하거나 아예 데이터를 확보하지 못하는 경우도 있었다. 그 이유로는 해당 사건이 중차대하지 않아 이슈화되기 어렵거나, 피해자 정보 유출에 따른 2차 피해를 방지하기 위해 언론에 노출되지 않았기 때문이다. 이렇듯 제안방안을 통해 수집된 범죄사건의 개수가 경찰서에서 작성한 통계치에 못 미치더라도 두 데이터간의 개수에 대한 상관관계가 존재한다면 수집된 데이터는 공공 데이터를 대신하여 사용될 수 있음을 입증하게 된다. 그렇기에 두 집단 간의 독립표본 t-검정을 통하여 공공 데이터와 수집 데이터 간의 차이를 검정한다.

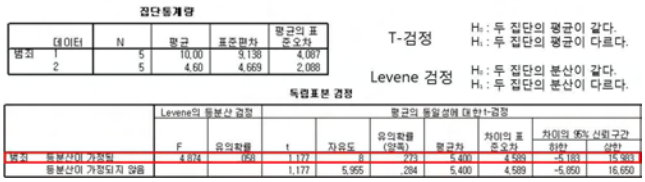


그림 5. 독립표본 t-검정 결과

독립표본 t-검정 결과는 집단통계량, 레벤(Levene)의 등분산 검정, t-검정 순으로 해석이 진행되며, 데이터 1은 공공 데이터로, 데이터 2는 수집 데이터로 정의한다. 집단통계량에선 수집 데이터가 공공 데이터에 비해 통계적으로 유의하게 낮다고 말할 수 있지만 레벤의 등분산 검정과 t-검정의 가설이 진행돼야 결론을 도출할 수 있다. 레벤의 등분산 검정의 유의확률은 0.058이므로 H_0를 채택, 등분산이 가정되므로 그림 5에서 적색 구간을 활용한다. 마지막으로 t-검정에서의 유의확률은 0.273, 등분산 검정처럼 H_0를 채택하게 되고, 정리하면, 공공 데이터와 수집 데이터의 분산이 같다는 가정 하에 두 데이터 집단의 평균이 같으므로 두 집단의 차이는 없음을 결론지을 수 있다. 이러한 결과로부터 수집 데이터가 공공 데이터와의 차이가 없다는 통계적 검정 결과를 바탕으로 수집 데이터가 공공 데이터를 대신할 수 있다고 판단할 수 있다.

2.4 메타 데이터 추출

쿼리 검색결과에서 중복된 URL 주소와 중복된 내용을 포함하는 뉴스기사를 필터링한 후에 각 기사의 메타 정보들이 데이터베이스에 저장된다. 메타 정보는 뉴스기사를 설명하는 속성(attribute) 정보의 집합으로 뉴스기사의 ID, 날짜, 언론사명, 뉴스제목, 동읍면, 살인/절도/강도/폭력/성범죄의 유무, 뉴스기사 파일의 저장 위치 등의 정보들이다. 뉴스기사의 ID는 시스템으로부터 자동 생성되며, 데이터베이스 테이블의 프라이머리키(primary key)로 사용된다. 뉴스기사 원문으로부터 날짜, 언론사명, 뉴스제목 등을 쉽게 추출할 수 있다. 또한 27개 군산시 행정구역 중에서 어떤 행정구역에 대한 단어(예: 미룡동)가 뉴스기사 본문에 있다면, 그 뉴스기사는 그 행정구역의 범죄 사건을 다룬다는 것을 알 수 있다. 또한 살인/절도/강도/폭력/성범죄 등의 각 범죄에 대해 연관어를 미리 정의한다. 예를 들면, 성범죄=<성폭행, 성추행, 성매매>과 같이 성범죄는 성폭행, 성추행, 성매매 등의 연관어를 가진다. 만일 '미룡동'과 '성폭행'이라는 단어가 어떤 뉴스기사 본문에 있다면, 그 뉴스기사는 미룡동에서 발생한 어떤 성범죄 사건에 대한 기사일 것으로 짐작할 수 있다.

2.5 시각화 방안

마지막으로 메타 데이터를 맵에 시각화 하여 군산 지역의 범죄 현황을 사용자에게 서비스하는 것이 필요하다. 이 절에서는 편리성과 신뢰성을 고려하여 사용자에게 가

장 효과적인 시각화 방안을 제안한다. 먼저 개인용 컴퓨터에서 사용하는 웹 브라우저와 스마트폰의 앱을 기반으로 하는 시각화 방안으로 나눈다. 웹 브라우저를 통한 시각화 방안은 그림 6에서 보여준다. 검색 기능을 통해, 사용자는 검색하기를 희망하는 군산시의 특정 동읍면을 입력하고, 검색을 수행한다. 검색결과로는 우축에 네이버 맵이 나타나고 해당 동읍면이 활성화된다. 그와 동시에 해당 동읍면의 5대 범죄에 대한 통계와 시간에 따른 범죄의 증감추세를 보여준다. 그리고 웹 브라우저의 좌측에는 관련 뉴스기사들이 최신 순서대로 뉴스기사의 제목이 나타난다. 해당 뉴스기사를 클릭하면 그 뉴스기사의 원문을 볼 수 있다.

그림 6. 웹 브라우저를 위한 범죄통계 시각화 프로토타입

스마트폰에서는 모바일 앱을 통해 데이터를 시각화 한다. 사용자가 모바일 앱을 활성화 시켜 놓으면, 현재 사용자가 위치해있는 장소의 위도와 경도를 계산하여 그 지점이 어떤 동읍면에 속하는지를 알아낸다. 그리고 그 동읍면의 범죄에 관한 메타 데이터를 가져와서 시각화한 다음에 모바일 앱 화면에 범죄 현황에 대한 통계 자료가 팝업으로 뜨게 된다. 또한 가장 빈번히 발생한 범죄 사건을 공지하여 사용자에게 범죄 예방에 대한 환기를 준다. 예를 들면, 미룡동의 최근 2년간 최고 범죄는 폭력이었음으로, 만일 사용자가 미룡동에 있다면 폭력이 빈번히 발생한 지역임으로 폭력 사건에 휘말리지 않도록 주의를 주는 메시지가 전송되거나 알람 또는 음성 메시지로 서비스 하게 된다.

3. 관련연구

해외에서는 미국 LA 지역을 대상으로 빅데이터 기반의 범죄 예측시스템(predictive policing system)을 운영하여 실시간으로 범죄가 일어나기 쉬운 지역을 예측하고 경찰력을 유동적으로 배치함으로써 치안 향상과 효율적인 인력 운용을 보여주었다[3]. 또한 크라임 리포트(crime report)라는 웹 사이트에서는 범죄지도 서비스를 제공하여 성범죄, 차량정보, 강도 등의 각종 사건의 기록, 범죄자의 사진, 주소지, 신상명세, 범죄 전력 등을 제공함으로써 많은 범죄 예방 효과를 거두었다[1,4].

국내에서는 여성가족부와 법무부가 운영하고 있는 성범죄자 알림e 서비스를 통해 현대사회에서 급격하게 발생하고 있는 성범죄 범죄자들의 신상정보를 공개하고 성범죄 예방에 기여하고 있는 시스템을 개발했으며, 부산시의 경우에는 공공 데이터와 빅데이터 기반으로 하는 범죄예방 시스템을 구축하여 경찰 범죄지도와 아동안전지도를 제공하여 관할하고 있는 행정구역이 이를 적극 활용할 예정이다[2].

기존연구는 이미 확보된 공공 데이터를 활용하며 빅데이터 분석 기법을 사용해서 과거의 데이터로부터 정보를 추출하여 범죄 예방 서비스를 제공하지만, 본 연구의 제안 방안은 실제 확보하기 쉽지 않은 공공 데이터를 웹에서 주기적으로 수집하고 정제한 후에 범죄 지도를 구성하여 서비스하는 자동화 시스템을 개발하는 것이 목적이며 누구나 쉽게 활용할 수 있도록 프레임워크를 개발하는 점이 기존연구와의 차이점이다.

4. 결론

범죄 데이터와 같이 공공 데이터 확보가 어려운 환경에서 웹에 있는 데이터를 수집하고 정제하여 실제 공공 데이터의 양과 질 측면에서 유사한 데이터를 확보하는데 필요한 주요 기술을 제안하였다. 그리고 얻어진 데이터로부터 군산 지역의 범죄 통계를 효과적으로 알리고 서비스하기 위하여 인터넷과 모바일 환경에 맞는 시각화 방안을 제안하였다. 더욱이 제안방안을 프로토타입으로 개발하고 프레임워크 형태로 개방하기 위해 OpenAPI를 지원함으로써 군산 지역의 누구라도 제안 시스템 위에 새로운 서비스를 확장할 수 있는 토대를 마련한다. 끝으로 본 논문에서 제안한 시스템을 사용한다면, 범죄예방과 범죄율 감소에 큰 기여를 할 것으로 예상된다.

향후 연구계획으로는 본 논문에서 제안한 세부 시스템들을 통합하고 자동화하여 프로토타입 시스템을 개발하는 것이다. 또한 프레임워크 형태로 개방하기 위해 OpenAPI를 개발하는 것이 필요하다. 그리고 좀 더 우수한 머신러닝 기술을 적용함으로써 데이터 분석을 통해 다양하고 지능적인 서비스를 할 수 있을 것이다.

참고문헌

[1] 박명규, “GIS의 공간분석을 활용한 범죄예측지도의 구현”, 경희대학교 석사학위논문, 2013

[2] 부산광역시, “빅데이터를 활용한 부산 범죄예방 시스템 구축에 관한 연구”, 부산시, 2013, pp. 12~18

[3] 이상철, 최진영, “해외 범죄지도 서비스 사례 분석을 통한 국내 생활안전지도에 관한 연구”, 한국정보과학회 추계학술대회, 2014

[4] 임승옥, “머신러닝을 활용한 범죄예측 시스템 구축 사례”, 2015 빅데이터 분석에 기반을 둔 머신러닝과 인공지능 워크숍, 2015, pp. 4~14