범죄 데이터의 전산처리를 위한 정규화 메트릭 설정 방안

임선영*, 박은영**, 박영호***
*숙명여자대학교 멀티미디어과학과
**협성대학교 시각디자인학과
***숙명여자대학교 멀티미디어과학과, 교신저자
e-mail : {sunnyihm, yhpark}@sm.ac.kr, parkey@uhs.ac.kr

A Normalization Matrics for Computational Processing of Crime Dataset

Sun-Young Ihm*, Eun-Young Park, Young-Ho Park*
*Dept. of Multimedia Science, Sookmyung Women's University
**Dept. of Visual Design, Hyupsung University

요 약

최근 데이터의 양이 급격하게 증가하면서 빅데이터의 시대가 도래했다. 빅데이터는 형식이 없는 비정형 데이터이므로 기존의 정형 데이터 처리 방법으로는 분석 및 데이터 처리가 불가능해졌다. 또한, 범죄예방에 대한 관심이 증가하면서, 범죄 데이터 분석의 수요가 증가하고 있다. 본 연구에서는 비정형 범죄 데이터를 분석, 예측 등의 전산처리를 하기 위한 정규화 메트릭을 설정하는 방안을 제안하고자 한다.

1. 서론

최근 데이터의 양이 급격하게 증가하면서 빅데이터 시대가 도래했다. 빅데이터는 대용량의 데이터로 정형화된 데이터 하지만, 텍스트, 동영상과 같은 형태의 비정형 데이터들은 기존의 데이터 처리 방법으로는 분석 및 처리가 불가능하다. 따라서 기존의 방법들로 처리하기 위해서는 비정형 데이터를 정형화하는 것이 중요하다.

또한, 최근 범죄예방에 대한 관심이 높아지면서, 범죄 데이터를 분석하는 응용[1-4]이 많아지고 있다. 예를 들어, 범죄에 대한 판결문을 분석하고자 할 때, 범죄의 중함 정도를 알 수 없으며, 종류에 따라 분류하거나 레벨링을 하는 것이 불가능하다. 따라서, 법에 명기된 범죄 형량 기준에 따른 정량적 수치화가 선행되어야 한다. 또한, 범죄의 종류에 따른 다양한 처리가 가능하기 때문에 범죄 종류와 그 범행의 위험 정도에 따른 구분 및 레벨링을 할 수 있게 된다.

본 연구에서는 비정형 데이터를 정형화된 데이터로 정규화하는 정규화 메트릭 기법을 연구하고자 한다. 데이터를 정규화하기 위해서는 먼저 데이터의 특성을 분석하여 정규화 기준을 생성해야 한다. 본 연구에서는 텍스트로 구성된 범죄 데이터를 정량적 수치로 정규화하고자 한다. 본 논문의 구성은 다음과 같다. 제 2 장에서는 범죄 데이터의 특성을 분석하고, 제 3 장에서는 정규화 메트릭을 설정하는 방법을 제안한다. 제 4 장에서는 결론과 향후 연구를 소개한다.

2. 범죄 데이터의 특성 분석

비정형 데이터를 정규화하기 위해서는 먼저 데이터의 특성을 분석해야 한다. 본 연구에서는 정규화할 대상을 범죄 데이터로 설정하고, 실제 데이터를 수집하였다. 범죄 데이터에는 범죄 명, 위치 정보, 가해자 정보, 피해자 정보, 판결 정보 등 다양한 속성이 존재한다. 이 중에서 본 연구에서는 범죄 종류, 피해자 수, 피해자 나이, 판결 결과, 가해자 나이, 연관된 범죄의 수와 같이 총 6 개의 속성에 대하여 정규화를 수행하고자 한다. 표 1 은 본 연구에서 정규화할 속성을 설명하고 있다. 먼저 crime_type 속성은 범죄의 종류를 의미하며, 범죄의 종류로는 강간, 폭행, 화재, 절도 등이 있다. victim_no 속성은 피해자의 수를 뜻하며, victim_age 속성은 피해자의 나이를 뜻한다. judgement 속성은 범죄의 판결 결과를 의미하며, 예를 들어 징역 1 년, 집행유예 3 년 등의 값을 가진다. attacker_age 속성은 가해자의 나이를 뜻하고, related_crime 속성은 연관된 범죄의 수를 의미한다.

<표 1> 범죄 데이터 속성

속성 명	설명
crime_type	범죄의 종류
victim_no	피해자의 수
victim_age	피해자의 나이
judgement	판결 결과 (ex.징역 1 년 …)
attacker_age	가해자의 나이
related_crime	연관된 범죄의 수

3. 정규화 메트릭 설정

본 장에서는 정규화 메트릭을 설정하는 방법을 설명한다. 텍스트 형태의 비정형 데이터를 정량적 수치의 정형화된 형태로 정규화하기 위해서는 먼저 정규화 기준이 생성되어야 한다. 정규화 기준은 모든 데이터에 대하여 변환이 가능한 명확한 기준이어야 한다. 먼저 범죄의 종류를 정규화하는 기준으로는 대한민국 형법[5]에 정의된 범죄의 정도와 기준 형량에 따라 기준을 정한다. 표 2 는 범죄의 종류에 따른 정규화 기준을 나타내고 있다.

<표 2> 범죄의 종류의 정규화 기준

속성 값	범죄의 종류
5	강간, 살인, 폭행 강간, 미성년자 강간
4	미성년자 추행, 강도, 특수 절도
3	강제 추행, 폭행, 절도, 주거침입
2	성적 수치심 유발, 공공장소 추행
1	폭행 예비행위, 지속적 괴롭힘

다음으로는 피해자 수와 연관된 범죄의 수 속성의 정규화 기준을 생성한다. 피해자 수와 연관된 범죄의 수는 많을수록 위험한 범죄이므로 속성 값을 크게 설정한다. 그리고 판결 결과도 마찬가지로 형량이 클수록 위험한 범죄이므로 속성 값을 크게 설정한다. 마지막으로 피해자와 가해자 나이는 통계청[6]에서 제공하는 범죄 통계에 기반하여 정규화 기준을 생성한다. 범죄 피해자가 많은 연령대와 범죄 가해자가 많은 연령대를 범죄 발생 가능성이 크므로 속성 값을 크게 정한다. 표 3 은 피해자와 가해자의 나이에 따른 정규화 기준을 나타내고 있다.

<표 3> 피해자와 가해자 나이의 정규화 기준

속성 값	피해자 나이	가해자 나이
5	15 ~ 29	15 ~ 24
4	9 ~ 14	25 ~ 29
3	30 ~ 49	30 ~ 49
2	0 ~ 8	0 ~ 15
1	50 ~100	50 ~ 100

4. 결론 및 향후 연구

본 논문에서는 비정형 데이터를 정량적 수치의 정형화된 데이터로 정규화하는 기법을 연구하였다. 정규화 대상 데이터로는 범죄 데이터를 사용하였으며, 범죄의 종류, 가해자의 나이, 피해자의 나이, 판결 결과, 피해자의 수, 연관된 범죄의 수로 총 6 개의 속성에 대하여 정규화를 진행하였다. 범죄 데이터를 정규화하기 위하여 먼저 정규화 기준을 생성하였으며, 표 4 는 3 장의 정규화 기준에 따라 범죄 데이터를 정규화 한 결과이다.

텍스트 형태의 비정형 범죄 데이터를 정량적 수치로 변환함으로써 데이터를 분석하거나 처리하는 것이 용이해졌다. 향후 연구로는 정규화된 데이터를 분석하는 연구를 하고자 한다.

사사문구

이 논문은 2012 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2012003797)

참고문헌

[1] Aziz Nasridinov, Young-Ho Park, "Combining Unsupervised and Supervised Machine Learning to Analyze Crime Data", International Journal of Applied Engineering Research, Vol.9, No.23, pp.18663-18669, 2014.12.

[2] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang Yi Qin, Michael Chau, “Crime data mining: a general framework and some examples”, Computer, Vol.37, No.4, pp.50-56, 2004.04.

[3] Sun-Young Ihm, Wu-In Jang, So-Hyun Park, Aziz Nasridinov and Young-Ho Park, "A Study on a Real-Time Crime Prevention for Residential Environments of South Korea", In Proceeding of the 3rd FTRA International Conference on Ubiquitous Computing Application and Wireless Sensor Network (UCAWSN 2014), Jeju Island, South Korea, July 7-10, 2014.

[4] Lawrence D. Chu, Jess F. Kraus, “Predicting Fatal Assault Among the Elderly Using the National Incident-Based Reporting System Crime Data”, Homicide Studies, Vol.8, No.2, pp.74-95, 2004.05.

[5] 국가법령정보센터, http://www.law.go.kr/main.html.

[6] 통계청, www.kostat.go.kr.

<표 4> 범죄 데이터 정규화 결과

id	crime_type	victim_no	victim_age	judgement	attacker_age	related_danger
1	3	1	1	4	2	0
2	5	1	3	3	5	3
3	3	1	5	3	5	3
4	3	1	5	3	5	3
5	5	1	5	3	5	3
6	3	1	5	4	2	2
7	3	1	5	4	2	2
8	2	1	5	4	2	2
9	4	1	3	2	5	0
10	5	1	4	3	3	1