경찰청 112 신고 데이터의 미분류 유형에 대한 LDA 토픽모델링 기반 잠재적 신고유형 분석

김용진[1], 유덕산[2], 방준성[3,*], 장광호[4]

과학기술연합대학원대학교[1], 충남대학교[2], 한국전자통신연구원[3], 경찰대학 치안정책연구소[4]

kyjwj92@ust.ac.kr, 201502692@o.cnu.ac.kr, *hjbang21pp@etri.re.kr, pathfinder@police.go.kr

Latent Crime Reporting Types Analysis Based on LDA Topic Modeling for Unclassified Types of 112 Crime Reporting Data in Police

Yongjin Kim[1], Deoksan Yoo[2], Junseong Bang[3,*], Gwangho Jang[4]

UST[1], ChungNam National Univ.[2], *ETRI[3], Police Science Institute[4]

요 약

본 논문은 서울시 경찰청에서 특정기간 동안 수집된 112 신고 데이터를 기반으로 미분류 신고유형을 선별하여 LDA(Latent Dirichlet Allocation) 토픽모델링 과정을 거쳐 잠재된 신고유형의 분석을 목표로 한다. 최적의 토픽 수를 선정하기 위하여 Coherence의 값을 기준으로 하이퍼 파라미터를 설정하였고, 실험결과를 바탕으로 기존 신고유형에 속하지 않는 새로운 신고유형을 추론한다.

Ⅰ. 서 론

통계청 사회조사 결과 국민이 사회적으로 가장 불안감을 느끼는 주된 요인은 '범죄 발생'으로 조사되었다. 이러한 국민의 불안감을 해소하기 위한 정부와 경찰의 노력으로 총 범죄발생량은 2010년 186만 건에서 2018년 158만 건으로 감소하였다. 하지만 이에 따라 경찰의 치안활동 영역이 확대되어 업무가 가중되고 있다. 112 신고 접수 업무를 살펴보면 총 112 신고 접수 건수가 2010년 856만 건에서 2018년 1,873만 건으로 10년 사이 2배 이상 증가하였고, 신고유형의 경우 2013년 42종에서 2018년 53종으로 5년간 11종이 증가하였다. 가중되는 112 신고 접수 업무의 효율성 및 정확한 대응을 위하여 112 신고 시스템에 관한 연구가 진행되어야 한다.

관련 연구로 경찰력의 효율적인 운용을 위해 긴급신고 접수 시 신고내용에 적합한 출동 대응을 하는 경찰의 차별적 대응의 중요성에 관한 연구[1]와 112 신고 통계 데이터를 긴급신고와 비 긴급신고 데이터로 나누어 비교 분석한 결과를 토대로 효율적인 긴급신고체계에 대한 운영 기준과 규정 마련, 교육훈련의 개선방안을 제시한 연구[2]가 있다.

본 논문에서는 112 신고 접수 시 기존 신고유형으로 분류되지 않았던 미분류 유형의 신고내용 데이터를 입력하여 LDA 토픽모델링[3]을 진행한다. 이를 통하여 기존 분류체계에 속하지 않는 잠재적 신고유형에 대한 추론이 가능하다.

Ⅱ. LDA 토픽모델링 기반 잠재적 신고유형 분석

본 논문의 전체적인 실험 구조는 아래의 그림 1과 같다.

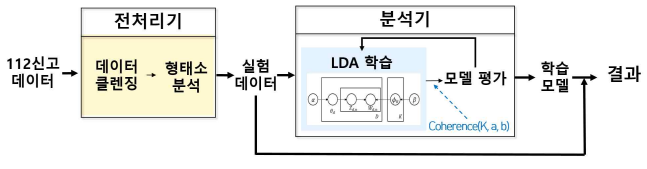


그림 1. 실험 구조도

실험 순서는 먼저 수집된 112 신고 데이터를 전처리기에 입력하여 필요한 칼럼추출, 정규표현식 및 불용어처리 과정을 거쳐 데이터 클렌징을 하고, 형태소 분석기를 거쳐 필요한 형태소로 실험데이터를 구축한다. 구축된 실험데이터를 LDA 모델링과 모델 평가과정을 거쳐 최적의 하이퍼 파라미터 기반의 학습모델을 만든다. 학습된 모델을 기반으로 실험데이터의 토픽과 토픽 내 주요 단어를 확인한다.

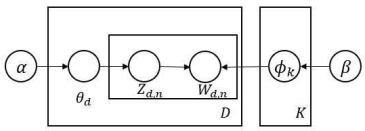
1. 데이터 전처리

서울시 경찰청에서 20○○년 ○월 ~ 20○○년 ○월 까지(1년) 수집된 112 신고 데이터는 총 36개의 필드 정보와 53종의 신고유형으로 총신고수는 424만여 건으로 이루어져 있고, 개인정보에 대해 비식별화를 거친 데이터이다. 이 중 '종결 시 사건코드'가 'None'으로 분류된 '신고내용' 칼럼 데이터를 추출하고, 112시스템 상 자동으로 입력되는 '[접수번호]', '[문자내용]' 등과 같은 대괄호 내용과 '신고내용'이 비어있는 값 등 분석에 필요하지 않은 내용은 정규표현식 및 불용어 처리를 거쳐 삭제하였다. 데이터 클렌징을 마친 '신고내용' 데이터를 형태소 분석하여 주요특징을 가지고 있는 명사만 추출하였다. 이 과정에서 사용된 형태소 분석기는 치안 분야 데이터에 특화된 자체 치안 형태소 분석기를 사용하였다. 데이터 전처리를 마친 실험데이터는 총 758,607건의 신고 수, Bag of Words는 112,144개, 총 Token의 수는 4,725,180개이다.

2. LDA 토픽모델링

LDA 모델은 확률적 토픽모델링 기법으로, 문서 내 단어들에 대한 디리클레 분포를 분석함으로써 잠재되어있는 K개의 토픽을 추출하는 모델이다. 디리클레 분포는 K차원의 실수 벡터 중 벡터 요소가 양수이며 모든 요소의 합이 1인 경우에 대해 확률값이 정의되는 분포이다. 전 처리된 실험데이터를 Python "gensim"패키지의 LDA 모델에 학습시킨다. 아래의 그림 2는 LDA 분석과정을 도식화한 것이다. α는 디리클레 분포를 따르

며 θ_d의 파라미터이다. $Z_{d,n}$은 문서 d의 n번째 단어에 대한 토픽이며 θ_d를 통해 정해진다. 문서 d의 n번째 단어인 $W_{d,n}$은 전체 토픽과 $Z_{d,n}$으로부터 도출된다. 도식화의 사각형은 내부 내용이 반복됨을 의미한다.



D: 문서집합의 수
d: 개별문서의 수
K: 토픽의 수
ϕ_k: k번째 토픽에 대한 단어들의 빈도
θ_d: 문서d에 대한 토픽 비율
$Z_{d,n}$: 문서d의 n번째 단어에 대한 토픽 할당
$W_{d,n}$: 문서d에서 관찰된 n번째 단어
α: 양의 K벡터 모수
β: 토픽 하이퍼파라미터

그림 2 LDA 분석과정

3. 모델 선택

LDA 모델은 하이퍼 파라미터 α, β, K를 설정하여 모델의 성능을 조절할 수 있다. α는 문서 내 토픽 밀도, β는 토픽 내 단어 밀도, K는 토픽 개수를 의미한다. 하이퍼 파라미터들을 조절하여 여러 LDA 모델을 학습시킨 후 각 모델의 *Coherence*[4]와 *Perplexity*[5]의 수치를 이용하여 성능을 평가한다. *Perplexity*는 확률모델의 예측 정확도를 판단할 때 사용되는 지표로 학습이 잘 된 정도를 나타내지만, 학습의 정도와 사람의 해석이 비례하지 않다는 연구 결과[6]를 참고하여 본 실험모델의 평가지표로 *Coherence*를 사용하였다. *Coherence*는 토픽의 의미론적 일관성을 나타내는 지표로 특정 토픽의 단어 간 유사도를 합하여 나타낸다. *Coherence*를 측정하는 여러 가지 측정 방법 중 *UMASS*[7] 측정법을 사용하였다. *UMASS* 수치는 단어 쌍의 동시 발생 문서 수에 의해 결정된다.

$$score(v_i, v_j, \epsilon) = \log\frac{D(v_i, v_j) + \epsilon}{D(v_j)} \quad (1)$$

$D(v_i, v_j)$는 단어 v_i, v_j가 동시에 포함된 문서의 수이고, $D(v_j)$는 단어 v_j가 등장하는 문서의 수이다. 하이퍼 파라미터 K = 3, 4, 5, 6, 7, 8, 9, 10, α = 0.01, 0.31, 0.61, .0.91, asymmetric$(1/K)$, auto, β = 0.01의 조합으로 구성된 모델 48개를 생성하고, 각 모델에 대해 *UMASS* 측정법을 기준으로 *Coherence*를 계산하여 모델의 성능을 평가하였다.

표 1. 모델의 하이퍼 파라미터 및 Coherence

모델	K	α	*Coherence*
모델 1	**5**	**0.01**	**-3.59789**
모델 2	4	asymmetric	-3.656
모델 3	5	asymmetric	-3.69118
모델 4	6	asymmetric	-3.71785
모델 5	4	0.31	-3.74983
모델 6	6	0.31	-3.76684

표 1의 내용은 총 48개의 모델 중 *Coherence* 수치로 선정한 상위 8개 모델의 하이퍼 파라미터값과 *Coherence* 값이다. 이 수치를 참고하여 각각 모델의 결과를 확인하고 설명력이 가장 높은 모델 1을 선정하였다.

4. 실험결과

아래의 표 2는 선정한 모델 1의 실험결과 도출된 토픽과 각 토픽을 구성하는 단어 중 토픽 간 상대적 베타성을 가지는 단어를 선택하여 토픽 주제를 추론한 결과이다. 토픽은 총 5가지로 분류되었고 각각 토픽의 내용을 보면 토픽 2와 토픽 3, 토픽 4는 기존 112 신고유형 분류체계에 속할 수 있는 교통 불편 및 교통 문의, 미귀가자 및 가정 발생 신고, 택시 관련 분실물 신고의 토픽이라 판단되고, 잠재적 신고유형으로 볼 수 토픽 1과 토픽 5의 경우 신고내용이 명확하진 않지만, 긴급으로 현장출동을 요청하는 신고 토픽과 최근 급증하고 있는 보이스피싱 피해 및 신고에 대한 토픽으로 추론된다. 이러한 결과를 바탕으로 기존 112 신고유형 체계에 새로운 유형 추가를 검토하거나, 교육훈련의 자료로 활용될 수 있다.

표 2. 토픽 주제 및 해당 토픽의 상위 단어

토픽	토픽 주제	토픽 내 단어 구성
1	**기타 긴급출동 요청**	**신고, 출동, 사람, 집, 확인, 말, 문, 필요, 전화, 연락, 현장, 도착**
2	교통 불편 및 교통사고	차량, 문의, 주차, 안내, 교통사고, 주차장, 불법주차, 오토바이, 불법, 번호, 견인
3	미귀가자 및 가정 발생 신고	문자, 정보, 위치, 연락, 위치추적, 남자, 아이, 남편, 아들, 친구, 엄마, 여자, 가정폭력
4	택시 및 분실물	택시, 파출소, 핸드폰, 지갑, 분실, 소리, 카드, 휴대폰, 택시기사, 손님, 분실신고, 버스, 가방
5	**보이스피싱 피해**	**보이스피싱, 주소, 성명, 피해, 번호, 명의자, 사이버, 통장, 사칭, 대포통장**

Ⅲ. 결 론

본 논문에서는 경찰청 112 신고 데이터에서 기존 분류체계에 속하지 않는 미분류 유형의 데이터를 LDA 토픽모델링으로 분석하여 잠재적 신고유형을 도출하였고, 최적의 토픽 수를 선정하기 위하여 Coherence 값을 기준으로 LDA 토픽모델을 선택하였다. 연구 결과 기존 112 신고유형 체계에 추가 할 수 있는 긴급 현장출동을 요청하는 토픽과 보이스피싱 피해 토픽을 도출하였다. 본 연구를 바탕으로 증가하는 치안 수요에 대한 치안 업무활동 경감과 112 신고 데이터 품질 향상을 기대한다.

ACKNOWLEDGMENT

본 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No.2018- 0-00440, 위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발].

참 고 문 헌

[1] 석청호 (2008). 112 신고에 대한 차별적 경찰 대응방안에 관한 연구. 한국공안행정학회보, 17(4), 213-246.

[2] 노성훈, 조준택. (2016). 112 긴급신고시스템 운용상의 문제점 실증분석 및 개선방안. 경찰학논총, 11(4), 7-38.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[4] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

[5] Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent dirichlet allocation. In Advances in neural information processing systems (pp. 1081-1088).

[6] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

[7] Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952-961).