빅데이터 활용 및 사업을 위한 비식별화 전략

Big Data Utilization and De-identification Strategy for Business

저자 (Authors)

김정범, 임채원, 하재현, 김문기, 박연지, 신진슬, 김유지, 이단비, 이진형, 하송미, 김지현, 김은석
Kim JeongBeom, Lim ChaeWon, Ha JaeHyub, Kim MoonKi, Park YeonJi, Shin JinSeol, Kim Yooji, Lee Danbi, Lee Jinhyung, Ha Songmi, Kim Jihyon, Kim Enseok

출처 (Source)

Proceedings of KIIT Conference , 2019.6, 616-619(4 pages)

발행처 (Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08750135

APA Style

김정범, 임채원, 하재현, 김문기, 박연지, 신진슬, 김유지, 이단비, 이진형, 하송미, 김지현, 김은석 (2019). 빅데이터 활용 및 사업을 위한 비식별화 전략. Proceedings of KIIT Conference, 616-619

빅데이터 활용 및 사업을 위한 비식별화 전략

김정범*, 임채원**, 하재현**, 김문기**, 박연지**, 신진슬**, 김유지**, 이단비**,
이진형**, 하송미**, 김지현**, 김은석**

Big Data Utilization and De-identification Strategy for Business

Kim JeongBeom*,Lim ChaeWon**,Ha JaeHyub**,Kim MoonKi**,Park YeonJi**,Shin JinSeol**Kim Yooji**,Lee Danbi*Lee Jinhyung**, Ha Songmi**, Kim Jihyon**, and Kim Enseok**

요 약

빅데이터를 사용할 때 가장 중요한 요소 중 하나는 비식별 전략입니다. 개인 정보를 식별 할 수없는 식별 불가능한 정보는 빅데이터 분석 및 출력으로 만 사용할 수 있습니다. 비식별 조치는 대용량 데이터 수집을 위한 개인 정보 적용 정책에 따라 적절하게 수행되어야합니다. 비식별전략은 큰 데이터 세트에서 개인을 식별할 수있는 요소 전부 또는 일부의 삭제, 대체 등을 통해 개인을 식별하는 것을 불가능하게 하는 정책입니다. 비식별 정보는 전략 수립을 통해 개인 정보 이외의 정보로 추정되므로 정보 주체의 동의없이 제 3 자에게 사용 또는 제공 할 수 있습니다. 따라서 빅데이터 분석과 결과의 활용에 있어서 가장 중요한 전략입니다. 식별되지 않은 결과는 비 개인 정보로 간주되지만 새로운 바인딩 기술이 나타나거나 결합 될 수있는 정보가 다시 식별 될 수 있으므로, 필수적인 관리 및 기술 안전장치를 구현해야합니다.

Abstract

One of the most important factors in using big data is the de-identification strategy. Non-identifiable information that does not identify personal information can only be used as a big data analysis and output. De-identification measures should be appropriately performed in accordance with the personal information application policy for the collection of big data types. De-identification is a policy that makes it impossible to identify an individual through deletion, substitution, etc., of all or some of the elements that can identify an individual in a big data set. Since the de-identification information is estimated as information other than personal information by establishing strategy, the information can be used or provided to a third party without consent from the information subject. Therefore, in the analysis of big data and utilization of the result it is the most important strategy. Although non-identified outputs are assumed to be non-personal information, essential management and technical safeguards should be implemented, as new binding techniques may appear or information that can be combined may be re-identified. Through this paper, we will examine the related strategies and implementation example.

Key words

big data, personal information, de-identification measure, de-identification policy, combine technology

* 남서울대학교 대학원 빅데이터인공지능학과 주임교수/사업단장
** 남서울대학교 대학원 빅데이터인공지능학과 석사과정학생

Ⅰ. 서 론

1990년대 중반부터 대기업을 중심으로 일반 직원들이 회사 경비를 사용할 때는 법인카드를 많이 사용하였으며, 이를 회계처리 하기 위해서 일반 직원들은 법인카드전표와 사용내역을 경리부서 직원들에게 전달하였고, 경리부서 직원은 이를 전표기표하고 카드매입 전표를 보관하였다. 그러나 현재는 정보기술의 발달 및 무증빙 시스템의 도입으로 카드사용내역이 카드사로부터 익일 전산시스템으로 자동 전달되며, 경비를 지출한 직원이 직접 회계시스템을 이용하여 본인이 사용한 비용을 신청하는 형태로 업무가 바뀌었다. 이 경우 비용을 신청하는 직원은 사용한 비용에 대한 회계계정을 선택해야 하는데, 실무에서 회계 계정코드에 익숙하지 않는 일반 직원들이 회계 계정코드를 선택할 때 많은 실수를 하여 경리부서로부터 반려되는 경우가 부지기수다. 본 연구에서는 이러한 실수를 최소화하기 위하여 카드사용 내역의 Text 정보를 활용하여 회계계정코드를 분류할 예정이며, 실무에서 활용하면 업무효율화 및 계정오류로 인한 Risk를 줄일 수 있다. 또한, 법인카드에 국한하지 않고 적요사항을 포함하는 모든 회계 전표에 응용할 수 있을 것이다.

Ⅱ. 이론적 배경

2.1 회계 계정 분류에 대한 선행 연구

국내외 논문을 조사한 결과, K-ICT 빅데이터센터에서 공개한 2016년 글로벌 융합 사례집(2015년 빅데이터 시범사업 및 국내외 사례를 중심으로)[11]에서 국산 ERP벤더 중 하나인 D社의 선행 연구가 있었다. 선행 연구는 빅데이터 기반 자동 기장 회계프로그램에 대한 연구로 판단되며, 선행 연구에서는 전통적인 기계학습 알고리즘인 SVM(Support Vector Machine)을 사용했으나, 본 연구는 CNN (Convolutional Neural Network)을 기반으로 할 것이며, 다수의 회계 기표 사례 중에서 신용카드 경비로 한정해서 전표 기표에 사용되는 계정코드를 분류할 예정이다. 또한, 형태소 분석이 분류의 정확도에 영향을 줄 것인지도 추가로 비교 분석할 것이다.

CNN 알고리즘을 이용하여 Text를 분류한 대표적인 선행연구로는 Yoon Kim, “Convolutional Neural Network for Sentence Classification [1] 과 조휘열, 김진화, 윤상웅, 김경민, 장병탁의 “컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술” [2] 을 들 수 있다. [1]에서는 word2vec을 사용했으나 [2]에서는 word2vec이 성능 향상에 큰 도움이 되지 않아 룩업 테이블을 사용하였다.

2.2 Convolutional Neural Network (CNN)

CNN 알고리즘은 이미지의 특징을 추출하여 유사성을 찾는 것으로, 이미지 분석에 주로 사용된다. 원본 이미지를 적당한 크기의 필터를 이용하여 슬라이딩 하면서 매번 슬라이딩 할 때 마다 원본 이미지의 매트릭스를 이용해, Convolution Value를 결정하게 되고, 이를 전체 이미지로 반복하여 Convolution의 매트릭스를 구한다.

그림 1은 일반적인 CNN을 Text 분류에 응용한 것으로 A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification [3] 에 기초한다. 초기에 단어를 벡터로 임베딩하는 과정이 선행되어야 하며, 이후는 이미지 분류와 유사하다.

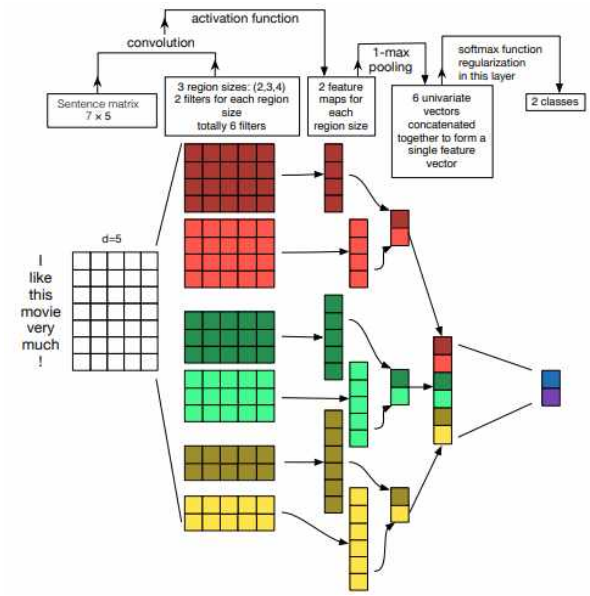

그림 1. 문장 분류를 위한 CNN 아키텍처
Fig. 1. Illustration of a CNN architecture for sentence classification

3개의 필터 사이즈 5x4, 5x3 ,5x2를 각 두 개씩 총 6개를 문서 매트릭스에 합성곱을 수행하고 Convolution Matrix 즉, Feature Map을 생성한다. 이후 각 Map에 대해 Max Pooling을 진행하여 각 Feature Map으로부터 가장 큰 값을 찾는다. 이들 6개 Map에서 Single Feature 벡터를 형성하며, Softmax Function의 Input이 되며, Softmax Function은 최종적으로 문장을 분류한다.

III. 연구 절차 및 평가 방법

3.1 연구절차

연구 순서는 데이터의 수집, Feature 정의 (형태소분석 포함), CNN 모델 구현, 모델평가 순서로 진행하였다.

3.2 데이터 수집 및 구성

본 연구에서 사용되는 입력데이터는 A社의 2018년 한해동안 사용한 법인카드 총 6,227건을 사용하였다. A社 의 경우, 기업의 직원이 경비를 지출할 때 법인카드를 사용하면 늦어도 다음날, 카드社로부터 사용내역이 A社 전산시스템으로 전달된다.

카드社로부터 전달되는 데이터는 사업자등록번호, 가맹점명, 업종명, 승인일, 승인번호, 부가세정보 등이 있다. 이외에도 A社 직원이 직접 경비 신청할 때 추가되어지는 정보로써 부서명, 직급명, 사용목적(적요) 등이 있으며, 분류에 사용되는 회계 계정코드는 비용 계정 15개이다.

3.3 학습을 위한 Feature 및 모델 종류

본 연구는 입력데이터를 총 4가지 Case로 나누어 분류 모델을 구현하고자 한다. Domain 관련 지식이 있는 A社 현업 전문가의 추천에 따라서, 카드社에서 전달하는 가맹점명, 카드업종명과 개인이 경비 신청할 때 입력하는 신청자 직급, 신청자 부서명, 사용목적을 추가하는 등 분류의 정확도를 높이기 위해서 다양한 시도를 하였다.

Case 1은 수집된 정보 중에서 사용목적만을 Feature로 하고, 학습하는 모델이다. 사용목적에 대해서 형태소분석 없이 띄어쓰기 기준으로 잘라내어 사전을 만들 것이며, CNN 학습의 입력 자료로 사용하기 위해서, 사전의 Index를 이용하여, 사용목적 Text를 벡터화 하였다.

Case 2는 Case 1과 유사하게 사용목적만을 Feature로 하고, 학습하는 모델이다. 다만, 사용목적 데이터를 형태소분석을 하여 사전을 만든다. 형태소분석결과의 품사 중에서 어말어미, 조사, 부호, 접두사, 접미사는 사전 작성 시 제거하였다. CNN 학습시는 형태소분석이 이루어진 사용목적 Text를 벡터화 하여, 입력 자료로 사용한다.

Case 3은 가맹점명, 카드업종명, 신청자 직급, 신청자 부서, 사용목적을 Feature로 하고, 학습하는 모델이다. 분류의 정확도를 높이기 위해서 가맹점명은 [A], 카드업종명은 [B], 신청자 직급은 [C], 신청자 부서는 [D], 사용목적은 [E]의 값으로 각 Text에 접두어 형태로 추가하여, 결합 Text를 완성한다. 완성된 문서에 대해 형태소 분석 없이 띄어쓰기 기준으로 잘라내어 사전을 만들었으며, CNN 학습의 입력 자료로 사용하기 위해서 결합 Text를 벡터화 한다. Case 4는 Case 3의 Feature를 형태소분석하여 사전을 만들며, 나머지는 Case 3와 동일하다.

3.1에서 언급한 각 Case에 공통적으로, word2vec 형태의 임베딩 기법보다는 조휘열, 김진화, 윤상웅, 김경민, 장병탁의 “컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술” [2] 처럼 룩업 테이블형태로 Text를 Embedding 한다.

3.4 모델의 파라미터 설정 및 평가 방법

계정분류를 위한 모델을 만들 때, 가장 성능이 좋은 모델을 확보하기 위해서는 여러 분류모형을 구축하고, 그 모델들 중 어떤 모델이 가장 뛰어난 분류를 보이는지에 대해 비교하여 최종적으로 평가해야 한다. 본 연구의 CNN 모델에서는 Embedding Dimension은 256 Filter Size는 3, 4, 5로 하였으며, 각 Filter 수는 256, Epoch은 2,000으로 하였을 때, 분류의 정확도 가장 높았다.

본 연구에서는 계정분류를 원하는 범주에 정확히 분류했는지 평가하기 위하여 Confusion Matrix를

사용하며, 평가 척도로써 분류의 Accuracy(정확도), Precision, Recall, F1 Score를 사용하였다.

Ⅳ. 결론 및 향후 연구

최종 학습모델을 평가한 결과, Case 4가 가장 우수한 것으로 나타났으며, 이는 사용목적으로만 학습시키는 것보다 가맹점명, 카드업종명, 신청자 직급, 부서 등 Feature를 추가한 것이 평가에 영향을 주었다는 것을 의미한다. 또한, Features를 단순히 결합하는 것보다, A/B/C/D/E의 구분자를 두어 결합하는 것이 효과적이었다는 것을 의미한다. 이외에도 회계계정코드 분류에서는 Text를 단순 띄어쓰기 형태로 사전을 만드는 것보다는 형태소 분석한 결과가 조금 더 우수하다는 것을 확인하였다.

기업의 경리부서 직원이 아닌 일반부서 직원이 직접 기표할 수 있는 부분은 법인카드 외에 다수 존재하며, 본 연구의 결과로, 해당 분야에 쉽게 적용할 수 있을 것이며, 경리부서 직원이 입력하는 전용 시스템에도 Text 기반으로 회계 계정코드를 추천할 수 있을 것이다.

표 1. Case 별 테스트 결과 요약 비교
Table 1. Comparison of case-by-case test result summary

Case	Accuracy	Precision	Recall	F1-score
Case 1	85.6%	86.1%	85.5%	85.6%
Case 2	86.8%	87.3%	86.7%	86.7%
Case 3	89.5%	89.7%	89.5%	89.5%
Case 4	91.2%	91.2%	91.2%	91.1%

다만, 본 연구는 CNN 알고리즘에 한정되어 연구한 부분에 아쉬운 점이 있으며, 향후 추가로 연구할 부분으로 기존 적용한 알고리즘이외에도 Character Level CNN/RNN, TF-IDF 등 다른 알고리즘을 이용하여 분류 정확도를 향상하고자 한다.

참 고 문 헌

[1] Yoon Kim, “Convolutional Neural Network for Sentence Classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014

[2] 조휘열, 김진화, 윤상웅, 김경민, 장병탁, “컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술”, 2015년 동계학술발표회 논문

[3] Ye Zhang, Byron Wallace, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification

[4] Rajaraman, Anand, and Jeffrey D. Ullman. "Mining of massive datasets," Vol. 77. Cambridge: Cambridge University Press, pp. 1-17, 2012.

[5] Lai, Siwei, et al. "Recurrent convolutional neuralnetworks for text classification." 29th Association for Advanced of Artificial Intelligence Conference on Artificial Intelligence. 2015.

[6] Johnson, Rie, and Tong Zhang. "Effective Use of Word Order for Text Categorization with Convolutional Neural Networks," North American Chapter of the Association for Computational Linguistics, 2015

[7] Turian, Joseph, Lev Ratinov, and Yoshua Bengio. “Word representations: a simple and general method for semi-supervised learning,” In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 384394, 2010

[8] Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space". In Proceedings of Workshop at International Conference on Learning Representations, 2013

[9] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series," In M.A. Arbib (Ed.), The handbook of brain theory and neural networks, Cambridge, MA: MIT Press, pp. 255258, 1995

[10] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International Conference on Machine Learning, 2015

[11] K-ICT 빅데이터센터에서 공개한 2016년 글로벌 융합 사례집 (2015년 빅데이터 시범사업 및 국내외 사례를 중심으로)