자동 기장 서비스를 위한 키워드 중심 품목 분류

[1] 최동빈, [*2] 박용범, [3] 조인수
[1,제 1 저자] 단국대학교, 72180417@dankook.ac.kr
[2*,교신저자] 단국대학교, ybpark@dankook.ac.kr
[3] 단국대학교, aor1@naver.com

Keyword-centric item classification for automatic bookkeeping services

[1]Daong-Bin Choi, [2]Young B. Park, [3]Insu Jo
[1, First Author]Dankook University, 72180417@dankook.ac.kr
[*2,Corresponding Author]Dankook University, ybpark@dankook.ac.kr
[3]Dankook University, aor1@naver.com

요 약

인공지능을 비롯하여 머신러닝의 기법들이 발전함에 따라서 많은 분야에 활용되고 있다. 특히 인건비 등의 문제로 세무서에서는 인력을 대체할 자동 기장서비스를 머신러닝을 활용하여 구축하고 있다. 자동 기장서비스의 기반이 되는 머신러닝 및 딥러닝을 활용기 위해서는 이를 학습시킬 데이터가 확보되어야한다. 하지만, 그러한 데이터는 현재 많이 존재하고 있지 않으며, 그 형태의 다양성으로 인하여 자동화 되지 않아, 많은 인력을 동원하여 생성되는 경우가 대부분이다. 세무분야 역시 마찬가지로 기존에 존재하는 많은 데이터를 머신러닝에 상용가능하게 해주는 전처리 단계가 필요하다. 하지만, 기존 데이터 즉 장부에서 품목 항목은 그 형태가 다양하여 전처리가 복잡하다. 이러한 전처리를 단순화 하기 위해 본 논문은 키워드 중심으로 품목을 1차 분류하는 알고리즘을 제안하여, 필요한 전처리의 과정을 단순화 하였다

Abstract

Machine learning techniques, including artificial intelligence, are being used in many areas as they develop. In particular, due to labor costs and other problems, the tax office is deploying automatic bookkeeping services using machine learning to replace the workforce. In order to utilize machine learning and deep learning, which are the basis for automatic bookkeeping service, data must be available to learn. However, such data do not currently exist and are not automated due to its type of diversity, which is often generated by the use of a large number of people. The tax sector also needs a preprocessing step that enables many existing data to be commercialized in machine learning. However, the existing data, or items in the books, have various forms of preprocessing complex. To simplify such preprocessing, this paper proposed an algorithm for first classifying items based on keywords, simplifying the process of required preprocessing.

Keywords: *Preprocessing, Machine learning, Classification, Automatic bookkeeping, services, Keyword-centric*

I. 서론

* Corresponding Author
Received: March. 27, 2019, Revised: July. 04, 2019, Accepted: July. 06, 2019

세무대리서비스에 대한 인건비의 문제는 지속적으로 제기되고 있다[1]. 세무서는 인건비에 대한 부담으로 기존 기장서비스를 자동화를 시도하고 있다. R.Deepa lakshmi[2]는 taxation analysis 에 여러 머신러닝 기법을 테스트하여, 그 효용성을 보여주었다. 인공지능을 비롯한 머신러닝을 활요하기 위해서는 해당 task 와 관련된 학습 데이터가 필수적이다. 세무분야의 경우 전산화를 통하여 기존에 사용되던 정보를 활용하여 머신러닝에 활용이 가능한 학습데이터를 생성하고 있다.

기존의 세무분야에서 활용하고 저장한 정보는 사람 기준으로 작성되어있기 때문에 바로 머신러닝에 활용하기엔 부적합하다. 따라서, 기존 장부 기록을 활용하여 학습데이터를 생성하기 위해서는 어느 정도의 전처리 과정이 필요하다.

기존 데이터에서, 큰 전처리 없이 바로 활용할 수 있는 데이터가 있는가 하면 품명 항목과 같이 다양하게 표현되어 있어 복잡한 전처리가 필요한 데이터도 있다.

본 연구는 품명 항목과 같이 복잡한 데이터의 전처리를 단순화하기 위해서 키워드 중심의 품목 분류 알고리즘을 제안한다. 키워드를 중심으로 1 차 분류를 시켜, 머신러닝 및 딥러닝에 활용할 수 있는 학습데이터 생성을 가능하게 한다.

본 연구의 구성은 다음과 같다. 2 장에서는 기존 관련 연구에 대해서 살펴보며, 3 장에서는 본 연구의 알고리즘에 대한 설명, 4 장에서는 제안한 알고리즘을 실제 장부에 사용하여 생성한 학습데이터를 머신러닝에 학습시켜 그 결과를 제시한다. 5 장에서는 본 연구의 한계점과 이를 개선할 방안에 대하서 논의하고자 한다.

II. 기존 연구

Zhang et al[5]은 기존 키워드에 대한 연구를 아래와 같이 조사하였다.

2.1. Statistics Approaches

통계를 기반으로 키워드를 추출하는 방식은 특별한 훈련데이터가 필요하지 않다. Cohen[7]은 N-Gram 통계 정보를 사용하였다. 다른 통계 방법으로는 Word frequency[8], TF*IDF[9], word co-occurrences[10], PAT-tree[11]등이 있다.

2.2 Linguistics Approaches

언어학적 특징을 이용한 접근법은 lexical analysis[12], syntactic analysis[13], discourse analysis[14,15]등이 있다. 이러한 접근방식은 단어, 문장, 문서의 언어학적 특징을 활용한다.

2.3 Machine Learning Approaches

Naïve Bayes[16], SVM[17], Bagging[13] 등의 머신러닝 기법을 활용하여 키워드 추출하는 기법들이 존재한다. 머신러닝 기법들 중 지도학습을 사용하기에 이에 맞는 학습 데이터가 요구된다.

2.4 Other Approaches

J. B. Keith Humphreys[18]의 연구와 같이 위에 제시된 방식들을 결합하거나 일부 휴리스틱 지식을 사용하여 추출하는 방식이 있다.

III. 제안 알고리즘

3.1. 키워드 추출

위의 기법들을 활용하여 키워드를 추출하기 위해서 제안된 여러 알고리즘[3, 4, 5]들이 있으나, 본 논문은 통계기반중 단어의 빈도를 바탕으로 키워드를 추축 하였다. 품명의 형태는 그림 1 과 같이 다양한 형태로 되어 있다. 품명에 사용되는 단어들의 빈도를 파악하면 다양한 품명에서도 특정 단어들이 빈번하게 발생하는 사실이 확인 가능하다. 이러한 특정 단어은 상품을 특성을 대변하는 특정 단어들로 구성되어 있다. 따라서 특정 반복 단어를 키워드로 선정이 가능하다.

품명
2017년 03월분 요금
물품구매
;03:05;카스(중)
물티슈
수수료
장류 외
전기요금
2017년 02월분 요금
003205611500일반전화
;03:05;카스(중)
수수료
355사이다외
장류 외
전기요금
2017년 01월분 요금
003205611500일반전화
[거래수수료][기계]커뮤니케이션미디어학전공/모니터(1)
[거래수수료][기계]MOOC센터/모니터(1)
여행용캐리어가방
복합기
전자레인지
전자레인지
프린터
빔프로젝트
CLP-426

Figure 1 1 Various types of name value

즉 패턴 별 단어의 빈도를 확인하여 가장 높은 빈도를 가지는 단어를 키워드로 선정하며 이를 통해서 품명이 가지는 특성을 대변하는 키값으로 사용이 가능하다. 이러한 키값으로 선정된 단어들을 바탕으로 그림 2 와 같이 1 차 분류 카테고리의 분류 기준으로 활용된다.

3.2 카테고리 생성

S. Vijayarani et al.[6]에서 추출한 텍스트를 하나의 카테고리로 묶어 활용도를 높인 것처럼, 선정된 키워드를 대표할 수 있는 카테고리로 묶어 1 차 분류의 기준으로 삼았다. 선정된 카테고리와 해당하는 키워드는 그림 2 와 같다.

관리비, 수도료, 수수료에서 요금과 같이 중복되어 나타나는 키워드는 비용이라는 상위 카데고리에 키워드를 매칭하였다.

모든 비용 관련 카테고리를 비용 항목에 통합 시키지 않은 이유는 1 차 분류 이후 세금의 계정과목 선정에 있어 활용할 수 있는 특징을 없애지 않기 위해서 이다.

식품	상품	관리비	수수료	수도료	통신비	비용	기타
장류	물티슈	관리비	수수료	수도료	통신	요금	기타
사이다	Office	과세	대여	도시가스	전화	보도자료	증모
반찬류	매출할인	주차권	사용분	lpg	이너텔	홈페이지제작	
카스	Display	일 반 관 리 비	이용료	LP	인터넷	사용료	
탁주	Monitor	관 리 비	일시납	가스	LTE	렌탈료	
음료	65LX541HW	임차료	수리비	LP.G	bizmeka	스티커	
농산물	PC	임대	임원변경	LN2	MTM	광고서비스	
식용유	서울여대	차임	조명공사		loT	광고	
베이킹파우더	CMP	월세	지점설치		광랜	공사	

Figure 2 Part of category

3.3 키워드 기준 1 차 분류

그림 2 에 제시한 매칭 테이블을 활용하여, 품명을 1 차 분류를 하면 다음과 같이 분류가 가능하며, 이를 활용하여 학습데이터 생성이 가능하다.

품명	카테고리
2017년 03월분 요금	비용
;03;05;카스(중)	식품
물티슈	상품
장류 외	상품
전기요금	비용
2017년 02월분 요금	비용
003205611500일반전화	통신비
;03;05;카스(중)	식품
355사이다외	상품
장류 외	상품
전기요금	비용
2017년 01월분 요금	비용
003205611500일반전화	통신비
[거래수수료][기계]커뮤니케이션	수수료
[거래수수료][기계]MOOC센터/	수수료
CLP-426	수도료

Figure 3 Example of 1st classification

그림 3 에 보이듯 복잡한 품명은 1 차 분류 카테고리에 키워드 중심으로 매칭이 되었으며, 이렇게 매칭된 카테고리는 one-hot encoding 등의 방식을 통해서 학습데이터로 활용이 가능하다.

IV. 알고리즘 적용하여 생성한 학습데이터 활용

자동 기장 서비스에서 세금 계정과목을 선정하는 모델에 본 논문이 제안한 알고리즘을 적용하여 학습데이터를 생성, 학습한 결과는 표 1 과 같다. 자동 기장 서비스를 구현하는데 사용된 머신러닝 기법은 랜덤 포레스트로 구성되었다.

학습 데이터는 세무서에 저장된 2017, 2018 년도 데이터를 사용하였다. 주로 개인 사업자로 구성되어 있으며, 총 4 가지의 형태로 거래가 이루어졌다. 전자 세금 계산서, 전자 계산서, 신용카드, 현금영수증으로 4 가지의 거래 형태에 대한 증빙으로 사용되었으며, 실제 장부에 기록된 정보는 증빙 자료를 기반으로 작성되었다,

이중 전자 세금 계산서, 전자 계산서가 품명에 대한 정보를 사용하였다. 전자 세금 계산서와 전자 계산서의 양 또한 다른 거래 형태 보다도 많아서 학습데이터 활용에 적합하였다. 전자 세금 계산서와 전자 계산서에 저장된 품명을 본 논문의 알고리즘을 사용하여, 2017 년도 장부 기록을 기반으로 생성한 2017 data set 과 2018 년도 장부 기록을 기반으로 생성한 2018 data set 를 활용하였다.

Train set 과 test set 은 각 년도 data set 에서 랜덤으로 75%, 25%분포로 나누어 구성하였다.

Table 1 Acc for machine learning

Train data set	Train acc	*Test acc*
2017 data set	0.98	0.94
2018 data set	0.97	0.94

표 1 의 결과와 같이 본 논문의 알고리즘을 사용하여 만들어진 data set 은 머신러닝 기법에 충분이 활용이 가능하다.

V. 결론

머신러닝 및 딥러닝을 활용하기 위해선 학습 데이터를 생성해야 할 필요성이 있다. 회계장부의 경우 특별한 전처리 없이 사용이 가능한 항목이 있는 경우도 있으나, 품명과 같이 복잡한 형태로 되어 있어 전처리 과정이 필요한 항목이 존재한다.

이러한 복잡한 형태를 지닌 품명의 전처리를 위해서 본 논문은 키워드 중심의 1 차 분류를 제안하였으며 이를 활용하여 생성된 학습데이터를 활용하여 머신러닝에 적용해 보았으며, 충분한 성과를 냈다.

다만, 키워드를 생성하는데 있어 다소 자동화에 대한 문제점이 있으며, 생성된 키워드와 카테고리 간의 매칭 문제 역시 지속적으로 개선해야 할 필요성이 있다.

VI. 감사의 글

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음" (IITP-2019-2017-0-01628)

VI. 참고문헌

[1] Jung Hwa Hong, Mi Young Lee. (2009). A Study on the Determinant Standards of Taxation Representation Service Fees. ACCOUNTING INFORMATION REVIEW, 27(1), 91-111*:*

[2] Lakshmi, R. D., & Radha, N. , "Machine Learning Approach for Taxation Analysis using Classification Techniques.," International Journal of Computer Applications, 12(10). , 2011.

[3] Huan et al., 2006 Huan C., Tian Y., Zhou Z., Ling C.X., Huang T. Keyphrase extraction using semantic network structure analysis, Proceedings of the sixth international conference on data mining (2006), pp. 275-284

[4] Hulth, 2003 Hulth A. Improved automatic keyword extraction given more linguistic knowledge, Proceedings of the conference on empirical methods in natural language processing (2003), pp. 216-223.

[5] Zhang et al., 2008 Zhang C., Wang H., Liu Y., Wu D., Liao Y., Wang B. Automatic keyword extraction from documents using conditional random fields, Journal of Computational Information Systems, 4 (3) (2008), pp. 1169-1180

[6] Vijayarani, S., Ilamathi, M. J., and Nithya, M., 2015. Preprocessing Techniques for Text Mining: An Overview. International Journal Computer Science and Communication Network, 5, 7-16.

[7] J. D. Cohen. Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting. Journal of the American Society for Information Science, 1995, 46(3): 162-174.

[8] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[9] G. Salton, C. S. Yang, C. T. Yu. A Theory of Term Importance in Automatic Text Analysis,

Journal of the American society for Information Science, 1975, 26(1): 33-44.
[10] Y. Matsuo, M. Ishizuka. Keyword Extraction from a Single Document Using Word Co-ocuurrence Statistical Information. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.
[11] L F. Chien. PAT-tree-based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the 20$^{\text{소}}$ Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR1997), Philadelphia, PA, USA, 1997: 50-59.
[12] G. Ercan, I. Cicekli. Using Lexical Chains for Keyword Extraction. Information Processing and Management, 2007, 43(6): 1705-1714.
[13] A. Hulth. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216-223.
[14] S. F. Dennis. The Design and Testing of a Fully Automatic Indexing-searching System for Documents Consisting of Expository Text. In: G. Schecter eds. Information Retrieval: a Critical Review, Washington D. C.: Thompson Book Company, 1967: 67-94.
[15] G. Salton, C. Buckley. Automatic Text Structuring and Retrieval –Experiments in Automatic Encyclopaedia Searching. In: Proceedings of the Fourteenth SIGIR Conference, New York: ACM, 1991: 21-30.