한국컴퓨터정보학회 동계학술대회 논문집 제25권 제1호 (2017. 1)

word2vec을 이용한 한약재 추천 시스템 연구

안주언O*, 김연주*, 김헌성*, 김우제*, 이윤호*
*서울과학기술대학교 SW분석설계학과
e-mail: llckybbang@naver.comO

A study on medical herb recommendation system using word2vec

Joo-Eon AhnO, Yeon-Ju Kim*, Hun-Sung Kim*, Woo-je Kim*, Yunho Lee*
O*Dept. of Software Analysis and Design, SeoulTech

● 요 약 ●

여러 약재의 복합적인 작용으로 치료를 행하는 한의학의 특성으로 여러 처방과 약재 조합들을 기억하고 있어야 하는 한의사의 어려움을 줄이고 환자에게 보다 높은 질의 의료 서비스를 제공할 수 있는 환경을 만드는 것이 목적이다.
다양하고 복합적인 약재의 조합으로 증상을 치료하는 한의학의 특성 때문에 셀 수 없이 많은 약재의 조합이 존재하며 한의사가 이 모든 조합을 기억하기는 어렵기 때문에 한의사들이 환자를 처방함에 있어 조금이라도 보탬이 될 수 있는 처방 지원 시스템을 개발할 필요가 있다.
word2vec을 이용하여 처방과 약재의 조합을 추천해주며 분석을 통해 산출된 약재의 조합과 그 조합이 실제 의서에 존재하는지의 여부를 함께 알려주어 한의사가 보다 더 주의하여 환자에게 처방할 수 있다.

키워드: 추천시스템(recommendation system), 한약재(medical herb), word2vec

I. Introduction

양의학이 특정 성분을 이용해 치료를 하는 방향으로 발전해왔다면 한의학은 여러 약재들의 성분 조합으로 증상을 치료하는 방식으로 발전해왔다. 기원전 약 300년에는 247개의 약재와 150개 처방이 있었지만 현재는 1800여종의 약재와 6만개의 처방이 존재한다. 이러한 조합을 정리한 의서가 동의보감, 방약합편 등이 있다. 따라서 한의사들은 이러한 의서를 참고하여 처방을 내리게 되는데 한의사가 많은 조합을 기억하고 있더라고 모든 처방을 처방 시에 기억해 내기엔 쉬운 일이 아니다.

또한 한의학의 다른 특성은 한의사마다 처방을 내리는 방법이 다르다는 것이다. 한의사는 처방을 내릴 때 병증만을 고려하는 것이 아니라 현재 몸의 상태와 환자의 체질을 고려하여 처방을 내리게 된다.

II. 이론적 배경

1. 추천시스템

1.1 내용기반(Content-based) 추천시스템

내용기반 추천시스템은 사용자가 구매한 내역이나 사용자가 입력한 정보를 분석하여 유사성을 분석하는 방식이다. 내용기반 추천시스템은 크게 두 가지로 나뉘는데 사용자 기반 추천과 아이템 기반 추천이다. 두 방식의 차이점은 사용자를 기반으로 유사성을 파악하는 방식과 사용자가 선택한 아이템을 중심으로 유사성을 파악하는 차이이다. 이러한 방식은 분석이 용이하여 영화, 음악 도서 등에서 다양하게 사용이 되고 있다.

1.2 협업필터링 추천시스템

협업필터링은 추천 시스템 중에서 우수한 성능을 나타낸다고 알려져 많은 분야에서 사용이 되고 있다. 대표적으로 아마존, 넥플렉스 등 추천시스템으로 유명한 대표적인 기업들에서 사용이 되고 있다. 협업필터링의 특징으로는 단순히 사용자의 이력을 기반으로 하는 것이 아니라 사용 이력이 없는 사용자에게도 추천을 해줄 수 있다는 것이 가장 큰 특징이다. 협업필터링에도 여러 가지 방법이 있지만 근본적인 개념은 아이템을 벡터공간에 벡터화시켜 클러스터링 등 데이터마이닝 알고리즘을 통해 예측을 수행하여 추천을 해주는 방식이다.

2. word2vec

NLP(Natural Language Processing)은 '컴퓨터가 인간이 사용하는 언어를 이해하고, 분석할 수 있게 하는 분야'를 총칭하는 말이다. 컴퓨터가 어떠한 단어를 이해하기 위해서는 단어를 수치적으로 표현할 수 있어야한다. 수치화를 통해 단어의 의미차이를 부여하기가 어려워 기존에는 'One-Hot Encoding' 방식을 사용했다. 이러한 방법도 단어의 의미 자체를 벡터화하여 이용하기에는 어려움이 존재했고 많은 연구자들이 의미를 다차원 공간에서 벡터로 표현하는 방법을 연구하기 시작했다.

word2vec은 2013년에 구글에서 발표한 연구로서 NN(Neural Net)기반의 학습 방법이다. word2vec은 두 가지 학습 모델을 제시하였다.

2.1 CBOW(Continuous Bag-of-Words)

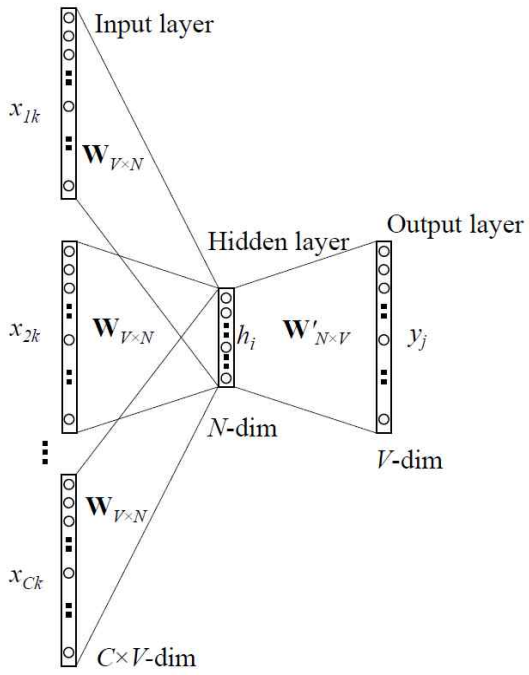

Fig 1. CBOW(Continuous Bag-of-Words)

CBOW(Continuous Bag-of-Words)의 기본적인 개념은 주변의 단어를 기준으로 단어를 예측하는 방식이다. 예를 들면 "나는 학교에 간다."라는 문장이 있다. CBOW 모델에 "나는 __ 간다." 라는 문장을 주면 "학교에"라고 예측을 하는 방식이다.

2.2 Skip-gram

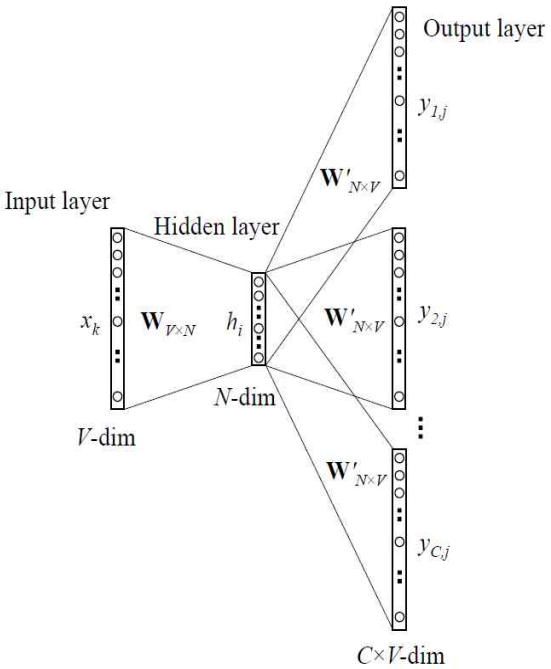

Fig 2. Skip-gram

Skip-gram의 기본적인 개념은 CBOW와 반대이다.

하나의 단어를 통해서 주변에 단어를 예측하는 방식이다. 예를 들면 "학교에"리는 단어를 주면 주변에 "나는", "간다"를 예측하는 방식이다. 본 연구에서는 Word2Vec의 Skip-gram 방식을 이용하여 약재를 추천하였다.

III. 약재추천 모형

3.1 데이터 수집

한약재 추천 연구에 사용된 데이터는 특허청의 재원으로 한국전통지식포탈 DB를 이용하여 데이터를 수집하였다. 한약재의 종류는 약 5000여개이고 사용된 처방은 약 2만개이다.

3.2 word2vec 알고리즘

word2vec의 Skip-gram 모델의 목적함수는

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^{\top} v_{w_I}\right)}$$

Log 확률 함수에 Softmax 함수를 적용한 함수를 사용한다.

학습은 방법은 역전파 학습방법을 사용한다. Skip-gram 모델을 통해 구한 값을 통해 오차율을 계산하여 최소화하는 가중치를 Skip-gram 모델에 업데이트 하여 학습을 진행한다. 여기서 최소화 방식은 SGD(Stochastic Gradient Descent)를 사용한다. SGD와 기존의 GD의 차이점은 모든 데이터의 가울기를 계산하는 것이 아닌 batch를 생성하여 가울기를 계산하여 업데이트를 한다는 점이다. 이러한 방식으로 학습한 약재는 fig 3과 같다.

Fig 3. 한약재 시각화

3.3 한약재 추천

일부 한약재를 선택하여 학습시킨 결과는 table 1.과 같다.

Table 1. 추천 결과

한약재	추천 약재
천궁(川芎)A [37.5 g]	천태오약(天台烏藥) [37.5 g] 하수오(何首烏)A [150 g] 천궁(川芎)A [56.25 g] 소목(蘇木)A [37.5 g] 홍두(紅豆)A [11.25 g] 초과(草果)A [75 g] 진피(陳皮)A [37.5 g] 선퇴(蟬退) [18.75 g]
오미자(五味子)A [37.5 g]	석곡(石斛)A [37.5 g] 원지(遠志)A [37.5 g] 육종용(肉蓯蓉)A [56.25 g] 오미자(五味子)A [18.75 g] 적복령(赤茯苓) [112.5 g] 쇄양(鎖陽)A [56.25 g] 오미자(五味子)A [75 g] 저령(猪苓) [37.5 g]
백작약(白芍藥)A [5.625 g]	백작약(白芍藥)A [7.5 g] 백작약(白芍藥)A [3.75 g] 백작약(白芍藥)A [9.375 g]
	복령(茯苓) [3 g] 백개자(白芥子)A [5.625 g] 박하(薄荷)A [1.875 g] 백개자(白芥子)A [2.625 g] 초오(草烏)A [5.625 g]

IV. Conclusions

word2vec 알고리즘을 이용하여 한약재 추천 시스템을 연구하였다. 현재는 한약재만을 이용하여 한약재를 추천하였지만 향후 연구에서 처방명, 증상까지 포함하여 추천 알고리즘을 연구할 생각이다.

Acknowledgment

본 연구는 미래창조과학부의 2016년 고용계약형 SW석사과정 지원 사업을 지원받아 수행한 결과입니다.

이 논문은 2016년도에 정부(특허청)의 재원으로한국전통지식포탈의 데이터를 제공받아 수행된 연구입니다.

References

[1] QV LE, and T Mikolov, "distributed representations of words and phrases and their compositionality," ICML, 2014

[2] X Rong, "word2vec parameter learning explained", arXiv preprint arXiv:1411.2738, 2014

[3] Gwi-Im Jung, ”Construction of Personalized Recommendation System Based on Back Propagation Neural Network” Journal of the Korean Institute of Industrial Engineers, 292-302 (11 pages), 2007

[4] Ji-enun Son, ”Review and Analysis of Recommender Systems” Korea-Press, 185-208 (24 pages), 2015