Journal of Knowledge Information Technology and Systems(JKITS), Vol. 15, No. 5, pp. 863-874, October 2020
https://doi.org/10.34163/jkits.2020.15.5.028

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700 (Print), ISSN 2734-0570 (Online)

http://www.kkits.or.kr

Word Embedding Based Clustering Method for Query Expansion of Korean Medicine Symptoms

Sangjun Yea[1], Sanghun Lee[2], Jiwon Yun[2], Ho Jang[1*]

[1]*Intellectual Information Team, Korea Institute of Oriental Medicine*
[2]*Future Medicine Division, Korea Institute of Oriental Medicine*

ABSTRACT

In the information age, smart search is essential due to the enormous amount of accumulated information. However most of the query entered into search engines are general nouns with various meanings and are only about 2.4 words on average, making it difficult for search engines to grasp the exact search intention of users. To mitigate this situation, figuring out exact search intention of users is supported by query expansion. In order to develop the query expansion of Korean Medicine clinical decision support system (KM-CDSS), we suggest novel algorithm that consist of 4 steps; first, symptom names are extracted from the prescription information fetched by the initial query. second, extracted symptoms are embedded into a vector space. third, vector space are clustered and each cluster's quality is evaluated by silhouette coefficient. finally, each words which are laid nearest to the center is suggested as representative symptom. In the experiments, we examined relevance and comprehensiveness qualitatively by KM doctors and utility are analyzed quantitatively. The results of the evaluation of the three indicators were analyzed in an integrated manner. The proposed model showed good enough results in relevance and comprehensiveness test and best result in utility test. It turned out that the proposed model is most suitable as a query expansion model for KM-CDSS. If user search logs are collected in the future, it is expected that related search words will be provided in more sophisticated ways through improved query expansion.

KEYWORDS: Ontology, symptom, CDSS, query expansion, M-tf model

ARTICLE INFO: Received 16 September 2020, Revised 24 September 2020, Accepted 13 October 2020.

[*]Corresponding author is with the Intellectual Information Team, Korea Institute of Oriental Medicine, Korea Institute of Oriental Medicine, 1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054, Korea.
E-mail address: jh@kiom.re.kr

1. 서론

정보화 시대에는 막대하게 축적된 정보량으로 인해서 스마트한 정보 검색이 필수적이다. 이를 위해서 오랫동안 정보 검색 분야에서 다양한 연구가 수행되어 사용자의 편리한 검색을 돕고 있다[1]. 그러나 사용자가 자신의 검색 의도를 명확하게 인지하고 있으며 적합한 키워드를 사용하여 검색 엔진에 입력할 수 있을 때, 스마트한 정보 검색이 이루어지나 실상은 그렇지 않다[2, 3]. 인터넷 검색 엔진에 입력되는 키워드는 대부분 다양한 의미를 지닌 일반 명사이며 평균적으로 2.4 단어 정도에 불과하여 사용자의 정확한 검색 의도를 검색 엔진이 파악하기 힘들다. 이러한 사용자 검색 패턴을 보완하기 위해서 검색 엔진에서는 '연관 검색어'를 제시하여 사용자의 정확한 검색을 돕고 있다[4].

정보 검색 분야에서는 이와 관련하여 질의 확장(Query Expansion)에 관한 연구가 수행되고 있으며, 질의 확장은 유의미한 질의를 추가하여 최초 질의의 모호함을 제거함으로써, 사용자의 의도가 정확하게 반영된 검색 결과를 제공하는 것이 목적이다. 질의 확장 연구는 확장된 질의를 생성하는 방법, 정보 리소스, 응용 분야에 따라서 다양하게 구분된다. 최근에는 사용자의 검색 기록을 분석하여 연관 검색어를 제공하는 방법과 외부 리소스 또는 최초 검색 결과를 분석하여 질의를 확장하는 방법에 관해서 많이 연구되고 있다[3, 4]. 사용자의 검색 기록을 이용하는 접근 방법은 콜드 스타트 이슈로 신규 정보 시스템에 적용할 수 없는 단점이 있다. 외부 리소스 또는 최초 검색 결과를 이용하는 질의 확장에는 키워드 코퍼스를 벡터 공간으로 변형하거나 네트워크로 변형한 이후에, 다양한 클러스터링 기법을 적용하여 대표 키워드를 생성해서 질의를 확장한다[5].

한국한의학연구원에서는 2009년 ~ 2014년에 걸쳐서 전국 한의과대학에서 교재로 사용되는 텍스트에서 약재, 병증, 처방, 침구, 증상 등의 정보를 추출하여 한의학 온톨로지를 구축하였다[6, 7]. 또한 기 구축된 온톨로지를 활용하여 한의사가 사용할 수 있는 진료 지원 시스템 (Clinical Decision Support System, CDSS)을 개발하고 한방병원 및 한의원에서 사용하고 있는 전자차트와 결합하기 위해서 작업이 진행 중이다. 한의사의 처방 선정을 지원하기 위해서, 한의 온톨로지에서 관련 증상 또는 질병을 효과적으로 치료할 수 있는 처방 검색 기능이 진료 지원 시스템에 구현되어 있다. 한의 온톨로지에는 18,000개의 처방명 및 관련 정보가 구축되어 있어, 한의사의 의도가 정확하게 반영된 검색 결과를 제공하는 것이 진료 지원 시스템이 한의 진료 현장에 성공적으로 보급되기 위한 필수요소이다.

이에 본 논문에서는 한의학 온톨로지 기반의 처방지원 시스템에서 증상 질의의 모호함을 제거하여 한의사의 검색 의도가 정확하게 반영된 검색을 지원하기 위해서 질의 확장을 위한 알고리즘을 제안하고 사용자 평가를 수행하였다. 제안된 알고리즘은 최초 질의로부터 생성된 처방 정보에서 증상명을 추출하여 벡터공간으로 임베딩 및 클러스터링하고 각 클러스터의 품질을 평가하여 높은 순위의 증상명부터 확장 질의로 추천하는 방법이다. 제안된 알고리즘을 평가하기 위해서 한의사에 의한 정성적 평가와 검색 건수에 대한 정량적 평가 수행하였으며, 실험을 통해서 제안된 알고리즘이 우수한 질의 확장이 가능한 것으로 평가되었다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 한의학 증상 질의 확장 방법에 대해서 구체적으로 기술하고, 3장에서는 제안된 방법을 평가하기 위해서 정성적 및 정량적 지표를 설계하고 실험하였다. 4장에서는 실험 결과를 통합적으로 분석하고 제안된 방법이 동작하는 사례를 제시하고 5장에는 결론을 기술한다.

2. 한의학 증상 질의 확장 방법

2.1 데이터 셋

한의학 온톨로지는 전국 한의과대학에서 교재로 사용되는 텍스트에서 약재, 병증, 처방, 침구, 증상 등의 정보를 추출하여 구축되었다. 처방이란 한의학적 원리에 의해서 2종 이상의 한약재를 가공하여 조재한 약물로써, 특정 증상을 완화하거나 질병을 치료하는 효능을 가지며, 치료대상이 되는 증상 및 질병은 주치로써 표기된다. 한의학 온톨로지에는 처방명이 약 18,000개 및 증상명이 약 1만개 존재[6]하며, 한의 진료 지원 시스템에서 처방정보는 <그림 1>과 같이 처방명, 출전, 주치, 효능, 구성약재 등의 정보가 도시된다. 본 연구에서는 18,000개의 처방 중에서 주치 정보가 함께 구축된 처방을 대상으로 하였으며 구체적인 처방 및 증상 개수는 <표 1>과 같다.

그림 1. 한의 진료 지원 시스템에서의 처방정보 샘플
Figure 1. Prescription sample in KM-CDSS

표 1. 논문에서 사용된 온톨로지 데이터 셋
Table 1. Ontology dataset in the paper

구분		개수	구분		개수
처방	중복 포함	5,181	증상	중복 포함	10,917
	중복 제외	2,486		중복 제외	4,546

2.2 단어 임베딩 기반의 질의 확장 방법

본 논문에서 제안하는 한의학 증상명 질의 확장 방법은 <그림 2>와 같다. 첫 번째, 사용자가 입력한 증상으로 검색된 처방에서 주치 정보를 분석하여 증상에 해당하는 단어를 추출한다. 두 번째, 추출된 증상을 FastText를 활용하여 벡터공간으로 임베딩한다. 세 번째, 벡터 공간에 임베딩된 증상을 K-Means 기반으로 클러스터링하고 각 클러스터의 품질을 평가하여 랭크를 계산한다. 네 번째, 높은 랭크의 클러스터부터 중심에서 가장 가까운 증상명을 확장 질의로 추천한다.

2.2.1 주치 분석 및 증상 임베딩

질의 확장을 위해서 외부 리소스를 활용하는 연구와 검색 결과 문서를 활용하는 다양한 연구들이 수행되었다. 검색 결과 문서를 활용하는 잠정적 적합 피드백(Pseudo Relevance Feedback, PRF) 방법은 정보 검색 분야의 많은 연구에서 효과적인 것으로 파악되었다[8, 9]. 그리고 검색 결과 문서의 활용에는 전역적인 정보와 국지적 정보를 사용할 수 있으며, Vechtomova et al (2003)에 의하면 전역적인 정보는 일반화 경향이 두드러져 질의 확장에는 국지적 정보 활용이 더욱 적합하다고 밝혔다[10]. 한의학 용어는 증상, 질병, 진단법의 구분이 명확하지 않으며, 기 구축된 한의학 온톨로지 처방의 주치도 세 가지 정보가 함께 표기되어 있다. 이에 본 연구에서는 사용자가 입력한 증상으로 검색된 처방에서 주치 정보를 분석하여 증상명으로만 사용되는 용어를 추출하여 임베딩 하였다.

자연어 처리에서 임베딩은 단어를 계산 가능한 벡터 공간으로 맵핑하는 것으로 단어의 통계적 패턴 정보를 벡터 공간으로 투영하는 것이다. 단어 임베딩을 만드는 세 가지 주요 방법은 <표 2>와 같다[11].

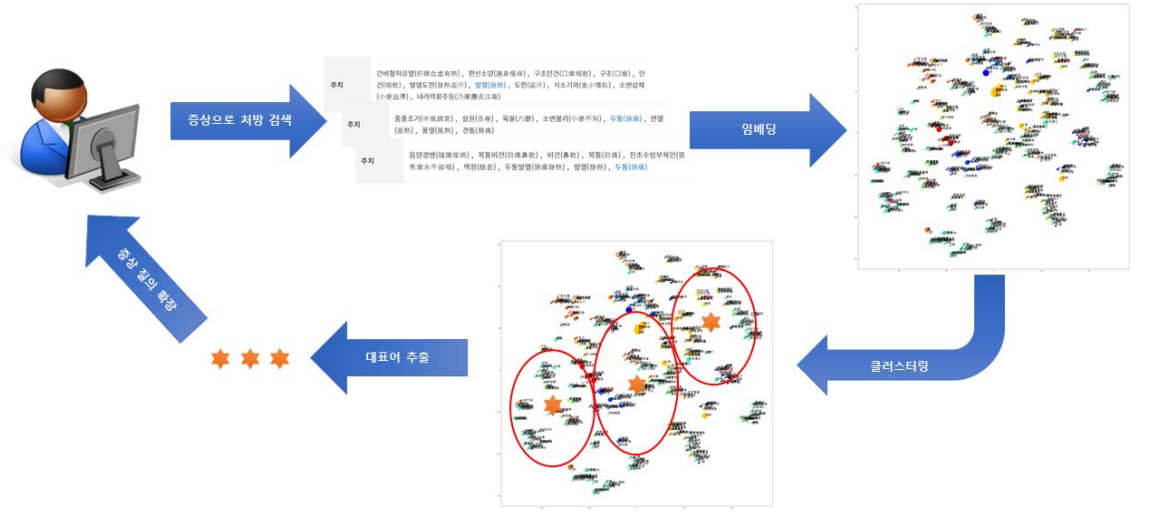

그림 2. 질의 확장을 위해 제안된 기법
Figure 2. The proposed query expansion method

본 연구에서는 윈도우내에서 동시에 등장하는 단어 쌍의 패턴을 점별 상호 정보량(Pointwise Mutual Information, PMI)으로 계량화하는 분포 모델 중에서 FastText를 채택하였다.

표 2. 임베딩을 만드는 세 가지 철학, Lee (2013)에서 인용[11]
Table 2. Three principles of embedding, quoted from Lee (2013)[11]

	백오브워즈 모델	언어 모델	분포 모델
가정	어떤 단어가 많이 쓰였는가?	단어가 어떤 순서로 쓰였는가?	어떤 단어들이 같이 쓰였는가?
대표	Data Clouds, DAN	ELMo, GPT	Word2Vec, FastText

FastText는 페이스북에서 2017년에 공개한 임베딩 기법으로, 널리 활용되고 있는 구글의 Word2Vec을 n-gram이 가능하도록 개선한 모델이다. 다른 임베딩 기법은 미등록 단어의 벡터를 추정할 수 없는 단점이 있으나, FastText는 오타나 미등록 단어도 강건하게 임베딩할 수 있는 장점이 있다. 수식1은 FastText가 최대화하는 로그우드 함수로써 포지티브 샘플 (tp, cp)과 네거티브 샘플 (tn, cn)을 학습해서 모델을 갱신한다[12].

$$L(\theta) = \log P(+|t_p, c_p) + \sum_{i=1}^{k} \log P(-|t_{n_i}, c_{n_i}) \quad (1)$$

2.2.2 증상 클러스터링 및 추천

고전적인 K-means는 여전히 다양한 현실세계 클러스터링에서 효과적으로 활용되고 있으며[13, 14], 수식2의 목적 함수를 최소화하는 클러스터 중심 C를 찾는 기법이다. 그러나 K-means는 고차원이나 빅데이터의 경우에는 계산 속도가 느려지는 단점이 있어, 이를 해결하기 위해서 미니 배치 기법을 이용하여 클러스터링 품질은 유지하면서 속도를 개선한 알고리즘이 제안되었다. mini-batch K-means 기법[15]에서는 학습 데이터 셋에서 임의로 추출된 미니 배치에 대해서 수식2의 K-means와 동일한 목적 함수를 최소화하고, 이러한 과정을 여러 번 반복하여 최선의 결과를 도출하게 된다. 본 연구에서는 고차원의 벡터공간에 존재하는 증상을 클러스터링하기 때문에, mini-batch K-means 기법

을 적용하여 계산량을 최소화하고 실시간 응답성을 확보하였다.

$$L(x) = \min \sum_{x INX} ||f(C,x) - x||^2 \quad (2)$$

클러스터링으로 생성된 증상명 군집의 품질을 평가하기 위해서 실루엣 지표(Silhouette Index)를 사용하였다. 실루엣 지표는 군집에 속한 모든 개체에 대해서 같은 군집에 속하는 개체와의 평균 거리와 가장 가까운 이웃 군집에 속하는 개체와의 평균 거리를 비교하는 지표로써, -1에서 1사이의 값을 가진다[16]. 실루엣 지표 값이 1에 가까울수록, 해당 클러스터가 동일한 특성을 공유하고 있다는 것을 의미하게 된다. 본 연구에서는 mini-batch K-means로 생성된 클러스터의 실루엣 지표를 계산하여 질의 추천의 우선순위로 사용하였다. <수식 2>에서 클러스터 중심 C는 해당 군집을 대표하므로, C에 가장 가까운 증상을 질의 확장을 위해서 사용하였다. 최종의 질의 확장을 위한 증상 추천은 실루엣 지표에 의해서 내림차순으로 제시되었다.

3. 한의학 증상 질의 확장의 실험 설계 및 결과

3.1 실험 설계

본 논문에서 제안된 방법의 성능을 평가하기 위해서, 2018 한국한의학 연감[17]의 다빈도 상병 급여 현황에서 기능성 소화불량, 감기, 복부 및 골반통증, 두통, 비염, 중풍 질병을 기준으로 하여 해당 질환과 관련 있는 증상을 선택하여 <표 3>의 질의 집합을 구성하였다. 검색 시스템에서 연관 검색어를 제시할 경우, 일반적으로 5개 ~ 10개의 키워드를 추천하는 것을 반영하여, 실험에서는 각 질의 증상에 대해서 모델별로 10개의 질의 확장을 가정하고 실험하였다.

표 3. 실험에 사용된 질의 집합
Table 3. Query set used in the experiments

순서	질의	
1	咳嗽	해수
2	頭痛	두통
3	發熱	발열
4	喘	천
5	腹痛	복통
6	虛勞	허로
7	小便不利	소변불리
8	惡寒	오한
9	心悸	심계
10	泄瀉	설사

Data Clouds 모델은 검색된 문서를 키워드 형태로 요약하는 모델로써, 질의 확장에 효과적으로 활용되고 있어[18], 본 논문에서 제안된 방법을 평가하기 위한 비교 모델로 선정하였다. 그리고 제안된 단어 임베딩 기반 질의 확장은 증상을 벡터 공간으로 임베딩할 때, 증상명의 출현 빈도를 반영하지 않는 모델(M-wo)과 증상명의 출현 빈도를 반영하는 모델(M-tf)의 두 가지 형태로 평가하였다.

앞에서 언급되었듯이, 질의 확장의 목적이 한의 진료지원 시스템에서 한의사가 환자의 증상에 효과적인 처방 검색 지원이므로, 질의 확장 평가에서 많이 사용되고 있는 지표를 목적에 적합하도록 수정하였다[3, 19]. 첫째, 사용자가 인지하지 못하는 관련 증상을 제시하여 검색의 정밀도를 향상할 수 있으므로 연관성(Relevance) 지표를 평가하였다. 둘째, 사용자에게 다양한 한의학적 증상을 추천하여 검색의 재현성을 확장할 수 있으므로 포괄성(Comprehensiveness) 지표를 평가하였다. 셋째, 한의사에게 제시되는 검색 결과가 각 처방의 상세 정보를 자세히 검토할 수 있을 정도의 건수인지를 유용성(Utility) 지표를 통해서 평가하였다. 연관성

과 포괄성은 한의사 6명이 정성적 평가를 통해서 1점(미흡) ~ 3점(우수)을 세 가지 평가모델에 대해서 상대적으로 부여하였다. 유용성 지표는 질의확장 전후의 검색결과 건수를 비교하여 모델에서 확장 질의로 제시되는 증상 용어를 정량적으로 평가하였다.

3.2 실험 결과

실험에 사용된 10개의 질의에 대해서 세 가지 모델에서 도출된 질의 확장 결과는 별첨 <표 S1>과 같다. 추천된 10개의 증상에 대해서 상위에 있는 증상이 우선순위가 높다. 예를 들어서 질의어 咳嗽(해수)에 대해서 Data Clouds 모델은 鼻塞(비색)을, M-wo 모델은 手足心熱(수족심열)을, M-tf 모델은 喘息(천식)을 최우선으로 추천하였다. 각 모델별로 상이한 증상으로 질의 확장하는 경우가 많지만, 질의어 喘(천)에 대해서 上氣(상기)는 모든 모델에서 공통으로 나타나며 Data Clouds 모델은 2번째, M-wo 모델은 4번째, M-tf 모델은 3번째로 추천하였다.

"입력된 증상용어와 얼마나 관련 있는 증상용어를 추천하는가?" 에 대한 연관성 평가 결과는 <그림 3>과 <표 4>에서 확인할 수 있다.

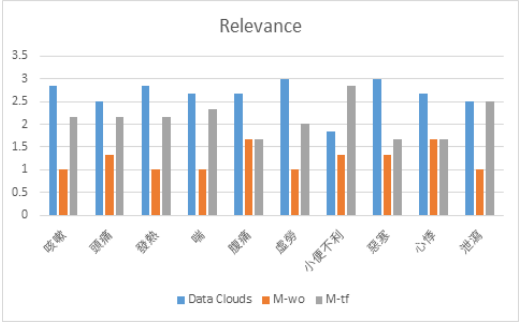

그림 3. 연관성 실험 결과
Figure 3. Relevance experiment results

실험에 참가한 한의사는 제시된 10개의 질의에 대해서 Data Clouds 모델이 가장 연관성 있는 질의 확장 결과를 도출했다고 평가하였으며, 그 다음으로 M-tf 및 M-wo 순이었다. 구체적으로 Data Clouds 모델은 3점 만점에 2.65점을 획득하였으며 총 60개(10 질의어 * 6명) 중에서 3점이 41개였다. M-tf 모델은 2.12점을 획득하였고 16개의 3점과 35개의 2점을 획득하였다.

표 4. 연관성 실험 결과
Table 4. Relevance experiment results

	Data Clouds	M-wo	M-tf
평균 점수	2.65	1.23	2.12
3점 개수	41	3	16
2점 개수	17	8	35
1점 개수	2	49	9

"추천된 증상용어가 얼마나 다양한 한의학적 증상을 포함하는가?" 에 대한 포괄성 평가 결과는 <그림 4>와 <표 5>에서 확인할 수 있다. M-wo 모델이 가장 포괄적인 질의 확장 결과를 도출했다고 평가하였으며, 그 다음으로 M-tf 및 Data Clouds 순이었다. 구체적으로 M-wo 모델은 3점 만점에 2.68점을 획득하였으며 총 60개 중에서 3점이 46개였다. M-tf 모델은 1.92점을 획득하였고 13개의 3점과 29개의 2점을 획득하였다.

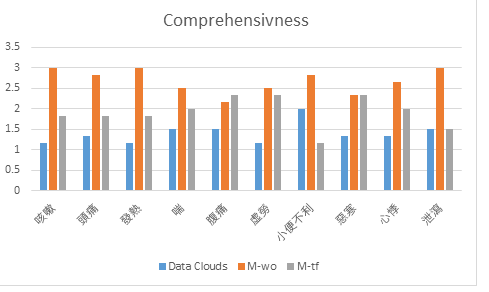

그림 4. 포괄성 실험 결과
Figure 4. Comprehensiveness experiment results

표 5. 포괄성 실험 결과
Table 5. Comprehensiveness experiment results

	Data Clouds	M-wo	M-tf
평균 점수	1.40	2.68	1.92
3점 개수	1	46	13
2점 개수	22	9	29
1점 개수	37	5	18

"입력 증상과 첫 번째 추천 증상 용어를 사용(AND)해서 검색된 처방이 얼마나 유용한가?"에 대한 유용성 평가 결과는 <표 6>에서 확인할 수 있다. 질의 확장 전에는 평균적으로 67.5개의 처방이 제시되어 한의사가 각각의 처방을 확인하기에는 불가능하다는 것을 알 수 있다. M-tf 모델에서 추천하는 첫 번째 확장 질의를 포함해서 검색할 경우, 평균 9.1개의 처방이 검색되어 한의사가 각각의 처방의 상세 정보를 확인하고 환자의 증상에 효과적인 처방을 선택할 수 있을 것으로 파악되었다. Data Clouds 모델은 14.7개, M-wo 모델은 1.6개의 처방이 검색되어 한의사가 구체적인 처방 정보를 확인하고 활용하기에 많거나 너무 적은 검색 결과를 보여주었다.

표 6. 질의 확장 전/후의 검색결과 건수 비교
Table 6. Comparison fetched results between before and after query expansion

	질의 확장 전	질의 확장 후		
		Data Clouds	M-wo	M-tf
咳嗽	143	16	2	3
頭痛	100	27	3	6
發熱	80	20	1	27
喘	74	15	1	6
腹痛	66	9	1	5
虛勞	54	5	1	5
小便不利	48	2	1	2
惡寒	43	20	4	5
心悸	36	24	1	23
泄瀉	31	9	1	9
평균	**67.5**	**14.7**	**1.6**	**9.1**

4. 논의

앞에서 밝혔듯이 한의 진료 지원 시스템에서 한의사에게 유용한 연관 검색어를 제안하는 질의 확장 모델을 평가하기 위해서, 실험에서는 연관성, 포괄성, 유용성의 세 가지 지표를 평가하였다. 모든 지표에서 가장 좋은 결과를 보여준 모델은 없었다. 이는 연관성과 포괄성은 모델의 상반되는 측면을 평가하는 지표이기 때문이다. 세 가지 지표에 대한 평가 결과를 통합적으로 분석하여, M-tf 모델이 한의 진료 지원 시스템의 연관 검색어 제시를 위한 질의 확장모델로 가장 적합한 것으로 드러났다. M-tf 모델은 연관성 및 포괄성 모두에서 적합한 평가 결과를 보여주었으며 유용성에서는 가장 좋은 결과를 보여주었다. 반면, Data Clouds 모델은 연관성에서는 가장 좋은 결과를 보여주었으나 포괄성에서는 미흡한 것으로 나타났다. 또한 M-wo 모델은 연관성에서는 미흡한 결과를 보여주었으나 포괄성에서는 가장 우수한 결과를 보여주었다. Data Clouds와 M-wo 모델은 특정 부문에서는 장점이 있으나 다른 부문에서는 미흡한 점이 드러나, 연관 검색어 추천을 위한 질의 확장 모델로 부족한 것으로 판명되었다.

본 논문에서 제안한 M-tf 모델의 클러스터링 및 추천 증상명 생성 결과를 외감질환인 咳嗽(해수)와 내상질환인 腹痛(복통)에 대해서 분석해보았다. 먼저 해수는 기침을 심하게 하는 증상으로 한의 온톨로지에서 중복을 포함하여 332개의 증상이 검색되었고, 해당 증상을 50차원 벡터공간에 임베딩하고 클러스터링하였다. 50차원의 벡터공간은 t-SNE를 사용하여 <그림 5>와 같이 도시되었고, 하단의 그림에서 점의 색깔은 클러스터 구분, 점의 크기는 증상의 중복, 붉은색 텍스트는 클러스터 중심에서 가장 가까운 추천 증상이다. 10개의 클러스터 중에서 특징적인 몇 가지 살펴보면 다음과 같다. ① 클

러스터는 潮熱(조열)만으로 구성되어 있으며 해수-객혈의 陰虛(음허)변증과 연관성이 높은 유사 증상이어서 추출된 것으로 보인다. ② 클러스터는 喘(천)과 上氣(상기)로 구성되어 있고 기침이 천식의 숨이 차는 증상이고 기가 위로 치미는 증상과 유사하여 추출된 것으로 보인다. ③ 클러스터는 18개의 증상으로 구성되어 있고 중심에는 咯血(객혈)이 있으며, 陰虛火動(음허화동)으로 인한 해수로 발현되는 증상들이다. 동의보감의 잡병편에 의하면 해수는 발열, 각혈, 오후에서 밤까지 열이 나는 것, 얼굴과 입술이 붉은 것, 소변이 붉고 잘 나오지 않는 증상들이 있는 것으로 기술되어 있다. ④ 클러스터는 13개의 증상으로 구성되어 있고 중심에는 鼻塞(비색)이 있다. 특이할 점은 코막힘과 쉰 목소리, 목의 통증이 같이 발현되는 비율이 높은 것인데 이는 바이러스로 인한 상부 기도의 감염으로 인한 증상과 유사점이 있다. 감염에 의해 상기도인 후두, 인두, 부비동 등의 염증으로 인해 부어올라 정상적인 호흡이 곤란해지며 쉰 목소리와 기침 등을 보이는 증상이다. 변증으로 보면 해당 증상은

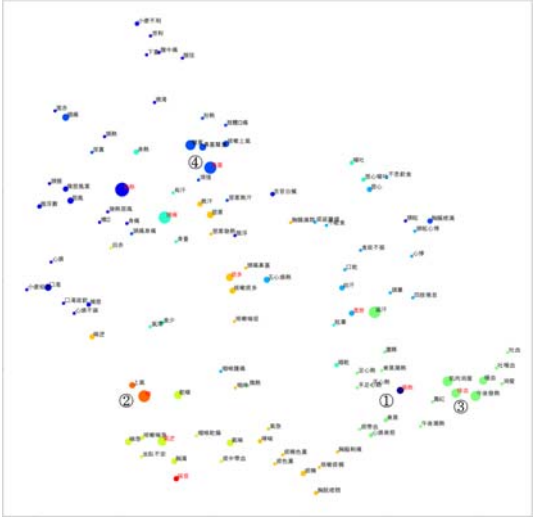

그림 5. 咳嗽(해수)에 대한 임베딩 및 클러스터링 결과
Figure 5. The embedding and clustering of symptom cough

濕(습), 風(풍), 寒(한)으로 인한 경우가 많으며 그로 인한 감기 증상도 같이 동반되어 그룹화된 것으로 판단된다.

그림 6. 腹痛(복통)에 대한 임베딩 및 클러스터링 결과
Figure 6. The embedding and clustering of symptom colic

복통은 한의 온톨로지에서 중복을 포함하여 188개의 증상이 검색되었고, 해당 증상을 50차원 벡터 공간에 임베딩하고 클러스터링한 결과가 〈그림 6〉과 같이 도시되었다. 10개의 클러스터 중에서 특징적인 몇 가지 살펴보면 다음과 같다. ① 클러스터는 4개의 증상으로 구성되어 있고 중심에는 不思飮食(불사음식)이 있으며, 비위가 허약하여 생긴 복통으로 볼 수 있다. 이와 연관 증상으로는 팔다리가 싸늘하며 힘이 없고 권태로우며 설사하는 것을 들 수가 있다. 동의보감 내경편에 의하면 설사의 종류도 허설(虛泄)에 속하는 오경설사(五更泄瀉)에 속한다고 분석된다. ② 클러스터는 8개의 증상으로 구성되어 있고 중심에는 腸鳴(장명)이 있다. 장명은 음기가 넘치는 상태, 脾虛(비허) 또는 대장에 한기가 있는 경우 나타난다. 따라서 허증, 한증에 속하는 증상인 脫肛(탈항), 喜按(희안), 便溏(변당), 四肢

不溫(사지불온)의 증상과 동반된다고 볼 수 있다. ③ 클러스터는 13개의 증상으로 구성되어 있고 중심에는 裏急(이급)이 있다. 복통의 표현 중 뱃속이 당기는 증상인 이급은 뒤가 묵직한 증상과 같이 발현하는 경우가 많다. 원인이 화기에 속하는 경우는 渴欲飮水(갈욕음수), 手足煩熱(수족번열), 口臭(구취), 小便短赤(소변단적)의 열증 증상과 같이 출현하게 된다.

5. 결 론

한국한의학연구원에서는 전국 한의과대학에서 교재로 사용되는 텍스트에서 약재, 병증, 처방, 침구, 증상 등의 정보를 추출하여 한의학 온톨로지를 구축하였으며, 이를 활용하여 한의 진료 지원 시스템을 구축하여 임상현장 보급하기 위한 작업이 진행 중이다. 약 18,000개의 처방명 중에서 한의사의 검색 의도가 정확하게 반영된 검색 결과를 제공하기 위해서, 질의 확장을 통한 연관 검색어 제공을 위한 새로운 알고리즘을 제안한다. 제안된 알고리즘은 최초 질의로부터 생성된 처방 정보에서 증상명을 추출하여 벡터공간으로 임베딩 및 클러스터링하고 각 클러스터의 품질을 평가하여 높은 순위의 증상명부터 확장 질의로 추천하는 방법이다. 본 논문에서 제안된 방법의 성능을 평가하기 위해서, 2018 한국한의학 연감에서 10개의 증상을 선택하고 한의사에 의한 연관성과 포괄성에 대한 정성 평가 및 검색 결과 유용성에 대한 정량 평가를 실시하였다. 실험 결과에서 연관성 지표에서는 Data Clouds, M-tf, M-wo의 순서를 나타내었고 포괄성 지표에서는 M-wo, M-tf, Data Clouds의 순서를 나타내었다. 유용성 지표 평가에서는 M-tf가 가장 좋은 결과를 보였다. 세 가지 지표에 대한 평가 결과를 통합적으로 분석하여, M-tf 모델이 한의 진료 지원 시스템의 연관 검색어 제시를 위한 질의 확장모델로 가장 적합한 것으로 드러났다. 그러나 본 연구에서 사용한 한의학 온톨로지는 증상에 대한 계층구조가 없어 모든 레벨의 증상어가 동시에 임베딩 되는 제한점과 클러스터 개수를 사용자가 선정해야 하는 단점이 있다. 향후, 이러한 제한점이 개선되고 한의 진단 지원 시스템이 임상 현장에서 활용되어 사용자 검색 기록이 수집된다면, 더욱 개선된 질의 확장을 통한 연관 검색어 제공이 가능해질 것으로 기대된다.

References

[1] P. Suthira, and J. Q. Gan, *A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search,* Artificial Intelligence Research Vol. 4, No. 2, pp. 119-125, 2015.

[2] C. Fei, and M. D. Rijke, *Query auto completion in information retrieval*. Universiteit van Amsterdam, Host, 2016.

[3] C-U. Kwak, H-G. Yoon, and S-B. Park, *Query expansion based on word sense community,* Journal of KIISE, Vol. 41, No. 12, pp. 1058-1065, 2014.

[4] A. H. Kumar, and A. Deepak, *Query expansion techniques for information retrieval: a survey,* Information Processing & Management Vol. 56, No. 5, pp. 1698-1735, 2019.

[5] O. Jessie, H. Qin, X. Ma, and S. C. Liew, *A survey of query expansion, query suggestion and query refinement techniques,* 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS). IEEE, 2015.

[6] H-C. Jang, J-H. Kim, S- K. Kim, C. Kim, S-H. Bae, A-N. Kim, D-M. Eom, and M-Y.

Song, *Ontology for medicinal materials based on traditional Korean medicine*, Bioinformatics, Vol. 26, No. 18, pp. 2359-2360, 2010

[7] H-C. Jang, Y-T. Oh, A-N. Kim, and S-K. Kim, *User guiding information supporting application for clinical procedure in traditional medicine.* HIMI/HCII, pp 100-109, 2013.

[8] Y. Xu, G. J. F. Jones, and B. Wang, *Query dependent pseudo-relevance feedback based on wikipedia*, Proc. of the 32nd SIGIR, pp. 59-66, 2009.

[9] Z. Liu, S. Natarajan, and Y. Chen, *Query expansion based on clustered results,* Proc. of the 37th VLDB, Vol. 4, No. 6, pp. 350-361, 2011.

[10] O. Vechtomova, S. Robertson, and S. Jones, *Query expansion with long-span collocates*, Information Retrieval, Vol. 6, No. 2, pp. 251-273, 2003.

[11] K-C. Lee, *Sentence embeddings using korean corpus*, acornPub, 2019

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, *Enriching word vectors with subword information,* Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135-146, 2017.

[13] S-J. Park, and J-H. Kim, *Paragraph-based k-means clustering by using meaning-based paragraph division,* Journal of Knowledge Information Technology and System, Vol. 12, No. 1, pp. 157-164, 2017.

[14] H-S. Jang, *Data-mining based anomaly detection in document management system*, Journal of Knowledge Information Technology and System, Vol. 10, No. 4, pp. 465-473, 2015.

[15] S. David, *Web-scale k-means clustering,* Proceedings of the 19th international conference on World wide web. 2010.

[16] S. Artur, and A. Krzyżak, *Performance evaluation of the silhouette index,* International Conference on Artificial Intelligence and Soft Computing. Springer, Cham, 2015.

[17] Publication Committee of Yearbook of Traditional Korean Medicine, 2018 Yearbook of Traditional Korean Medicine, 2020.

[18] G. Koutrika, Z. M. Zadeh, and H. Garcia-Molina, *Data clouds: summarizing keyword search results over structured data,* Proc. of EDBT, pp. 391-402, 2009.

[19] M. Zhongrui, Y. Chen, R. Song, T. Sakai, L. Jiaheng, and J. R. Wen, *New assessment criteria for query suggestion,* Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 1109-1110, 2012.

한의학 증상 질의 확장을 위한 단어 임베딩 기반의 클러스터링 기법

예상준[1], 이상훈[2], 윤지원[3], 장호[4]

[1] *한국한의학연구원 지능화추진팀 책임연구원*
[2] *한국한의학연구원 미래의학부 책임연구원*
[3] *한국한의학연구원 미래의학부 선임연구원*
[4] *한국한의학연구원 지능화추진팀 선임연구원*

요 약

정보화 시대에는 막대하게 축적된 정보량으로 인해서 스마트한 정보 검색이 필수적이다. 이를 지원하기 위해서 질의 확장을 통해서 사용자의 정확한 검색을 지원하고 있다. 본 연구에서는 한의 진료 지원 시스템의 질의 확장을 위해서 최초 질의로부터 생성된 처방 정보

에서 증상명을 추출하여 벡터공간으로 임베딩 및 클러스터링하고 각 클러스터의 품질을 평가하여 높은 순위의 증상명부터 확장 질의로 추천하는 방법을 제안하고 한의사에 의한 정성평가 및 검색 결에 대한 정량평가를 실시하였다. 실험에서는 연관성, 포괄성, 유용성의 세 가지 지표를 평가하였으며, 세 가지 지표에 대한 평가 결과를 통합적으로 분석하여, 제안된 모델이 한의 진료 지원 시스템의 연관 검색어 제시를 위한 질의 확장모델로 가장 적합한 것으로 드러났다. 향후 사용자 검색 기록이 수집된다면, 더욱 개선된 질의 확장을 통한 연관 검색어 제공이 가능해질 것으로 기대된다.

감사의 글

본 연구는 한국한의학연구원 'AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN2012110)' 과제의 지원을 받아 수행되었습니다. 그리고 알고리즘 평가에 도움을 주신 한의사에게 감사드립니다.

Sangjun Yea received the Ph.D. in knowledge service engineering from KAIST, Korea in 2018. Since 2008, he is a principal researcher at Korea Institute of Oriental Medicine. The current research interests is biomedical data science.
E-mail address: tomita@kiom.re.kr

Sanghun Lee He received his M.S. degree and his Ph.D. degree from Department of Krean Medicine at Wonkwang University in 2007 and 2011, respectively. After graduation, he has been researched on Modernization and Standardization of Korean medicine devices and informatics in Korea Institute of Oriental Medicine.
E-mail address: ezhani@kiom.re.kr

Jiwon Yoon received the Ph. D. in Korean Medicine from Kyung Hee University, Korea in 2013. Since 2016, she is a senior researcher at Korea Institute of Oriental Medicine. The current research interests is future medicine and prescriptionology.
E-mail address: yoonjw@kiom.re.kr

Ho Jang received his Ph.D. degree from School of Electrical Engineering and Computer Science at Gwangju Institute of Science and Technology in 2017. He has been working for Korea Institute of Oriental Medicine since 2018. His research interests include computational biology and machine learning.
E-mail address: jh@kiom.re.kr

별첨

표 S1. 실험에 사용한 질의확장 결과
Table S1. Query expansion results in the experiment

입력 증상	질의 확장		
	Data Clouds	M-wo	M-tf
咳嗽	鼻塞	手足心熱	喘息
	聲重	腹中痛	潮熱
	喘急	發熱惡風	喘
	發熱	惡心	頭痛
	頭痛	氣逆	咯血
	氣喘	午後潮熱	鼻塞
	咽痛	哮喘	發熱

	盜汗	無汗	氣逆
	潮熱	盜汗	痰多
	咯血	虛勞	虛勞
頭痛	發熱	頭痛身疼	頭痛身疼
	身熱	手足厥冷	惡寒
	惡寒	眼昏	咳嗽
	無汗	氣逆	發熱
	身痛	頭痛目眩	身熱
	咳嗽	子宮下垂	惡風
	口渴	言語低微	惡寒無汗
	小便不利	神昏	項强
	惡風	聲重	口渴
	鼻塞	無汗	嘔吐
發熱	惡寒	心煩不寐	頭痛
	頭痛	鼻塞流涕	乾嘔
	惡寒發熱	腹脹泄瀉	咳嗽
	口渴	噴嚏	惡寒
	咳嗽	少腹急結	咽乾
	無汗	虛勞	盜汗
	腹痛	煩渴引飮	惡寒發熱
	小便不利	泛惡	神昏
	發熱惡寒	咳嗽	腹痛
	胸悶	口渴	口渴
喘	咳嗽	心下硬	喘促
	上氣	乾咳無痰	喘急
	喘急	咽喉不利	上氣
	喘促	上氣	上氣喘促
	上氣喘急	癲狂	頭痛
	上氣喘促	咳嗽喘急	咳嗽
	汗出	鼻塞	汗出
	氣逆	不思飮食	口苦口乾
	頭痛	嗽	氣逆
	鼻塞	口渴	四肢厥逆
腹痛	泄瀉	氣脹	不思飮食
	腸鳴	裏急	不渴
	不思飮食	手足厥冷	腸鳴
	不渴	耳聾	裏急
	身熱	身反不惡寒	惡寒
	手足厥冷	五更泄瀉	泄瀉
	裏急	四肢厥冷	口燥
	發熱	泛惡	發熱
	嘔吐	月經不調	身熱
	手足厥逆	口燥	神疲
虛勞	自汗	吐唾血	自汗
	少氣	咽乾	少氣
	咳嗽	腹中時痛	咳嗽

	盜汗	心動悸	腹中時痛
	潮熱	盜汗	骨蒸
	裏急	心煩	吐唾血
	遺精	腹痛	腰痛
	骨蒸	腰痛	乾咳
	乾咳	寒熱	盜汗
	倦怠	骨蒸潮熱	腹痛
小便不利	黃疸	心煩不寐	水腫
	頭痛	汗出惡風	乾嘔
	煩渴	腹中痛	腹滿
	發熱	身痛	腹中痛
	腹痛	肢體沈重	泄瀉
	泄瀉	惡寒	腹滿便秘
	水腫	寒熱往來	煩渴引飮
	腹脹	足冷	發熱
	下痢膿血	肢體浮腫	往來寒熱
	腰痛	口舌乾燥	肢體浮腫
惡寒	發熱	項强	胸悶
	惡寒發熱	少氣懶言	無汗
	頭痛	口脣靑紫	發熱惡寒
	無汗	自汗	脈浮
	身熱	泛惡	惡寒發熱
	惡風	微惡風寒	頭痛
	胸悶	目眩	發熱
	鼻塞	聲重	身熱
	肢冷	胸悶不舒	鼻塞
	發熱惡寒	頭痛鼻塞	惡風
心悸	頭眩	大便乾	寒熱
	寒熱	不眠	頭眩
	頭眩心悸	頭眩心悸	頭眩心悸
	心悸怔忡	心悸失眠	嘔吐
	怔忡	食少	心悸怔忡
	失眠	口燥	大便乾
	短氣	氣短懶言	多夢
	健忘	失眠多夢	怔忡
	氣短	夜寐不安	失眠
	多夢	心悸健忘	盜汗
泄瀉	腹痛	消化不良	腹痛
	嘔吐	反胃	食欲不振
	腸鳴	腸鳴泄瀉	嘔吐
	腹脹	噯腐吞酸	小便不利
	小便不利	下痢	頭痛
	食欲不振	噯氣	脘腹脹痛
	發熱	神昏	腸鳴
	脘腹脹痛	寒熱	腹脹
	水腫	煩渴引飮	不思飮食
	頭痛	頭痛	反胃