제28회 한글 및 한국어 정보처리 학술대회 논문집 (2016년)

단어 의미 표현과 질병 중심 의학 문서 클러스터 기반 의학 문서 검색 기법

조승현○, 이경순
전북대학교 전자정보공학부
{jackaa, selfsolee}@chonbuk.ac.kr

Method of Document Retrieval Using Word Embeddings and Disease-Centered Document Clusters

Seung-Hyeon Jo○, Kyung-Soon Lee
Division of Computer Science and Engineering, CAIIT, Chonbuk National University

요 약

본 논문에서는 임상 의사 결정 지원을 위한 UMLS와 위키피디아를 이용하여 지식 정보를 추출하고 질병 중심 문서 클러스터와 단어 의미 표현을 이용하여 질의 확장 및 문서를 재순위화하는 방법을 제안한다. 질의로는 해당 환자가 겪고 있는 증상들이 주어진다. UMLS와 위키피디아를 사용하여 병명과 병과 관련된 증상, 검사 방법, 치료 방법 정보를 추출하고 의학 인과 관계를 구축한다. 또한, 위키피디아에 나타나는 의학 용어들에 대하여 단어의 효율적인 의미 추정 기법을 이용하여 질병 어휘의 의미 표현 벡터를 구축하고 임상 인과 관계를 이용하여 질병 중심 문서 클러스터를 구축한다. 추출한 의학 정보를 이용하여 질의와 관련된 병명을 추출한다. 이후 질의와 관련된 병명과 단어 의미 표현을 이용하여 확장 질의를 선택한다. 또한, 질병 중심 문서 클러스터를 이용하여 문서 재순위화를 진행한다. 제안 방법의 유효성을 검증하기 위해 TREC Clinical Decision Support(CDS) 2014, 2015 테스트 컬렉션에 대해 비교 평가한다.

주제어: 임상 의사 결정 지원, 단어 의미 표현, 문서 클러스터링, 질의 확장, 재순위화

1. 서론

최근 환자들은 본인이 왜 아픈지 궁금해 하며 아픈 경우에는 어떤 방식으로 관리해야 하는지를 알고 싶어 한다. 이 경우 환자의 증상과 관련된 의학 문서를 빠르고 정확하게 찾을 수 있다면 임상 의사 결정에 도움을 줄 수 있다. 또한, 의사들은 환자의 증상에 대하여 임상 의사 결정을 내릴 때 해당 환자와 증상이 비슷한 환자들을 다룬 의학 문서를 이용한다면 임상 의사 결정에 큰 도움을 줄 수 있게 된다.

최근 의학 문서 처리 연구는 정보 처리 분야에서 많은 연구가 이루어지고 있다. TREC(Text REtrieval Conference)에서 2014년부터 진행 중인 Clinical Decision Support(CDS) Track[1]에서는 환자가 겪고 있는 증상들을 질의로 구성하여 해당 증상이 발생했을 시 병을 진단하거나, 검사 방법 또는 치료 방법에 관하여 서술된 논문을 검색하는 방법에 대한 연구들이 진행 중이다. 또한, NTCIR에서 2013년부터 진행 중인 Medical Natural Language Process(MedNLP) Task[2]에서는 문서에서 증상이나 진단과 관련된 정보를 추출하고 icd-10 code를 이용하여 증상이나 진단에 관련된 정보를 보편화하는 연구들이 진행 중이다. 또한 2009년부터 진행 중인 ACL의 워크샵 중 하나인 BioNLP(Biomedical Natural Language Processing) Shared Task에서는 개체 추출 및 표현, 이벤트 및 관계 추출과 관련된 연구가 진행 중이다. 정보 검색 연구에서 질의 확장은 검색 결과의 정확률과 재현율을 모두 향상시킬 수 있는 방법이며, 이를 이용하여 환자들이 겪고 있는 증상과 관련된 추가 증상들을 찾아내어 임상 의사 결정 지원에 도움을 줄 수 있다.

본 연구에서는 단어 의미 표현을 이용하여 의미 표현 벡터를 구축하고 의학 문서 검색에 사용할 임상 의미 지식을 이용하여 질병 중심 문서 클러스터를 구축하는 방법을 제안한다. 임상 의미 지식은 의학 사전 중 하나인 UMLS와 위키피디아에 있는 의학 데이터를 이용하여 구축하며 병명-증상 지식, 병명-검사 지식, 병명-치료 지식 등을 구축하고, 인공 신경망을 이용하여 단어 표현을 구축한다. 또한, 구축한 임상 의미 지식과 의학 사전 중 하나인 MeSH의 병 카테고리 정보를 이용하여 해당 병과 유사한 병명을 찾아 의학 문서들을 그룹화 하는데 사용한다. 이후, 임상 의미 지식을 이용하여 질의와 관련된 병명을 탐지하고 인공 신경망을 이용하여 구축한 단어 표현을 이용하여 질의 확장을 하고 그룹화한 의학 문서들의 정보를 이용하여 잠정적 적합 피드백, 문서 재순위

화를 진행하여 검색 향상 여부를 확인한다. 제안 방법의 유효성을 검증하기 위해 TREC CDS 2014, 2015 테스트 컬렉션에 대해 비교 평가한다.

2. 관련 연구

임상 인과 관계 추출 관련 연구에서 [3]은 종속성 그래프를 이용하여 병명-증상 관계를 추출하기 위해 SPARE라는 관계 추출을 위한 구문 패턴을 제안하였다. SPARE를 통해 병명-증상 패턴을 학습시킨 뒤, 증상이 묘사되어 있는 패턴을 주고 병명-증상 관계를 추출하는 연구를 진행하였다. [4]는 병명-치료 방법 관계를 추출하기 위하여 rich function을 적용한 최대 엔트로피 모델을 이용하였다. 본 연구에서는 UMLS(Unified Medical Language System)와 위키피디아를 이용하여 임상 인과 관계를 구축한다.

의학 문서 클러스터링 관련 연구에서 [5]는 MeSH를 이용하여 암, 바이러스, 눈병에 관련된 개념을 추출하고 온톨로지를 기반으로 하여 개념의 가중치를 계산한 뒤 코사인 유사도와 K-means 알고리즘을 이용하여 의학 문서를 클러스터링 하였다. [6]은 그래프 클러스터링 알고리즘을 이용하여 단어 클러스터를 생성하고 생성된 단어 클러스터를 이용하여 의학 문서들을 클러스터링하였다. 본 연구에서는 구축한 임상 인과 관계와 의학 사전인 MeSH(Medical Subject Headings)를 이용하여 의학 문서를 클러스터링한다.

인공 신경망을 이용한 단어 표현 관련 연구에서 [7]에서는 의학 문서에서 단어 표현 특징을 이용하여 약의 이름을 인지하는 연구를 진행하였다. [8]에서는 단어의 효율적인 의미 추정 기법을 이용하여 질의와 유사한 단어를 선택한 후 질의 확장을 하였으며 [9]에서는 단어 표현을 이용하여 질의와 관련된 의학 요약어를 찾는 연구를 진행하였다. 본 연구에서는 단어의 효율적인 의미 추정 기법을 이용하여 위키피디아에서 추출한 의학 어휘들에 대하여 의미 표현 벡터를 구축하였다. 병명의 위키피디에 페이지의 요약 부분과 내용 부분을 말뭉치로 사용하였다.

3. 임상 인과 관계와 질병 중심 의학 문서 클러스터 구축

이 장에서는 의학 자원 (UMLS, 위키피디아)을 이용하여 임상 인과 관계를 구축한 뒤, 하나의 병에 대하여 비슷한 내용의 문서들을 하나의 클러스터로 묶어 질병 중심 클러스터를 구축하는 방법에 대하여 설명한다. 또한, 단어의 효율적인 의미 추정 기법을 이용하여 의학 용어들의 단어 표현 벡터를 생성하는 방법에 대하여 설명한다. 그림 1은 임상 인과 관계와 질병 중심 클러스터를 생성하기 위한 절차이다.

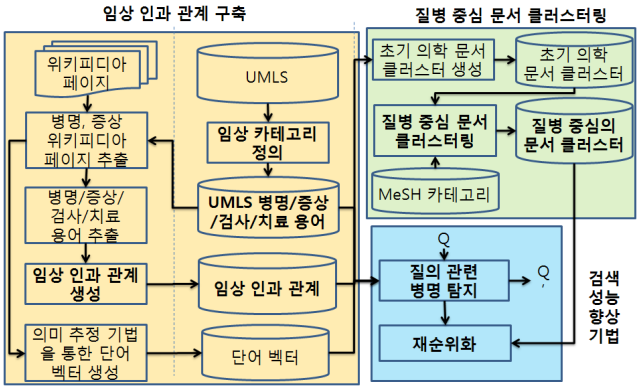

그림 1. 임상 인과 관계와 질병 중심 의학 문서 클러스터 구축 절차

3.1 의학 자원을 이용한 임상 인과 관계 구축

이 장에서는 의학 사전인 UMLS(Unified Medical Language System)와 위키피디아의 의학 페이지에서 의학 용어(병명, 증상, 검사, 치료)를 추출하고 이를 이용하여 임상 인과 관계를 구축하는 방법을 제안하고자 한다.

의학 사전인 UMLS에는 의학 용어에 대한 의미 형태가 포함되어 있으며, 의미 형태를 이용하여 해당 의학 용어가 어떤 정보(병, 증상, 치료 방법 등)와 관련이 되어 있는지 알 수 있다. 본 연구에서는 UMLS가 가지고 있는 133개의 의미 형태 중 의학 분야와 관련된 27개의 의미 형태를 선택한 뒤 선택된 의미 형태를 4개의 임상 카테고리(병명(disease), 증상(symptom), 치료(treatment), 검사(test))로 정의하였다. 병명 카테고리에서 추출한 의학 용어의 수는 610,356개, 증상 카테고리에서 추출한 의학 용어의 수는 1,224,254개, 검사 카테고리에서 추출한 의학 용어의 수는 296,161개, 치료 카테고리에서 추출한 의학 용어의 수는 609,675개이다.

위키피디아에는 해당 병과 관련된 증상, 검사, 치료 방법과 같은 의학 정보가 있다. 따라서, 위키피디아의 의학 정보를 이용한다면 의학 용어들의 인과 관계를 생성할 수 있을 것이다. 추출한 병명, 증상 카테고리의 UMLS 용어를 이용하여 의학 관련 위키피디아 페이지를 추출하였다. 추출된 의학 관련 위키피디아 페이지는 60,892개이다. 이후 위키피디아 페이지의 제목에서 병명을 추출한다. 추출된 병명의 개수는 18,442개이다.

위키피디아 페이지는 제목(title), 초록(abstract), 내용물(contents)로 구성되어 있다. 또한, 내용물은 정보를 포함하고 있는 필드(field)들로 구분되어 있다. 내용물에는 여러 개의 필드가 존재한다. 이 중 임상 카테고리와 관련된 7개의 필드('징후와 증상(Signs and symptoms)', '진단(Diagnosis)', '특징

(Characteristics)', '합병증(Complications)', 검사(Screening)', '치료(Treatment)', '관리(Management)')를 선택하고 임상 카테고리 별로 구분하였다. 선택한 7개의 필드에서 임상 카테고리의 UMLS 용어를 이용하여 증상, 검사, 치료 용어를 추출하였다. 이때, '징후와 증상', '진단', '특징', '합병증' 필드에서는 증상 용어를, '진단', '검사' 필드에서는 검사 용어를, '치료', '관리' 필드에서는 치료 용어를 추출하였다. 추출한 증상 용어는 70,312개, 검사 용어의 수는 36,048개, 치료 용어의 수는 57,625개이다.

위키피디아에서 추출한 의학 용어를 이용하여 임상 인과 관계를 구축한다. 임상 인과 관계의 형식은 다음과 같으며 생성된 인과 관계의 수는 18,442개이다.

1) 증상-병명 관계: < 증상$_i$: 병명$_{i1}$, 병명$_{i2}$, ... >
2) 병명-증상 관계: < 병명$_j$: 증상$_{j1}$, 증상$_{j2}$ … >
3) 검사-병명 관계: < 검사$_k$: 병명$_{k1}$, 병명$_{k2}$ ... >
4) 병명-검사 관계: < 병명$_l$: 검사$_{l1}$, 검사$_{l2}$ … >
5) 치료-병명 관계: < 치료$_m$: 병명$_{m1}$, 병명$_{m2}$ ... >
6) 병명-치료 관계: < 병명$_n$: 치료$_{n1}$, 치료$_{n2}$ … >

3.2 질병 어휘의 의미 표현 벡터 구축

이 장에서는 단어의 효율적인 의미 추정 기법(word2vec)을 이용하여 위키피디아에서 추출한 의학 어휘들에 대하여 의미 표현 벡터를 구축하였다. word2vec은 2013년 구글에서 발표된 연구로, Tomas Mikolov라는 사람을 필두로 여러 연구자들이 모여서 만든 연속적인 의미 표현(Continuous Word Embeddings) 학습 모형이다. word2vec의 학습 방법은 두 종류(CBOW, skip-gram)가 있다. 두 종류 모두 입력 층, 투사 층, 출력 층으로 이루어져 있다. CBOW(Continuous Bag of Words)방식은 주변 단어가 만드는 맥락을 이용하여 타겟 단어를 예측하는 방법이며, skip-gram은 한 단어를 기준으로 주변에 올 수 있는 단어를 예측하는 방법이다. 본 연구에서는 word2vec의 skip-gram 방법을 이용하여 의학 용어들의 단어 표현을 구축하였다. 병명의 위키피디에 페이지의 요약 부분과 내용 부분을 말뭉치로 사용하였다.

3.3 질병 중심 의학 문서 클러스터 구축

임상 인과 관계를 이용하여 병과 관련된 문서를 검색할 3가지 유형의 질의를 생성한다.

1) '진단(diagnosis)' 유형 질의: 병명 용어 1개, 증상용어 여러 개
2) '검사(test)' 유형 질의: 병명용어 1개, 검사용어 여러 개
3) '치료(treatment)' 유형 질의: 병명용어 1개, 치료용어 여러 개

생성한 3가지 유형의 질의를 이용하여 문서를 검색한다. 이후, 그림 2에서 A~D에 포함되는 문서를 클러스터 후보 문서로 선택한다. 이후, 병명으로 문서를 검색한다. 클러스터 후보 문서 중 병명으로 검색한 결과에 포함된 문서가 있다면 해당 문서를 클러스터로 묶는다. 생성된 초기 의학 문서 클러스터의 개수는 18,442개이다.

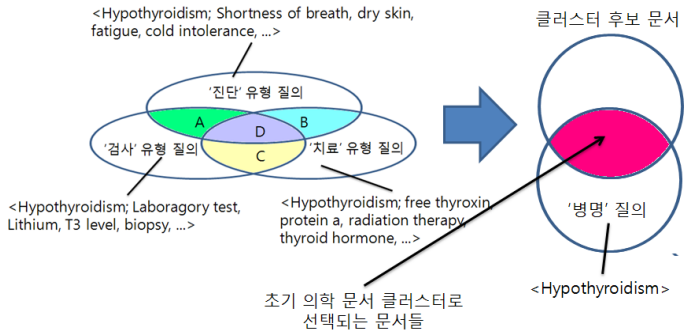

그림 2. 초기 의학 문서 클러스터 선택 방법

이후 MeSH를 이용하여 질병 중심 의학 문서 클러스터링을 진행한다. MeSH는 미국 국립의학도서관이 정하는 의학 분야의 주제명이다. MeSH의 최상위 수준 카테고리에는 질환 카테고리가 존재하며, 본 연구에서는 해당 카테고리를 사용하여 질병 중심 의학 문서 클러스터링을 하게 된다. 만약 병명이 MeSH에 존재한다면 MeSH 카테고리의 2계층 카테고리를 기준으로 병명을 클러스터링하고 클러스터링 된 병명의 초기 의학 문서 클러스터에 포함된 문서들을 묶어서 질병 중심 클러스터를 생성한다.

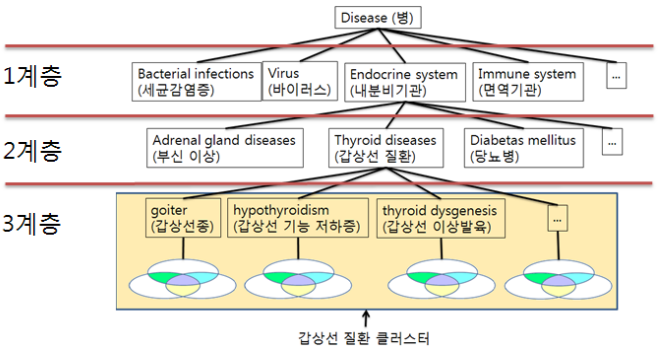

그림 3. MeSH를 이용한 병명 중심 의학 문서 클러스터링

4. 임상 의미 지식, 단어 표현 벡터, 질병 중심 의학 문서 클러스터를 이용한 문서 검색

이 장에서는 임상 의미 지식을 이용하여 질의와 연관된 병명을 탐지하는 방법에 대하여 설명한다. 또한, 단어 표현 벡터를 이용하여 질의 확장을 하는 방법과 질병 중심 클러스터를 이용하여 문서 재순위화를 하는 방법에 대하여 설명한다.

4.1 임상 의미 지식과 단어 표현 벡터를 이용한 질의와 연관된 병명 탐지

본 연구에서 사용하는 질의들은 기본적으로 증상을 포함하고 있다고 가정한다. 그 후, 질의에서 UMLS를 이용하여 증상들을 추출하고 이 증상들을 구축한 증상-병명 쌍을 이용하여 질의와 관련된 병명을 추출한다. 질의에서 추출한 증상 중 3개 이상이 증상-병명 쌍 정보에 매칭이 되면 해당 질의는 이 병명과 관련이 있다고 한다. 질의와 관련된 병명을 추출한 후, 해당 질의와 관련된 병명들을 확장 어휘로 추출한다. 또한, 3.2에서 구축한 의미 표현 벡터를 이용하여 탐지한 병명과 관련된 확장 어휘를 선택하고 질의 확장을 한다. 이 때, 탐지한 병명들의 단어 유사도 값을 모두 더해 가장 높은 값이 나온 단어를 확장 단어로 선택하게 된다.

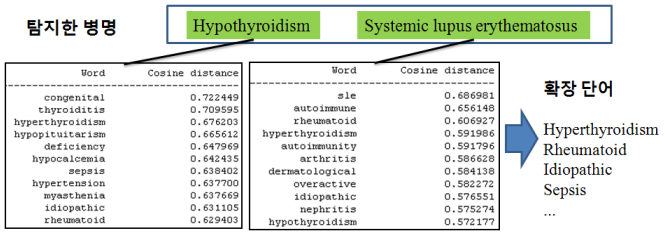

그림 4. 단어 표현 벡터를 이용한 확장 단어 선택

4.2. 질병 중심 의학 문서 클러스터를 이용한 문서 재순위화

문서 검색 시 낮은 점수를 받은 문서가 적합 문서일 경우에는 검색 성능이 저하된다. 이 때, 낮은 점수를 받은 적합 문서에 추가 점수를 부여하여 상위권으로 끌어올릴 경우 검색 성능을 향상시킬 수 있다. 4.1에서 질의와 연관된 병명을 선택했다면 해당 질병 중심 문서 클러스터에 포함된 문서들을 상위 문서로 올려 재순위화를 한다면 성능 향상에 도움을 줄 수 있을 것이라 가정하였다.

$$Rerank(Q,D) = \lambda \cdot QL(Q,D) + (1-\lambda)\frac{1}{C}\sum_{i=1}^{C} QL(Q', C_i) \quad (1)$$

Q는 원 질의이며, Q' 은 3.2.1)에서 구축한 질의다. QL(Q, D)는 초기 질의로 검색한 결과이며, QL(Q′ , C_i)는 3.3에서 검색했을 때의 결과이다. |C|는 질병 중심 의학 문서 클러스터에 포함된 병명의 수이다. 즉, 해당 문서의 초기 검색 결과가 높으면서 3.3에서 구축한 질의로 검색된 결과가 높은 경우 상위 문서로 될 가능성이 높아진다.

5. 실험 및 평가

5.1 실험 집합

제안 방법의 유효성을 검증하기 위해 TREC CDS 2014, 2015 테스트 컬렉션을 사용하여 실험하였다. TREC CDS 2014, 2015 테스트 컬렉션은 총 30개의 질의로 구성되어 있으며, 해당 질의는 Description 파트와 Summary 파트로 나뉜다. 본 논문에서는 TREC CDS 2014의 질의를 학습 질의로, TREC CDS 2015의 질의를 테스트 질의로 사용하였다.

언어모델(LM)과 적합모델(RM)에 대한 실험 결과는 인드리(Indri-5.7)[10] 시스템을 사용하였다. 각 모델에 대해 학습 질의를 이용하여 파라미터를 학습한 후 테스트 질의에 대해 적용하여 성능을 평가하였다. 초기 질의에 대한 가중치($\lambda \in \{0.1, 0.2, \ldots, 0.9\}$)로 실험하였다. 확장 질의의 수($W \in \{5, 10, 15, 20, 25, 50, 75, 100\}$)는 학습 질의를 이용하여 결정한 뒤 테스트 질의에 적용하였다.

5.2 비교 실험 결과

성능 평가 방법은 TREC CDS 2014, 2015에서 사용된 문서 관련성 등급(graded relevance scale)를 이용한 추정된 NDCG(inferred Normalized Discounted Cumulative Gain), 상위 10개 문서에서의 정확률 (P@10), R-precision (R-prec)이다. 추정된 NDCG는 상위 100개의 문서에 대한 NDCG값이다.

- LM: 언어 모델
- QE_LM: 탐지한 병명으로 질의 확장
- FDUDMIIP[11]: MeSH를 이용하여 질의와 관련된 MeSH 용어를 추출하여 질의 확장
- QE_WE: 의미 추정 기법을 이용하여 질의 확장
- ECNU[12]: MeSH를 이용하여 질의 확장 뒤 SVM을 이용한 문서 재순위화
- Re-ranking: 질병 중심의 문서 클러스터를 이용한 재순위화
- Re-ranking + WE: 질병 중심 문서 클러스터를 이용한 재순위화 + 의미 추정 기법을 이용하여 질의 확장

비교 방법	추정된 NDCG	P@10	R-prec
LM	0.2179	0.3733	0.1863
QE_LM	0.2316	0.3867	0.1971
FDUDMIIP	0.2469	0.3900	0.1847
QE_WE	0.2513	0.4133	0.2173
ECNU	0.2680	0.4533	0.2157
Re-ranking	0.2769	0.5067	0.2553
Re-ranking + WE	**0.2831**	**0.5167**	**0.2602**

표 1. 비교 실험 결과

표 1을 통해 우리의 제안 방법이 테스트 컬렉션에 대하여 LM 및 TREC에 참가한 시스템에 비해 성능이 향상된 것을 알 수 있다. 또한 의미 표현을 사용했을 경우 그렇지 않은 경우에 비해 성능이 향상된 것을 알 수 있다.

6. 결론

본 논문에서는 UMLS와 위키피디아를 이용하여 임상 인과 관계를 구축하고 임상 인과 관계와 MeSH를 이용하여 질병 중심 클러스터를 구축하는 방법을 제안하였다. 또한 의미 추정 기법을 이용하여 질병 어휘의 의미 표현 벡터를 구축하는 방법을 제시하였다.

비교 실험을 통해 의미 추정 기법을 이용하여 질의 확장을 할 경우 문서 검색 성능을 향상시킬 수 있음을 알 수 있었다. 또한, 임상 인과 관계와 질병 중심의 의학 문서 클러스터를 구축하여 이용할 경우 문서 검색 성능을 향상시킬 수 있음을 보였다.

참고문헌

[1] K. Roberts and M. S. Simpson, “Overview of the TREC 2015 Clinical Decision Support Track”. In Proceedings of the Text Retrieval Conference 2015, 2015.

[2] E. Arimaki, M. Morita, Y. Kano, and T. Ohkuma, “Overview of the NTCIR-11 MedNLP-2 Task”. In Proceedings of the 11th NTCIR Conference, 2014.

[3] M. Hassan, O. Makkaoui, A. Coulet, and Y. Toussaint, “Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs”. In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). pp 71-80, 2015.

[4] L. Yao, C. J. Sun, X. L. Wang and X. Wang, "Relationship extraction from biomedical literature using Maximum Entropy based on rich features". In Proceedings of the Ninth Internation al Conference on Machine Learning and Cybernetics(ICMLC’10). pp 3358-3361, 2010.

[5] S. Logeswari and K. Premalatha, "Biomedical Document Clustering Using Ontology based Concept Weight". In proceedings of 2013 International Conference on Computer Communication and Informatics(ICCCI’13), 2013.

[6] R. Prasath and P. O’Reilly, “Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis”. In proceedings of the Medical Information Retrieval Workshop at SIGIR 2014(MedIR’14), pp 12-15, 2014.

[7] W. Yonghui, X. Jun, Z. YaoYun and X. Hua, “Clinical Abbreviation Disambiguation Using Neural Word Embeddings”, In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing, pp.171-176, 2015.

[8] W. Yonghui, X. Jun, J. Min, Z. Yaoyun and X. Hua, “A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text”, In Proceedings of the AMIA Annual Symposium Proceedings. pp.1326-1333, 2015.

[9] L. Yue, G. Tao, S. M. Kusum, J. Heng and L. M. Deborah, “Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion”, In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing, pp.92-97, 2015.

[10] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, “Indri: A language model-based search engine for complex queries”, In Proceedings of the International Conference on Intelligence Analysis, http://www.lemurproject.org/indri. 2005.

[11] R. You, Y. Zhou, S. Peng, and S. Zhu, “FDUMedSearch at TREC 2015 Clinical Decision Support Track”, In Proceedings of the 24th Text Retrieval Conference, 2015.

[12] Y. Song, Y. He, Q. Hu, and L. He, “ECNU at 2015 CDS Track:Two Re-ranking Methods in Medical Information Retrieval”, In Proceedings of the 24th Text Retrieval Conference, 2015.

[13] 조승현, 이경순. “의학 문서 검색을 위한 UMLS 개념 정보와 위키피디아 정보를 이용한 질의 확장”, 2015 한국컴퓨터종합학술대회, pp 672-674, 2015.

[14] 조승현, 이경순. “의학 문서 검색을 위한 지식 추출 및 LDA 기반 질의 확장”, 제 26회 한글 및 한국어 정보처리 학술대회, pp. 31-34, 2015.