의학 문서 검색을 위한 지식 추출 및 LDA 기반 질의 확장

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

Query Expansion based on Knowledge Extraction and Latent Dirichlet Allocation for Clinical Decision Support

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 질의 유형 정보를 이용한 LDA 기반 질의 확장 방법을 제안한다. 질의로는 해당 환자가 겪고 있는 증상들이 주어진다. UMLS와 위키피디아를 사용하여 병명과 병과 관련된 증상, 검사 방법, 치료 방법 정보를 추출한다. UMLS와 위키피디아를 사용하여 추출한 의학 정보를 이용하여 질의와 관련된 병명을 추출한다. 질의와 관련된 병명을 이용하여 추가 증상, 검사 방법, 치료 방법 정보를 확장 질의로 선택한다. 또한, LDA를 실행한 후, Word-Topic 클러스터에서 질의와 관련된 클러스터를 추출하고 Document-Topic 클러스터에서 초기 검색 결과와 관련이 높은 클러스터를 추출한다. 추출한 Word-Topic 클러스터와 Document-Topic 클러스터 중 같은 번호를 가지고 있는 클러스터를 찾는다. 그 후, Word-Topic 클러스터에서 의학 용어를 추출하여 확장 질의로 선택한다. 제안 방법의 유효성을 검증하기 위해 TREC Clinical Decision Support(CDS) 2014 테스트 컬렉션에 대해 비교 평가한다.

주제어: 임상 의사 결정 지원, UMLS, 위키피디아, LDA, 질의 확장

1. 서론

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. 최근 병원에서는 환자들의 진료 기록을 이용하여 환자들의 증상에 대하여 임상 의사 결정 지원을 하고 있다. 하지만, 환자 기록 데이터들은 환자들의 지속적인 관찰을 요하기 때문에 많은 양을 얻기가 힘들며 이를 보완하기 위해 웹 페이지나 의학 사전 등의 의학 지식 정보를 활용하기도 한다. 정보 검색 연구에서 질의 확장은 검색 결과의 정확률과 재현률을 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 지식 정보 중 위키피디아[3]와 의학 사전인 UMLS(Unified Medical Language System)[4]을 이용하였으며, 접근 방법은 다음과 같다. 첫째, UMLS에서 출현하는 의학 용어를 개념 정보를 기반으로 매핑한다. 둘째, 개념 정보를 기반으로 매핑된 의학 용어 정보를 이용하여 위키피디아에서 병명 및 해당 병명과 관련된 의학 용어들을 추출한다. 셋째, 질의에서 UMLS를 이용하여 증상과 관련된 의학 용어를 추출한 뒤, 위키피디아를 이용하여 얻은 증상 정보들과 매칭시켜 질의와 관련된 병명을 찾는다. 넷째, LDA를 이용하여 질의와 관련된 후보 확장 어휘들을 추출하고 질의 유형 정보를 이용하여 확장 어휘를 선택한 뒤 질의 확장을 한다. 제안 방법의 유효성을 검증하기 위해 TREC CDS 2014 테스트 컬렉션에 대해 비교 평가한다.

2. 관련 연구

환자들의 진료 기록을 이용하여 환자들의 증상에 대하여 임상 의사 결정 지원을 하는 연구 중 Lu Liu[5]는 약 170,000개의 환자 기록지를 이용하여 당뇨병과 관련된 합병증과 치료 패턴을 모으는 모델을 구현하였으며, Corey[6]는 13,028개의 환자 기록지를 이용하여 뇌암 환자 기록에서 시간에 따라 환자 상태가 어떤 식으로 변하는지를 파악하는 모델을 제시하였다.

지식 정보를 활용하여 임상 의사 결정 지원을 하는 연구 중 Isabelle[7]는 크라우드소싱과 위키피디아를 이용하여 증상과 관련되는 질의를 모아 진단 의학 용어들의 우회적인 표현을 매칭 시키는 연구를 진행하였으며, Ryen[8]은 웹 페이지에서 사람들의 행동 양식 정보 등을 이용하여 질병 발생이나 지속성을 예측하였다.

질의 확장을 통해 임상 의사 결정에 도움을 주는 연구로는 의학 사전 정보를 이용하여 연관 관계 등을 추출하여 질의 확장 단어로 선택한 연구가 있다.

3. UMLS와 위키피디아를 이용한 의학 용어 정보 쌍 구축

본 연구의 기본 접근은 질의에 증상 정보가 있으면 증상에 대한 전문 지식 정보를 활용하여 주어진 증상 정보들에 대한 병명을 알아내고 추가 증상이나 치료 방법과 관련된 어휘를 이용하여 질의 확장을 하여 성능 개선을 하는 것이다.

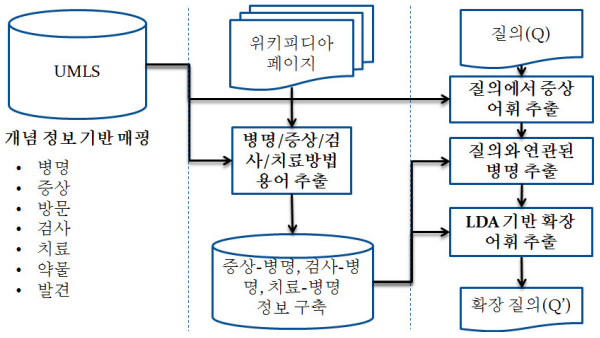

그림 1. 시스템 구조

본 연구에서는 지식 정보(UMLS, 위키피디아)를 이용하여 의학 용어들의 정보 쌍을 구축한다.

3.1 UMLS에서 개념 정보 기반 의학 용어 매핑

의학 사전인 UMLS에는 의학 용어에 대한 개념 정보가 포함되어 있으며, 개념 정보를 이용하여 해당 의학 용어가 어떤 정보(병, 증상, 치료 방법 등)와 관련이 되어 있는지 알 수 있다. 예를 들어, “Acute repiratory distress"라는 의학 용어는 ” Disease or Syndrome"이라는 개념 정보를 가지고 있으며, 이를 통해 해당 의학 용어는 병명과 관련이 있다는 것을 알 수 있다. 본 연구에서는 UMLS의 개념 정보를 총 7개(병명(Dx), 증상(Sx), 검사(Test), 치료 방법(Tx), 약물(Mx), 방문(Visit), 발견(Finding))로 분류했으며 분류한 의학 관련 개념 정보와 의학 용어들을 매핑시켰다. 그 중, 병명과 증상으로 분류된 개념 정보들은 아래와 같다.

UMLS 개념 정보	병명
Anatomical Abnormality	해부학적 이상
Congenital Abnormality	선천성 이상
Acquired Abnormality	취득 이상
Pathologic Function	병리학적 기능
Disease or Syndrome	질병 또는 증후군
Mental or Behavior Dysfunction	정신 또는 행동 장애
Neoplastic Process	종양 프로세스

표 1. UMLS에서 병명과 관련된 개념 정보

UMLS 개념 정보	증상
Signs and Symptoms	징후 또는 증상
Pathologic Function	질병 또는 증후군
Disease or Syndrome	병리학적 기능
Mental or Behavior Dysfunction	정신 또는 행동 장애
Neoplastic Process	종양 프로세스
Finding	발견

표 2. UMLS에서 증상과 관련된 개념 정보

3.2 위키피디아를 이용한 의학 용어 정보 쌍 구축

위키피디아는 많은 양의 데이터를 포함하고 있으며, 이 중 의학 관련 용어에 대한 정보도 포함되어 있다. 또한, 위키피디아에서 병명을 검색 시, 해당 병명에 관한 상세 정보를 알 수 있다.

본 연구에서는 위키피디아를 이용하여 3.1에서 개념 정보를 기반하여 매핑된 의학 용어들에 대한 페이지 정보를 추출한다. 추출한 위키피디아 페이지에 해당 의학 용어의 정보가 존재한다면 병명의 증상("Signs and symptoms", "Diagnosis", "Characteristics", "Complications"), 검사 방법("Diagnosis",

"Screening"), 치료 방법("Treatment", "Management")과 관련 있는 Field의 정보를 추출한다. 만약 해당하는 Field의 정보가 없다면 해당 페이지의 개요 부분의 정보를 추출한다.

병명과 관련된 페이지에서 추출한 정보를 UMLS를 이용하여 증상, 검사 방법, 치료 방법과 관련된 의학 용어를 추출한다. 이 때, 의학 용어를 추출하기 위해 UMLS에서 개념 정보(증상, 검사, 약물, 치료)를 이용한다. 의학 용어 추출 시, 1~3음절의 단어를 UMLS와 매칭시켜 용어를 추출하였다. 추출한 의학 정보를 이용하여 증상-병명, 검사-병명, 치료-병명 쌍을 구축한다.

4. 의학 용어 정보와 LDA를 이용한 질의 확장

구축된 의학 용어 정보를 이용하여 질의와 관련된 병명을 추출했을 때, 해당 병명의 추가 증상, 검사 방법, 치료 방법과 관련된 의학 용어들은 질의와 적합한 문서를 찾는데 도움을 줄 수 있을 것이다. 4장에서는 3장에서 구축한 의학 용어 정보 쌍을 이용하여 질의 확장을 하는 방법에 관하여 설명한다.

4.1 의학 용어 정보를 이용한 질의 관련 의학 용어 추출

질의가 주어졌을 때 질의의 내용이 증상 정보를 내포하고 있다면 해당 증상과 관련된 병명을 확장 질의로 주었을 때 더 높은 검색 성능을 보일 수 있다.

본 연구에서 사용하는 질의들은 기본적으로 증상을 포함하고 있다고 가정한다. 그 후, 질의에서 UMLS를 이용하여 증상들을 추출하고 이 증상들을 3.2에서 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 정보들을 이용하여 추가 증상, 검사 방법, 치료 방법에 대한 정보를 추출한다.

4.2 Latent Dirichlet Allocation(LDA) 기반 질의 확장

토픽 모델은 어휘나 문서들을 주제 별로 묶어주는 역할을 한다. 이를 이용한다면 해당 질의에 연관되는 문서나 용어들을 더 자세히 추출할 수 있을 것이다. 본 논문에서는 토픽 모델 중 LDA[9]를 이용하여 실험하였다.

TREC CDS의 문서들을 LDA를 이용하여 분류하였다. 분류한 LDA 정보에서 Topic-Word 클러스터의 상위 N개 어휘 중 질의에 포함된 어휘가 S개 이상 포함된 클러스터를 추출하고, Topic-Document 클러스터의 상위 M개 문서 중 초기 검색 결과의 상위 D개에 포함된 문서가 T개 이상 포함된 클러스터를 추출하였다. 이 때, 추출한 Topic-Word 클러스터와 Topic-Document 클러스터가 같은 묶음이라면 해당 클러스터는 질의와 관련 있는 클러스터라고 판단할 수 있으며, 해당 클러스터에 나타나는 의학 용어들은 질의와 관련 있다고 판단할 수 있다. Topic-Word 클러스터와 Topic-Document 클러스터가 같은 번호를 가지고 있다면 해당 Topic-Word 클러스터에 출현한 의학 용어들을 추출한다. 이 때, 클러스터는 여러 개가 출현 가능하며 더 많은 클러스터에서 출현한 의학 용어의 가중치를 추가적으로 부여한다. 추출한 의학 용어를 바탕으로 W개의 확장 어휘 선택 후 질의 확장을 한다.

5. 실험 및 평가

5.1 실험 집합

제안 방법의 유효성을 검증하기 위해 TREC CDS 2014 테스트 컬렉션을 사용하여 실험하였다. TREC CDS 2014 테스트 컬렉션은 총 30개의 질의로 구성되어 있으며, 해당 질의는 Description 파트와 Summary 파트로 나뉜다. Description 파트에서 질의는 상세한 정보로 이루어져 있으며, Summary 파트에서는 Description 파트의 정보가 요약되어 있다. 본 논문에서는 이 중 20개의 질의를 학습 질의로, 10개의 질의를 테스트 질의로 사용하였다. 실험 집합에 대한 구성은 표 4와 같다.

문서 개수	학습 질의 개수	테스트 질의 개수
733,138	20	10

표 4. TREC Clinical Decision Support 2014 실험 집합

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 테스

트 질의에 대해 적용하여 성능을 평가하였다. 초기 질의에 대한 가중치(λ∈{0.1, 0.2, ..., 0.9})로 실험하였다. LDA 기반 확장 어휘 추출 시 파라미터(N∈{5, 10, 15, ..., 50}, M∈{10, 20, ..., 100}, S∈{5, 7, 10, ..., 25}, T∈{3, 4, ..., 10}, D∈{50, 100, 200, ..., 500})는 실험을 통해 가장 좋은 결과를 나타내는 값을 사용하였다. 확장 질의 W의 수는 학습 질의를 이용하여 결정한 뒤 테스트 질의에 적용하였다.

5.2 비교 실험 결과

제안 방법과 이전 연구[11]를 비교하여 성능을 평가하였다. 성능 평가의 척도는 TREC CDS 2014에서 사용된 문서 관련성 등급(graded relevance scale)를 이용한 infNDCG(inferred Normalized Discounted Cumulative Gain)이다.

- **Baseline: 언어모델(LM)**
  질의가 특정 문서에서 발생할 확률을 계산하여 그 확률이 가장 큰 문서를 적합한 문서로 하고 상위에 순위화
- **이전 연구(KCC2015)[11]: 병명을 이용한 질의 확장**
  질의와 관련된 병명을 추출하여 확장 질의로 선택
- **방법 1: 질의 관련 의학 용어를 이용한 질의 확장**
  병명을 이용하여 질의와 관련된 추가적인 증상, 검사 방법, 치료 방법 용어를 추출하여 확장 질의로 선택
- **방법 2: LDA를 이용한 질의 확장**
  방법 1과 LDA를 이용하여 병명, 증상, 검사 방법, 치료 방법 용어를 추출하여 확장 질의로 선택

평가방법	LM	KCC2015	방법 1	방법 2
infNDCG	0.1713	0.2039	0.2107	0.2137

표 6. Description 파트 실험 결과

평가방법	LM	KCC2015	방법 1	방법 2
infNDCG	0.1887	0.2145	0.2203	0.2237

표 7. Summary 파트 실험 결과

표 6, 7에서와 같이 제안 방법이 언어모델에 비해 성능이 향상됨을 알 수 있었다. 비교 실험 결과, 질의와 연관된 의학 용어를 추출하여 질의 확장을 하는 방법이 유효함을 알 수 있었으며, LDA를 기반으로 하여 질의 확장 시 더 좋은 성능을 보인다는 것을 알 수 있었다.

5. 결론

본 논문에서는 지식 정보를 통하여 지식 쌍을 구축하고, 이를 이용하여 질의에서 나타나는 증상들과 매칭시켜 관련 있는 병명을 얻어낸 후, 해당 질의와 관련된 병명, 추가 증상, 검사 방법, 치료 방법 등의 어휘를 LDA를 이용하여 추출한 뒤 질의 확장을 하는 방법을 제안하였다. 구축한 의학 정보를 이용하여 질의와 연관되어 있는 병명을 추출하는 것이 가능함을 보였으며, 제안 방법이 언어모델보다 향상됨을 보임으로서 LDA를 사용할 경우 질의 확장에 도움을 줄 수 있음을 확인할 수 있었다.

참고문헌

[1] http://trec-cds.appspot.com/2014.html
[2] http://mednlp.jp/ntcir11/
[3] http://en.wikipedia.org
[4] Olivier Bodenreider. “The Unified Medical Language System(UMLS): intergrating biomedical terminology”. Nucleic Acids Res. 2004;32:D267-D270.
[5] Lu Liu, Jie Tang, Yu Cheng, Ankit Agrawal, Wei-keng Liao, Alok Choudhary. "Mining diabetes complication and treatment patterns for clinical decision support". CIKM '13 Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. Pages 279-288.
[6] Corey Arnold, William Speier. "A topic model of clinical reports". SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Pages 1031-1032.
[7] Isabelle Stanton, Samuel Ieong, Nina Mishra. "Circumlocution in diagnostic medical queries". SIGIR '14 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. Pages 133-142.
[8] Ryen W. White, Eric Horvitz. "Studies of the onset and persistence of medical concerns in search logs". SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Pages 265-274.
[9] David M. Blei, Andrew Y. Ng, Michael I. Jordan. "Latent dirichlet allocation". The Journal of Machine Learning Research. Volume 3, 3/1/2003. Pages 993-1022.
[10] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, “Indri: A language model-based search engine for complex queries”. In Proc. International Conference on Intelligence Analysis. http://www.lemurproject.org/indri. 2005.
[11] 조승현, 이경순. “의학 문서 검색을 위한 UMLS 개념 정보와 위키피디아 정보를 이용한 질의 확장”. 2015 한국컴퓨터종합학술대회.