NLM Medical Text Indexer를 활용한 우리나라 의학문헌의 MeSH Semi Indexing 방안

MeSH Semi Indexing of the Korean Biomedical Literature, using NLM Medical Text Indexer

정소나, 가톨릭대학교 여의도성모병원 도서실, sona@catholiclac.kr
이춘실, 숙명여자대학교 문헌정보학과, cslee@sookmyung.ac.kr
Sona Jeong, The Catholic University of Korea Yeouido St. Mary's Hospital Library
Choon Shil Lee, Dept. of library and Information Science, Sookmyung Women's University

본 연구에서는 PubMed에 등재되었으나 Medical Subject Headings(MeSH)가 부여되지 않은 국내 의학학술지의 문헌을 대상으로 미국국립의학도서관 (NLM: National Library of Medicine)의 Medical Text Indexer(MTI)를 활용하여 MeSH 용어를 추천받은 후, PubMed 레코드의 유사주제문헌 (Relation Citations, PRC)에 부여된 MeSH와의 일치여부를 분석하였다. 또한 논문의 저자가 부여한 키워드(저자키워드)와 PRC MeSH의 일치여부도 비교하였다. PRC MeSH와 MTI MeSH 추천어의 일치율은 주표목이 21.1%였고, 체크태그는 18.1%, 부표목은 16.5%로 나타났다. 우리나라 의학논문에 나타난 저자키워드의 중요한 특징은 MeSH 주표목 위주이고, 체크태그와 부표목은 거의 사용하지 않는 것이다. 따라서 저자키워드와 PRC MeSH 주표목과의 일치율은 23.4%에 이르지만, 체크태그와 부표목의 일치율은 각각 1%, 2.1%였다. 색인전문가가 통제어휘를 사용하여 색인하는 과정에서 PRC와 MTI의 MeSH 주표목과 저자키워드가 일치하는 용어를 주표목으로 부여하고, PRC와 MTI가 추천하는 체크태그와 부표목을 활용하는 등 국내 의학문헌의 MeSH 용어 부여 작업을 반자동화(semi-indexing)하면, 정확하고 신속한 MeSH 부여 작업이 가능할 것이다.

1. 서론

1.1 MeSH 색인의 필요성

주제색인은 문헌의 내용 또는 주제를 표현하기 위하여 문헌을 적절히 표현할 수 있는 색인어나 분류기호를 부여하여 검색시에 관련 문헌이 누락됨이 없이 찾을 수 있도록 작성해야한다.

정보기술의 발달과 의학학술커뮤니티의 연구활동이 활발해지면서 의학문헌이 폭발적으로 증가하고 2000년이후 근거중심의학(evidence based medicine, EBM) 연구 활동이 이슈화되어 저자 자신의 논문을 대표할 수 있으면서 검색에 유용하게 쓰이는 주제어의 중요성이 강조되고 있다.

또한 효율적으로 많은 양의 문헌을 빠른 시간 내에 일관성있게 주제를 정확하게 기술하

기 위한 자동색인이 필요하게 되었고 색인전문가(Indexer)가 색인할 수 있는 문헌의 양이 한정됨을 보완하여 색인 작성이 원활해 질 수 있도록 지원하는 반자동색인(semi indexing), 자동색인시스템(automatic indexing)이 적용되고 있다.

미국국립의학도서관(NLM: National Library of Medicine)은 Medical Subject Headings (MeSH)라는 통제어휘를 사용하여 색인전문가가 NLM의 데이터베이스에 주제색인을 하고 있고 2002년 이후 문헌의 내용을 MeSH로 추천하는 Medical Text Indexer(MTI)를 NLM의 데이터베이스의 색인과 관리를 위한 Data Creation Maintenance System(DCMS)상에서 활용하고 있다.

국내의학문헌의 경우 MeSH를 활용한 주제색인은 한국학술총람, Korean Index Medicus가 있는데 1993년까지 색인전문가에 의해 작성되었다. 그러나 1993년 이후 간헐적인 색인작성이 이루어지기는 하였으나 MEDLINE 등재학술지를 제외하고 지속적인 색인 작성이 이루어지지 않고 있다.

대한의학학술지편집인협의회에서 운영하고 있는 KoreaMed는 현재 학술지평가를 통과한 163종의 의학학술지를 수록하고 있는 데이터베이스로 2010년 7월말 현재 약 164,000건의 데이터가 수록되어 있다. KoreaMed 학술지중에서 17종(10%)의 학술지는 NLM의 색인전문가에 의하여 MeSH가 부여되는 MEDLINE/PubMed학술지이다. 그러나 최근에 PubMedCentral (PMC)의 등재를 통해 PubMed에 레코드가 생성된 14종의 학술지에는 MeSH가 부여되지 않고 있다.

KoreaMed는 PubMed와 동일하게 영어로 작성되는 데이터베이스로 최근 종별(species : humans, animals), 성별(gender), 연령별(age groups), 특정유형의 동물명 등 연구대상을 기술하는 MeSH의 체크태그(check tag)를 부여하여 제한기능(limit option)으로 구현하였다. 그리고 임상근거를 체계적인 의학문헌의 고찰을 통해 획득하려는 EBM 환경하에서는 특히 임상연구(clinical trials)의 검색이나 체계적 문헌고찰(systematic reviews), 특정질병의 원인(etiology), 진단(diagnosis), 예후(prognosis) 그리고 치료(therapy)의 검색에 대한 KoreaMed 이용자들의 요구를 충족시키고 검색의 재현율을 높이는 동시에 정확률을 높이는 검색어 집합을 제공하기 위해서 MeSH에 의한 색인 작성이 불가피한 상황이다. 따라서 MEDLINE 학술지를 제외한 146종의 학술지는 색인전문가가 직접 색인을 해야 하는데 MeSH를 신속하게 체계적이고 효율적으로 입력하는 자동화하는 방법이 요구된다.

본 연구의 목적은 색인전문가가 통제어휘를 사용하여 색인하는 과정에서 색인전문가를 효율적으로 지원할 수 있는 반자동색인방법을 모색하고자 한다.

구체적으로 MTI에서 추천한 MeSH 용어를 사용하여 기관, 질병, 질병의 원인, 치료와 결과 등의 임상논문의 주제와 치료유형, 치료기관, 시설, 의료인의 유형 등을 나타내는 보건관련 주제 등 문헌의 내용을 MeSH 주표목(main heading)으로 부여할 수 있는지와 주표목의 특정 관점을 표현하는 부표목(subheading, NLM의 공식용어는 qualifier임)과 PubMed 레코드의 유사주제문헌 (Relation Citations, PRC)를 참조하여 MeSH 주표목에 적합한 부표목을 부여할 수 있는지, 논문의 제목과 초록에 나타난 체크태그 관련 용어들을 MTI에서 어느 정도 추천하고 일치하고 있는지를 분석하였다.

이를 위하여 PubMed에 등재된 국내 의학학술지중에서 MEDLINE 학술지가 아니라서 국내의 색인전문가가 MeSH를 부여해야 하는 학술지를 대상으로 다음과 같은 비교분석을

하였다.

첫째, 각 문헌의 제목과 초록을 MTI에 입력하여 MeSH 용어를 추천받았다.

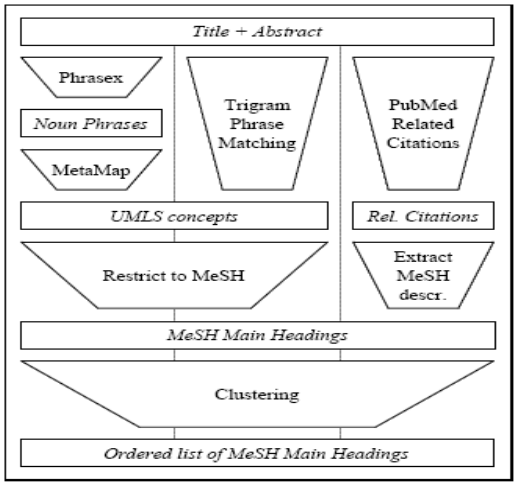
둘째, 논문의 주제를 나타내는 주표목과 연구대상을 기술하는 체크태그, 주표목의 특정 관점을 표현하는 부표목 등 MeSH 용어의 유형에 따라 MTI MeSH 추천어와 유사도가 가장 높은 PRC에 부여된 MeSH, 그리고 논문의 저자가 부여한 키워드(저자키워드)를 분류하였다.

셋째, MeSH 용어의 유형에 따라 MTI MeSH 추천어와 저자키워드를 PRC MeSH에 비교하여 일치여부를 조사하였다.

1.2 NLM Medical Text Indexer와 PubMed Related Citations

1) Medical Text Indexer

MTI는 NLM Indexing Initiative에서 개발한 시스템으로 NLM의 2015 색인 프로젝트의 일환으로 진행되고 있는 색인추천 프로그램이다. MTI 시스템 도식도는 <그림 1>과 같다 (Aronson 2008).



<그림 1> NLM Medical Text indexer

MeSH 주표목은 MetaMap과 Trigram에서 논문의 제목과 초록을 대상으로 추출한 UMLS 용어를 MeSH 용어로 한정한 용어집합과 PRC에서 MeSH를 추출하는 과정을 통해 생성된 용어집합으로 구성된다.

MetaMap은 비정형문에서 UMLS 메타시소러스 개념을 찾아주는 자연어 처리 기반의 툴이다. 문장을 명사구로 파싱(parsing)을 한 후, 명사구의 철자변이, 약어, 동의어, 어형변화 등의 변형집합을 생성하여 이 변형집합에 포함된 후보 용어들을 만들고 매핑의 강도를 점수화하여 후보용어들 중에서 문장의 구와 가장 잘 매핑된 용어를 결과로 제시한다. UMLS에서 용어가 주어지면 의미론적으로 가장 가까운 MeSH 용어를 한정하여 추천한다.

2) PubMed Related Citations

PubMed의 PRC는 텍스트 전처리과정을 거쳐 유사도가 높은 문헌순으로 유사주제문헌을 제공한다.

텍스트 전처리 과정은 다음과 같다.

① 불용어 리스트를 구축한다.

② 문헌에서는 제목, 초록, MeSH에서 용어를 추출한다. 제목의 단어는 초록의 단어에 2배의 가중치를 부여하여 Local weight scheme에 활용한다. MeSH용어의 경우 부표목을 포함하는 용어는 부표목을 포함한 주표목과 부표목을 제외한 주표목으로 2배의 가중치를 부여한다. 단, PubMed 레코드의 MeSH 용어에 "*"로 표기하는 용어 즉, 기관, 생물체, 질병, 화합물, 치료등을 표현하는 논문의 주요 관점을 나타내는 Major Topic은 적용하지 않는다.

PubMed의 유사주제문헌을 추출하는 과정에는 Global 가중치와 Local 가중치를 부여하는데 Global 가중치는 데이터베이스 전체 용어에 가중치를 부여하는 것으로 특정용어에

대한 Global 가중치는 저빈도 용어보다 크게 적용한다. Local가중치는 특정문헌내에서의 고빈도 용어의 중요도를 반영하는 것이다.

문헌간 유사도는 (Local가중치*1 × Local 가중치*2 × Global 가중치) 로 산출한다 (NLM 2008). 따라서 PubMed 레코드의 유사 주제문헌은 유사도가 높은 논문이고 PRC에 부여된 MeSH를 색인전문가가 참조할 수 있다.

MTI의 최종단계에서 수행되는 순위점수 (RankScore)는 동시출현한 용어들과 용어가중치의 곱에 대한 합과 PRC의 용어 산출값과 용어가중치의 곱에 대한 합, 그리고 용어가 MetaMap이나 Trigram 그리고 PRC의 용어인 경우 2, 그렇지 않은 경우 1의 값을 곱하여 산출한다(Aronson 2008).

2. 연구대상과 방법

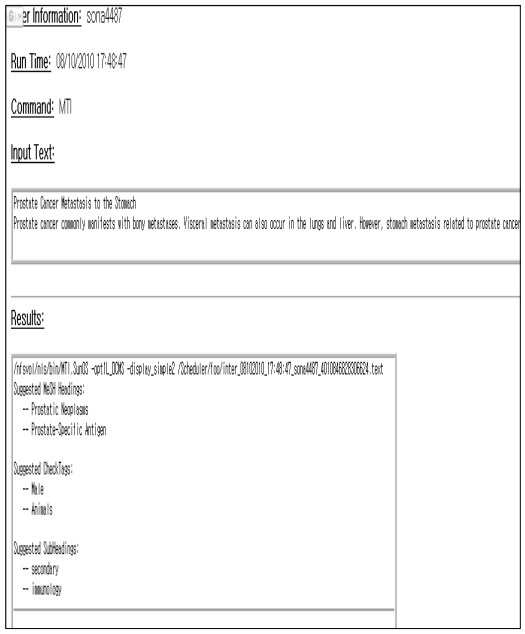
2.1 연구대상

본 연구에서는 PubMed 학술지중에서 MeSH가 부여되지 않은 학술지인 Korean Journal of Urology(KJU) 2010년 1월-7월호 95편을 대상으로 저자키워드가 부여되지 않은 논설(editorial) 2편을 제외한 93편(이하 연구대상문헌)을 선정하였다. KJU는 2010년부터 PMCI에 등재되는 과정을 통해 PubMed에 레코드가 생성되는 학술지로 PubMed 레코드에 PRC를 활용할 수 있으므로 MTI에서 추천한 MeSH 용어와의 비교가 가능하다.

2.2 연구방법

1) Medical Text Indexer의 MeSH 추천어 생성과정

MTI에 문헌의 제목과 초록을 입력하여 MeSH 주표목, 체크태그, 부표목을 추천받았다. MTI의 MetaMap과 PRC에 Path 가중치는 디폴트 값인 7:2로 설정하였다. 유사도가 높은 10개의 PRC 문헌을 대상으로 Major Topic과 MeSH 용어를 디폴트값인 1:0.8로 설정하여 MeSH 용어를 추출하였다. 결과 MTI MeSH 추천어 집합을 생성하였다. <그림 2>는 MTI에 문헌의 제목과 초록을 입력하여 추천받은 결과의 예이다.



<그림 2> MTI MeSH 추천어 결과 예

2) PubMed Related Citations MeSH 검색 과정

PubMed에서 연구대상 논문을 "korean j urol"[Journal] AND Limits: Publication Date from 2010으로 검색하여 PubMed 레코드의 상세화면에서 제시되는 PRC의 첫 번째 문헌을 MeSH가 포함되어있는 MEDLINE 형식으로 저장하여 파일을 생성하였다. <그림 3>은 PRC 첫 번째 서지레코드를 파일로 저장한 예이다.

PRC은 유사주제문헌 순으로 출력된다.

TI-[(NeurologicalCPC-59).A65-year-oldmanwithahistoryofgastriccancerwhopresentedprogressivelossofvision,memorylossandconsciousnessdisturbance]
PG-1041-51
PT-CaseReports
PT-ClinicalConference
PT-EnglishAbstract
PT-JournalArticle
PL-JAPAN
MH-Adenocarcinoma/*pathology
MH-Aged
MH-Diagnosis,Differential
MH-Humans
MH-LiverNeoplasms/secondary
MH-Male
MH-NeoplasmRecurrence,Local
MH-StomachNeoplasms/*pathology
MH-VisionDisorders/etiology
MH-WaldenstromMacroglobulinemia/complications/*pathology

<그림 3> PRC 서지레코드의 예

PRC의 첫 번째 문헌이 비교하고자 하는 문헌과 가장 유사한 문헌으로 두 문헌에는 서로 유사한 MeSH가 부여될 것이다. 첫 번째 문헌에 MeSH가 없는 경우 유사도 순위에 의해 문헌을 채택하였다.

이후 PRC MeSH에 대한 MTI MeSH 추천어, 저자키워드의 일치여부를 비교분석하였다. 일치여부는 MeSH의 유형별로 PRC 주표목과의 완전일치, 부분일치(우선어, 단·복수표기, 약어, 도치어, 상위어, 하위어)여부, 부표목과 체크태그로 구분하였다. <그림 4>는 연구의 진행과정을 도식화 한 것이다.

<그림 4> 연구진행과정

3. 연구결과

3.1 PRC MeSH와 MTI MeSH 추천어, 저자키워드의 MeSH 유형별 분포

연구대상문헌을 PRC MeSH와 MTI MeSH 추천어, 저자키워드에 의해 분석한 결과 연구대상논문에 대한 전체 MTI MeSH 추천어는 514개이고 이중에서 306개가 주표목, 122개가 부표목, 86개가 체크태그였다. <표 1>과 같이 PRC에서는 논문 1편당 평균 16.9개의 MeSH를 부여하였고 MTI의 간략정보에서는 평균 5.5개의 MeSH 용어를 추천하였다.

<표 1> MTI MeSH, PRC MeSH, 저자키워드의 MeSH 유형별 분포

(n=93)

	주표목	체크태그	부표목	계
MTI	306	86	122	514
%	59.5	16.7	23.7	100
평균	3.3	0.9	1.3	5.5
PRC	825	404	340	1,569
%	52.6	25.7	21.7	100
평균	8.9	4.3	3.7	16.9
저자 키워드	285	6	14	305
%	93.4	2.0	4.6	100
평균	3.1	0.1	0.2	3.3

저자는 평균 3개의 저자키워드를 부여하였는데, 93.4%가 주표목이었다. PRC와 MTI를 비교했을 때 MTI에서는 체크태그보다는 부표목을 PRC는 부표목보다 체크태그를 상대적으로 많이 부여하였다. 연구대상을 기술하는 체크태그의 경우 문헌의 “연구대상과 방법”이나 본문의 “표”에 기술된다. 따라서 문헌의 제목과 초록을 MTI에 입력하였으므로 체크태그의 비율이 부표목에 비해 낮게 나타났다. PRC의 경우 NLM의 색인전문가에 의해 작성되는 MeSH이므로 체크태그가 충분히 부여되었다. 따라서 색인작성시 체크태그의 정확한 색인을 위해서는 MTI 입력시 본문의 “연구대상과 방법”, 표의 캡션을 활용할 필요성이 있다.

3.2 PRC MeSH에 대한 MTI MeSH 추천어와 저자키워드의 일치도 분석

1) PRC MeSH와 MTI MeSH 추천어 일치도

PRC MeSH와 MTI MeSH 추천어를 비교하였을 때 <표 2>에서와 같이 MTI 주표목 174개(21.1%)가 PRC MeSH와 일치한 주표목이었다. 174개중 19개는 하위어, 도치어, 약어 등 부분일치어이다. 체크태그는 73개(18.1%)이고 부표목은 56개(16.5%)이다. MTI에서 추천한 부표목이 체크태그에 비해 많았음에도 불구하고 PRC와의 일치는 부표목보다는 체크태그가 높았다.

<표 2> PRC MeSH와 MTI MeSH 추천어의 유형별 일치도

	MTI		
	주표목	체크태그	부표목
PRC MeSH	825	404	340
MTI MeSH	306	86	122
PRC 일치도	174(19)	73	56
(%)	21.1	18.1	16.5

( ) 부분일치어임

2) PRC MeSH와 저자키워드 일치도

연구대상문헌의 저자들이 부여한 305개의 저자키워드와 PRC MeSH를 비교한 결과는 <표 3>과 같다.

<표 3> PRC MeSH와 저자키워드의 유형별 일치도

	저자		
	주표목	체크태그	부표목
PRC MeSH	825	404	340
저자키워드	285	6	14
PRC 일치도	193(25)	4	7
(%)	23.4	1.0	2.1

( ) 부분일치어임

MeSH 주표목이 285개(93.4%)를 차지하고 있는데 이중 부분일치어 25개를 포함한 193개(23.4%)가 PRC MeSH와 일치하였다. MTI의 주표목 일치율인 21.1%보다 높음을 알 수 있다. 저자키워드에 나타난 체크태그와 부표목은 각각 6개(1.0%)와 14개(2.1)이다.

4. 결론

저자키워드와 MTI의 MeSH 추천어를 PRC MeSH와 비교하여 일치도를 분석한 결과 다음을 확인할 수 있었다.

주표목의 경우 MTI MeSH의 일치율 21.1%에 비해 저자키워드의 일치율이 23.4%로 높았다. 체크태그의 경우 저자키워드에는논문의 주제를 3-5개로 부여하기 때문에 잘 반영되지 않았고 저자키워드의 부표목도 마찬가지였다. MTI 부표목은 MTI 체크태그보다 용어 수는 많은 반면 체크태그 일치율 18.1%에 비해 16.5%로 낮은 일치율을 보였다.

색인전문가는 주표목의 관점과 의미를 가장 잘 반영할 수 있도록 부표목과 체크태그를 색인해야 하는데 MTI를 통해 추천된 부표목과 체크태그를 활용할 수 있다.

또한 PRC의 유사도가 높은 논문에 부여된 MeSH 용어가 MTI에서 추천하는 용어와 주표목에서는 21.1%, 체크태그에서는 18.1%, 부표목에서는 16.5%로 일치하였다. PRC에서는 평균 16.9개의 MeSH 용어(8.9개의 주표목과 4.3개의 체크태그, 3.7개의 부표목)를 부여하고 있어 활용성이 높음을 알 수 있었다.

색인전문가가 통제어휘를 사용하여 색인하는 경우 PRC와 MTI의 MeSH 주표목과 저자키워드가 일치하는 용어를 주표목으로 부여하고, PRC와 MTI가 추천하는 체크태그와 부표목을 활용하는 등 국내 의학문헌의 MeSH 용어 부여 작업을 반자동화하면, 정확하고 신속한 MeSH 부여 작업이 가능할 것이다.

또한 체크태그의 경우 full text에서 연구대상을 나타내는 단락을 검토하여 부여하는 방안과 문헌에서의 주표목의 관점과 의미를 잘 반영할 수 있도록 부표목을 정확하고 신속하게 부여할 수 있는 방법을 제안하는 것이 향후 과제로 남아있다.

5. 참고문헌

Aronson A. R. (2008). "NLM Medical Text Indexer : a tool for automatic and assisted indexing." National Library of Medicine.

Interactive Medical Text Indexer
http://skr.nlm.nih.gov/interactive/mti.shtml

Koreamed
http://koreamed.org/

MeSH Browser (2010 MeSH)
http://www.nlm.nih.gov/mesh/MBrowser.html

Névéol A, Shooshan SE, Mork JG, Aronson AR.(2007) Fine-grained indexing of the biomedical literature: MeSH subheading attachment for a MEDLINE indexing tool. AMIA Annu Symp Proc. 553-7

NLM (2008). PubMed related citations algorithm
http://ii.nlm.nih.gov/MTI/related.shtml

MTI Processing flow explained.

PubMed.gov
http://www.ncbi.nlm.nih.gov/pubmed/

PubMedCentral
http://www.ncbi.nlm.nih.gov/pmc/

권애경(2001) MeSH를 확장한 한국보건의료정보분야 주제어연구: 대한의료정보학회

지를중심으로. 연세대학교 보건대학원.
김수영(2007) MeSH 색인에서검색까지. 한국의학도서관협의회.
대한의학학술지편집인협의회 http://www.kamje.or.kr/
문혜원(1999) 한국의학학술논문의 저자선정 주제어와 MeSH용어의 비교분석연구
이춘실(2010). 의학논문데이터베이스 검색및 활용의 실제. 대한의사협회지. 53(8):668-686.
한국의학도서관협의회. Korean Index Medicus.