Journal of the Korea Academia-Industrial cooperation Society
Vol. 18, No. 5 pp. 596-607, 2017

https://doi.org/10.5762/KAIS.2017.18.5.596
ISSN 1975-4701 / eISSN 2288-4688

검색용 MeSH 필터와 단어인접탐색 기법을 활용한 KoreaMed 검색 효율성 향상 연구

정소나[1], 정지나[2*]
[1]가톨릭대학교 성의교정 도서관, [2]전주대학교 보건관리학과

A Study on the Retrieval Effectiveness of KoreaMed using MeSH Search Filter and Word-Proximity Search

So-Na Jeong[1], Ji-Na Jeong[2*]
[1]**Medical Library, Catholic University of Korea**
[2]**Department of Health management, Jeonju University**

요 약 의학학술문헌에는 해부학적 조직이나 기관명이 종양, 질환 또는 감염 용어들과 서로 조합하여 사용되는 언어적 특성을 가지고 있다. 의학학술문헌을 검색할 때 데이터베이스가 제공하는 통제어휘도구인 Medical Subject Headings (MeSH)를 활용하면 합성어, 동의어, 그리고 관련어를 추가로 검색할 수 있어 검색효율이 높다. 본 연구에서는 위암(Stomach Neoplasms) 어휘군을 검색용 필터로 추가하는 방법과 동시출현용어의 거리를 측정하여 단어인접탐색 기법으로 검색효율성을 향상시키는 연구를 수행하였다. 검색용 MeSH에 추가할 어휘군을 결정하기 위해 실험데이터로 PubMed에서 중심주제어가 “Stomach Neoplasms"인 2007년~2016년 논문 8,625편을 내려 받아 논문제목으로부터 Stomach와 Neoplasms 관련 용어의 동시출현여부를 분석하였다. 검색효율성은 KoreaMed에서 검색되는 MEDLINE 학술지를 대상으로 “Stomach Neoplasms”가 MeSH로 색인되어 있는 277편으로 검증하였는데 MEDLINE MeSH, MeSH on Demand, 그리고 KoreaMed MeSH Indexer의 “Stomach Neoplasms” 색인어 추출여부와 검색용 필터로 어휘군을 적용했을 때, 그리고 동시출현 용어의 단어인접검색 기법을 적용했을 때 “Stomach Neoplasms”의 매칭여부를 비교하였다. 가장 출현빈도가 높은 용어는 “Gastric Cancer”로 2,780회 출현하였다. “Gastric Adenocarcinoma”, “Gastric MALT Lymphoma” 등과 같이 “Stomach” 용어와 “Neoplasms" 관련 조직학적 용어가 조합된 경우는 7,376개(88.51%)였다. 동시출현 거리가 2단어인 용어는 “Stomach”와 “Neoplasms”의 합성어로 5,234개(70.95%)였다. 연구 결과 MeSH용어를 제외하고 973개의 용어를 후보어휘군으로 선정하였다. MEDLINE MeSH와 KoreaMed MeSH Indexer의 MeSH 매칭률은 209편(75.5%)이었는데 검색필터를 적용한 결과 263편(94.9%)으로, 동시출현 용어의 13단어 단어인접탐색 기법을 적용한 경우 268편(96.7%)으로 매칭률이 향상되었다. 본 연구를 통해 자연어 검색에 있어서 검색효율을 향상시키는 수단으로 검색용 시소러스를 사용하면 색인비용에 대한 부담이 적고, 통제어의 망라적 장점과 자연어가 가지는 용어의 특정성을 유지할 수 있음을 증명하였다. 또한 불리안 검색보다는 단어인접탐색 기법을 활용하면 정확률을 높일 수 있어 검색 효율성이 향상됨을 알 수 있었다.

Abstract This study examined the method for adding related to "stomach neoplasms" as filters to the Medical Subject Headings (MeSH) for search as well as a method for improving the search efficiency through a word-proximity search by measuring the distance of co-occurring terms. A total of 8,625 articles published between 2007 and 2016 with the major topic terms "stomach neoplasms" were downloaded from PubMed article titles. The vocabulary to be added to the MeSH for search were analyzed. The search efficiency was verified by 277 articles that had "Stomach Neoplasms" indexed as MEDLINE MeSH in KoreaMed. As a result, 973 terms were selected as the candidate vocabulary. "Gastric Cancer" (2,780 appearances) was the most frequent term and 7,376 compound words (88.51%) combined the histological terms of "stomach" and "neoplasm", such as "gastric adenocarcinoma" and "gastric MALT lymphoma". A total of 5,234 compounds words (70.95%), in which the co-occurring distance was two words, were found. The matching rate through the MEDLINE MeSH and KoreaMed MeSH Indexer was 209 articles (75.5%). The search efficiency improved to 263 articles (94.9%) when the search filters were added, and to 268 articles (96.7%) when the 13 word-proximity search technique of the co-occurring terms was applied. This study showed that the use of a thesaurus as a means of improving the search efficiency in a natural language search could maintain the advantages of controlled vocabulary. The search accuracy can be improved using the word-proximity search instead of a Boolean search.

Keywords : Medical Subject Headings, Retrieval Efficiency, Stomach Neoplasms, Co-word Analysis, Word-proximity Search, KoreaMed

[*]Corresponding Author : Ji-Na Jeong(Jeonju University.)
Tel: +82-63-220-2506 email: naji2004@hanmail.net
Received March 31, 2017 Revised (1st April 17, 2017, 2nd April 27, 2017)
Accepted May 12, 2017 Published May 31, 2017

1. 서론

1.1 연구의 필요성

의료분야에서 활용되는 바이오 텍스트마이닝은 의학 학술문헌의 생의학적 문맥과 상호작용관계를 파악하여 개념들 간의 유의미한 관련성을 찾아내고자 자연언어처리 및 문서처리 기술을 적용한다. 바이오 텍스트마이닝 분야에서 많이 활용되고 있는 데이터베이스는 PubMed로 미국국립의학도서관(National Library of Medicine, 이하 NLM)에서 구축하고 있다. 현재 약 2,700백만 건의 문헌이 수록되어 있고 하루에도 수천편의 문헌들이 업데이트되고 있는 의학 분야에서 가장 독보적인 의학학술문헌 데이터베이스이다.

과거에도 의학 분야 연구자들은 PubMed에서 추출된 정보들로부터 이전에 발견되지 않았던 특정 질병 또는 질환과 관련된 새로운 지식들을 밝혀내어 왔다. 의학 분야 연구자들은 의학문헌 데이터베이스 중에서 PubMed를 가장 많이 이용하고 PubMed 플랫폼에서 검색이 되는 MEDLINE 학술지를 가장 많이 읽는다. 또한 연구자 자신의 연구성과를 학술지 논문을 통해 발표하면서 MEDLINE 학술지를 주로 참고문헌으로 인용한다[1]. MEDLINE은 NLM의 학술지 선정위원회 평가를 거쳐 선정된 핵심 학술문헌 데이터베이스이다. 4,600여종 학술지 약 1,100만 편이 PubMed에서 검색되는데 의학분야의 대표적인 시소러스(Thesaurus)라 할 수 있는 Medical Subject Headings(이하 MeSH)에 의해 색인을 하고 있다는 것이 가장 큰 특징이다[2].

MeSH 색인은 15-20개 정도의 MeSH 용어를 사용하여 논문의 내용을 완전하고 정확하게 표현하는 것이다. 학술 문헌에 양질의 주제 색인이 되어 있으면 검색에 소요되는 시간을 단축시킬 뿐만 아니고 잠재적인 관계들을 발견하는데 아주 중요한 역할을 한다. MEDLINE 학술지는 1960년대 이후로 현재까지 NLM 색인전문가가 MeSH로 색인하고 있다.

MeSH 검색은 하나의 개념에 여러 가지의 동의어로 사용되어 논문에 표현하는 것을 검색해 준다. 저자가 어떤 어휘를 사용했고, 탐색자가 어떤 단어로 검색을 했는가에 관계없이 개념을 모두 검색해내므로 민감도가 높은 검색 방법이 MeSH 검색이다. 또한 검색어별로 상위어로 확장하거나 하위어로 제한하여 검색할 수 있어 포괄적인 검색을 하면서도 특정적인 용어에 대해서는 특이도가 높은 검색을 할 수 있다[3]. 따라서 MeSH 시소러스 기반인 MeSH Database 검색 툴을 활용해서 PubMed 검색을 하면 MeSH로 문헌이 색인되어 있어 탐색자의 요구에 가장 최적화된 검색결과를 얻을 수 있다.

미국의 NLM에서는 150여명의 색인전문가가 매년 50만 편의 문헌에 색인을 하고 있다[2]. 색인전문가는 문헌을 가장 잘 표현할 수 있는 정확한 MeSH를 부여하기 위하여 문헌에 기술되는 어휘들의 유형과 패턴을 분석하고 MeSH를 갱신하고 발전시키는 전문가이다. 따라서 그 역할이 분명히 필요하고 중요하다. 하지만 색인을 해야 하는 문헌의 양은 폭발적으로 증가하고 있는 반면 정작 색인 전문가를 양성하여 색인하는 데 소요되는 비용과 시간이 부족한 것이 현실이고 문제점이다. PubMed에는 MeSH로 색인이 되지 않은 채 검색이 되는 PubMed Central (PMC) 학술지 논문 500만 여편과 출판 전 온라인으로 제공(Epub of ahead)되는 논문도 함께 제공된다. 따라서 망라적으로 검색을 하려면 자연어 검색도 추가로 해야 하는 데 탐색자가 직접 동의어, 유사어, 상하위어 등을 모두 자연어로 입력해야만 한다. 그리고 색인과 검색에 동일한 MeSH를 사용하고 있어 좀 더 다양한 어휘의 반영이 쉽지 않다.

우리나라에서는 KoreaMed가 MeSH에 의하여 색인되고 검색이 되는 유일한 의학학술문헌 데이터베이스이다. KoreaMed에는 대한의학편집인협의회(의편협)의 국내 의학학술지 평가를 통해 선별된 핵심학술지 260여종(23만여 편)의 의학문헌이 수록되어 있다. 2013년에 KoreaMed MeSH Indexer를 개발하여 MeSH로 자동색인을 하고 있고 MeSH에 의한 검색기능을 제공하고 있다. 이러한 KoreaMed의 중요성을 인지해 근거중심의학(Evidence-Based Medicine, 이하 EBM)의 대표적인 연구기관인 Cochrane에서는 Crowd Sourcing Project일환으로 MeSH를 포함한 검색전략을 사용해서 KoreaMed의 무작위 대조군 연구(Randomized controlled trial) 문헌을 수집해 가고 있다[4,5].

1990년대 시작된 EBM 환경 속에서 의사들은 입수가능한 체계화된 연구로부터 얻은 과학적인 근거를 바탕으로 의사 개인의 임상 경험을 접목시켜 환자를 치료하려는 움직임이 확산되고 있다[6]. 따라서 정보 수집의 가치가 점점 더 높아지고 있고, 의학학술문헌 데이터베이스를 통한 문헌 검색이 급증하고 있으며, 의학 분야 연구자들의 정보 활용능력이 강조되고 있다. 정보 검색 분야

에 있어서도 특정 개념이 항상 같은 용어로 검색할 수 있는 MeSH로 색인하고 검색함으로써 민감도와 특이도를 동시에 높일 수 있도록 하는 강력한 검색기법들이 개발되고 있다[7].

그러나 정작 의학 분야 연구자들은 PubMed 검색 시에 불리언(Boolean) 연산을 이용한 질의어 형성(Query formation)을 가장 많이 사용하고 있다. 하이퍼링크를 활용한 두루살핌도구(Navigation tool)이나 제한검색(Search Limit), 그리고 MeSH 검색 방법을 많이 사용하지 않는다[8]. MeSH 검색의 장점에도 불구하고 MeSH 검색기법을 잘 활용하지 못하고 있어 아쉽다[9]. 따라서 탐색자가 MeSH에 대한 지식이나 검색방법을 모르는 경우라도 검색시스템이 탐색자와 탐색어의 특성을 고려해 맞춤화된 검색결과를 제공해야 한다. 의학 세부 주제 분야에 대해서도 대량의 학술 문헌에서 출현하는 어휘나 구문의 패턴을 텍스트마이닝 기술을 활용하여 학습한 후 검색용 MeSH 시소러스의 어휘군으로 추가하여 검색에 편의성과 효율성을 제공하는 것이 필요하다.

같은 문장 또는 문단에 함께 나타나는 동시출현(co-occurrence) 용어는 의미 있는 관계가 있는 용어들이라고 가정할 수 있다. 더욱이 동시출현 횟수가 많으면 그 관계성은 확실하다고 볼 수 있다. 하지만 같이 동시 출현하는 것으로만 판단하기 때문에 용어간의 직접적, 간접적 관계성을 파악할 수 없는 없다. 따라서 텍스트마이닝을 통해 의학문헌에서 중요 키워드를 자동 추출해내고, 관련 어휘군을 작성하여 MeSH 색인용이 아닌 검색필터(filter)로 추가한 후 탐색자의 질의어에 대하여 자동 수행케 하면 민감도와 특이도가 같이 향상되는 검색결과를 얻을 수 있다.

의학 학술문헌의 검색용으로 MeSH를 활용하면 검색에만 사용하는 탐색도구이므로 색인비용에 대한 부담이 없고 통제어의 장점과 자연어가 가지는 용어의 특정성을 유지할 수 있다. 즉, 동의어, 유사어, 상•하위어 및 관련어로 검색이 가능하고 추가로 자연어로 기술되는 다양한 용어를 반영할 수 있다. 따라서 탐색자의 탐색행위를 지원할 수 있는 검색용 MeSH 시소러스의 적용이 아쉽기만 하다.

단어인접탐색(Word-proximity Search)은 두 개 이상의 탐색어가 한 문장 내에서 근접하여 동시 출현하는 문헌을 탐색하는 방법이다. 탐색자가 단어인접탐색을 하는 것은 탐색어간에 들어갈 수 있는 단어의 수를 지정할 수 있어 불리언 검색을 통해 질의어 형성하는 것보다는 정확한 검색이 된다. 용어간의 관계를 추출하는 경우 문장의 의미와 구문을 파악하여 원하는 정보를 추출해내는 자연어 처리기법(Natural Language Processing)이 검색분야에서 진행되고 있는 연구 방법이지만 여러 문장으로 표현되는 관계정보를 완벽하게 추출하는 데는 아직 어려움이 있다. 따라서 문장 내에서 의미 있는 두 용어 간에 단어인접탐색기법이 유용함을 인지하여 많은 검색시스템이 단어인접탐색기법을 시스템에서 제공하고 탐색자 스스로 단어인접탐색 기법을 잘 활용할 수 있도록 교육하고 안내하는 것이 가장 현실적으로 검색 효율성을 향상시킬 수 있는 방법이다.

1.2 연구의 목적

텍스트마이닝 기술이 발전하여 자동분류하고 패턴을 인식하는 기술이 점차 지능화되고 있지만 결국 보다 자동색인을 통해 검색효율이 높일 수 있는 정교한 시소러스의 개발은 인간의 지적 해석능력 및 전문성에 의존할 수밖에 없다. PubMed에는 의학 학술용어의 특성이 잘 반영되어 있는 대량의 문헌집합으로 본 연구에서는 PubMed를 활용하여 생의학적 개념들 간의 유의미한 관련성을 찾고자 한다. PubMed에서 다양한 어휘들을 추출해내고 이를 패턴화하여 MeSH의 검색용 필터로 반영한다면 MeSH로 색인된 용어와 함께 검색이 이루어져 망라적이면서도 정확한 검색결과를 얻을 수 있을 것이다.

또한 우리나라 의학분야 연구자들이 작성한 논문이 수록된 KoreaMed에는 국내 의학 분야 연구자들만이 독특하게 주로 사용하는 어휘와 구문이 있을 수 있다. 따라서 본 연구에서는 “Stomach Neoplasms”와 관련하여 주로 사용되는 언어 패턴(어휘, 표현 등)을 분석하기 위해 PubMed 문헌을 실험 데이터로 내려 받아 검색의 효율을 높일 수 있는 어휘군을 선정하고자 한다. 그리고 KoreaMed에서 검색되는 MEDLINE 학술지를 대상으로 선정된 어휘군을 검색 필터로 적용해보고, 단어인접탐색 기법을 적용하는 검증 실험을 통해 “Stomach Neoplasms”의 매칭률이 향상되었는지를 분석하고자 한다.

2. 연구 방법

본 연구에서는 우리나라 질병중 사망률이 가장 높은

암(cancer)분야 중에서 위암(이하 “Stomach Neoplasms”)을 대상으로 하여[10] 검색용 MeSH에 추가할 어휘군을 결정하기 위해 실험문헌으로 PubMed에서 중심주제어(Major Topic)가 “Stomach Neoplasms"인 2007년~2016년 논문 8,625편을 내려 받았다. 그리고 논문제목으로부터 Stomach와 Neoplasms 관련 용어의 동시출현여부를 분석한 후 “Stomach Neoplasms”를 의미하는 어휘를 수집하여 검색필터로 사용할 어휘군을 작성하였다. 또한, KoreaMed에서 검색되는 MEDLINE 학술지중 “Stomach Neoplasms”가 색인이 되어있는 277편을 대상으로 MEDLINE MeSH, MeSH on Demand, 그리고 KoreaMed MeSH Indexer의 “Stomach Neoplasms” 색인어 추출여부와 검색용 필터로 어휘군을 적용했을 때, 그리고 동시출현 용어의 단어인접검색 기법을 적용했을 때 “Stomach Neoplasms”의 매칭여부를 비교하여 검증하였다. 본 연구방법을 도식화하면 다음과 같다(Fig. 1).

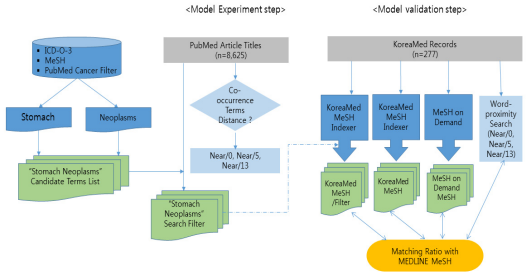

Fig. 1. Experiment and validation process of using MeSH Search Filter and Word-Proximity Search

2.1 “Stomach Neoplasms” 검색필터 작성방법

MeSH 용어에는 의학의 학문적 특성상 합성어(Compound words)가 많다. 즉, 신체의 기관(Organ), 생물체(Organisms), 질병(Diseases), 신생물(Neoplasms), 감염(Infection)관련 용어들이 전조합(Precoordinated)되어 빈번하게 함께 사용된다. 예를 들면 Stomach Diseases, Stomach Neoplasms, Dog Diseases, Staphylococcal Infections처럼 두개 이상의 분리된 단어의 모임이 하나의 단어 구실을 한다. 따라서 인접한 두개 이상의 단어를 각각의 MeSH로 색인하거나 검색하는 것보다 전조합된 용어 즉, 합성어로 색인하거나 검색하는 것이 더 구체적이고 정확하다. “Stomach Neoplasms”을 표현하는 의미 있는 용어로 출현하는 개념은 위(이하 Stomach), 암(이하 Neoplasms) 그리고 전조합된 합성어 “Stomach Neoplasms”이다.

2.1.1 Stomach 어휘군 작성

MeSH는 의학용어를 16개 범주로 분류하고 각 범주 내에서 하위어들이 11개 단계까지 확장하는 계층구조로 되어 있다. 예를 들면 MeSH에서 “Stomach”는 해부(Anatomy) 범주의 소화기시스템(Digestive system)의 하위어로 본문(Cardia)이외에 7개의 하위어로 더 세분화할 수 있다[11]. 또한 Stomach 용어는 Internaitonal Classification of Diseases for Oncology(ICD-O)[12]에서 해부학적 부위(Topography)의 분류번호 C16이다. 이를 비교한 결과는 다음의 Table 1과 같다.

Table 1. Comparison of International Classification of Diseases for Oncology 3rd edition and MeSH 2017 for Stomach

MeSH	MeSH entry term	ICD-O	ICD-O Number
·Stomach	·Stomach	·Stomach	C16.0
·Cardia	·Cardia	·Cardia	C16.0
·Esophageal Sphincter, Lower	·Esophageal Sphincter, Lower ·Lower Esophageal Sphincter	·Cardio-oesophageal junction	
·Esophagogastric Junction	·Esophagogastric Junction ·Gastroesophageal Junction	·Gastro-oesophageal junction	C16.0
		·Oesophagus and stomach	
·Gastric Fundus	·Gastric Fundus	·Fundus of stomach ·Gastric fundus	C16.1
		·Body of stomach	C16.2
·Pyloric Antrum	·Pyloric Antrum ·Antrum, Pyloric ·GastricAntrum	·Pyloric antrum	C16.3
·Pylorus	·Pylorus ·PyloricSphincter	·Pylorus	C16.4

ICD-O: International Classification of Diseases for Oncology

따라서 Stomach 관련 후보 어휘군은 Stomach, Cardia, “Esophageal Sphincter, Lower”, “Gastroesophageal Junction”, "Gastric Fundus", “Pyloric antrum”, Pylorus로 확정하였다.

2.1.2 Neoplasms 어휘군 작성

Neoplasms 어휘군을 작성하는데 참고한 첫 번째 용어집은 MeSH이다. MeSH의 Neoplasms 디스크립터의 기입어에는 Benign Neoplasms, Cancer, Malignancy,

Neoplasia, Neoplasm, Neoplasms, Benign Tumors가 있다. 더 세부적인 Neoplasms와 관련된 조직학적 유형 MeSH는 Neoplasms 범주 하에 Neoplasms by Histologic Type [C04.557]용어에서 참고하였는데 기입어를 포함하여 총 522개 용어이다. 두 번째 용어집은 ICO-O-3로 "Stomach Neoplasms"과 연관된 형태코드는 8000-8152/8154-8231/8243-8245/8250-8576/8940-8950/8980-8990 이다. 세 번째 참고자료원은 NLM과 National Cancer Institute가 Neoplasms 관련 주제 분야의 검색을 위해 공동으로 작성한 Cancer subsets Filter로 745개의 Neoplasms 조직형 용어가 있다[13]. 이 세종류의 용어를 분석한 결과 동시출현용어 중 "Stomach Neoplasms"의 조직학적 형태 용어인 Neoplasms의 후보군은 Cancer, Malignancy, Neoplasia, Neoplasm, Neoplasms, Tumors, Adenocarcinoma, Leiomyosarcoma, Carcinoid, Lymphoma, Schwannoma, Polyps, Myeloma, Plasmacytoma, Carcinoid, Leiomyoma이다.

2.2 Stomach–Neoplasms 동시출현용어의 거리 분석방법

MeSH에는 디스크립터와 기입어 즉, 이형동의어나 유사어, 그리고 여러 개의 단어로 구성된 합성어, 구(phrase)로 이루어진 용어가 있어 디스크립터로 색인되고 기입어까지 모두 검색된다. 예를 들면 "Stomach Neoplasms"의 기입어는 Cancer of Stomach; Cancer of the Stomach; Gastric Cancer; Gastric Cancer, Familial Diffuse; Gastric Neoplasms; Neoplasms, Gastric; Stomach Neoplasms; Stomach Cancer이다. 이러한 MeSH의 기입어는 논문에서 동시출현하는 빈도가 높을 가능성이 있다. 따라서 본 연구에서는 "Stomach Neoplasms" 기입어의 동시출현 빈도를 분석하였다.

"Gastroesophageal Junction"에서 발생하는 암은 식도-위 경계에서 발생하거나 또는 식도-위 경계로부터 5cm인 위에서 발생하여 식도-위 경계를 가로지르는 종양으로 식도 샘암종 TNM 체계를 사용하여 병기 설정하는 바 식도암(Esophageal Neoplasms)으로 색인되어야 한다. 하지만 위식도 역류의 결과로 식도의 샘암종 빈도가 증가하는 관점에서 보면 구분이 쉽지 않다. 분문(Cardia)은 식도이행부(esophagogastric junction)와 바로 맞닿아 있으며 음식물이 식도에서 위로 지나가는 통로로 여기에 생긴 종양은 주로 "Stomach Neoplasms"로 분류한다[14]. 즉, type 2와 3만 "Stomach Neoplasms"이다. MeSH 색인 시에 색인전문가가 정확히 본문을 읽으면서 주의깊게 색인해야 하는 용어로 본 연구에서는 이러한 용어의 패턴도 분석하였다.

"Stomach"와 "Neoplasms"이 한 문장 내에서 동시 출현하는 경우 서로 가까이 위치하는 거리에 따라 검색의 효율성이 달라질 것이다. 동시출현 용어간의 위치를 판별한 후 가중치에 따라 검색결과를 제공하면 검색의 효율성이 높아진다. 또한 초록 전체에서 동시 출현하는 용어를 검색하는 불리안 검색보다는 논문 제목에서의 동시출현 용어가 더 의미 있는 관계성을 나타낼 수 있다. 본 연구에서는 PubMed 실험문헌 8,625편의 논문 제목에서 동시 출현하는 "Stomach Neoplasms" 관련 용어에 대하여 단어 간의 거리(path)를 측정하여 관사, 전치사 등을 포함하여 몇 단어 사이에 두 단어가 출현하는지를 분석하였다.

2.3 검색 효율성 검증

검색 필터로 적용한 어휘군과 "Stomach"와 "Neoplasms" 단어 간에 거리를 분석한 결과를 검증하기 위하여 우리나라 의학 분야 연구자들의 논문이 주로 수록되는 KoreaMed에서 검증문헌을 선정하였다. KoreaMed 학술지이면서 PubMed에서도 검색이 가능한 MEDLINE 학술지 11종이 대상이다. "Stomach Neoplasms"이 MeSH로 주제 색인되어 있는 277편(2007년~ 2016년분)을 KoreaMed에서 내려 받았다.

검증문헌 277편은 Korean Journal of Gastroenterology가 136편(49.1%)으로 가장 많았다. Gut and liver에 37편(13.4%), Journal of Korean Medical Science에 33편(11.9%) 순이었다.

검증문헌 277편은 모두 MEDLINE 학술지 논문으로 NLM의 색인전문가가 부여한 MeSH 색인어(이하 MEDLINE MeSH)가 색인되어 있다. 따라서 277편 논문을 대상으로 KoreaMed의 MeSH 자동 Indexer에서 추출한 MeSH (이하 KoreaMed MeSH) 그리고 NLM이 개발한 MeSH 자동추천 프로그램인 MeSH on Demand[15]에서 "Stomach Neoplasms"를 색인어로 추출했는지를 비교하였다. 전술한 바와 같이 색인어로 "Stomach Neoplasms"이 추출되었다는 것은 통제어 검색을 통해 검색된 문헌의 총 문헌수 중에 포함되었음을 의미한다. 따라서 MEDLINE MeSH와 MeSH on

Demand 그리고 KoreaMed MeSH에서 추출한 MeSH를 비교하여 매칭률을 측정하였다. 이후 검색용 필터로 “Stomach Neoplasms"를 추가한 경우 매칭되는 비율과 동시출현용어의 단어인접탐색 결과 매칭되는 비율을 측정하였다. “Stomach Neoplasms”용어의 추출 전과 후의 검색효율성은 SPSS for Windows (Version 14; SPSS Inc., Chicago, Illinois, USA)를 사용하여 Wilcoxon 검증을 실시하였다.

3. 결과 분석

3.1 “Stomach Neoplasms” 검색필터 분석

PubMed를 통해 수집된 실험문헌 8,625편의 논문 제목은 평균 13.85 단어(최댓값: 44, 최소값: 1)이다. 1,248편(14.46%)에는 논문 제목에 “Stomach" 관련어와 ”Neoplasms” 관련어가 동시에 출현하지 않았음에도 불구하고 MEDLINE MeSH 색인전문가에 의해 “Stomach Neoplasms”이 색인된 경우이다. 7,376편(85.51%)의 경우 명확하게 논문 제목에 “Stomach Neoplasms”을 의미하는 용어가 기술되어 있었다. 이는 초록이나 저자키워드 혹은 본문보다 논문 제목에 그 문헌의 주제를 나타내는 중요 키워드가 집중되어 있음을 의미한다.

MeSH 용어인 “Stomach Neoplasms"와 기입어 12개(단·복수 포함)를 제외하고 973개 유형이 검색필터로 적용할 수 있는 후보 어휘군이었다. Table 2는 빈도 순위별로 15위에 해당하는 후보 어휘군으로 7위인 ”Gastro-enteropancreatic neuroendocrine tumor“는 MeSH의 supplementary concept 용어로 “Intestinal Neoplasms”, “Pancreatic Neoplasms”, “Stomach Neoplasms”, “Neuroendocrine Tumors”를 모두 MeSH로 색인하는 용어이다. 따라서 “Stomach Neoplasms”의 검색필터 어휘군으로 포함하였다. 전술한 바와 같이 “Esophagogastric Junction”에서 발생한 암은 Type 2와 3만이 위에서 발생한 암이다. 분석결과 실험문헌 8,625편중 21개의 유형, 97개(1.12%)의 용어가 “Esophagogastric Junction”에서 발생한 암이었다. 11편(0.13%)만이 Type 1에 해당하여 검색필터로 추가할 어휘군 용어로 채택하였다.

Table 2. Candidate Terms of “Stomach Neoplasms”

Rank	Term Types	Freq
1	Gastric Cancer*	2,780
2	Gastric Adenocarcinoma	921
3	Gastric Carcinoma	757
4	Gastric Lymphoma	78
5	Adenocarcinoma of the Stomach*	70
6	Carcinoma of the Stomach	68
7	Gastroenteropancreatic Neuroendocrine Tumors	63
8	Gastric Carcinoid	60
9	Gastric MALT Lymphoma	57
10	Gastric Metastasis	46
11	Gastric Cancers*	46
12	Gastrointestinal Stromal Tumor of the Stomach	44
13	Gastric Cardia Adenocarcinoma	44
14	Gastric Mucosa-associated Lymphoid Tissue Lymphoma	43
15	Gastroesophageal Adenocarcinoma	39

*MeSH entry terms

출현빈도 순위로는 2,780회 출현한 용어인 “Gastric Cancer”가 1위이고 921회 출현한 “Gastric Adenocarcinoma”가 2위이다. "Gastric Cancer"는 MeSH의 “Stomach Neoplasms”의 기입어이다. 후보 어휘군에는 MeSH 기입어를 포함하여 12개 용어(단·복수 포함)가 MeSH 용어였고 무려 2,894회(34.24%)나 출현하였다(Table 3).

Table 3. MeSH Types and Frequency of Occurrence of “Stomach Neoplasms”

No	MeSH Terms	MeSH Types*	Frequency of Occurrence
1	Stomach Neoplasms	MH	0
2	Cancer of Stomach	ET	3
3	Cancer of the Stomach	ET	7
4	Gastric Cancer	ET	2780
5	Gastric Cancers	ET_P	46
6	Gastric Cancer, Familial Diffuse	ET	0
7	Gastric Neoplasms	ET	24
8	Gastric Neoplasm	ET_S	7
9	Neoplasms, Gastric	ET	0
10	Neoplasms, Stomach	ET	0
11	Stomach Cancer	ET	23
12	Stomach Cancers	ET_P	4

*MH: MeSH Heading, ET: entry term, ET_P: entry term plural-nouns, ET_S: entry term singular-nouns

반면, 1회 출현하는 용어들은 무려 733개(8.50%)의 유형이 있었고, 2회 출현한 용어들은 75개(0.87%) 유형이 있었다(Fig. 2).

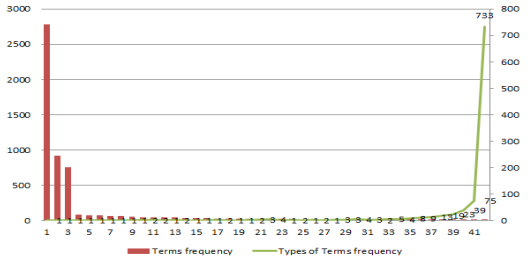

Fig. 2. Frequency of Terms and Frequency of Terms Types

3.2 Stomach-Neoplasms 동시출현용어의 거리분석

동시 출현하는 "Stomach Neoplasms" 용어의 관계성 파악을 위해 해부학적 부위와 조직학적 유형 단어사이에 출현하는 단어 수가 의미가 있는지 PubMed 실험문헌 8,625편의 논문 제목을 대상으로 구문을 분석하였다. 두 단어사이의 평균 거리는 4.62 단어였고 "Stomach"과 "Neoplasms" 관련 용어가 동시 출현한 문헌은 7,376편으로 두 단어의 동시출현 거리는 최소 2단어에서 최대 16단어였다(Fig. 3).

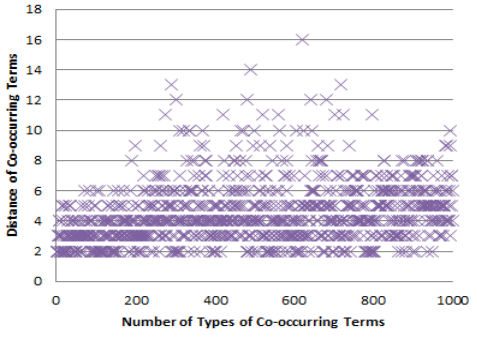

Fig. 3. Distance of Co-occurring Terms and Types of Co-occurring Terms

동시출현 거리가 2단어인 용어는 5,234개(70.95%)로 해부학적 부위 용어와 조직학적 유형 용어의 합성어 조합이었는데 Gastric과 Adenocarcinoma, Carcinoma, Lymphoma, Carcinoid, Cancers, Metastasis의 합성어였다. 검색필터로 작성한 어휘군중 출현빈도가 높은 용어인 "Gastric Cancer", "Gastric Adenocarcinoma" 등과 일치하였다. 동시출현 거리가 3단어인 용어는 203개 유형 762개(10.3%) 용어로 이 중에서 35개 용어가 "Tumor of Stomach"와 같이 전치사 구를 포함한 구문으로 구성되어 있었다.

3.3 검색효율성 분석

3.3.1 검색효율성 측정

KoreaMed 검증문헌 277편은 NLM의 색인전문가에 의해 MEDLINE MeSH로 "Stomach Neoplasms"가 부여된 문헌이다. 논문 제목, 초록과 저자키워드에서 검색용 필터로 적용할 어휘군의 용어가 매칭되는지 그 비율을 측정한 결과 MEDLINE MeSH에는 "Stomach Neoplasms"가 주제어로 부여되었으나 KoreaMed MeSH Indexer MeSH와 매칭이 되지 않은 문헌이 277편중에서 68편(24.5%)이었다. 어휘군을 작성하여 검색용 필터로 적용한 경우 최대 54개 문헌(94.9%)에 "Stomach Neoplasms"이 매칭되었다(Table 4).

Table 4. Matching Ratio of "Stomach Neoplasms" by MeSH Indexing Systems

MeSH Indexing Systems		Match	Non-Match	Matching (%)
MEDLINE MeSH		277	0	100
MeSH on Demand		245	32	88.4
KoreaMed MeSH Indexer		209	68	75.5
	Stomach Neoplasms Search Filter	263	14	94.9
	Co-occurrence Terms	268	9	96.7

3.3.2 "Stomach Neoplasms" 검색필터 및 Stomach-Neoplasms 동시출현 용어의 검색효율성 분석

구체적으로 973개의 어휘를 검색용 필터로 추가하였을 때 매칭되는 비율을 비교하여 측정하였다. 결과 921회(20.5%) 출현한 용어는 "Gastric Adenocarcinoma" 로 한 종류였다. 검색필터로 "Gastric Adenocarcinoma"만 적용해도 KoreaMed Indexer의 매칭률 75.5%였던 것에 비해 1.39% 향상되어 76,89%가 되었다. 후보 어휘군의 50%에 해당하는 12개 용어를 검색필터로 추가하면 83.39%로 검색 효율이 향상되었다. 99개의 용어(75%)의 경우 89.5%로 향상되었다. 973개의 후보 어휘군을

모두 필터로 적용하면 KoreaMed MeSH Indexer에 의한 75.5%의 매칭률이 19.4%나 향상되어 94.9%가 되었다.

탐색자가 단어인접탐색을 하는 경우 동시출현용어의 거리를 입력하게 되는데 검색효율성을 측정하기 위해 Near/0, Near/5 그리고 Near/13의 경우 매칭률을 측정하였다. “Stomach Neoplasms” 관련 용어로 논문의 제목, 초록, 그리고 저자키워드에서 동시출현한 거리를 분석한 결과 Near/0 인 경우 즉, 합성어인 경우 83.8%로, Near/5의 경우 95.3%까지 매칭률이 향상되었다. Near/13인 경우는 검증문헌에서 1편이 있었는데 이 경우 59편의 문헌(96.7%)이 매칭되었다(Table 5).

Table 5. Matching Ratio of “Stomach Neoplasms” Search filter and Co-occurrence Terms by Word-Proximity Search

Types	Number of Words(%)	Match	Non-Match	Matching (%)
“Stomach Neoplasms” Search Filter	1(20.5)	213	64	76.89
	12(50)	231	46	83.39
	99(75)	248	29	89.5
	982(100)	263	14	94.9
Co-occurrence Terms by Word-Proximity Search	Near/0*	232	45	83.8
	Near/5*	264	13	95.3
	Near/13*	268	9	96.7

*Near/(n): n means distance of words.

“Stomach Neoplasms” 검색용 필터에 용어가 없어서 매칭에 실패한 경우는 14편이다. 5편은 초록에서 13단어 내에 “Stomach Neoplasms”관련 용어가 동시 출현하였다. 9편은 초록이 없거나 논문의 제목, 초록, 저자키워드에 “Stomach Neoplasms”관련 근거를 찾을 수 없는 경우였다. 초록이 없는 경우가 3편, 6편은 초록에서도 관련 용어가 동시출현하지 않는 경우였다.

3.3.3 검색효율성 검증

KoreaMed Indexer MeSH와 “Stomach Neoplasms” 검색용 필터를 적용했을 때의 매칭률이 유의적인 차이가 있는지 알아보기 위하여 비모수통계방법인 Wilcoxon 검증을 실시하였다. 결과 통계적으로 유의한 차이가 있었다($Z = 7.5077$, $p < 0.0001$). 4분위로 나누어 검색필터로 적용한 경우 1개 용어(20.5%)를 적용한 경우 매칭률은 76.89% 향상되었지만 통계적으로 유의한 차이가 있다고 볼 수 없었다. 그러나 50%에서 100%로 어휘군을 필터로 추가하는 경우 통계적으로 유의한 차이가 있었다. 따라서 검색시스템의 성능과 추가할 어휘군의 규모와 특성에 따라 50 ~ 75%의 어휘군을 추가하는 것을 제안한다.

동시출현 용어에 대하여 Near/0, Near/5, Near/13의 단어인접탐색 기법을 적용한 경우 모두 통계적으로 유의한 차이가 있었다(Table 6).

Table 6. “Stomach Neoplasms” Search filter and Co-occurrence Terms by Word-Proximity Search Test: Results of the Significance Tests by Applying Wilcoxon Test

	No of Terms/ Distance of words	Test Statistics	Test result
		Z(W+)	P= 0.05
“Stomach Neoplasms” Search Filter	1	0.5263	P>0.1
	12	3.0534***	P<0.001
	99	8.0258***	P<0.001
	982	7.3529***	P<0.0001
Co-occurrence Terms by Word-Proximity Search	Near/0	3.2121**	P<0.001
	Near/5	7.6625***	P<0.0001
	Near/13	8.2043***	P<0.0001

4. 고찰

MeSH에 관한 연구는 특정 개념이 항상 같은 용어에 의해 색인되어 있으면 필요한 문헌을 정확하고 누락 없이 검색할 수 있어 검색의 효율성이 증대되고 문헌 분석에 소요되는 시간도 단축시킬 수 있다는 점을 강조해 MeSH를 활용한 자동색인에 대한 연구가 주를 이루어왔다. 문헌에 부여된 MeSH와 연동될수록 개체 간에 관계가 명확하다는 것을 증명하여 Subject-heading weight를 정의한 연구[16]는 MEDLINE 학술지에 부여된 MeSH 주제 색인어를 활용하여 텍스트마이닝을 연구한 대표적인 사례이다. 의학학술문헌의 논문 제목과 초록에서 후보 주제어를 추출하여 MeSH 색인어와 저자 주제어의 연관성을 분석하면 MeSH를 함께 활용했을 때 저자 주제어와 연관성이 더 높다[17]. MeSH에 기반하여 주제색인을 하면 정확성과 망라성이 높은 장점이 있다. 이를 활용하여 MeSH는 자동색인 및 자동분류를 위한 기계학습에 적용되고 있다.

NLM은 2002년 자동색인기인 Medical Text Indexer (MTI)를 개발하였다. 논문의 제목과 초록을 대상으로 MetaMap과 Trigram으로 추출한 용어를 Unified Medical language System (UMLS)의 용어와 매칭하여

MeSH로 변환하는 과정이 핵심이다. 거기에 PubMed의 유사주제문헌(PubMed Related Citations) 알고리즘을 통해 추출한 의학용어의 집합을 추가해 MeSH를 추출한다. 이 과정을 거쳐 추천된 MeSH를 NLM의 색인전문가가 참고하여 데이터 관리프로그램인 Data Creation Maintenance System(DCMS) 상에서 논문을 읽으면서 일일이 MeSH로 색인한다[3]. 최근 색인전문가가 1차적으로 MTI를 통해 자동 추출된 색인어를 참고 (MTI as the First-Line Indexer)하여 색인을 하는 반자동색인시스템으로 색인방법을 변경하였다. 현재 NLM 색인전문가가 전체 1일 색인 하는 양의 2,500편(58%)정도가 MTI 추천용어를 참고하여 색인되고 있다[18,19].

우리나라의 경우 KoreaMed와 Synapse 데이터베이스가 유일하게 MeSH에 의한 색인과 검색이 이루어지는 의학학술 문헌 데이터베이스이다. KoreaMed MeSH 반자동 색인 시스템은 Korea Med 레코드의 논문 제목, 초록, 저자키워드에 출현하는 용어에 대하여 토큰을 생성한 후 MeSH의 기입어와 일치하는 경우 MeSH를 부여하는 자동 MeSH 매핑 색인 시스템이다. 또한, 자동으로 용어를 매핑한 후 MeSH 파싱 리스트에 MeSH 표목을 삭제 또는 추가하는 5가지 유형의 어휘집을 필터로 적용하여 보다 정확한 MeSH 색인이 되도록 정제하는 기능을 갖춘 반자동 색인 시스템이다[3].

일본의 경우 의학논문 데이터베이스인 의중지 웹(이하 ICHUSI Web)[20]에서도 MeSH에 기초하여 색인전문가에 의한 MeSH 색인을 하고 있고 CHUSI Web에서 MeSH에 의한 검색이 가능하다.

PubMed의 MeSH Database, KoreaMed, CHUSI Web 모두 MeSH 색인과 검색에 동일한 MeSH를 적용하고 있다. MEDLINE MeSH나 ICHUSI Web처럼 색인전문가에 의해 색인된 후 검색하는 경우라면 색인을 하는데 소요되는 시간이 길다는 단점에도 불구하고 동일한 시소러스를 색인과 검색에 사용하면 된다. ICHUSI Web의 경우 영문 논문의 경우 MeSH 색인에 드는 비용이 5,000원이고 미국 색인전문가의 급여가 약 GS 9($42,000) ~ GS 12($61,000)[21]이다. 또한 1명의 색인전문가를 양성하는 비용과 노력이 쉬운 것은 아니다.

텍스트마이닝 기법을 활용하여 특허의 특정 패턴을 찾아내고 내용기반에 근거한 특허정보검색시스템을 개발한 연구[22]는 MeSH 주제색인처럼 국제특허분류코드(IPC)를 활용하여 검색식을 작성한 후 대량의 특허문서 집합을 검색해내는 연구이다. 이처럼 우리나라 의학 학술 문헌 데이터베이스에서도 동의어 및 유사어와 계층언어 및 관련어의 통제어의 장점과 자연어가 가지는 용어의 다양성을 반영할 수 있는 검색용 MeSH 시소러스의 개발과 활용이 절실한 상황이다.

PubMed의 MeSH Database에서는 MeSH 검색은 탐색자가 검색어를 입력한 후 관련 MeSH를 선택하여 체크하면 탐색식 형성창(Search Builder)에 질의어가 자동생성되고 곧바로 PubMed와 연동하여 검색결과를 얻을 수 있다. 통제어인 MeSH로 검색을 하는 것은 MeSH로 색인되어 있는 레코드를 찾아내는 작업으로 탐색자가 MeSH의 기입어로 검색을 해도 MeSH 디스크립터와 해당 기입어 그리고 그 하위개념까지를 모두 검색해낸다. 따라서 체계적 문헌고찰(Systematic Reviews)이나 임상가이드라인 등의 연구를 위한 민감도(Sensitivity) 높은 검색에 매우 효율적이다. 하지만 PubMed에는 MeSH로 색인되지 않는 문헌도 존재하기 때문에 망라적으로 검색하려면 자연어 검색도 같이 수행해야한다.

“Stomach Cancer”와 같은 합성어를 검색하는 경우 탐색자는 두 개념 즉, “Stomach” 관련 용어와 “Cancer” 관련 용어 두 단어를 AND 조합하여 검색한다. “Stomach Cancer”를 의미하는 용어는 어휘나 구문 등의 형태로 다양하게 표현되기 때문에 “Stomach Cancer” 관련 합성어를 모두 검색해내려면 “Stomach”라는 기관명과 MeSH 용어인 “Neoplasms”의 조직학적 형태의 용어까지를 포함하여 입력하는 것이 필요하다. 1986년 의료분야에서 사용되는 각종 용어체계를 공통개념으로 통합하기 위해 구축한 150여개의 시소러스 집합체인 UMLS에서 “Stomach Neoplasms”을 나타내는 용어는 총 64개로 다양한 유사어와 구문들이 사용되고 있음을 알 수 있다[23].

종양을 주제로 다룬 문헌에서는 해부학적 부위와 조직형 용어를 합성하여 “Gastric Adenocarcinoma”와 같이 표현한다. 혹은 정관사나 부정관사를 포함한 전치사구 등의 어구로 “Adenocarcinoma of the Stomach”와 같이 신생물을 표현하기도 한다. 즉, 학술문헌에서 저자가 서술하는 용어들은 사실상 그 어휘나 구문상의 유형이 MeSH나 UMLS에서 제공되는 기입어보다 더 다양하다. MeSH 색인 원칙에 의하면 신생물의 색인은 신생물의 발생부위에 대하여 해부학적 부위 명에 "Neoplasms"을 붙여 합성어로 부여하고 종양 자체의 특성을 신생물의

조직형 용어로 추가하여 색인하도록 되어 있다. 즉, 위암종(Gastric carcinoma)의 경우는 "Stomach Neoplasms"과 "Carcinoma"로 2개의 MeSH를 색인하는 것이 원칙이다. 그러나 자동색인의 경우 논문의 제목, 초록, 저자 키워드에 기입어이외의 용어로 기술되어 있으면 매칭이 일어나지 않으므로 “Carcinoma of the Stomach”의 경우 “Stomach Neoplasms”이 아닌 “Carcinoma”와 “Stomach"로 색인되고 검색된다. 탐색자가 “Carcinoma of the Stomach”를 입력해야만 검색된다. 문헌에 “Stomach Neoplasms”을 의미하는 단어가 없거나 문장에서 어구로 표현한 경우 “Stomach Neoplasms”으로 색인되지 않을 뿐만 아니라 검색도 되지 않는다.

본 연구 결과 MeSH의 디스크립터인 “Stomach Neoplasms”은 정작 실험문헌인 PubMed의 논문 제목에서 한 번도 출현하지 않은 용어였다. 또한 기입어인 “Cancer of Stomach”나 “Stomach Cancer”, “Gastric Neoplasms”은 복수형을 포함하고도 논문 제목에서 출현하는 빈도순이 높지 않았다. 오히려 논문 제목에 종양의 해부학적 발현 부위인 Stomach와 Neoplasms의 조직형 용어가 결합된 합성어인 “Gastric Adenocarcinoma”, “Gastric Carcinoma”, “Gastric Lymphoma” 등이 “Stomach Neoplasms”보다 더 유의미하게 사용됨을 알 수 있었다. 혹은 “Adenocarcinoma of the stomach”나 “Carcinoma of the stomach”와 같이 구문으로 기술되고 있었다. “Gastric Cancer”는 MeSH의 기입어이지만 “Gastric Tumor”는 MeSH의 기입어가 아니다. MeSH의 색인은 통제어를 사용하기 때문에 실제 학술논문에서 사용하는 용어들의 반영이 늦을 수밖에 없는데 자연어 검색 시 탐색자가 “Gastric Tumor”를 검색어로 입력하지 않으면 검색에 누락될 수밖에 없다.

본 연구결과 의학 학술문헌의 검색용으로 MeSH를 사용하고 특정 어휘군을 검색필터로 적용함으로써 색인비용에 대한 부담이 없이 통제어의 장점과 자연어가 가지는 용어의 특정성을 유지할 수 있음을 확인하였다. 통제어인 MeSH가 동의어 및 유사어로 사용할 수 있는 어휘가 제한되어 용어선택에 대한 융통성이 없고, 새로 생성된 개념에 대한 반영이 늦다는 단점을 극복하고 데이터베이스에서 제공하는 MeSH를 검색용 필터로 활용하면 동의어, 유사 동의어, 동형이의어를 통제하거나 관련어들을 연결시켜줌으로써 잘못된 검색어 조합이나 애매한 용어관계로 인해 발생하는 검색실패를 줄일 수 있음을 확인하였다.

본 연구 결과 검색용 MeSH에 추가 필터를 활용하여 MeSH 검색의 효과를 극대화하였다. 특정 개념을 표현하는 용어간의 의미 있는 관계를 파악하여 어휘군을 작성한 후 검색용 MeSH 필터로 적용하였더니 KoreaMed MeSH Indexer를 통해 색인어로 추출된 용어(75.5%)보다 좀 더 정확하고 많이 검색(94.9%)되어 효율성이 향상되었다.

본 연구결과 단어인접탐색 기법에 적용한 두 단어사이 거리는 Near/0 즉, 합성어가 KoreaMed MeSH Indexer에 의한 매칭률 75.5%에 비해 83.39%로 검색 효율성이 향상되었다. Near/0이 Near/13보다 정확한 검색결과를 얻을 수 있지만 “Stomach Neoplasms"를 의미하는 많은 구문패턴의 논문이 검색에서 누락될 수 있다. 두 단어사이에 13 단어를 허용하는 Near/13은 망라적으로 검색은 되지만 탐색자가 원하는 정확한 검색은 아닐 수 있다. 또한 문장에서 단어의 위치 정보를 함께 색인파일에 구축하는 것은 검색시스템의 성능을 떨어트릴 수 있다. 하지만 초록 전체에서 AND로 조합하여 검색하는 경우보다는 정확한 결과를 얻을 수 있다는 장점이 있다. 탐색자가 검색하고자 하는 주제 분야의 유형과 주제를 표현하는 용어들의 표현 패턴들을 파악하여 적절한 단어인접탐색기법을 적용할 필요가 있다.

대량의 학술 문헌에서 출현하는 어휘나 구문의 패턴을 텍스트마이닝 기술을 활용하여 분석한 후 검색필터로 적용했을 때 MeSH 시소러스에 의한 검색보다 검색의 효율성이 향상되었다. 또한 적용할 두 단어사이 거리를 측정하여 의학문헌에 사용되는 어휘나 구문적 특징을 분석한 결과 동시 출현 거리가 2단어인 합성어가 실험문헌 8,625편중에서 5,234(70.95%)개나 차지하므로 불리언 검색보다 단어인접탐색 기법이 보다 특이도가 높은 정확한 검색임을 증명하였다.

5. 결론

본 연구에서는 의학 분야에 대표적인 시소러스인 MeSH로 색인한 의학문헌에 대하여 “Stomach Neoplasms"관련 어휘군을 작성하여 검색필터로 적용하였다. 또한 단어사이 이내에 순서에 상관없이 동시 출현하는 문헌을 검색하는 단어인접탐색기법을 활용하였다.

그 결과 매칭률이 향상됨을 증명하였다.

본 연구에서 제안한 검색필터 작성방법은 의학 분야의 용어 특징인 전조합용어의 어휘군 작성에 적용할 수 있다. 또한 탐색자가 자연어 검색 시에 검색어로 활용할 수도 있다. KoreaMed에서 자동색인을 통해 추출된 MeSH 색인어와 특정 주제관련 검색필터를 활용하여 탐색자가 스스로 검색한다면 재현율과 정확률을 향상시키는 최적의 검색이 될 것이다.

탐색자는 불리안 검색으로 동시 출현한 용어를 검색하는 것보다 검색시스템에서 제공하는 단어인접탐색기법을 활용하면 보다 더 만족스러운 결과를 얻을 수 있다. 따라서 탐색자는 검색에 소요되는 시간과 노력을 줄이고 검색결과에 대한 만족도를 높이기 위하여 데이터베이스가 통제어색인과 검색을 지원하고 있는 시스템인지의 여부를 확인하고 검색 시에 사용할 수 있는 검색기법의 종류와 방법 등을 인지한 후 검색을 실행해야 한다.

또한, 논문 작성 시 저자가 사용하는 용어에 의해 검색에서 누락될 수 있음을 인지하여 논문 제목, 초록 그리고 저자키워드의 경우 신중하게 용어를 선택하여 사용하고 오탈자가 없도록 논문작성 시 주의를 기울여야 할 필요가 있다.

References

[1] S. L. De Groote, M. Schultz, D. D. Blecic, "Information-seeking behavior and the use of online resources: a snapshot of current health sciences faculty", Journal of the Medical Library Association, vol. 102, no. 3, p. 169, 2014.
DOI: https://doi.org/10.3163/1536-5050.102.3.006

[2] US National Library Medicine. Fact Sheet Bibliographic Services Division,(BSD) 2017. [cited 2017 Mar 2], Available From: https://www.nlm.nih.gov/archive/20050322/pubs/factsheets/bsd.html.(accessed Mar., 31, 2017)

[3] S. N. Jeong, C. S. Lee, "MeSH Semi Indexing of the Korean Biomedical Literature, using NLM Medical Text Indexer", in, *Korea Society for Information Management*, pp. 21-28, 2010.

[4] Cochrane Library. How CENTRAL is created [cited 2017 Mar 31], Available From: http://www.cochranelibrary.com/help/central-help.html.(accessed Mar., 31, 2017)

[5] Cochrane Library. Cochrane Crowd [cited 2017 Mar 31], Available From: http://crowd.cochrane. org/index.html. (accessed Mar., 31, 2017)

[6] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, W. S. Richardson, "Evidence based medicine: what it is and what it isn't", BMJ, vol. 312, no. 7023, pp. 71-72, 1996.
DOI: https://doi.org/10.1136/bmj.312.7023.71

[7] C. S. Lee, "Medical Database Search", Journal of the Korean Medical Association, vol. 53, no. 8, pp. 668-686, 2010.
DOI: https://doi.org/10.5124/jkma.2010.53.8.668

[8] M. Macedo-Rouet, J. F. Rouet, C. Ros, N. Vibert, "How do scientists select articles in the PubMed database? An empirical study of criteria and strategies", Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, vol. 62, no. 2, pp. 63-72, 2012.
DOI: https://doi.org/10.1016/j.erap.2012.01.003

[9] N. Baumann, "How to use the medical subject headings (MeSH)", International Journal of Clinical Practice, vol. 70, no. 2, pp. 171-174, 2016.
DOI: https://doi.org/10.1111/ijcp.12767

[10] Korean Statistical Information System National Statistical Office. Cancer occurrence and death status. 2017 [cited 2017 Mar 2], Available From: http://kosis.nso.go.kr.(accessed Mar. 31, 2017)

[11] US National Library of Medicine. Medical Subject Headings 2017. Available From: https://meshb.nlm.nih.gov/#/fieldSearch. (accessed Mar., 31, 2017)

[12] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D. M. Parkin, S. Whelan, International classification of diseases for oncology, World Health Organization, 2000.

[13] US National Library Medicine. Search Strategy Used to Create the Cancer Subset on PubMed. 2017 [cited 2017 Mar 2], Available From: https://www.nlm.nih.gov/bsd/pubmed_subsets/cancer_strategy.html.(accessed Mar., 31, 2017)

[14] C. C. Compton, D. R. Byrd, J. Garcia-Aguilar, S. H. Kurtzman, A. Olawaiye, M. K. Washington, "AJCC cancer staging atlas", pp. 143-153, Springer, New York, 2012.
DOI: https://doi.org/10.1007/978-1-4614-2080-4

[15] US National Library of Medicine. MeSH on Demand. Available From: https://www.nlm.nih.gov/mesh/MeSHonDemand.html.(accessed Mar., 31, 2017)

[16] D. R. Swanson, N. R. Smalheiser, V. I. Torvik, "Ranking indirect connections in literature based discovery: The role of medical subject headings," Journal of the American Society for Information Science and Technology, vol. 57, no. 11, pp. 1427-1439, 2006.
DOI: https://doi.org/10.1002/asi.20438

[17] S. Y. Bong, K. B. Hwang, "A Method for Author Keyphrase Recommendation for Bioinformatics Papers Using Assigned MeSH Terms", *The HCI Society of Korea*, pp. 236-238, 2011.

[18] J. G. Mork, A. J. Jimeno-Yepes, A. R. Aronson, "The NLM Medical Text Indexer System for Indexing Biomedical Literature", in *BioASQ@ CLEF*, 2013.

[19] A. Jimeno-Yepes, J. G. Mork, D. Demner-Fushman, A. R. Aronson, "A one-size-fits-all indexing method does not exist: automatic selection based on meta-learning", Journal of Computing Science and Engineering, vol. 6,

no. 2, pp. 151-160, 2012.
DOI: https://doi.org/10.5626/JCSE.2012.6.2.151

[20] ICHUSI Web. 2017 [cited 2017 Mar 2], : Available From http://www.jamas.or.jp/index.html. (accessed Mar. 31, 2017)

[21] US National Library Medicine. How can I become an indexer? 2017 [cited 2017 Mar 2], Available From: https://www.nlm.nih.gov/bsd/indexfaq.html# translator.(accessed Mar., 31, 2017)

[22] G. S. Go, W. K. Jung, Y. G. Shin, S. S. Park, "A Study on development of patent information retrieval using textmining", Journal of the Korean Academia-Industrial cooperation Society, vol. 12, no. 8, pp. 3677-3688, 2011. DOI: http://doi.org/10.5762/KAIS.2011.12.8.3677

[23] US National Library of Medicine. Unified Medical Language System (UMLS). Available From: https://www.nlm.nih.gov/research/umls/ index.html. (accessed Mar., 31, 2017)

정 소 나(So-Na Jeong) **[정회원]**

- 1994년 8월 : 숙명여자대학교 문헌정보학과 (문헌정보학 석사)
- 2013년 8월 : 숙명여자대학교 문헌정보학과 (문헌정보학과 박사)
- 1992년 6월 ~ 현재 : 가톨릭대학교 성의교정 도서관 사서

<관심분야>
의료정보, 문헌정보

정 지 나(Ji-Na Jeong) **[정회원]**

- 1989년 2월 : 연세대학교 간호학과 (석사)
- 2013년 2월 : 원광대학교 보건행정학과 (보건학 박사)
- 2015년 9월 ~ 현재 : 전주대학교 보건관리학과 교수

<관심분야>
보건행정, 의료정보