인용 정보를 고려한 미발견 공공 지식 추출: Swanson의 ABC 모델 재현 및 확장

Detection of Hidden Knowledge Using a Citation-Based Approach Based on Swanson's ABC Model

함정은 (Jung Eun Hahm)*
송 민 (Min Song)**

초 록

많은 연구들 가운데 살펴볼 가치가 있는 대상을 찾아 제시해주는 문헌기반 발견의 접근법은 연구자들에게 매우 유용할 것이다. 문헌기반 발견 연구의 대표 이론인 Swanson의 ABC 모델은 기존에 검증되지 않은 개체들의 관계를 연구할 것을 제안해 준다. 본 연구는 Swanson의 ABC 모델에 인용 정보를 고려하여 유의한 관계에 있는 개체들을 더 효율적으로 찾아내고자 하였다. 수집 논문들의 참고문헌 목록에서 인용 정보를 확인하고 논문의 표제와 초록을 대상으로 텍스트 마이닝 기법으로 중요한 단어들을 추출하였다. Swanson의 연구들 중 어유와 레이노드 질병 및 증상의 관계를 재현하였으며 기존의 접근법으로 확인되는 개체들과 어떤 차이가 있는지 분석하였다.

ABSTRACT

It is useful to find something valuable for researching through literature based discovery. Swanson's ABC model, known as literature based discovery, suggests the relationship between entities undiscovered yet. This study tries to find the valid relationship between entities by referring to citation which connects articles on similar topic. We collect citation from references in articles, and extract important concepts in titles and abstracts through text mining techniques. We reproduce the relationship between fish oil and Raynaud's disease, which is known as one of Swanson's works, and compare the results with entities identified from traditional approach.

키워드: 텍스트마이닝, 개체계량학, 문헌기반 발견, ABC 모델
text mining, entitymetrics, literature based discovery, ABC model

* 연세대학교 문헌정보학과 대학원(jungeunhahm@yonsei.ac.kr) (제1저자)
** 연세대학교 문헌정보학과 부교수(min.song@yonsei.ac.kr) (교신저자)

■ 논문접수일자 : 2015년 5월 27일 ■ 최초심사일자 : 2015년 5월 27일 ■ 게재확정일자 : 2015년 6월 11일
■ 정보관리학회지, 32(2), 87-103, 2015. [http://dx.doi.org/10.3743/KOSIM.2015.32.2.087]

1. 서 론

1.1 연구의 배경 및 목적

빅 데이터 시대인 지금은 넘쳐나는 데이터로 인해 연구자들이 학문의 변화 속도를 따라가기가 갈수록 어려워지고 있다. 연구자들은 학문 활동으로 사회에 공헌할 사명을 가졌으며 많은 문헌들을 읽으면서 이미 진행된 연구들 사이에서 새로운 지식을 찾아야 한다. 그동안 연구자들의 개별 노력으로 이루어진 이러한 연구 활동이 데이터 증가 추세를 감안하면 앞으로는 더욱 어려워질 것으로 예상된다. 이러한 상황에서 앞으로 정보학이 할 수 있는 일 중 하나는 문헌에서 아직 확인되지는 않았지만 연구할만한 가치가 있는 개념들을 찾아서 가설 설정을 위한 정보를 제공해주는 것이라 판단된다.

문헌기반 발견(LBD: Literature Based Discovery) 연구는 이러한 맥락에서 중요하다. 아직 증명되지는 않았지만 유의한 관계를 형성하는 것으로 추정되고 있는 개념들을 실제 실험이 아니라 관련 문헌들을 살펴보는 것을 통해 찾아내는 것으로 실험 과정에서 소모되는 자원들을 아낄 수 있어 비용 효과적이다. 특히 비정형적인 텍스트에서 중요한 개념들을 추출할 때 사용되는 텍스트마이닝 기법은 앞으로 계속 발전할 가능성이 크다.

본 연구에서 참조하는 Swanson의 ABC 모델은 문헌기반 발견의 대표적 이론인데 그의 기본 가정은 A 개념과 B 개념의 관계를 연구한 문헌 집합과 B 개념과 C 개념의 관계를 연구한 문헌 집합이 존재한다면 A 개념과 C 개념의 관계를 연구한 문헌 집합이 존재하지 않더라도 A 개념과 C 개념 사이에는 검증될만한 관계가 형성되어 있을 수 있다는 것으로 여러 개체들의 상관관계가 ABC 모델의 접근 방법으로 밝혀지고 있다.

본 연구는 인용 정보를 고려한 Ding et al(2013)의 개체계량학(Entitymetrics) 이론에 근거해 Swanson의 연구를 새롭게 재현하고자 한다. 논문에서 인용은 대부분이 자신의 의견을 뒷받침하기 위해 유사한 다른 연구들이 존재한다는 것을 밝히기 위한 의도로 이루어진다. 인용 정보를 활용하면 문헌에 등장하는 개체들의 유사성이 고려되어 관련 있는 개체들의 관계를 발견하는 것이 훨씬 용이할 것이다.

2. 이론적 배경

2.1 개체계량학

개체계량학 이론은 Ding, Song, Han, Yu, Yan, Lin, Chambers(2013)의 논문에 처음 소개되었는데 기본 가정은 논문 A가 논문 B를 인용한다면 논문 A에 등장하는 개체들은 논문 B에 등장하는 개체들과 인용 정보가 고려되어 더 의미 있는 관계일 것이라는 생각이다. PubMed 및 PubMed Central(PMC) 데이터베이스에서 당뇨병을 치료하는 물질인 메타포민(Metformin)과 관련된 논문들을 대상으로 인용 정보를 고려하여 관련 개체들의 네트워크를 형성하였다. 네트워크를 분석하기 위해 여러 중심성 지수를 근거로 20개의 상위 개체 쌍을 추출하였다. 성능 평가를 위해 CTD 데이터베이스를 참고하여 메타포민과 관련된 것으로 추측되는 질

병 및 유전자 개체들이 실제로 유의미한지 확인하였고 메타포민과 관련된 다른 치료물질에 대한 정보는 drug-interaction checker를 참고하였다.

Song, Han, Kim, Ding, Chambers(2013)는 개체계량학 이론을 기반으로 인용 정보를 고려하여 암과 관련된 유전자 개체들의 네트워크를 형성하였다. 인용 정보를 고려해서 논문에서 등장한 개체들로 형성한 네트워크를 단순 동시출현 기반 개체들이 형성하는 네트워크와 비교하였는데 다양한 네트워크 중심성 지수를 근거로 25개의 상위 유전자 쌍을 추출하고 BioGRID를 참고하여 유의미한 관계를 확인하였다. 결과적으로 인용 정보가 고려된 네트워크는 전통적인 방법으로 형성한 네트워크보다 의미 있는 것으로 확인되는 관계의 수는 적었지만 더 많은 유전자 쌍을 발견할 수 있었다는 것에 의미가 있었으며 두 네트워크에서 도출된 유전자 쌍을 결합하니 전통적인 방법에서 발굴한 것보다 유의하다고 밝혀진 유전자 쌍이 더 많이 확인되었다.

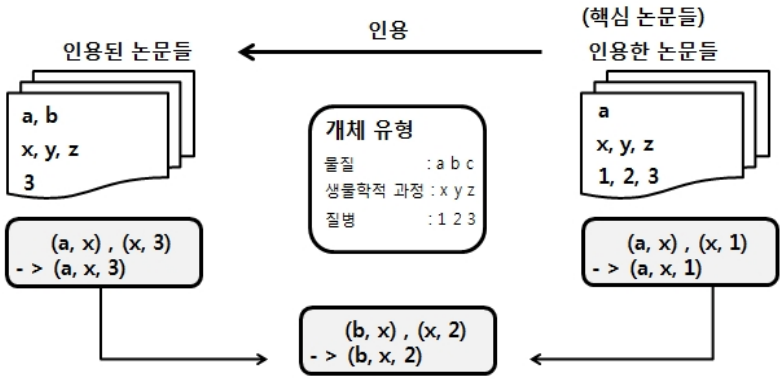
Yu, Ding, Song, Song, Liud, Zhange(2015)는 개체계량학 이론에 근거해 생의학 분야 데이터베이스 간의 영향력을 측정하려고 데이터베이스 링크 네트워크를 만들어 분석하였다. 수집한 20,861개의 논문과 참고문헌에서 확인된 논문의 인용정보를 고려하여 1,512개의 데이터베이스가 만들어낸 네트워크에서 위상 정보와 주 경로(main path) 분석으로 중요한 데이터베이스를 확인하였다.

3. 연구 설계

3.1 이론적 가정

인용 정보가 고려된 ABC 모델에서 기본 가정은 다음과 같다. A 논문이 B 논문을 인용하는 관계이고 동시에 두 논문이 공통적으로 다루는 개체 X 용어가 존재한다면 A 논문에서 X 용어와 직접적인 관련이 있는 개체들은 B 논문에서 X 용어와 직접적인 관련이 있는 다른 개체들과 인용 관계가 고려되어 더 의미 있는 관계일 가능성이 높을 것이다. 본 연구에서는 개체계량학 이론을 적용한 ABC 모델이 전통적인 동시출현 기반 접근법과 비교할 때 어떤 차이가 있는지 확인하기 위해 같은 논문을 대상으로 두 가지 기법을 적용한 결과를 비교하고자 한다.

〈그림 1〉과 같이 우측에 특정 주제어와 관련된 핵심 논문들이 존재하고 좌측에 핵심 논문들이 인용한 다른 논문들이 존재한다. 기존의 ABC 모델의 접근법으로 단어 쌍을 형성하는 방법은 인용 정보를 고려하지 않기 때문에 핵심 논문들이나 그것들이 인용한 다른 논문들의 표제 또는 초록에서 개체들이 동시에 출현한다면 유의미한 관계로 보고 하나의 단어 쌍을 형성한다. 이렇게 형성된 단어 쌍들은 어떤 두 개의 단어 쌍이 특정 개체를 공유하고 있다면 그것을 매개로 삼중의 단어 쌍을 생성한다. 본 실험에서 다루는 삼중의 단어 쌍은 “물질-생물학적 과정-질병”의 패턴을 이루는 것으로 제한하여 Swanson의 연구와 관련이 있는 개체들만을 분석 대상으로 한다. 한편 인용 정보를 고려하는 개체계량학 이론을 접목한 ABC 모델은 표제나 초록에서 개체들이 동시에 출현하고 특정

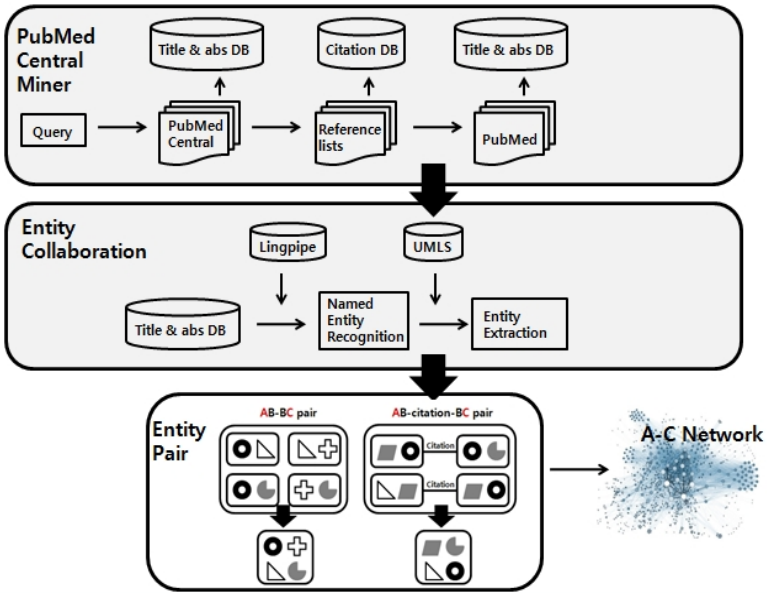


〈그림 1〉 동시출현과 인용정보 기반 단어 쌍 형성과정

개체가 중복되는 두 개의 단어 쌍이 존재하더라도 그 두 단어 쌍이 인용 관계로 연결되어 있지 않다면 삼중의 단어 쌍은 생성될 수 없다. 예를 들어 〈그림 2〉에서 전통적인 ABC 모델 기반으로 생성된 삼중의 단어 쌍은 우측에서 (a,x)와 (x,1) 두 개의 단어 쌍으로 연결되는 (a,x,1)이다. 좌측의 (a,x,3)도 같은 원리로 형성된다. 그러나 개체계량학 이론을 적용한 ABC 모델은 좌측에서 등장한 (b,x)와 우측에서 등장한 (x,2)에 인용 관계가 존재해야 그림 하단에서 보이는 것처럼 (b,x,2)로 삼중의 단어 쌍을 형성할 수 있다. 이와 같이 두 가지 접근법으로 삼중의 단어 쌍을 형성하고 나면 개체들을 매개하는 역할을 하는 B 개념으로 어떤 것들이 있는지를 분석하고 Swanson의 연구에서 발견된 관계가 두 접근법에서 어떻게 다르게 확인되는지 차이를 비교해볼 예정이다.

3.2 시스템 개요

본 연구는 〈그림 2〉와 같은 과정으로 진행된다. 첫 번째로 분석대상 데이터의 수집을 위해 생명과학(life science) 및 생의학(biomedical) 분야 논문들의 전문(Full text)을 제공하는 PubMed Central 데이터베이스(http://www.ncbi.nlm.nih.gov/pmc/)에서 특정 주제어를 질의로 날려 관련된 것으로 검색된 논문들의 원문을 XML 형식으로 다운받는다. 다음으로 다운받은 핵심 논문들(Seed articles)에서 인용 정보를 수집하기 위해 참고문헌 목록을 분석하여 핵심 논문들이 인용한 논문들의 서지 정보를 확인하고 이들을 인용 정보를 수록하는 데이터베이스에 저장한다. 인용 관계가 확인된 논문들은 추가적으로 PubMed 데이터베이스에서 표제로 검색하여 표제 및 초록 정보를 가져와 핵심 논문들과 함께 표제 및 초록 정보를 수록하는 데이터베이스에 저장한다. 이 때 PubMed Central 데이터베이스가 아닌 PubMed 데이터베이스에서 데이터를 수집하는 이유는 전문(full-text) 수준의 데이터를 제공하는 PubMed Central 데이터베이스보다 초록 수준의 데이터를 제공하는 PubMed 데이터베이스가 더 많은 데이터를 보유하고 있어 자료 수집이 용이하기 때문이다.



〈그림 2〉 실험 개요

다음으로 생의학 개체를 추출하기 위해서 Lingpipe에서 제공하는 개체명 인식기(Named Entity Recognizer) 중에서 TokenShapeChunker를 사용하였다(http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html). TokenShapeChunker를 사용하기 위해서는 정답이 규정된 데이터(training data)를 통해 학습하는 과정이 선행되어야 하며 본 연구에서는 Medline 데이터베이스가 수록한 논문의 초록 2,000개에서 추출한 400,000개 이상의 생의학 단어들과 100,000개 이상의 주석들로 구성되어 있는 Genia Corpus (2003)에서 중요한 개체들을 학습을 시켰다. 첫 번째 단계에서 수집한 논문들의 표제 및 초록을 대상으로 개체명 인식을 수행하고 다음으로 통제 어휘인 UMLS(The Unified Medical Language System)를 참고하여 추출된 단어의 생물학적 개체 여부를 추가적으로 확인하였다.

세 번째 단계는 두 번째 단계를 거쳐 확인된 개체들로 단어 쌍을 형성하는 단계이다. 동시출현 기반인 전통적인 ABC 모델은 확인된 개체들이 하나의 논문에서 동시에 출현한다면 서로 관련이 있는 것으로 간주하였고 인용정보를 고려한 새로운 접근법은 인용관계를 형성하는 두 논문에서 등장하는 개체들이 서로 관련이 있는 것으로 간주하였다.

3.3 데이터 수집

본 연구에서는 Swanson의 대표적인 연구 중 하나인 어유와 레이노드 증상의 관계를 재현하고자 하였으며 전문(Full text)을 제공하는 PubMed Central 데이터베이스에서 "어유(Fish Oil)"나 "레이노드(Raynaud)"와 관련된 논문들을 검색하였다. PubMed Central은 미

국 국립의학도서관(NLM)에서 운영하는 생의학분야 전문 논문 데이터베이스로 2014년 2월 기준 290만개의 자료를 보유하고 있다. 질의어를 "레이노드"로 설정한 이유는 "레이노드 증상"이나 "레이노드 질병" 등 다양한 단어들이 "레이노드"와 관련이 있었기 때문이다. 한편 저자명 필드에서 저자의 이름으로 "레이노드"를 포함하고 있는 경우가 있어서 "레이노드" 질의어와 관련된 논문을 수집할 경우에는 질의어가 저자명 필드에서 등장하는 경우는 제외하도록 검색 조건을 설정하였다. 선행 연구들을 살펴보면 어떤 연구들은 논문 검색을 수행할 때 Swanson이 연구를 발표하기 이전의 논문들로 출판 연도를 제한해서 수집하는 경우가 있었는데(Weeber, Vos, Klein, Aronson, & Molema, 2001; Pratt & Yetisgen-Yildiz, 2003; Cameron, Bodenreider, Yalamanchili, Danh, Vallabhaneni, Thirunarayan, & Rindflesch, 2013) 본 연구는 인용 정보의 수집을 위해 전문의 참고문헌 목록이 필요하고 PubMed에 비해 PubMed Central에서 제공하는 전문 논문의 개수가 많지 않아 출판 연도를 고려하지 않고 관련 논문을 모두 수집하였다. Swanson의 연구 이후 A 개념과 C 개념 사이의 암묵적이었던 관계가 밝혀져 A 개념과 C 개념이 하나의 논문에서 동시에 출현하는 경우도 존재하는데 이러한 논문들이 분석 결과에 미치는 영향을 고려하여 선행연구(Hristovski, Peterlin, Mitchell, & Humphrey, 2005)에서 수행한 것처럼 "어유"나 "레이노드" 단어가 하나의 논문에 함께 출현하는 경우는 수작업으로 제거하였다. 최종적으로 2,596개의 핵심 논문이 PubMed Central 데이터베이스에서 수집되었으며 이들의 참고문헌 목록을 통해 추가적으로 61,817개의 논문이 수집되어 총 64,405개의 논문을 분석 대상으로 선정하였다.

3.4 생물학적 개체 추출 및 네트워크 형성

선행연구 중에서 UMLS의 의미론적 유형(Semantic type)에 따라 후보 단어를 추출한 경우가 존재하는데 135개의 의미론적 유형 중에서 연구 주제와 관련이 있을 것으로 판단되는 유형을 선정하여 그 유형에 해당하는 단어들만을 추출하도록 하였다(Pratt & Yetisgen-Yildiz, 2003; Srinivasan, 2004; Weeber et al., 2001). 본 연구에서도 Swanson의 어유와 레이노드 질병 및 증상과 관련해서 등장한 관련 단어들의 의미론적 유형을 참고해서 생물학적 개체의 유형을 한정하였다.

추출한 생물학적 개체들을 대상으로 단어 쌍을 형성하고 네트워크를 형성하기 위해 Gephi(Bastian, Heymann, & Jacomy, 2009)라는 사회 연결망 분석 도구를 사용하였다. 선행연구(Ding et al., 2013; Song et al., 2013)들을 참고하여 사회 연결망 분석에 쓰이는 여러 중심성 지수들 중에서 연결정도 중심성(Degree centrality)으로 영향력 있는 단어들을 살펴보았다. Otte, Rousseau(2002)에 의하면 연결정도 중심성은 특정 노드가 몇 개의 다른 노드들과 연결되어 있는지에 대한 정보를 나타낸다.

3.5 평가

많은 연구들이 문헌기반 발견 연구를 수행하면서 그들이 고안한 방법의 성능을 평가하기

위해 Swanson의 연구를 재현하는데 Swanson의 연구에서 발견된 관계들이 유사하게 확인된다면 성능이 좋은 것으로 간주하고 있다. Yetisgen-Yildiz, Pratt(2009)는 연구자들이 상위 순위에 있는 단어들에 관심을 가질 것이라는 가정을 하고 평가 과정에서도 후보 목록의 단어 전체보다 상위에 존재하는 단어들을 분석하는 것이 더 의미 있는 작업이라고 주장하고 있다. Wren, Bekerediian, Stewart, Shohet, Garner(2004)의 연구에서도 단어들을 순위로 정렬하고 상위에 있는 관계들을 분석하는 것이 더 적절하다고 본다. 본 연구에서도 동시출현과 인용정보 기반 두 가지 다른 접근법으로 확인되는 단어들을 상위 목록을 중심으로 살펴보고자 한다.

4. 결과 분석

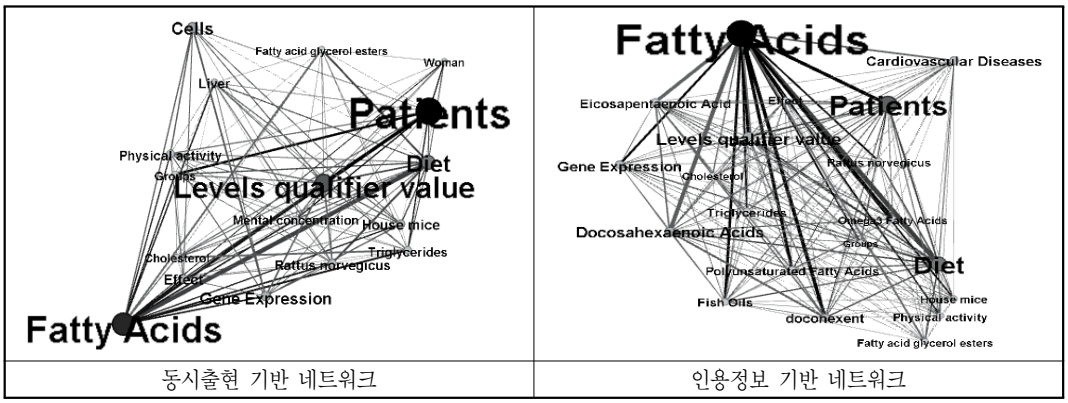
4.1 모든 개체들 대상 네트워크 분석

우선 분석대상 논문에 등장하는 모든 개체들을 대상으로 형성된 단어 쌍을 분석하였다. 전통적인 ABC 모델의 동시출현 기반 접근방법과 본 연구에서 제안하는 인용정보 기반 접근방법에서 발견된 단어들이 차이가 있는지를 확인하였다. 동시출현 기반 접근방법으로 형성한 단어 쌍은 총 4,164,321개이며 인용정보 기반 접근방법으로 형성한 단어 쌍은 총 2,133,653개이다. 〈표 1〉은 연결정도(degree) 중심성 값에 따라 상위 20위에 있는 단어들을 목록으로 정리하여 제시한 것이다. 의미론적 유형에 따라 출현한 단어들을 살펴보면 인용정보 기반보다 동시출현 기반 네트워크에서 더 다양한 유형의 단어들이 영향력 있는 것으로 나타난다. 상위 20개의 단어들 중에서 어유와 관련된 단어는 동시출현 기반 네트워크에 3개, 인용정보 기반 네트워크에 8개가 등장한다. 레이노드 질병 및 증상과 관련된 단어는 양쪽 네트워크 모두 상위 목록에서는 발견되지 않는다. 그 이유는 데이터 수집 과정에서 레이노드 질병 및 증상 관련 논문의 개수가 어유 관련 논문의 개수와 비교할 때 현저하게 적은 데서 비롯되는 것으로 파악된다. 레이노드 질병 및 증상이 아닌 다른 질병과 관련된 단어는 동시출현 기반 네트워크에서는 하나도 등장하지 않지만 인용정보 기반 네트워크에서는 2개가 등장한다. 〈표 1〉에서 어유와 관련된 단어는 블록으로 표시하였고 질병과 관련된 단어는 진하게 나타내었다.

〈그림 3〉은 네트워크를 Gephi로 시각화하여 가시성을 위해 연결정도가 20 이상인 노드들로만 제한하여 표시한 것이다. 동시출현 기반 네트워크에서는 Patients, Fatty Acids, Levels qualifier value, Diet, Cells 등의 단어들이 중심이 되어 다른 단어들과 밀집하게 연결되어 있다. 인용정보 기반 네트워크에서는 Fatty Acids, Patients, Diet, Levels qualifier value 등의 단어들을 제외하고도 대부분의 단어들이 다른 단어들과 밀집하게 연결되어 있다. 이들 중 대부분은 Eiconoicsapentae acid, Triglycerides, Docosahexaenoic Acids 등 어유와 관련 있는 단어들이다. 동시출현 기반 네트워크는 17개의 단어가 나타났으며 이중에서 어유와 관련된 단어는 3개이고 14개의 다른 개체들이 등장한다. 인용정보 기반 네트워크에서는 20개의 단어가 등장하며 어유 관련 9개의 단어를 제외하고 11개의 다른 개체들을 확인할 수 있다.

〈표 1〉 연결정도(Degree) 중심성 지수에 따른 상위 20위 단어들 목록(모든 개체)

	동시출현 기반		인용정보 기반	
	출현 단어	의미론적 유형	출현 단어	의미론적 유형
1	Patients	Patient or Disabled Group	Fatty Acids	Lipid
2	Fatty Acids	Lipid	Patients	Patient or Disabled Group
3	Levels qualifier value	Qualitative Concept	Diet	Food
4	Diet	Food	Levels qualifier value	Qualitative Concept
5	Cells	Cell	Docosahexaenoic Acids (3)	Lipid Pharmacologic Substance Biologically Active Substance
6	Gene Expression	Genetic Function	doconexent (2)	Lipid Pharmacologic Substance
7	Physical activity	Daily or Recreational Activity	Eicosapentaenoic Acid (3)	Lipid Pharmacologic Substance Biologically Active Substance
8	House mice	Mammal	Gene Expression	Genetic Function
9	Rattus norvegicus	Mammal	Cardiovascular Diseases	Disease or Syndrome
10	Effect	Qualitative Concept	Triglycerides (2)	Lipid Biologically Active Substance
11	Mental concentration	Mental Process	Fish Oils (3)	Lipid Pharmacologic Substance Biologically Active Substance
12	Groups	Idea or Concept	Physical activity	Daily or Recreational Activity
13	Liver	Body Part, Organ, or Organ Component	Effect	Qualitative Concept
14	therapeutic aspects	Functional Concept	Cholesterol (2)	Steroid Biologically Active Substance
15	Inflammation	Pathologic Function	Polyunsaturated Fatty Acids (2)	Lipid Biologically Active Substance
16	Triglycerides (2)	Lipid Biologically Active Substance	Rattus norvegicus	Mammal
17	Woman	Population Group	Fatty acid glycerol esters (2)	Lipid Pharmacologic Substance
18	Cholesterol (2)	Steroid Biologically Active Substance	Disease	Disease or Syndrome
19	Fatty acid glycerol esters (2)	Lipid Pharmacologic Substance	House mice	Mammal
20	Homo sapiens	Human	Liver	Body Part, Organ, or Organ Component



동시출현 기반 네트워크 | 인용정보 기반 네트워크

〈그림 3〉 연결정도가 20 이상인 노드들로 제한한 네트워크 시각화(모든 개체)

추가적으로 레이노드 질병 및 증상이나 어유 관련 단어가 어떤 다른 단어들과 쌍을 형성하고 있는지에 대해 확인해 보았다. 레이노드 질병 및 증상 관련 단어와 동시 출현한 단어는 동시출현 기반 네트워크에서 총 1,195개, 인용정보 기반 네트워크에서 총 371개가 등장하였다. 〈표 2〉와 〈표 3〉에 제시된 것처럼 상위 12개의 단어를 살펴보았을 때 인용정보 기반 네트워크에서만 레이노드 질병 및 증상 관련 단어가 대표적인 B 개념들과 단어 쌍을 형성하고 있는 것을 확인할 수 있었다.

다음으로 어유가 어떤 다른 단어들과 관계를 형성하는지 살펴보았다. 어유 관련 단어들과 단어 쌍을 형성하는 경우는 제외하고 다른 개체들과 단어 쌍을 형성하는 경우를 살펴보았다. 인용정보 기반 네트워크에서만 질병 관련 단어들이 높은 순위에 등장한다. 상위 12위를 기준으로 하였을 때 각 네트워크에서만 등장하는 단어들을 블록으로 처리하였다(〈표 4〉, 〈표 5〉 참조).

〈표 2〉 레이노드 질병 및 증상과 쌍을 형성하는 단어들(모든 개체)

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Patients	Patient or Disabled Group	135
2	Systemic Scleroderma	Disease or Syndrome	68
3	Retinitis Pigmentosa	Disease or Syndrome	40
4	Scleroderma	Disease or Syndrome	30
5	Disease	Disease or Syndrome	27
6	Therapeutic aspects	Functional Concept	26
7	Lupus Erythematosus	Disease or Syndrome	22
8	Blood Vessels	Body Part, Organ, or Organ Component	18
9	Ulcer	Pathologic Function	16

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
10	Woman	Population Group	16
11	Encephalitis S tLouis	Disease or Syndrome	16
12	Congenital Abnormality	Congenital Abnormality	15

〈표 3〉 레이노드 질병 및 증상과 쌍을 형성하는 단어들(모든 개체)

인용정보 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Patients	Patient or Disabled Group	23
2	Migraine Disorders	Disease or Syndrome	20
3	Hypertensive disease	Disease or Syndrome	16
4	Twin sibling person	Family Group	15
5	Cardiovascular Diseases	Disease or Syndrome	12
6	Status	Qualitative Concept	12
7	Child	Age Group	11
8	Arteriopathic disease	Disease or Syndrome	11
9	Rheumatoid Arthritis	Disease or Syndrome	11
10	Vascular occlusion	Disease or Syndrome	11
		Vascular constriction function	
11	Vascular constriction function	Organ or Tissue Function	11
12	Vasospasm	Pathologic Function	11

〈표 4〉 어유와 쌍을 형성하는 단어들(모든 개체)

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Diet	Food	784
2	Patients	Patient or Disabled Group	523
3	Dietary Supplementation	Therapeutic or Preventive Procedure	416
4	Effect	Qualitative Concept	395
5	Levels qualifier value	Qualitative Concept	378
6	Rattus norvegicus	Mammal	376
7	Groups	Idea or Concept	337
8	Gene Expression	Genetic Function	227
9	Mental concentration	Mental Process	221
10	House mice	Mammal	220
11	Inflammation	Pathologic Function	212
12	Cholesterol (2)	Steroid Biologically Active Substance	211

〈표 5〉 어유와 쌍을 형성하는 단어들(모든 개체)

인용정보 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Diet	Food	892
2	Patients	Patient or Disabled Group	824
3	Levels qualifier value	Qualitative Concept	674
4	Effect	Qualitative Concept	516
5	Gene Expression	Genetic Function	504
6	Rattus norvegicus	Mammal	439
7	Physical activity	Daily or Recreational Activity	404
8	Groups	Idea or Concept	383
9	House mice	Mammal	377
10	Disease	Disease or Syndrome	355
11	Dietary Supplementation	Therapeutic or Preventive Procedure	352
12	Cardiovascular Diseases	Disease or Syndrome	350

4.2 특정 패턴을 이루는 개체들 대상 네트워크 분석

이번에는 Swanson의 연구에서 밝혀진 '물질-과정-질병' 패턴을 이루는 단어들을 대상으로 삼중의 단어 쌍을 형성하였다. '물질-과정' 및 '과정-질병'이라는 두 단어 쌍이 있을 때 가운데 '과정'이라는 개체가 같다면 '물질-과정-질병'으로 연결하였다. 각 개념별로 제한한 의미론적 유형은 〈표 6〉과 같다. '물질-과정-질병' 삼중의 단어 쌍에서 가운데 매개 단어를 제거하고 '물질-질병'으로 단어 쌍을 정제하였다.

〈표 7〉은 연결정도 중심성 값에 따라 상위 20위에 있는 단어들을 목록으로 정리하여 제시한 것이다. 의미론적 유형을 참고하여 출현 단어들을 살펴보면 이제 동시출현 기반 네트워크에서도 어유 관련 단어가 많이 등장한다. 어유와 관련된 단어는 동시출현 기반 네트워크에서 10개, 인용정보 기반 네트워크에서 11개 등장

〈표 6〉 각 개념별로 한정한 의미론적 유형

		의미론적 유형
A 개념	물질	Lipid Pharmacologic Substance Biologically Active Substance
B 개념	과정	Cell Function Physiologic Function Clinical Attribute Organ or Tissue Function
C 개념	질병	Disease or Syndrome Phenomenon or Process

〈표 7〉 연결정도(Degree) 중심성 지수에 따른 상위 20위 단어들 목록(특정 패턴)

	동시출현 기반		인용정보 기반	
	출현 단어	의미론적 유형	출현 단어	의미론적 유형
1	Fatty Acids	Lipid	Fatty Acids	Lipid
2	Disease	Disease or Syndrome	Fish Oils (3)	Lipid Pharmacologic Substance Biologically Active Substance
3	Eicosapentaenoic Acid (3)	Lipid Pharmacologic Substance Biologically Active Substance	Eicosapentaenoic Acid (3)	Lipid Pharmacologic Substance Biologically Active Substance
4	Lipids	Lipid	Docosahexaenoic Acids (3)	Lipid Pharmacologic Substance Biologically Active Substance
5	Docosahexaenoic Acids (3)	Lipid Pharmacologic Substance Biologically Active Substance	Polyunsaturated Fatty Acids (2)	Lipid Biologically Active Substance
6	Diabetes	Disease or Syndrome	Disease	Disease or Syndrome
7	Fatty acid glycerol esters (2)	Lipid Pharmacologic Substance	Cardiovascular Diseases	Disease or Syndrome
8	Cardiovascular Diseases	Disease or Syndrome	Omega-3 Fatty Acids (3)	Lipid Pharmacologic Substance Biologically Active Substance
9	Atherosclerosis	Disease or Syndrome	Lipids	Lipid
10	Triglycerides (2)	Lipid Biologically Active Substance	Fatty acid glycerol esters (2)	Lipid Pharmacologic Substance
11	Fish Oils (3)	Lipid Pharmacologic Substance Biologically Active Substance	Triglycerides (2)	Lipid Biologically Active Substance
12	Antioxidants	Pharmacologic Substance	Oils	Lipid
13	Polyunsaturated Fatty Acids (2)	Lipid Biologically Active Substance	Antioxidants	Pharmacologic Substance
14	Reactive Oxygen Species (2)	Biologically Active Substance Element, Ion, or Isotope	Phospholipids (2)	Lipid Biologically Active Substance
15	Omega-3 Fatty Acids (3)	Lipid Pharmacologic Substance Biologically Active Substance	Hypertensive disease	Disease or Syndrome
16	Hypertensive disease	Disease or Syndrome	Atherosclerosis	Disease or Syndrome
17	Phospholipids (2)	Lipid Biologically Active Substance	Linoleic Acid (3)	Lipid Pharmacologic Substance Biologically Active Substance
18	Nitric Oxide (3)	Pharmacologic Substance Inorganic Chemical Biologically Active Substance	Heart Diseases	Disease or Syndrome
19	Pharmaceutical Preparations	Pharmacologic Substance	Chronic disease	Disease or Syndrome
20	Linoleic Acid (3)	Lipid Pharmacologic Substance Biologically Active Substance	Metabolic Syndrome X	Disease or Syndrome

한다. 레이노드 질병 및 증상이 아닌 다른 질병과 관련된 단어는 동시출현 기반 네트워크에서 4개, 인용정보 기반 네트워크에서 7개가 등장한다. 또한 어유를 제외하고 새로운 물질 관련 단어들이 등장하는 것을 확인할 수 있는데 동시출현 기반 네트워크에서 5개, 인용정보 기반 네트워크에서 2개가 등장한다.

추가적으로 레이노드 질병 및 증상이나 어유 관련 단어가 어떤 다른 단어들과 쌍을 형성하고 있는지에 대해 확인해 보았다. 레이노드 질병 및 증상 관련 단어와 동시 출현한 단어는 동시출현 기반 네트워크에서 총 1,909개, 인용정보 기반 네트워크에서는 총 729개였다. 〈표 8〉과 〈표 9〉에 제시된 것처럼 상위 12개의 단어를 살펴보았을 때 레이노드 질병 및 증상 관련 단어와 동시 출현한 어유 관련 단어는 동시출현 기반 네트워크에서 5개, 인용정보 기반 네트워크에서 9개가 등장하여 인용정보 기반 네트워크에서 레이노드 질병 및 증상과 어유의 관계가 더 잘 발견되는 것을 확인할 수 있었다.

다음으로 어유가 어떤 다른 단어들과 관계를 형성하는지 살펴보았다. 어유 관련 단어들과 단어 쌍을 형성하는 경우는 제외하고 다른 개체들과 단어 쌍을 형성하는 경우를 살펴보았다. 동시출현 기반과 인용정보 기반 네트워크에서 어유와 관계를 형성하는 단어들이 다른 것으로 확인된다(〈표 10〉, 〈표 11〉 참조).

〈표 8〉 레이노드 질병 및 증상과 쌍을 형성하는 단어들(특정 패턴)

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Disease	Disease or Syndrome	44
2	Fatty Acids	Lipid	40
3	Systemic Scleroderma	Disease or Syndrome	39
4	Diabetes	Disease or Syndrome	37
5	Heart Diseases	Disease or Syndrome	35
6	Fish Oils	Lipid	35
		Pharmacologic Substance	
		Biologically Active Substance	
7	Fatty acid glycerol esters	Lipid	35
		Pharmacologic Substance	
8	Docosahexaenoic Acids	Lipid	35
		Pharmacologic Substance	
		Biologically Active Substance	
9	Omega-3 Fatty Acids	Lipid	34
		Pharmacologic Substance	
		Biologically Active Substance	
10	Skin Specimen	Body Substance	33
11	Cerebrovascular accident	Disease or Syndrome	33
12	Diabetes Mellitus	Disease or Syndrome	33

〈표 9〉 레이노드 질병 및 증상과 쌍을 형성하는 단어들(특정 패턴)

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Eicosapentaenoic Acid	Lipid	16
		Pharmacologic Substance	
		Biologically Active Substance	
2	Hypertensive disease	Disease or Syndrome	16
3	Arteriopathic disease	Disease or Syndrome	16
4	Cardiovascular Diseases	Disease or Syndrome	16
5	Fish Oils	Lipid	16
		Pharmacologic Substance	
		Biologically Active Substance	
6	Fatty Acids	Lipid	16
7	Omega-3 Fatty Acids	Lipid	16
		Pharmacologic Substance	
		Biologically Active Substance	
8	Oils	Lipid	16
9	Polyunsaturated Fatty Acids	Lipid	15
		Biologically Active Substance	
10	Triglycerides	Lipid	15
		Biologically Active Substance	
11	Fatty acid glycerol esters	Lipid	15
12	Docosahexaenoic Acids	Disease or Syndrome	15

〈표 10〉 어유와 쌍을 형성하는 단어들(특정 패턴)

동시출현 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Disease	Disease or Syndrome	172
2	Diabetes	Disease or Syndrome	131
3	Pharmaceutical Preparations	Pharmacologic Substance	108
4	Heart Diseases	Disease or Syndrome	106
5	Diabetes Mellitus Non Insulin Dependent	Disease or Syndrome	103
6	Diabetes Mellitus	Disease or Syndrome	103
7	Syndrome	Disease or Syndrome	87
8	Skin Specimen	Body Substance	84
9	Heart failure	Disease or Syndrome	79
10	Arteriopathic disease	Disease or Syndrome	76
11	Cerebrovascular accident	Disease or Syndrome	74
12	Septicemia	Disease or Syndrome	64

〈표 11〉 어유와 쌍을 형성하는 단어들(특정 패턴)

인용정보 기반 네트워크			
	출현 단어	의미론적 유형	출현빈도
1	Disease	Disease or Syndrome	213
2	Cardiovascular Diseases	Disease or Syndrome	205
3	Antioxidants	Pharmacologic Substance	156
4	Heart Diseases	Disease or Syndrome	149
5	Hypertensive disease	Disease or Syndrome	144
6	Atherosclerosis	Disease or Syndrome	139
7	Metabolic Syndrome X	Disease or Syndrome	138
8	AntiInflammatory Agents	Pharmacologic Substance	138
9	Diabetes	Disease or Syndrome	136
10	Chronic disease	Disease or Syndrome	136
11	Rheumatoid Arthritis	Disease or Syndrome	127
12	Metabolite	Biologically Active Substance	125

4.3 매개단어 분석

이번에는 어유 및 어유 관련 단어들과 레이노드 질병 및 증상의 관계를 매개하는 단어들을 살펴보았다. 동시출현 기반 네트워크에서 레이노드 질병 및 증상과 관계를 형성하는 어유 및 어유관련 단어는 상위 12개의 단어를 기준으로 할 때 Fatty Acids, Fish Oils, Fatty acid glycerol esters, Docosahexaenoic Acids, Omega3 Fatty Acids이며 이 두 단어를 매개하는 단어는 대표 B 단어로 밝혀져 있는 Vasodilation 이다. 그러나 인용정보 기반 네트워크에서는 Response process가 매개단어가 되어 어유관련 단어인 Eicosapentaenoic Acid, Fish Oils, Fatty Acids, Omega3 Fatty Acids, Oils, Polyunsaturated Fatty Acids, Triglycerides, Fatty acid glycerol esters, Docosahexaenoic Acids 등이 확인되고 있다.

〈표 12〉 어유와 레이노드 질병 및 증상을 매개하는 단어

동시출현 기반 네트워크			
순위	출현 단어	매개 단어	출현 횟수
2	Fatty Acids	Vasodilation	40
6	Fish Oils	Vasodilation	35
7	Fatty acid glycerol esters	Vasodilation	35
8	Docosahexaenoic Acids	Vasodilation	35
9	Omega3 Fatty Acids	Vasodilation	34

〈표 13〉 어유와 레이노드 질병 및 증상을 매개하는 단어

인용정보 기반 네트워크			
순위	출현 단어	매개 단어	출현 횟수
1	Eicosapentaenoic Acid	Response process	16
5	Fish Oils	Response process	16
6	Fatty Acids	Response process	16
7	Omega3 Fatty Acids	Response process	16
8	Oils	Response process	16
9	Polyunsaturated Fatty Acids	Response process	15
10	Triglycerides	Response process	15
11	Fatty acid glycerol esters	Response process	15
12	Docosahexaenoic Acids	Response process	15

5. 결 론

본 연구는 Swanson의 ABC 모델을 인용 정보를 고려하는 새로운 시각에서 접근하고자 하였으며 전통적인 동시출현 기반 접근법과 차별되는 특징들을 확인하였다. 본 연구에서 확인된 특징들을 종합하면 인용정보 기반으로 형성하는 네트워크에서 발견되는 단어들은 동시출현 기반 네트워크에서 발견되는 단어들에 비해 관련 있는 개체들이 더 강하게 연결되어 있었으며 암묵적인 관계도 더 많이 잡아내고 있었다. 또 각 네트워크에서 차별적으로 등장하는 단어들이 존재하였는데 인용정보 기반 네트워크에서만 고유하게 등장하는 단어들이 더 많았다. 이는 인용정보 기반 네트워크에서 어유관련 단어들이 더 많이 확인되어 상대적으로 다른 단어들이 등장할 기회가 적다는 것을 고려할 때 흥미로운 특징이다.

본 연구가 지니는 한계점들은 다음과 같다. 첫째, 생물학적 개체를 확인하는 과정에서 수집한 논문들의 전문이 아니라 표제 및 초록을 대상으로 하였는데 전문을 대상으로 하면 더 많은 생물학적 개체가 확인될 것으로 예상된다. 한편 전문에 등장하는 생물학적 개체들은 서로 연관성이 더 떨어져 유의한 관계의 추정이 어려워질 가능성도 고려해야 한다. 둘째, 인용정보의 방향성을 고려하지 않았다. 인용을 하는 논문과 인용을 당하는 논문에 등장하는 개체들을 구분하여 단어 쌍을 형성하면 더 흥미로운 결과가 나올 가능성도 없지 않아 이들을 어떻게 구분하여 처리할 것인지에 대해 고려해보고 추후 연구에 반영할 예정이다.

참 고 문 헌

Bastian M., Heymann S., & Jacomy M. (2009). Gephi: An open source software for exploring and manipulating networks, In Proceedings of the 3rd International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media. CA, USA.

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., & Rindflesch, T. C. (2013). A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. Journal of biomedical informatics, 46(2), 238-251.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities, PloS one, 8(8): e71416.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics, 74(2), 289-298.

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science, 28(6), 441-453.

Pratt, W., & Yetisgen-Yildiz, M. (2003). October. LitLinker: Capturing connections across the biomedical literature, In Proceedings of the 2nd International Conference on Knowledge Capture. NY, USA

Song, M., Han, N. G., Kim, Y. H., Ding, Y., & Chambers, T. (2013). Discovering implicit entity relation with the gene-citation-gene network, PloS One, 8(12): e84639.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology, 55(5), 396-413.

Weeber, M., Klein, H., de Jong-van den Berg, L., & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology, 52(7), 548-557.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationship. Bioinformatics, 20(3), 389-398.

Yetisgen-Yildiz, M., & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. Journal of Biomedical Informatics, 42(4), 633-643.

Yu, Q., Ding, Y., Song, M., Song, S. J., Liud, J., & Zhange, B. (2015). Tracing database usage: Detecting main paths in databaselink networks. Journal of Informetrics, 9(1): 1-15.