텍스트 마이닝 기반의 그래프 모델을 이용한 미발견 공공 지식 추론*

Inferring Undiscovered Public Knowledge by Using Text Mining-driven Graph Model

허고은 (Go Eun Heo)**
송 민 (Min Song)***

초 록

정보통신기술의 발달로 학술 정보의 양이 기하급수적으로 증가하였고 방대한 양의 텍스트 데이터를 처리하기 위한 자동화된 텍스트 처리의 필요성이 대두되었다. 생의학 문헌에서 생물학적 의미와 치료 효과 등에 대한 정보를 발견해내는 바이오 텍스트 마이닝은 문헌 내의 각 개념들 간의 유의미한 연관성을 발견하여 의학 영역에서 상당한 시간과 비용을 줄여준다. 문헌 기반 발견 연구로 새로운 생의학적 가설들이 발견되었지만 기존의 연구들은 반자동화된 기법으로 전문가의 개입이 필수적이며 원인과 결과의 한가지의 관계만을 밝히는 제한점이 있다. 따라서 본 연구에서는 중간 개념인 B를 다수준으로 확장하여 다양한 관계성을 동시출현 개체와 동사 추출을 통해 확인한다. 그래프 기반의 경로 추론을 통해 각 노드 사이의 관계성을 체계적으로 분석하여 규명할 수 있었으며 새로운 방법론적 시도를 통해 기존에 밝혀지지 않았던 새로운 가설 제시의 가능성을 기대할 수 있다.

ABSTRACT

Due to the recent development of Information and Communication Technologies (ICT), the amount of research publications has increased exponentially. In response to this rapid growth, the demand of automated text processing methods has risen to deal with massive amount of text data. Biomedical text mining discovering hidden biological meanings and treatments from biomedical literatures becomes a pivotal methodology and it helps medical disciplines reduce the time and cost. Many researchers have conducted literature-based discovery studies to generate new hypotheses. However, existing approaches either require intensive manual process of during the procedures or a semi-automatic procedure to find and select biomedical entities. In addition, they had limitations of showing one dimension that is, the cause-and-effect relationship between two concepts. Thus, this study proposed a novel approach to discover various relationships among source and target concepts and their intermediate concepts by expanding intermediate concepts to multi-levels. This study provided distinct perspectives for literature-based discovery by not only discovering the meaningful relationship among concepts in biomedical literature through graph-based path interference but also being able to generate feasible new hypotheses.

키워드: 바이오 텍스트 마이닝, 문헌 기반 발견, 미발견 공공 지식, 그래프 모델
biotext mining, literature based discovery, undiscovered public knowledge, graph model

---

* 본 연구는 미래창조과학부 및 한국연구재단의 (재)유전자동의보감사업단(2013M3A9C4078138) 연구비 지원에 의해 수행되었음.

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

■ 논문접수일자 : 2014년 2월 20일 ■ 최초심사일자 : 2014년 2월 27일 ■ 게재확정일자 : 2014년 3월 13일
■ 정보관리학회지, 31(1), 231-250, 2014. 〔http://dx.doi.org/10.3743/KOSIM.2014.31.1.231〕

1. 서 론

정보의 가속화 및 기하급수적으로 증가하는 출판 정보들의 양과 복잡성에 따라 과거의 수작업을 통한 텍스트 처리에서 자동화된 접근법의 필요성이 중요한 문제로 대두되었다. 대량의 정보자원을 효율적으로 처리하기 위한 텍스트 마이닝은 방대한 양의 텍스트 데이터에서 유의미한 정보들을 자동적으로 발견하는 기술이다. 특히 생의학 영역에 적용되는 바이오 텍스트 마이닝은 생의학적 개념들 간의 유의미한 연관성을 자동적으로 발견하고 생의학적 문맥과 상호작용 관계를 이해하여 새로운 지식을 발견(knowledge discovery)해 내는 것을 목적으로 한다. 이는 가설을 발견하기 위한 생의학 연구의 연료가 되는 영역이라 일컬으며, "Conceptual Biology"라고 정의한다(Blagosklonny & Pardee, 2002). 대량의 디지털 자원으로 인해 등장하게 된 Conceptual Biology는 생물학을 보충해주는 역할로 생체실험이나 동물실험으로 인해 드는 추가적인 비용과 시간적인 소모를 절약하는 장점을 지니며 이러한 자동적인 분석 및 실험결과를 통해 분자생물학자나 임상의학자들에게 다방면으로 도움을 주고 있다. 특히 문헌 내에서 새로운 가설을 발견하는 문헌 기반 발견(Literature Based Discovery, LBD)의 선구적인 역할을 한 Don R. Swanson의 연구는 텍스트 마이닝을 기초로 ABC 모델을 제안하여 문헌을 통해 기존의 생의학적 가설을 검증해낼 뿐만 아니라 새로운 가설을 발견하고 예측해내는 미발견 공공 지식(Undiscovered Public Knowledge, UPK) 연구들을 수행했다. 이를 기반으로 시스템을 통해 생의학적 가설을 발견하기 위한 다양한 가설 기반 발견 시스템이 개발되었다. 하지만 기존 연구들의 문제점은 개념을 추출하거나 의미를 부여하는 작업에서 전문가에 의한 선택과 입력이 요구되는 반자동화(semi-automated)된 접근법이며 중간 개념을 원인과 결과의 한 가지 관계로만 확인한다는 제한점을 가진다.

따라서 본 연구에서는 Swanson의 연구에서 밝히지 못했던 전체적인 구조를 새로운 접근법인 그래프 모델로 경로를 표현하여 두 개체간의 숨겨진 관계성을 추론하고 새로운 규칙 기반의 가설을 자동적으로 발견하고자 한다. 특히 두 생의학 개체 사이에 존재하는 동사관계를 확인하여 유의미한 생의학적 가설 분석이 가능하도록 한다.

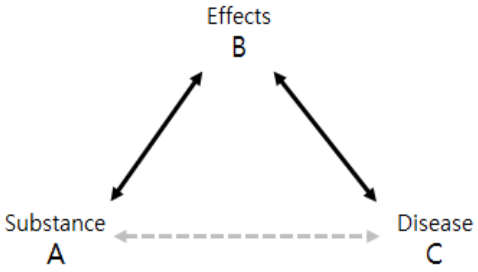
2. 이론적 배경

2.1 문헌 기반 발견

문헌 기반 발견은 학술 문헌 내의 정보들을 통해 이전에는 발견되지 않았던 잠재적인 관계들을 발견하여 새로운 지식을 밝혀내는 과정을 의미하며 생의학 영역에서 특정 질병과 관련된 치료제 및 증상 등을 발견하는데 적용된다. 이는 Swanson(1986a)에 의해 처음으로 제안되었으며 문헌을 통해 새로운 지식을 추론해낸다는 의미로 미발견 공공 지식 또는 ABC 모델이라고 칭한다.

ABC 모델은 〈그림 1〉과 같이 나타낼 수 있는데 우선 특정 질병이나 질환인 C를 확인하는 것에서 시작한다. B단어는 특정 질병에 대한 생리적인 조건이나 상태 또는 과정의 단어로

볼 수 있으며, A는 질병에 대한 치료제 또는 증상으로 나타낼 수 있다. 특정 문헌에서 단어 C와 B가 동시에 출현하고, 또 다른 문헌에서 A와 B가 동시에 출현할 경우 A와 C는 문헌 내에서 직접적인 연결성을 지니지 않지만 B라는 중간 개념을 통해서 연결성을 갖게 된다는 이론이다.



〈그림 1〉 ABC 모델

Swanson은 ABC 모델을 기반으로 MEDLINE에서 추출된 텍스트 데이터를 통해 특정 질병 또는 질환과 관련한 흥미로운 가설들을 발견하는 연구들을 수행하였다(Swanson, 1986b/1988/1990a; Swanson & Smalheiser, 1997; Swanson, Smalheiser, & Bookstein, 2001; Smalheiser & Swanson, 1994/1996a/1996b).

가장 초기 연구로 레이노병(Raynaud's Disease)과 어유(Fish Oils)의 관계에 대한 가설을 제안하였다(Swanson, 1986b). 레이노병은 혈관운동신경 장애에 기인하는 질병으로 MEDLINE의 문헌 데이터를 통해 레이노병이 혈소판 응집력(platelet aggregability), 혈관 수축(vasoconstriction), 혈액 점도(blood viscosity)에 의해 악화된다는 점을 발견하였고 이 현상들은 어유에 의해 감소될 수 있다는 점을 발견했다. 이러한 두 가지의 발견점을 확인하여 레이노병 환자들에게 어유가 유익한 치료제로 작용될 수 있다는 점을 상정하였다(Swanson, 1986b).

이 가설은 실제로 임상의학자들에 의해 유효성이 검증되었으며(DiGiacomo, Kremer, & Shah, 1989) 이러한 선구적인 연구들로 인해 많은 연구자들이 문헌 기반 발견 연구들에 관심을 가지고 다양한 연구들을 수행하게 되었다.

2.2 선행 연구

Swanson의 연구를 기반으로 시도된 다양한 문헌 기반 발견의 연구들은 중간 개념의 역할을 하는 B를 선정하기 위해 방법론적인 시각에 초점을 두어 시스템을 개발하였다. 본 연구에 기저가 된 선행 연구들은 크게 의미론적 접근을 수행한 연구, 술어를 선정하는 연구, 그래프 기반의 연구로 나눌 수 있다.

2.2.1 개념 기반의 의미론적 접근 연구

중간 개념 B를 선정하기 위한 분석 대상 단어 리스트가 방대하다는 문제점에 착안하여 단어들을 효율적으로 필터링 하기 위한 다양한 접근법이 시도되었다. Swanson의 가설에 대한 새로운 학술적인 가설 발견과 실험을 위한 두 가설 발견 접근 모델인 개방형 발견 과정(open discovery process)과 폐쇄형 발견 과정(closed discovery process)을 제안한 Weeber Klein, de Jong-van den Berg, Vos(2001)은 이 모델을 분석 단위로 UMLS 개념을 이용하여 NLP시스템인 DAD시스템에 적용했다. 레이노병과 어유, 편두통과 마그네슘 결핍에 대해 의미론적(semantic) 정보를 이용하여 실험했다. 또한

탈리도마이드(thalidomide)의 치료법을 발견하기 위해 개방형 발견 과정을 이용하여 UMLS 맵핑을 통해 면역학(immunologic) 요인의 의미 유형을 가진 단어들을 선정하였다(Weeber, Vos, Klein, Aronson, & Molema, 2003). Srinivasan (2004)은 Swanson과 Smalheiser에 의해 제안된 가설에 대해 MEDLINE의 MeSH 기반 메타 프로파일을 이용하여 토픽을 발견하는 알고리즘을 제안하였고, Swanson, Smalheiser, Torvik(2006)은 B리스트 순위화의 성능을 높이는 여러 기법을 제안하고 평가하기 위해 각 입력 레코드에 할당된 MeSH를 이용하였다. 각 단어는 동일한 MeSH 단어와 연동될수록 더 깊은 관계를 가진다고 보고 Subject-heading weight(sh-wt)를 정의하여 제목의 B리스트를 순위화 했다.

2.2.2 술어 선정 기법 연구

생의학 개체들과 이들의 관계를 발견하기 위해 Hristovski, Peterlin, Mitchell, Humphrey(2005)는 새로운 문헌 기반 발견 시스템인 BITOLA를 제안하여 질병과 유전자의 관계를 밝히는 과정에서 후보 관계(candidate relations) 수를 감소시키고, 후보 유전자의 염색체 상의 위치(chromosomal location)뿐만 아니라 질병의 염색체의 시작 위치에 유전학적 지식을 포함시켜 효율성을 높였다. 더불어 Hristovski, Friedman, Rindflesch, Peterlin(2006)은 의미론적 술어를 이용하는 방법을 제안하여 동시출현에만 의존했던 시스템을 강화하는 BITOLA 시스템을 제안하였다. 또한 새로운 치료 접근 방법인 문헌 기반 발견의 서술부 활용 방안에 대해 제안하여 전통적인 치료약 발견에 도움을 주었다(Hristovski, Rindflesch, & Peterlin, 2013). Kilicoglu, Shin, Fiszman, Rosembalt, Rindflesch (2012)는 PubMed Citations로부터 추출한 의미론적인 주어-동사-목적어 구조들을 저장하여 사용할 수 있는 레포지토리인 SemMedDB를 구축하여 생의학 영역의 가설 발견이 용이하도록 했다.

2.2.3 그래프 기반의 선행 연구

최근에는 ABC 모델의 가설 추론을 위해 각 단어나 개념간의 관계를 네트워크화 하여 시각적인 분석을 시도한 연구들이 진행되었다. Narayanasamy, Mukhopadhyay, Palakal, Potter (2004)는 ABC 모델에 대한 그래프 기반의 접근법을 시도한 초기 연구로 MEDLINE 데이터베이스의 학술 문헌을 마이닝하여 생물학적 개체간의 관계를 그래픽 기반으로 찾아내는 시스템인 TransMiner를 개발했다. 또한 Frijters, Heupers, van Beek, Bouwhuis, van Schaik, de Vlieg, Polman, Alkema(2008)은 마이크로어레이 데이터의 생물학적인 해석을 위해 MEDLINE 데이터베이스 정보를 이용하는 CoPub 시스템을 개발하였고, 후속 연구에서 새로운 생의학 관계를 독립적인 문헌을 통해 검증했다(Frijters, van Vugt, Smeets, van Schaik, de Vlieg, & Alkema, 2010). 또한 Liekens, De Knijf, Daelemans, Goethals, De Rijk, Del-Favero(2011)은 생의학 정보를 발견하고 탐구할 수 있는 데이터 통합 마이닝 플랫폼인 BioGraph를 제안하여 시스템의 우수성을 입증했다.

최근에는 기존의 ABC 모델에서 벗어나 A와 C사이에 존재하는 중간 개념인 B를 여러 개념으로 인식하기 시작하였다. Wilkowski, Fiszman, Miller, Hristovski, Arabandi, Rosemblat,

Rindflesch(2011)은 처음으로 B개념을 단일 개념이 아닌 여러 개의 개념 집합으로 확장하여 구체화했다. 지금까지 시도되지 않았던 Swanson의 가설에 대한 깊이 있는 분석 연구를 수행한 Cameron, Bodenreider, Yalamanchili, Danh, Vallabhaneni, Thirunarayan, Sheth, Rindflesch(2013)은 Swanson의 레이노병-어유 가설을 구조화된 배경 지식과 그래프 기반의 알고리즘을 통해 반자동적으로 의미론적 술어를 추출하여 분해하는 방법론을 제안하였다.

2.2.4 선행 연구 제한점 종합 분석

기존의 ABC 모델을 비롯한 문헌 기반 가설을 수행한 선행 연구들은 중간 개념인 B를 효율적으로 선정하기 위해 통계적 접근법이나 새로운 알고리즘을 제안하였고 MeSH나 UMLS를 이용하여 통합적인 개념을 통해 다양한 방법으로 B를 필터링하고자 했다. 그러나 기존의 연구들은 반자동화된 기법으로 생의학 개념 또는 생의학 개체의 발견과 선정 과정에 전문가에 의한 개입이 필수적이라는 문제점이 있다. 또한 네트워크 관점에서 ABC 모델을 접근한 연구들의 대부분은 시스템 내의 부분적인 한 기능으로 시각화를 적용하는데 그쳤고, A와 C 사이에 존재하는 중간 개념인 B선정을 한가지로만 제한하였기에 검색된 단어에 해당하는 관계 범위 내에서 표현된 네트워크는 각 개체와 관계성에 대한 전체적인 흐름을 파악하기 어려운 제한점을 지닌다. 특히, 최신의 연구인 Cameron, Bodenreider, Yalamanchili, Danh, Vallabhaneni, Thirunarayan, Sheth, Rindflesch(2013)은 ABC 모델에서 벗어나 개념들 간의 연관관계를 더 잘 설명할 수 있는 모델이 필요함을 주장하였고, AnC 모델을 제안하여 B개념의 다수준(multi-level) 모델에 대한 가능성을 입증했다.

따라서 이러한 기존 연구의 문제와 B개념의 확장 가능성에 착안하여 그래프 모델 기반의 다차원적인 개념 접근을 통한 가설 발견 자동화 시스템의 필요성을 확인하였다.

3. 그래프 모델 기반 ABC 모델 추론 시스템

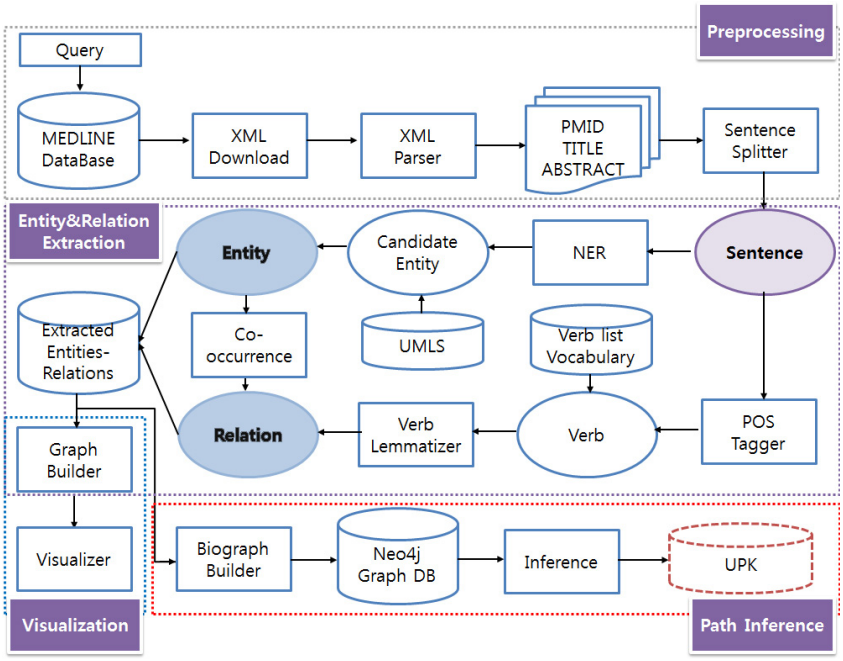
3.1 시스템 설계 및 구현

본 시스템은 그래프 모델을 통해 Swanson의 ABC 모델에서 밝혀주지 못했던 다양한 개체와 관계성을 규명하고 개체들 간의 경로를 확인하여 미발견 공공 지식을 추론하기 위한 시스템으로 전체 시스템 개요는 〈그림 2〉와 같다.

시스템 개요는 (1) 전처리, (2) 개체 추출, (3) 동사 추출, (4) 두 개체간의 관계성 추출, (5) 개체 시각화, (6) 경로 추론으로 구분 지을 수 있으며 이에 대한 상세 설명은 다음에서 순서대로 기술한다.

3.1.1 전처리

개체와 관계를 추출하기 위한 전처리 과정은 크게 두 단계로 나눌 수 있다. 우선 MEDLINE 데이터베이스로부터 XML 형식으로 문헌 집합을 다운받아 XML 파싱을 통해 원하는 정보를 추출하는 단계이다. 이 단계에서 미국 국립의학도서관(National Library of Medicine, NLM)에서 제공하는 XML 데이터 요소를 활용하였으며 46개의 상위 요소 중 본 연구에서 필요한 3가지의



〈그림 2〉 연구 개요

요소인 〈PMID〉, 〈ArticleTitle〉, 〈Abstract〉에 포함된 정보들을 가져오기 위해 SAX(Simple API for XML) 파서를 수행하였다.

두 번째 단계로 파싱된 텍스트 데이터에 LingPipe에서 제공하는 문장 경계 인식(sentence boundary detection)을 적용하여 문장단위로 분할을 수행하였다. 이는 텍스트 파일을 입력하여 토큰(tokens)과 공백(whitespace)의 배열로 리스트화한 후 문장 영역을 찾기 위해 MEDLINE 문장 모델을 적용하여 문장 영역을 잡아내는 학습 기반(supervised learning)의 문장 분할 기법이다.

3.1.2 개체 추출

전처리를 통해 분할된 문장을 대상으로 개체 추출을 수행하는 단계는 우선 문장에서 LingPipe 개체명 인식(Named Entity Recognition, NER)을 통해 생의학 개체 유형을 인식해낸다. LingPipe 개체명 인식 모듈은 MEDLINE 데이터베이스로부터 추출한 문헌들의 집합인 GENIA 말뭉치에 의해 학습된 모델로 문장에서 나타난 후보 개체들이 〈표 1〉과 같은 31가지 개체 유형으로 할당된다(Kim, Ohta, Tateisi, & Tsujii, 2003).

〈표 1〉 GENIA 말뭉치의 개체 유형

No.	Entity Type
1	ORGANIC_COMPOUND
2	OTHER_NAME
3	CELL_TYPE
4	PROTEIN_FAMILY
5	DNA_DOMAIN
6	PEPTIDE
7	PROTEIN_MOLECULE

No.	Entity Type
8	TISSUE
9	CELL
10	MULTI_CELL
11	MONO_CELL
12	VIRUS
13	LIPID
14	BODY_PART
15	ARTIFICIAL_SOURCE
16	PROTEIN_SUBUNIT
17	PROTEIN_DOMAIN
18	RNA_FAMILY
19	CARBOHYDRATE
20	INORGANIC
21	DNA_MOLECULE
22	AMINO_ACID
23	CELL_LINE
24	PROTEIN_SUBSTRUCTURE
25	DNA_SUBSTRUCTURE
26	ATOM
27	DNA_FAMILY
28	PROTEIN
29	POLYNUCLEOTIDE
30	NUCLEOTIDE
31	RNA_DOMAIN

추출된 후보 생의학 개체 유형들을 대상으로 UMLS(Unified Medical Language System)와 맵핑하여 존재하지 않는 개체들은 분석대상에서 제외시키고 메타시소러스(metathesaurus)에 포함된 CUI(Concept Unique Identifier)와 해당되는 대표 개념을 추출하며 의미망(semantic network)에 포함된 해당 개체에 대한 의미 유형(semantic type)을 확인한다. 메타시소러스의 개념 표현 중 CUI는 개념에 관한 모든 정보를 연계하고 속성을 규명하여 동일한 의미를 지니는 다양한 형태의 용어들을 하나의 CUI로 통합해주는 역할을 한다. 의미 유형은 메타시소러스에 표현된 모든 개념들과 일치하는 범주(categorization)를 제공하는 넓은 주제 범주 집합(broad subject categories)으로 133개의 의미 유형을 가진다. CUI는 추후 분석을 수행하기 위한 최종 개체로 설정이 되며 이러한 과정을 통해 생성된 개체 추출 결과 값은 〈표 2〉와 같다.

〈표 2〉 개체 추출 결과 값 예시

PMID	Candidate Entity	Entity Type	Semantic Type	CUI
14835125	Raynaud's disease	OTHER_NAME	Disease or Syndrome	C0034734
14835125	disease	OTHER_NAME	Disease or Syndrome	C0012634
14835125	children	MULTI_CELL	Age Group	C0008059
14835125	haematological	OTHER_NAME		
14835125	phenoxybenzamine	OTHER_NAME	Organic Chemical	C0031441
14835125	prostacyclin	OTHER_NAME	Eicosanoid	C0033567
3006901	cod-liver	BODY_PART		
3006901	intimal hyperplasia	OTHER_NAME	Pathologic Function	C0020507
3006901	vein grafts	OTHER_NAME	Biomedical or Dental Material	C0181074
3006901	bypass	POLYNUCLEOT	Therapeutic or Preventive Procedure	C0741847
3006901	Cod-liver oil rich	CELL	Lipid	C0009213
3006901	eicosapentaenoic acid	AMINO_AC	Lipid	C0000545
3006901	unsaturated fatty acid	LIP	Lipid	C0015684
3006901	platelet aggregation	OTHER_NAME	Cell Function	C0032176

이는 PMID가 14835125, 3006901인 두 문헌에 해당하는 후보 개체와 각 개체에 대한 개체명 인식을 통해 추출된 생의학 개체 유형, 이를 다시 UMLS에 맵핑하여 후보 개체에 해당하는 의미 유형과 CUI를 결과로 보여준다. 공란으로 표현된 두 후보 개체인 haematological과 cod-liver는 개체명 인식으로는 개체 유형을 정의하였지만 UMLS에는 해당되는 개체가 존재하지 않았으므로 의미 유형과 CUI는 NULL값을 가진다.

이와 같이 두 단계의 필터링 과정을 통해 생의학 개체를 추출하는 이유는 우선 개체명 인식을 통해 생의학 개체를 추출하여 문장 단위로 UMLS에 직접 맵핑하기 어려운 점을 해결하고자 하였으며, 개체명 인식을 통해 추출한 개체는 정확도가 낮고 개체 인식의 오류가 많은 문제점이 있기 때문에 용어체계를 공통개념으로 통합한 개념을 이용하여 용어의 의미를 올바르게 확장할 수 있도록 하고자 했다.

3.1.3 동사 추출

전처리 과정을 거쳐 파싱된 문장에서 각 문장의 부정여부를 판별하게 된다. 이는 Chapman, Bridewell, Hanbury, Cooper, Buchanan(2001)이 제안한 정규 표현식(regular expression) 알고리즘인 NegEx 알고리즘을 이용하였다. 이 알고리즘은 약 98%의 정확도를 가지는 알고리즘으로 총 272개의 부정 표현이 수록된 사전을 확인하여 각 문장의 부정 여부를 판별하는 과정을 거친다.

개체간의 유의미한 관계를 형성하기 위한 동사 추출은 두 단계로 나눌 수 있는데 우선 조건 랜덤 필드 기반의 품사 태깅을 수행하여 동사를 식별해 낸다. 조건 랜덤 필드는 Lafferty, McCallum, Pereira(2001)에 의해 정의된 구조적인 예측을 위해 사용되는 연속적인 데이터의 레이블링과 분할을 위한 확률적 모델이다.

품사 태깅을 통해 동사로 인식된 단어들의 표제어 복원을 수행하기 위해 캠브리지 대학의 컴퓨터학과 실험실에서 만든 어휘 획득(lexical acquisition) 기법인 영어 동사를 자동적으로 분류하기 위한 동사 리스트들을 사용하였다(Sun & Korhonen, 2009). 생의학 영역으로 분류된 399개의 동사 리스트를 참고하여 다양한 형태로 표현된 동사들을 표준 형태로 통일하는 동사의 표제어 복원을 수행하였다.

3.1.4 개체간 관계성 추출

개체 추출과 동사 추출을 통해 두 개체간의 관계성(relationship)을 확인하기 위해 기본적으로 한 문장 내에 존재하는 개체들 간의 동시출현 빈도 기반의 관계를 설정하며, 또 다른 유형은 추출된 동사를 기반으로 하는 관계를 설정한다.

우선 개체 추출을 통해 생성된 CUI를 기반으로 대표 개념을 최종적으로 선정하여 문장 내에서 개체와 개체간의 동시출현 빈도를 계산하였다. 〈표 3〉은 본 연구의 데이터인 레이노병과 어유를 질의어로 레코드를 받아 수행한 결과 값으로 총 38,878건의 동시출현 개체 쌍 중 빈도가 100이상인 8개의 개체 정보를 보여준다.

〈표 3〉 100회 이상의 동시출현 빈도를 갖는 개체 정보

No.	Entity	Entity	Frequency
1	Raynaud Phenomenon	Patients	604

No.	Entity	Entity	Frequency
2	Raynaud Disease	Patients	382
3	Patients	Systemic Scleroderma	250
4	Patients	Scleroderma	129
5	Eicosapentaenoic Acid	Arachidonic Acid	122
6	Diet	Fatty Acids	120
7	Fatty Acids	Phospholipids	120
8	Raynaud Phenomenon	Systemic Scleroderma	107

이 중 전 단계의 관계 추출을 통해 추출된 동사를 기준으로 구문 분석을 수행하여 좌측과 우측으로 가장 근접한 두 개체를 대표적인 개체로 연결한다. 거리 계산은 단어의 위치에 기반한 것으로 문장을 문자열로 표현하여 추출된 후보 개체명의 시작열과 종료열을 표시한다. 즉, 문장 내의 모든 단어를 리스트 형태로 불러들여온 후 동사를 기준으로 좌, 우측으로 가장 근접한 개체를 선택하여 개체-관계-개체 구조를 형성한다. 〈표 4〉는 레이노병-어유와 관련된 문헌 데이터에서 CUI에 해당하는 대표 개념과 이들 사이에 존재하는 동사를 보여준다. 동사를 기준으로 좌측에 존재하는 개체는 동사를 수행하는 주체이며 우측에 존재하는 개체는 동사 수행을 받는 객체가 된다. 더불어 개체간 동사 연결 관계가 긍정인지 부정인지의 여부를 판별하게 되며 긍정문은 POS, 부정문은 NEG로 표현된다.

3.1.5 개체 시각화

전 단계들을 통해 추출된 개체 쌍과 동시출현 빈도 값을 그래프로 표현하기 위해 XML기반의 파일 형식인 GraphML로 변환하였고 시각화 도구인 Gephi를 통해 개체간의 동시출현 연결성을 구조적으로 파악하고자 했다. 데이터의 크기와 속성에 따라 에지와 노드의 값을 필터링하여 네트워크를 분석하였다.

동시출현 빈도 값들을 기반으로 시각화를 하여 각 개체들은 네트워크상의 노드로 표현하고 동시출현 빈도수는 각 노드를 연결하는 에지의 가중치로 표현하였다. 네트워크상에서 중심성이란 네트워크상의 특정 노드가 중심과 가까이 위치하는지를 나타내는 정도로 중심성을 확인하는 대표적인 기법인 연결정도 중심성(degree centrality)은 네트워크상의 노드가 얼마나 많

〈표 4〉 개체와 개체간의 관계 표현

No.	Entity	Relation	Entity	Negation
1	Fatty Acids	inhibit	Platelet aggregation	POS
2	Organic	reverse	Phase	POS
3	Epoprostenol	support	Resistance(Psychotherapeutic)	POS
4	Rattus norvegicus	modify	Fatty Acids	POS
5	Voluntary Workers	reduce	Triglycerides	POS
6	Platelet aggregation	prolong	Skin Specimen	POS
7	Raynaud Disease	increase	Skin Temperature	POS
8	trimethyloxamine	react	Sodium Nitrite	POS
9	Immune system	occur	Histocompatibility	POS
10	Blood Platelets	convert	Eicosapentaenoic Acid	NEG

은 노드들과 직접적으로 연결되어 있는지를 측정하는 도구이다. 즉, 연결정도 중심성이 높은 개체는 타 개체와 연결성을 보이는 경우의 수가 많은 값이다.

3.1.6 그래프 모델 기반의 경로 추론

본 연구의 최종 단계로 그래프 데이터베이스인 Neo4j를 이용하여 그래프 모듈을 구축하여 개체 추출과 관계 추출을 통해 나타난 개체들과 관계들을 대상으로 두 개체 사이의 모든 경로를 구한다. Neo4j는 네오 테크놀로지(Neo Technologies)에서 개발한 NoSQL(Not Only SQL) 그래프 데이터베이스로 노드(node)와 관계(relationship), 속성(property)으로 구성된다. 노드를 이용하여 데이터를 보관하며 관계를 통해 두 노드 사이의 연결을 표현해준다. 관계는 두 노드의 방향성을 가지며 단방향성과 양방향성을 가질 수 있다. 본 연구의 관계 유형은 첫째, 개체와 개체간의 동시출현 빈도 유형이며 둘째, 개체와 개체간의 동사로 연결된 관계 유형이다. 개체 사이를 연결하는 Reduce, Increase와 같은 동사는 관계성을 통해 노드 사이의 연결을 표현하는 중요한 단서로 작용한다. 또한 노드와 관계는 〈키, 값〉으로 구성되는 속성을 통해 값을 가지게 되며 속성은 해당 노드에 대한 다양한 정보 값을 저장할 수 있다. 즉, 본 연구에서 속성은 노드와 노드 사이의 관계에 대한 동시출현 빈도 값이 된다. 소스 노드(source node)와 목표 노드(target node) 사이에 존재하는 경로를 확인하여 유의미한 관계성을 추론하고자 했다.

3.2 실험

본 연구의 실험에서는 ABC 모델의 그래프 추론을 위해 Swanson의 문헌 기반 발견 연구에서 초기 연구(Swanson, 1986b)로 진행된 레이노병과 어유의 관계를 폐쇄형 접근법을 통해 모사한다. 동시출현 개체 기반의 네트워크 시각화와 그래프 모델 기반의 경로를 파악함으로써 노드간의 관계성을 풍부하게 밝히는 사례 연구를 수행한다.

3.2.1 데이터 집합

데이터는 생의학 관련 문헌들의 서지정보가 수록된 데이터베이스인 MEDLINE을 접근 가능하게 해주는 검색 엔진인 PubMed에서 XML 형식으로 데이터를 다운 받았다. 폐쇄형 발견 과정에서의 개념 A와 C는 사전에 관계성이 알려져 있거나 정의된 용어로 검색 과정에서 양방향성을 지닌 상태로 동시에 수행된다. 따라서 상세 검색(advanced search)을 통해 OR 연산자로 “Raynaud’s Disease”와 “Fish Oils” 용어를 제목과 초록에서 포함하고 있는 모든 학술 문헌을 수집하였다. OR 연산자를 적용한 이유는 두 용어에 해당하는 데이터를 개별적으로 수집했을 때 발생하는 중복의 문제점을 피하기 위한 것이다.

본 연구의 데이터 집합은 크게 두 가지로 기본 집합(background set)과 검증 집합(test set)으로 나뉜다. 우선 실험분석 대상이 되는 기본 집합은 Swanson이 처음으로 발견한 레이노병과 어유의 관계에 대한 가설 발견 연구인 1986년의 서지학적 상황을 동일한 환경으로 적용하기 위한 데이터로 출판일(publication date)을

MEDLINE 데이터베이스가 활용되기 시작한 시점인 1960년부터 1986년까지로 설정하여 총 3,877건의 학술 문헌 데이터를 수집하였다. 또한 추후 검증을 위한 집합으로 1987년부터 현재까지의 "Raynaud's Disease"와 "Fish Oils" 용어를 포함한 총 24,704건의 데이터를 수집하였다. 이를 통해 기본 집합의 분석 결과에 대한 신뢰성을 검증하고자 하며 그래프 모델로 표현된 두 집합의 공통점을 비교 분석해보고자 한다.

3.2.2 데이터 통계

MEDLINE 데이터베이스를 통해 수집한 실험 데이터 집합을 가지고 본 연구에서 설계한 시스템 수행에 대한 결과값을 MySQL에 저장하였다. 두 데이터 집합에 대한 기초 통계 결과는 다음과 같다. 〈표 5〉는 기본 집합에 대한 기초 통계 값이며 〈표 6〉은 검증 집합에 대한 기초 통계 값이다.

〈표 5〉 레이노병-어유 질의어에 대한 기초 통계(1960-1986)

Sentence	Entity	Unique Entity	Relation
23,497	68,139	16,912	2,474

〈표 6〉 레이노병-어유 질의어에 대한 기초 통계(1987-현재)

Sentence	Entity	Unique Entity	Relation
21,380	77,734	31,517	4,019

4. 결과 분석

4.1 그래프 모델 기반의 레이노병-어유 경로 분석

4.1.1 Swanson의 베이스라인

본 연구에서는 Swanson(1986b)에 의해 발견된 세 개의 단어인 혈액 점도, 혈소판 응집력, 혈관 수축과 본 연구에서 제안하는 그래프 모델 기반의 경로 추론 결과를 비교 분석한다. 따라서 Swanson의 세 개의 단어를 베이스라인으로 설정하였고 본 연구의 기법에 의해 변경된 대표 개념과 각 대표 개념에 해당하는 개체 유형, 의미 유형, CUI에 대한 정보를 나타내면 〈표 7〉과 같다.

우선 두 개체간의 동시출현 분석 결과에 대한 시각화를 통해 전체적인 네트워크 구조를 파악하고자 하며 Swanson에 의해 제안된 세 가지 개념의 분석을 시도한다. 더불어 그래프 모델 기반의 경로를 단계별로 추론하여 각 개념의 대표 경로를 파악하고 각 경로에 대한 분석을 수행한다.

4.1.2 두 개체간의 동시출현 분석

레이노병-어유 데이터에서 총 38,878건의 개체 쌍이 나타났으며 전체 데이터의 노드는 4,991개, 에지는 35,341개의 구조를 보였다. 초기 데이터를 대상으로 시각화를 한 결과 노드와 에

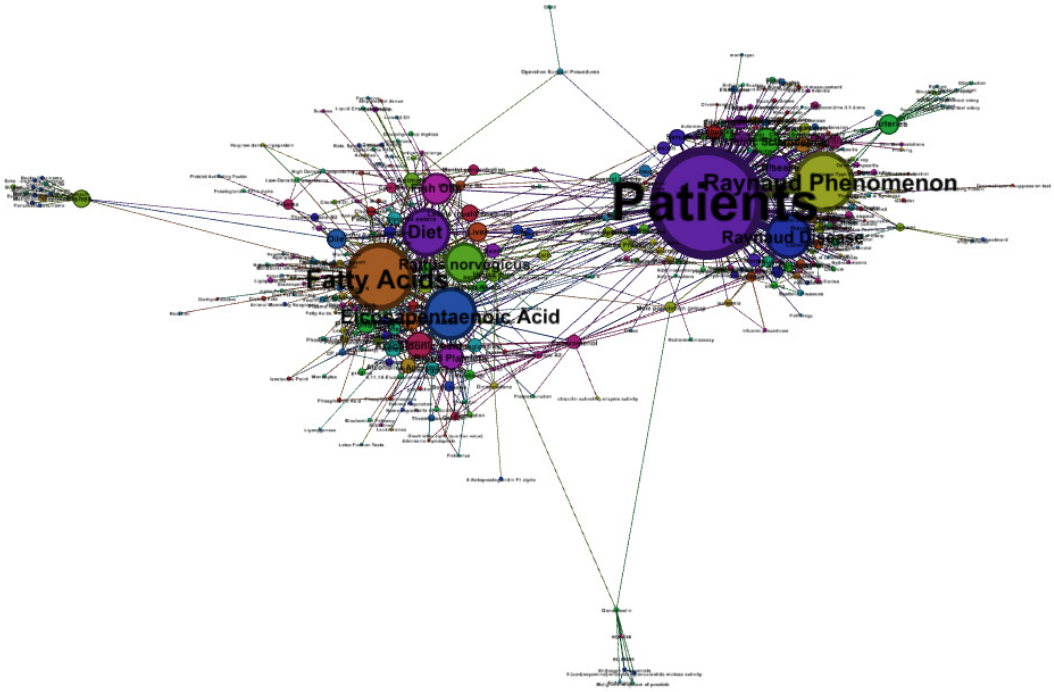
〈표 7〉 Swanson의 B단어에 대한 개념 정보

Concept	Entity Type	Semantic Type	CUI
Blood Viscosity	OTHER_NAME	Physiologic Function	C0005848
Platelet aggregation	OTHER_NAME	Cell Function	C0032176
Vascular constriction(function)	OTHER_NAME	Organ or Tissue Function	C0042396

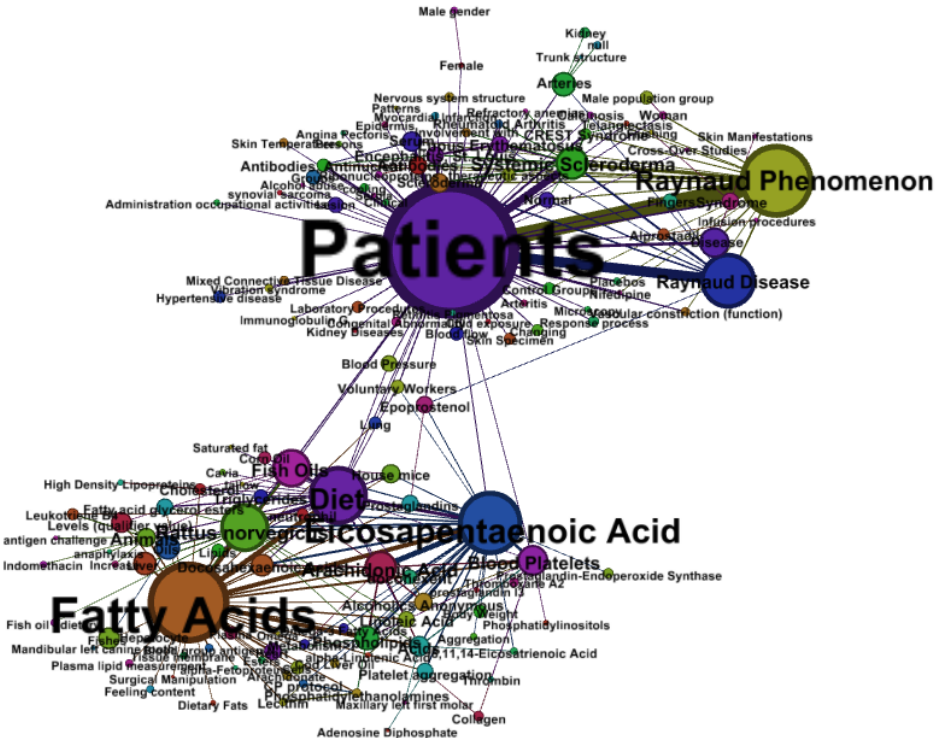
지의 수가 방대하여 노드간의 관계성 확인이 어려운 점을 고려하여 연결정도 중심성 값이 10 이상인 노드들로 3번의 정제화 과정을 거친 470개의 노드와 가중치가 20 이상인 907개의 에지에서 고립노드 131개를 제거하여 〈그림 2〉와 같은 339개의 노드와 907개의 에지에 대한 네트워크가 형성되었다. 이에 대한 평균 연결정도 중심성 값은 5.351이다.

가장 높은 연결정도 중심성 값을 가지는 지방산류인 Fatty Acids와 환자인 Patients는 크게 좌, 우로 나뉜 노드 집합에서 중심을 형성하고 있다. 본 사례 연구의 대표 개체로 선정된 Fish Oils는 연결정도 중심성 값이 234로 Fatty Acids와 연결성을 지니는 좌측 군집을 이루고, 또 하나의 대표 개체인 Raynaud Phenomenon, Raynaud Disease는 각각 중심성 값이 500, 362로 Patients 노드를 중심으로 우측 군집을 이루고 있다.

두 군집을 매개해주는 대표적인 노드들을 확인하기 위해 추가적으로 에지의 가중치 값이 20 이상인 282개의 에지들을 선별하였고 〈그림 3〉의 339개의 노드 중 에지를 삭제함에 따라 고립노드가 된 190개의 노드를 삭제하여 146개의 노드로 구성되었다. 최종적으로 평균 연결정도 중심성 값은 3.863이며 이에 대한 결과는 〈그림 4〉와 같다. 두 군집 사이에 존재하는 4가지의 매개 노드인 Blood Pressure, Voluntary Workers, Epoprostenol, Lung은 두 군집의 대표 노드인 Patients, Raynaud Disease, Diet, Eicosapentaenoic Acid, Fatty Acids와 같은



〈그림 3〉 노드 339, 에지 907개의 동시출현 빈도 시각화



〈그림 4〉 노드 146, 에지 282개의 동시출현 빈도 시각화

연결정도 중심성 값이 높은 노드들과 연결성을 지니는 점을 확인할 수 있었다.

Swanson의 세 가지 개념들에 대한 동시출현 정보로는 Platelet aggregation 개념이 179회로 총 179개의 개념들과 동시출현 연결성을 지녔다. 다음으로 Vascular constriction(function)은 121회, Blood Viscosity는 73회로 나타났다.

4.1.3 그래프 모델 기반의 경로 추론

두 대표 개념인 Raynaud Disease와 Fish Oils의 경로를 확인하기 위해 모든 개체 사이에 존재하는 중간 개념인 B를 다수준으로 확장해나가는 분석을 시도했다. 추출된 모든 개체와 관계에 대해 중간 개념의 수에 따라 경로 단계를 1단계부터 시작하여 3단계의 경로까지 확인했다. 추출된 모든 경로에 속한 대표적인 경로를 Swanson의 B단어에 해당하는 개념들과 비교 분석한다. 대표 경로의 선정 기준은 (1) Swanson의 세가지에 해당되는 B단어가 나타난 경로 (2) 이 중 개체 간의 동사 관계가 나타난 경로로 수행되었다. 단계가 높아짐에 따라 B개념이 다른 개념들과 어떠한 관계성을 보이는지에 대해 특징을 발견하고 Swanson의 연구에서 보여주지 못했던 각 개체와의 관계성을 발견하여 유의미한 경로를 추적하였다.

분석에서 각 경로에 표현된 화살표는 방향성을 나타낸다. 두 가지의 관계설정 유형 중 동시출현 유형에서의 두 개체는 방향성 측면에서 큰

의미를 갖지 않는 무방향 그래프(undirected graph)가 적합하지만 경로를 파악하기 위해 문장 내에서 좌측에 위치하는 순서대로 순방향(→)으로 표현하였다. 동사 유형의 경우 방향 그래프(directed graph)이기 때문에 동사를 행하는 개체에서 받는 개체로 향하도록 방향성을 설정하였다. 개체 사이에 존재하는 동사 연결은 "( )" 안에 표현하였으며 동사의 부정을 의미하는 not과 같은 표현이 함께 등장하였을 경우 이를 부정문으로 해석하여 함께 표현했다.

1) 1단계 경로

우선 두 대표 개체의 경로 단계를 1단계로 설정하였을 때 두 개념에서 나타난 중간 개념 B는 총 54개였으며, 이 중 중복되는 개념 21개를 삭제하여 〈표 8〉과 같이 총 33개의 고유한 B개념이 추출되었다.

〈표 8〉 33개의 B개념

No.	B concept
1	Groups
2	Heart
3	Heart Diseases
4	Human body
5	Hypertensive disease
6	Indomethacin
7	Ischemia
8	Kidney
9	Lesion
10	Levels (qualifier value)
11	Lung
12	Lupus Vulgaris
13	Macao
14	Male gender
15	Male population group
16	Mental concentration
17	Metabolite
18	Myocardial Infarction
19	Necrotizing vasculitis
20	Norepinephrine
21	Normal blood pressure
22	Observation parameter
23	Patients
24	Physical activity
25	Plasma
26	Platelet aggregation
27	Platelet Factor 4
28	Platelet function
29	Process of secretion
30	Prostaglandins
31	Response process
32	Thromboxanes
33	vascular reactivity

2) 2단계 경로

2단계의 B개념 경로를 확인한 결과 Platelet aggregation은 총 49회 출현하였으며 Prostaglandins, doconexent 개념과 각각 동사 유형 regulate, affect로 연결되었고 두 동사 유형에 대한 경로는 "Raynaud Disease → Prostaglandins → (regulate) → Platelet aggregation → Fish Oils", "Raynaud Disease → doconexent → (affect) → Platelet aggregation → Fish Oils"이다. 동사의 좌측에 존재하는 Prostaglandins, doconexent 개체는 소스 노드인 Raynaud Disease와 동시 출현하였고 동사의 우측에 존재하는 Platelet aggregation은 목표 노드인 Fish Oils와 동시 출현하였다.

3) 3단계 경로

3단계의 경로에서는 이 전에는 나타나지 않았던 Swanson의 B단어 중 하나인 Blood Viscosity

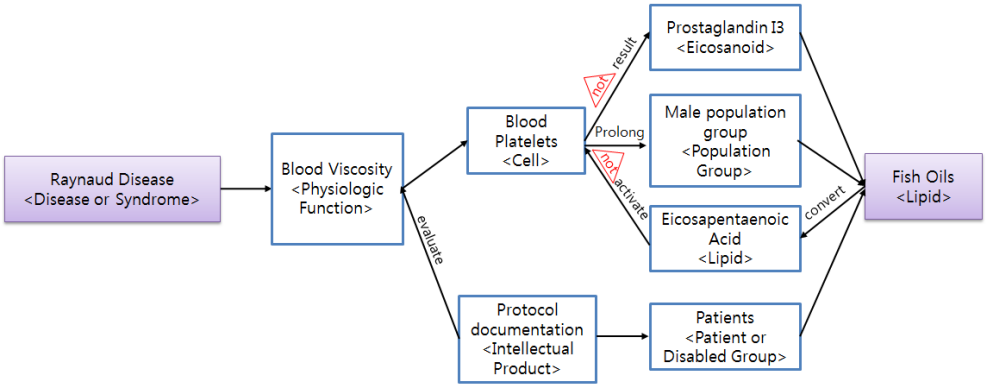
개념이 총 4회 출현하였으며 소스 노드에서 목표 노드까지 가는 전체 경로는 4가지로 〈그림 5〉와 같다. 각 개체명에 해당하는 의미 유형은 "〈 〉" 안에 표현하였다.

Raynaud Disease는 질병 또는 증후군 의미 유형에 속한 개념이며 Blood Viscosity는 생리적 기능을 의미하는 Physiologic Function 의미 유형에 속해 있다. 이 둘은 동시출현으로 연결성을 보이며, Blood Platelets는 Cell 의미 유형에 속해 있다. Blood Viscosity를 기준으로 좌측으로는 Raynaud Disease, 우측으로는 Blood Platelets이 3회, Protocols documentation이 1회 연결되며 이는 evaluate 동사를 통해 "Protocols documentation → (evaluate) → Blood Viscosity" 연결 구조를 보였다. 동사 유형 관계에서 1-2단계의 B개념과 연결된 evaluate 동사를 비롯하여 2-3단계의 B개념과 연결되는 3개의 동사인 not result, Prolong, not activate, 그리고 3단계와 목표 노드와 연결되는 동사인 convert의 연결 구조 경로를 확인할 수 있었다.

또한 2단계까지 나타나지 않았던 Swanson의 B단어인 Vascular constriction(function)은 3단계 수행동안 총 12회 출현하였으며 13개의 동시출현한 개념과 그에 해당하는 의미 유형은 〈표 9〉와 같다. 좌측으로는 1번 개념인 receptor가 총 12회 출현하였고 우측으로는 나머지 12개의 각 개념이 1회씩 출현하였다.

〈표 9〉 Vascular constriction(function)과 동시출현한 13개 개념과 의미 유형

No.	Concept	Semantic Type
1	receptor	Amino Acid, Peptide, or Protein
2	Lesion	Finding
3	Thromboxane A2	Eicosanoid
4	Norepinephrine	Organic Chemical
5	Platelet aggregation	Cell Function
6	Physical activity	Daily or Recreational Activity
7	Response process	Clinical Attribute
8	Thromboxanes	Eicosanoid
9	Prostaglandins	Eicosanoid
10	Scientific Study	Laboratory Procedure
11	Heart	Body Part, Organ, or Organ Component
12	Metabolite	Biologically Active Substance
13	Eicosapentaenoic Acid	Lipid



〈그림 5〉 Blood Viscosity 개념과 연결되는 4가지의 경로

4.2 결과 검증

분석 결과를 검증하기 위해 검증 데이터를 대상으로 동일하게 실험을 수행하여 세 가지 개념에 대한 동사 관계를 중심으로 검증 문헌에서도 개념간의 관계경로가 동일한 의미로 발견되는지를 비교 분석하였으며 대표적인 예를 한가지씩 기술한다.

4.2.1 Platelet aggregation 검증

기존 문헌 2단계 경로의 Platelet aggregation 관계에서 나타난 doconexent는 혈소판 응집에 영향을 미치는(affect) 관계가 형성되었다. 이는 검증 집합의 3단계 경로에서 "Raynaud Disease → Clinical action → (elicit) ← doconexent → (reduce) → Platelet aggregation → Fish Oils"와 동일한 의미로 나타났다.

즉, doconexent는 양방향으로 두 가지의 동사 연결을 보이는데 〈표 10〉에서 나타난 실제 문장의 후보 개체에서 소염작용을 의미하는 anti-inflammatory actions를 끌어내는(elicit) 역할을 하며, 혈소판 응집형을 감소시키는(reduce) 역할을 한다는 점을 확인했다. 즉, 기존 문헌에서 영향을 미치는 관계가 검증 문헌에서는 감소시키는 관계로 나타나 보다 명확한 관계를 확인할 수 있었다.

4.2.2 Blood Viscosity 검증

기존 분석의 3단계 경로에서 Blood Platelets와 연결되는 동사 유형 중 "Raynaud Disease → Blood Viscosity → Blood Platelets ← (not activate) ← Eicosapentaenoic Acid ← (convert) ← Fish Oils"에 대한 검증 결과로 "Raynaud Disease → Child → Blood Platelets ← (increase) ← doconexent ← Fish Oils" 경로가 나타났다.

기존 분석에서 Eicosapentaenoic Acid는 불포화 지방산으로 혈소판 활성화를 일으키지 않는다는 의미로 해석되었다. 이는 검증 데이터를 통한 분석에서 "doconexent → (increase) → Blood Platelets" 관계를 통해 불포화 지방산의 한 종류인 doconexent가 혈소판 작용을 증가시켜 혈소판 응집을 억제하는 것으로 해석될 수 있다.

4.2.3 Vascular constriction(function) 검증

3단계 경로의 "Eicosapentaenoic Acid → (reduce) → Vascular constriction(function)"인 불포화 지방산이 혈관 수축을 막는다는 사실에 대한 검증을 수행하기 위해 "Raynaud Disease → Blood Vessels ← (reduce) ← Eicosapentaenoic Acid → (reduce) → Gene Expression → Fish Oils" 경로를 발견하였다. 불포화 지방산은 혈관을 의미하는 Blood Vessels와 유전자에 의해 결정

〈표 10〉 doconexent와 동사 유형으로 연결되는 두 문장

PMID	Sentence
22765297	In adult humans, an EPA plus DHA intake greater than 2 g day^{-1} seems to be required to elicit anti-inflammatory actions, but few dose finding studies have been performed.
23390192	Although long-chain n3 polyunsaturated fatty acids [n3 PUFAs; eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA)] have been reported to reduce platelet aggregation, the available evidence on this is equivocal.

〈표 11〉 Eicosapentaenoic Acid와 동사 유형으로 연결되는 두 문장

PMID	Sentence
23246023	These results suggested that EPA but not pravastatin may reduce cardiac afterload by reducing vascular reflected waves and lowering C-SBP.
23361365	Taken together, our results demonstrate that DHA and EPA are able to reduce IL-6-induced CRP expression in HepG2 cells via an inhibition of STAT3 activation.

되는 형질이 표현형으로 나타나는 유전자 발현인 Gene Expression 두 개념에 대해 모두 reduce 동사 유형을 가지며 이에 해당하는 문장은 〈표 11〉과 같다. 즉, 불포화 지방산은 혈관 수축을 감소시키며 유전자 발현의 속도를 늦출 수 있다. 이는 기존 분석에서 불포화 지방산이 혈관 수축 기능을 감소시킨다는 의미와 일맥상통한 의미로 해석되며 특히 "Blood Vessels ← (reduce) ← Eicosapentaenoic Acid" 경로 관계에서 나타난 혈관을 수축한다는 의미는 혈관 수축 개념인 Vascular constriction(function)과 동일한 의미를 갖는다.

5. 결 론

Swanson에 의해 제안된 문헌 기반 발견의 연구들은 생의학 문헌 내에서 생의학 개체의 동시출현 값을 기반으로 중간 개념을 발견하여 다양한 가설들을 발견해내었다. 이러한 연구는 임상의학자들에 의해 의학적 실효성이 검증되었으며 많은 연구자들의 관심을 불러일으켰다. 하지만 기존 연구들의 문제점인 반자동화된 접근과 중간 개념인 B를 한가지로만 발견해낸다는 문제점에 착안하여 본 연구에서는 문장 내에서 발견된 동사를 추출함으로써 다수준의 다양한 상호작용 관계를 규명하고자 했다. 즉 그래프 데이터베이스인 Neo4j를 이용하여 소스 노드와 목표 노드 사이의 다양한 경로를 확인하며 이들 간의 관계성을 종합적으로 분석하여 미발견 공공 지식을 추론하고자 했다.

본 연구 결과로 동시출현 기반의 개체 시각화를 통해 개체간의 관계를 구조적으로 파악하였으며 궁극적으로 동사 연결을 포함한 경로를 Swanson의 세 가지의 B단어와 비교하였다. 중간 개념 B의 단계를 1단계부터 3단계까지 확장해나가며 경로 내 개체 간의 다양한 관계를 풍부하게 해석하였다. 또한 본 연구에서 제안한 그래프 기반의 경로 추론 기법의 분석 결과를 검증하기 위해 레이노병과 어유 가설이 제안된 1986년을 기준으로 두 문헌 집합을 나누어 분석 결과를 비교 분석하였으며 결과적으로 검증 집합에서도 기존 집합과 유사한 의미로 유의미한 경로를 형성하고 있음을 확인했다. 즉, 두 문헌 집합의 비교 분석을 통해 분석 결과의 신뢰성을 입증할 수 있었으며 생의학 개념들 간의 내재된 관계성을 효율적으로 파악할 수 있었다.

본 연구의 제한점으로 경로의 효율성 뿐만 아닌 정확성 측면에서 자동적으로 유의미한 경로를 순위화하는 과정이 필요하며 보다 정교한 구문분석을 통해 개체와 개체간의 관계성을 추출해야 하는 연구과제가 남아있다.

본 연구는 국내에서는 시도되지 않았던 문헌 기반 발견 연구에 대해 차별화된 방법론적 시

도를 보임으로써 가설 발견의 효율성을 높이고자 다양한 관계성을 파악하는데 주목적을 둔 연구로 시사점을 지닌다. 이는 문헌 기반 가설 연구의 새로운 가설 발견의 가능성을 제시할 수 있으며 문헌 검증에서 더 나아가 추후 임상 실험을 통한 현장 검증이 이루어진다면 본 연구의 활용가치가 더 높아질 수 있을 것이라 기대한다.

참 고 문 헌

Automatic Classification for English Verbs. (2013, July 1). Retrieved from http://www.cl.cam.ac.uk/~ls418/resource_release/

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A. P., & 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.

DiGiacomo, R. A., Kremer, J. M., & Shah, D. M. (1989). Fish oil dietary supplementation in patients with Raynaud's phenomenon: A doubleblind, controlled, prospective study. American Journal of Medicine, 8, 158-164.

Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., Polman, J., & Alkema, W. (2008). CoPub: a literature-based keyword enrichment tool for microarray data analysis. Nucleic Acids Research, 36(suppl 2), W406-W410.

Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J., & Alkema, W. (2010). Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Computational Biology, 6(9), 1-11. e1000943.

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. In AMIA Annual Symposium Proceedings, 349-353. American Medical Informatics Association.

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.

Hristovski, D., Rindflesch, T., & Peterlin, B. (2013). Using literature-based discovery to identify novel therapeutic approaches. Cardiovascular and Hematological Agents in Medicinal Chemistry, 11(1), 14-24.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28(23), 3158-3160.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics, 19(1), 180-182.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, 282-289.

Liekens, A. M., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., & Del-Favero, J. (2011). BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biology, 12(6), R57.

LingPipe: Named entity tutorial. (2013, July 1). Retrieved from http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html/

LingPipe: Sentence boundary detection. (2013, July 1). Retrieved from http://alias-i.com/lingpipe/demos/tutorial/sentences/read-me.html/

MEDLINE, PubMed XML element descriptions and their attributes. (2013, October 10). Retrieved from http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html/

Narayanasamy, V., Mukhopadhyay, S., Palakal, M., & Potter, D. A. (2004). TransMiner: Mining transitive associations among biological objects from text. Journal of Biomedical Science, 11(6), 864-873.

NegEx (2013, December 1). Retrieved from http://code.google.com/p/negex/

PubMed (2013, August 2). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/

Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. Neuroscience Research Communications, 15(1), 1-9.

Smalheiser, N. R., & Swanson, D. R. (1996a). Indomethacin and Alzheimer's disease. Neurology, 46(2), 583-583.

Smalheiser, N. R., & Swanson, D. R. (1996b). Linking estrogen to Alzheimer's disease: An informatics approach. Neurology, 47(3), 809-810.

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

Sun, L., & Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2, 638-647. Association for Computational Linguistics.

Swanson, D. R. (1986a). Undiscovered public knowledge. The Library Quarterly, 56(2), 103-118.

Swanson, D. R. (1986b). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine, 31(4), 526-557.

Swanson, D. R. (1990a). Somatomedin C and arginine: Implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine, 33(2), 157-186.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 91(2), 183-203.

Swanson, D. R., Smalheiser, N. R., & Bookstein, A. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. Journal of the American Society for Information Science and Technology, 52(10), 797-812.

Swanson, D. R., Smalheiser, N. R., & Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. Journal of the American Society for Information Science and Technology, 57(11), 1427-1439.

UMLS Reference Manual. (2013, October 10). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK9676/

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.

Weeber, M., Vos, R., Klein, H., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association, 10(3), 252-259.

Wilkowski, B., Fiszman, M., Miller, C., Hristovski, D., Arabandi, S., Rosemblat, G., & Rindflesch, T. (2011). Discovery browsing with semantic predications and graph theory. In AMIA Annual Symposium Proceedings.