Korean Chem. Eng. Res., 57(6), 781-789 (2019)
https://doi.org/10.9713/kcer.2019.57.6.781
PISSN 0304-128X, EISSN 2233-9558

반응 위험성분석 및 사고방지를 위한 스마트 합성경로 탐색시스템

정준수 · 김창완 · 곽동호 · 신동일†

명지대학교 화학공학과
17058 경기도 용인시 처인구 명지로 116
(2019년 4월 29일 접수, 2019년 7월 17일 수정본 접수, 2019년 8월 21일 채택)

Smart Synthetic Path Search System for Prevention of Hazardous Chemical Accidents and Analysis of Reaction Risk

Joonsoo Jeong, Chang Won Kim, Dongho Kwak and Dongil Shin†

Department of Chemical Engineering, Myongji University, 116, Myongji-ro, Cheoin-gu, Yongin-si, Gyeonggi-do, 17058, Korea
(Received 29 April 2019; Received in revised form 17 July 2019; accepted 21 August 2019)

요 약

연구실 실험, 파일럿 플랜트 및 반응기 운전 중 화학물질에 의한 안전사고가 발생하고 있다. 합성 실험을 시작하기전 사고예방을 위해 관련 정보들을 찾아볼 필요가 있으며, 공정설계 단계에서도 반응 폭주 예방을 위한 반응 정보의 확보는 필수적이다. 합성반응 관련 정보는 인터넷을 포함해 다양한 source가 존재하지만, 검색에 오랜 시간이 걸리고, 합성법마다 사용되는 물질도 달라 적정경로 선택의 어려움이 있다. 연구자들의 합성경로 검색시간 단축과 합성 시 존재할 수 있는 위험성 및 중간생성물질들의 확인에 도움을 주고자 본 연구는 스마트 합성경로 탐색시스템을 제안하였다. 제안한 탐색시스템은 Python 패키지인 Selenium을 사용한 Web scraping 및 Web crawling을 통해 인터넷에 존재하는 정보를 수집하여 DB를 자동으로 갱신한다. 경로 탐색 알고리즘은 depth-first search에 기반하여 목표 물질을 기준으로 탐색을 진행하고, 유해화학물질 등급, 수율 등을 구분하여, 제한된 경로 단계 수치내에 있는 모든 합성 경로를 제안한다. 또한 각자의 연구 목적에 맞게 연구원들이 가진 비공개 데이터를 형식을 맞춰 DB에 등록하여 확장할 수 있다. 시스템은 차후에 무료 사용이 가능하도록 open source로 공개할 예정이다. 개발 시스템은 연구자들이 제안된 경로를 참고하여 더 안전한 반응 방법을 찾고, 사고의 예방에도 도움을 줄 것으로 기대된다.

Abstract – There are frequent accidents by chemicals during laboratory experiments and pilot plant and reactor operations. It is necessary to find and comprehend relevant information to prevent accidents before starting synthesis experiments. In the process design stage, reaction information is also necessary to prevent runaway reactions. Although there are various sources available for synthesis information, including the Internet, it takes long time to search and is difficult to choose the right path because the substances used in each synthesis method are different. In order to solve these problems, we propose an intelligent synthetic path search system to help researchers shorten the search time for synthetic paths and identify hazardous intermediates that may exist on paths. The system proposed in this study automatically updates the database by collecting information existing on the Internet through Web scraping and crawling using Selenium, a Python package. Based on the depth-first search, the path search performs searches based on the target substance, distinguishes hazardous chemical grades and yields, etc., and suggests all synthetic paths within a defined limit of path steps. For the benefit of each research institution, researchers can register their private data and expand the database according to the format type. The system is being released as open source for free use. The system is expected to find a safer way and help prevent accidents by supporting researchers referring to the suggested paths.

Key words: Chemical Accident, Synthetic Path, Graph Algorithm, Web Scraping, Intelligent Search System

† To whom correspondence should be addressed.
E-mail: dongil@mju.ac.kr
‡이 논문은 서울과학기술대학교 김래현 교수님의 정년을 기념하여 투고되었습니다.

1. 서 론

안전사고 통계자료 중 2018년도 위험물 통계자료의 사고 현황을 확인해보면 2016년에 총 81건의 위험물 사고가 발생하였고, 2017년에 총 72건의 위험물 사고가 발생하였다(Fig. 1(a), Table 1)[1]. 그 중 절반에 가까운 수치가 화재 사고이며 나머지는 폭발사고와 누출사고의 비율이 비슷하다. Fig. 1(b), Table 2에서 전체 위험물 사고에 대한 인명피해 현황을 살펴보면 2016년에는 총 81건의 사고에 대해 총 47명의 인명피해가 있었으며, 그중 사망자는 6명, 중상자는 24명, 경상자는 17명이 있었다. 2017년에는 총 72건의 사고가 있었으며, 총 인명피해는 39명, 사망자는 7명, 중상자는 6명, 경상자는 26명으로 중상자의 비율은 줄었지만 사망자는 늘어났다. 이처럼 인명피해 발생 시 중상 이상의 피해를 입게 될 확률이 크다. 위험물 사고와 비슷하게 파일럿 플랜트, 반응기 운전 및 실험 진행 시 화학물질을 다룰 때에도 안전사고가 발생하게 되면 큰 피해가 생길 가능성이 있다.

안전사고의 한 가지 예로, 부타디엔 실험 중 폭발사고가 있다[2]. 유화중합 실험을 하는 도중에 밸브를 연 후 샘플 채취를 하고 밸브를 제대로 잠그지 않아 인화성물질이 누출되어 일어난 폭발사고이다. 사망 1명, 부상 1명의 인명피해가 있었으며, 건물 2층에서 발생한 화재로 8~10억의 물적 피해가 발생하였다. 유사한 사고 사례로 합성의 반응물과 중간 생성물과 관련된 유해성 및 위험성을 제대로 파악하지 못하여 실험 도중 위험성을 인지하지 못하여 발생한 사고, 시약 정리 도중 표시가 불분명한 시약병의 폭발사고 등도 존재한다. 이러한 사고 현황들을 종합하여 Fig. 2에 나타냈다. Fig. 2(a)를 보면 충돌 및 접촉사고가 49%로 가장 높고, 유해화학물질 노출/접촉이 19%로 두 번째로 높다. 그리고 Fig. 2(b)에서 실험기계/기구 사용 부주의가 34%로 가장 높고 유해화학물질 취급 부주의가 28%로 두 번째로 높다.

연구자들은 실험을 진행하기 전 필수적으로 합성 방법 및 반응물과 생성물에 대한 정보를 얻기 위해 검색을 진행한다. 이 때, 관련된 자료들은 인터넷, 논문, 특허나 문헌 등 다양한 경로를 통해서 얻을 수 있지만, 문제점은 검색 시 시간이 오래 걸린다는 점이고, 많은 자료들 중 어떠한 것들을 선택하여 합성 경로를 설정하고 실험을 진행할지 선택을 하는데 있어서 적절한 판단을 하는데 어려움을 겪는다. 연구자의 합성계획을 돕는 S/W중 한 가지 예로 Chematica가 존재한다[3]. Chematica는 수백만개의 화학물질과 관련 합성정보 및 규칙을 가지고 있으며, 특수한 알고리즘을 구축하여 긴 합성 경로를 더 경제적이고 간단한 합성 경로로 나타내도록 설계한다. 비슷한 S/W로 Elsevier의 Reaxys AutoPlan, CAS의 Scifinder가 존재하나, 모두 가격이 비싸다는 단점이 있어 사용이 제한된다.

웹상에 다양한 합성 정보가 존재하지만, 이러한 데이터를 무료로 사용할 수 있는 것은 아니다. 기업 또는 유료 툴과 같은 경우 정보를 비공개로 하고, 일정 금액을 지불해야 공개하기 때문이다. 반면, 정부사이트에서는 화학계의 발전을 위하여 정보를 공개하기도 하고, 부분적으로 정보를 공개하는 곳 또한 여러 사이트가 있다. PubChem,

Table 1. Accident status by type of hazardous material accidents [1]

	Total	Fire	Explosion	Leakage
2017	72	44	16	12
2016	81	46	14	21
Increase and decrease	-9	-2	+2	-9

Table 2. Status of victims of hazardous accidents [1]

	Total	Death	Severely injured	Injured
2017	39	7	6	26
2016	47	6	24	17
Increase and decrease	-8	+1	-18	+9

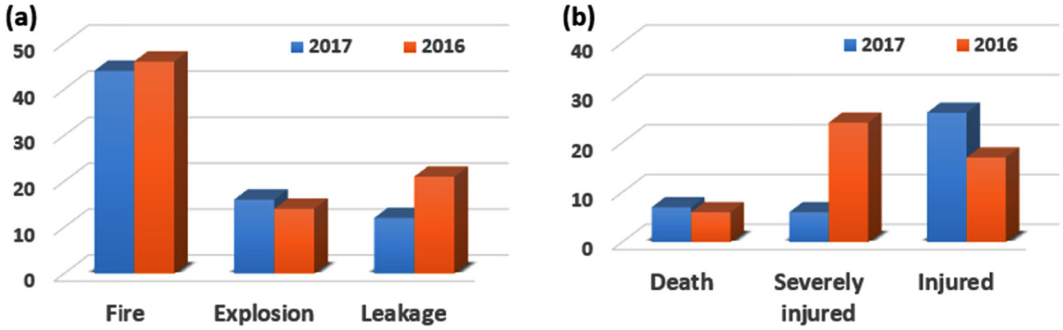

Fig. 1. (a) Accident status by type of hazardous material accident, (b) Status of victims of hazardous accident [1].

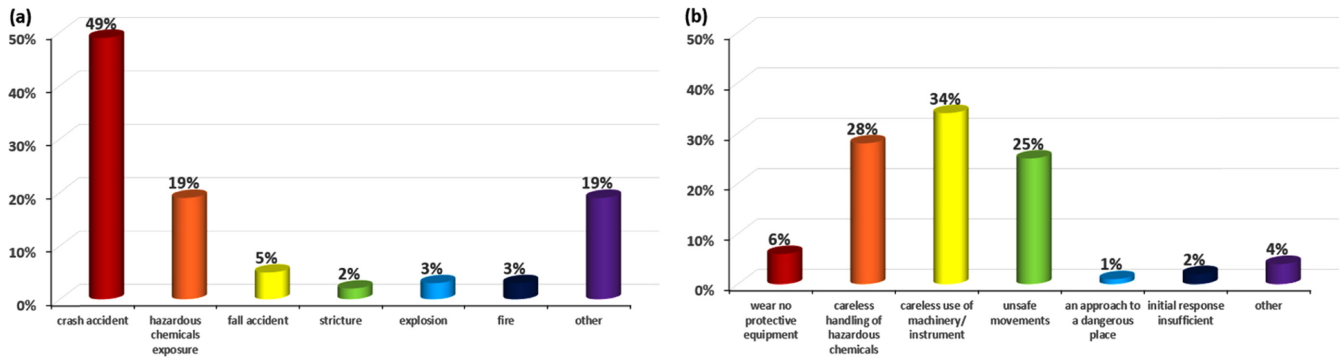

Fig. 2. Accident status [2]: (a) by occurrence, (b) by cause.

ChEMBL, Protein data bank (PDB) 등에서 화학물질 관련 정보를 얻을 수 있고, ChemSpider Synthetic page, MOLBASE, ChemSrc 등에서는 합성 관련 정보가 있다[4-9]. 이러한 공개 데이터를 장점으로 활용하여 다음과 같은 목표를 가지고 검색시간 단축 및 유해물질 확인에 도움을 주기 위한 합성 경로 탐색 시스템을 제안하고자 한다. 첫째, 무료 공개된 웹상의 데이터를 기반으로 합성경로를 나타내는 시스템을 설계한다. 둘째, 사용자가 경로 상 유해물질을 확인하고 위험성을 파악할 수 있도록 한다. 셋째, 누구나 사용할 수 있도록 open source로 공개한다.

2. 배경이론

2-1. 시스템 개발환경

시스템의 공개를 위해서 open source 프로그래밍 언어인 Python을 사용하여 개발을 진행하였다. Python은 프로그래밍 관련 지식이 많이 없은 화학공학자들에게도 접근이 쉽고, 인터프리터 방식의 언어이기 때문에 개발에도 용이하고, 개발자들 사이에서도 인기있는 언어로 많이 사용된다. 이후 사용되는 Selenium, bs4, RDKit 등과 같이 다른 개발자들에 의해 만들어져 있는 패키지들 또한 open source로 공개되어 있어 코드 작성 시 유용하게 사용할 수 있다[10]. Selenium, bs4는 자동으로 웹을 조작하고 빠르게 검색을 하는 Web Crawling과 웹 페이지에서 원하는 데이터를 추출하고 가공하는 Web Scraping을 쉽게 만들어주며, RDKit은 컴퓨터로 화학물질을 다루는데 있어서 화학물질 구조변환, MOLfile 생성, 식별자 변환 등 유용한 기능이 많다. 이후에 다른 패키지들처럼 시스템을 open source로 공개하면 사용자들은 무료 사용이 가능하고, 필요한 경우 기능을 수정하고 공유하여 지속적인 발전 또한 가능하다.

2-2. 데이터 수집 방법

데이터 수집을 위해서는 어떤 웹사이트에서 정보를 얻을 수 있고 어떠한 정보가 제공되는지 알 필요가 있다. 그에 따라서 어떠한 사이트들을 방문할 것인지, 어떠한 정보를 검색할 것인지 정해야 한다. 그 후에 Web Scraping 및 Web Crawling을 통해 데이터를 수집한다[11]. Web Scraping은 인터넷상의 데이터를 추출해내는 기법으로, 일반적으로 자동화된 프로세스를 말한다. 데이터 추출을 위해서는 일단 페이지 내용을 가져와야 하는데, 우선 Web Scraping S/W를 통해 Hypertext Transfer Protocol (HTTP)를 사용하여 직접 World Wide Web에 접속하거나, 웹 브라우저를 통하여 접속을 한다. 그리고 관련 페이지로 이동하여 소스 내용을 가져오는데, 가져오는 기능이 Web Crawling의 기능이다. 때문에 Web Crawling은 데이터 추출 시 사용되는 중요한 구성요소로 작용한다. 얻은 페이지 내용을 기반으로 필요한 내용만 선정하여 이외의 데이터는 버리고, 목적에 맞게 다듬어 정리하면 데이터 추출을 완료하는 것이다.

2-3. 화학물질 인식

현재 발견된 화학물질의 수만 해도 수천, 수만 가지가 존재하며, 명명법도 각각의 국가에서 그 나라의 언어 체계에 따라 조금씩 다르다. 예를 들면 영어로 methane인 물질이, 프랑스에서는 méthane, 독일에서는 methan, 인도네시아에서는 metan으로 부르며, 한국에서도 한국어 체계에 맞게 변경하여 사용하고 있다. 하나의 화학물질이 여러 가지 이름으로 불리기 때문에, 수많은 화학물질들에 대해서 이름만 가지고 정확하게 어떤 것을 나타내는지 알기가 쉽지 않다. 이 때문에 일관된 화학물질의 명칭 지정을 위해서 International Union of Pure and Applied Chemistry (IUPAC)에 의해서 만들어진 IUPAC Name과 같은 명명법이 존재한다. IUPAC Name 외에도 CAS No, InChI, SMILES 등의 다른 식별자들도 있다. 그 중에서 SMILES는 분자 구조를 문자열로 표현하기 위해 만들어 졌으며, 반응식 또한 나타낼 수 있다. SMILES는 몇 가지 규칙을 통해 분자 구조를 ASCII 문자열로 나타내는데, 생성 알고리즘은 이미 다양하게 개발되어 있다. 그리고 MOLFile, SDfile, RXNfile 등 특정 형식의 파일과 같이 원자 및 결합의 위치를 지정해주어 분자 구조를 나타낼 수 있다. 하지만 이런 형식의 데이터는 용량이 큰 편으로 SMILES와 같은 짧은 문자열로 화학물질을 나타낸다면, 더 작은 저장공간을 사용하여 많은 수의 화학물질을 저장할 수 있다. open source S/W인 Open Babel 이나 다른 S/W들 또는 Python의 RDKit 패키지를 사용하면 이러한 형식의 파일을 읽을 수 있고, 구조식과 다른 식별자들 사이의 상호변환이 가능하다[12].

2-3-1. IUPAC Name

International Union of Pure and Applied Chemistry (IUPAC)에 의해서 만들어졌으며, 어려운 명칭을 피하기 위해 특정한 법칙을 가지고 화학물질의 명칭을 체계적으로 지정하는 명명법이다.

2-3-2. CAS No

Chemical Abstracts Service (CAS)에 의해서 지정되며, 화학구조나 조성이 확정된 화학물질에 부여되는 고유번호이다. 컴퓨터를 이용해 검색을 하여 찾는 것은 편하지만, 사람이 번호만 보고 알기는 어렵다.

2-3-3. InChI

IUPAC에서 개발한 컴퓨터 알고리즘으로 생성한 문자열이다. 웹 검색 엔진에서 주어진 InChI를 검색하고 찾을 수 있도록 설계된 InChIKey라는 압축된 일정한 길이의 문자열도 있다. InChI 및 InChIKey는 웹이나 저널, 잡지, Database 등과 같은 화학 콘텐츠 소스들 사이에서 웹 기반으로 연결될 수 있도록 하기 위해 만들어졌다.

2-3-4. SMILES

분자와 반응을 입력하고 나타내기 위한 선형 표기법이다. 각 원소들의 연결성을 포함하며 수소 원자는 표현하지 않는다.

2-3-5. SYBYL Line Notation (SLN) [13]

SMILES를 바탕으로 입체 화학을 나타내기 위한 선형 표기법이다.

2-3-6. Smiles Arbitrary Target Specification (SMARTS) [14]

분자 패턴을 묘사하는 언어이다. SMILES에서 확장된 규칙을 사용하여 하위 구조를 지정할 수 있게 만든다. 하위 구조 지정은 컴퓨터 화학에서 탐색을 위한 중요한 기술이다.

2-3-7. A Reaction Transform Language (SMIRKS) [15]

SMILES를 기반으로 화학반응에서 사용되는 반응물/생성물을 표현한다.

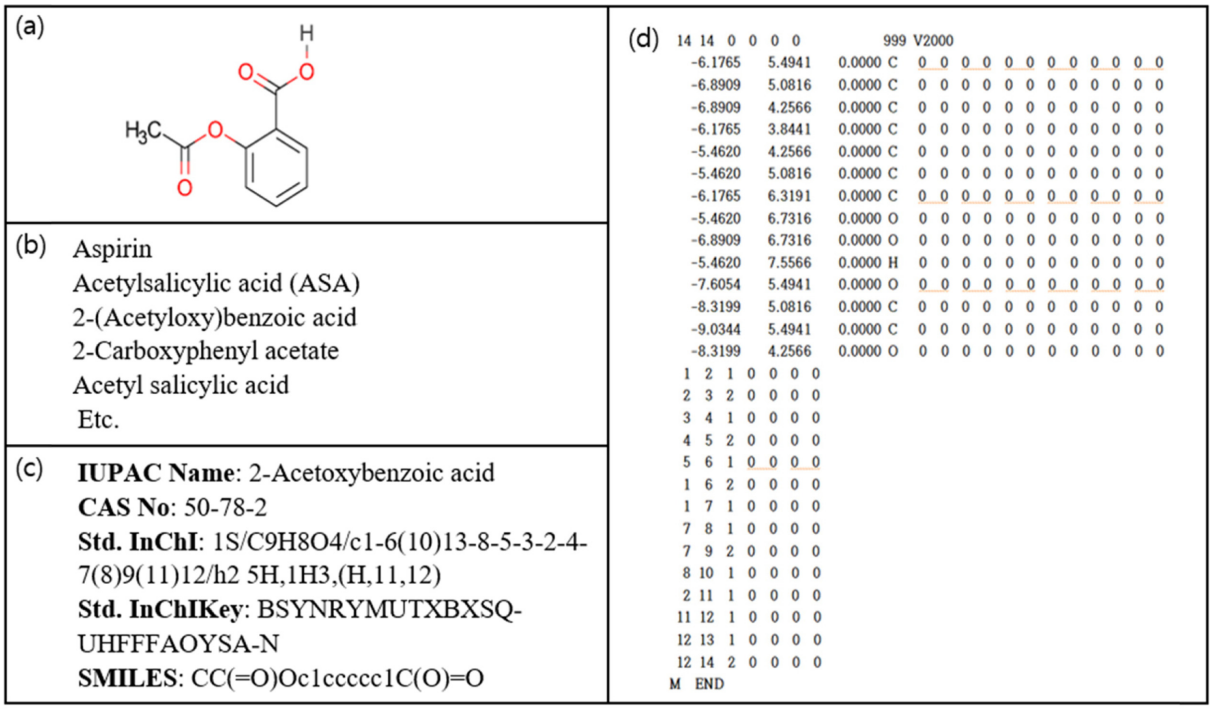

Fig. 3. A variety of notation types representing aspirin. (a) Two-dimensional aspirin structure, (b) other names for aspirin, (c) IUPAC Name, CAS No, Std.InChI, Std.InChIKey, (d) MOLfile for drawing the structure of aspirin.

2-4. 경로 탐색 문제

경로 탐색 문제는 다양한 분야에 존재한다. 단순하게는 큐브 풀이 문제에서부터 복잡하게는 공장 설비 배치나 GPS를 활용한 내비게이션 시스템 등에 사용되며 다른 분야에도 응용하여 활용된다. 최단 거리 문제는 단일 출발 최단 경로, 단일 도착 최단 경로, 단일 쌍 최단 경로, 전체 쌍 최단 경로 4가지 종류가 있다. 단일 출발 최단 경로는 단일 노드에서 출발하여 그래프 내의 다른 모든 노드에 도착하는 가장 짧은 경로를 찾는 문제이다. 단일 도착 최단 경로는 모든 노드에서 출발하여 그래프 내의 단일 노드에 도착하는 가장 짧은 경로를 찾는 문제이다. 단일 쌍 최단 경로는 주어진 노드 쌍에 대해 최단 경로를 찾는 문제이다. 전체 쌍 최단 경로는 그래프 내의 모든 노드 쌍들 사이의 최단 경로를 찾는 문제이다.

최단 거리 문제의 해결을 위해서는 다양한 경로 탐색 방법들이 존재한다. Breadth-First Search (BFS), Depth-First Search (DFS), Dijkstra algorithm, A* algorithm, Floyd-Warshall algorithm 등이 그 예이다 [16]. BFS와 DFS는 트리(Tree)와 그래프(Graph)를 탐색하는 Tree search algorithm에 속한다. 두 알고리즘의 차이점은 시작 노드로부터 BFS는 순차적으로 가까운 노드를 탐색하고, DFS는 순차적으로 깊게 탐색한다. 새로 만들어진 알고리즘들은 대부분 BFS 및 DFS를 기반으로 단점을 보완해서 생겨났다. Dijkstra algorithm 같은 경우, BFS 기반에 가중치를 추가하여 단일 출발 최단 경로 문제 해결에 적합하며, 응용을 통해 다른 문제들도 풀 수 있다. A* algorithm은 Dijkstra algorithm에서 휴리스틱 추정 값을 추가하여 확장한 기법이며, Floyd-Warshall algorithm은 재귀적인 공식을 적용하여 모든 노드 쌍에 관하여 최단 경로를 찾게 된다. DFS 알고리즘은 다음과 같이 간단히 나타낼 수 있다. 우선 시작 노드를 큐(Queue)에 배정한다. 그리고 큐가 비어 있다면 실패하고 멈추고, 큐의 첫번째 노드가 목표 노드라면 성공하고 멈춘다. 둘 다 해당하지 않는다면 큐에서 첫번째 노드를 제거하고 그 노드에 하위 노드들이 있다면 큐의 첫번째에 배정한다. 이후 같은 작업을 모든 노드들에 대해서 반복하여 결과를 얻는다.

3. 시스템 설계

3-1. 합성 경로 탐색 시스템 설계

Database 구축을 위해, 이후 나오는 사이트들에 대해서 사전 조사를 진행하였다. 우선 화학물질과 관련된 기본적인 정보는 The National Institutes of Health (NIH)의 open chemistry database인 PubChem에

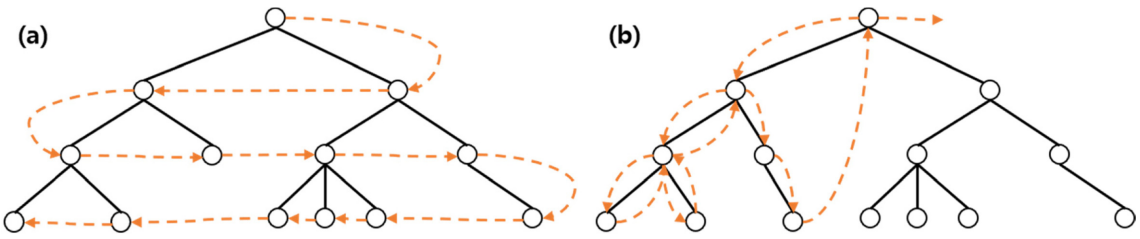

Fig. 4. Algorithm search order: (a) BFS, (b) DFS.

서 얻었다. PubChem에서는 다운로드를 지원하는데, 여기에 포함된 정보는 화학물질의 IUPAC Name, International Chemical Identifier (InChI), Simplified Molecular Input Line Entry System (SMILES), 분자식 등이 있다[17-19]. 하지만 CAS Registry Number (CAS No)는 포함되어 있지 않고, 웹페이지 상에 존재하여 Web Scraping을 통해 얻었다[20]. 합성과 관련된 정보는 ChemSrc나 기타 다른 사이트에서 확인할 수 있지만 PubChem처럼 다운로드 하는 것은 불가능하고, 웹페이지 내에서 확인하는 것만 가능하다. 합성 정보를 얻기 위해 Web Crawling및 Web Scraping을 같이 사용하여 자동으로 사이트를 이동하면서 데이터를 수집하였다.

Web Scraping 코드 구성 시 Selenium 패키지를 사용하였다. 이 패키지는 사람이 웹 브라우저에서 할 수 있는 것들을 동일하게 조작할 수 있도록 돕는다. 이는 자동조작중인 웹 페이지상에서 JavaScript 코드를 실행할 수 있도록 신호를 보내주기 때문이다. Web Crawling 코드 구성 시 bs4를 사용하였다. BeautifulSoup 패키지의 일부분으로 웹 페이지의 html을 읽어서 Python으로 전달한다. 전달된 html source에서 특정한 태그 및 속성을 찾아보고, 속성 중 합성과 관련된 내용은 추출하여 database 구성 시 사용하였다. 태그 및 속성 중 웹 조작과 관련된 내용들도 존재하는데 Python 내부에서는 이러한 정보를 활용하여 웹 조작을 하는 것이 따로따로 사용할 때 보다 훨씬 빠르다. 때문에 Web Crawling 및 Web Scraping이 동시에 진행되도록 구성하였다. 추출한 데이터를 정리하기 위하여 코드 구성 시 pandas, openpyxl, RDKit 등의 패키지를 사용하였다. 웹 조작 시 기반이 되는 페이지에서 따라서 화학물질을 나타내는 식별자 기준이 다르기 때문에 화학물질에 대한 여러 종류의 식별자를 상호 변환 및 MOLfile 형식의 데이터로 변환하였다. 이후 변환된 식별자 데이터를 CAS No, InChI, InChIKey, SMILES, IUPAC Name 5가지 종류로 분류하고, 차후 검색이 용이하도록 만들었다. MOLfile 형식의 데이터는 필요 시 2차원 구조 그림으로 역으로 변환한다. CAS No를 찾기 위해서는 NCI/CADD Group에서 지원하는 화학물질 식별자 변환 사이트를 기반으로 사용하였다[22]. 분류된 데이터는 스프레드시트 형식으로 배열하여 화학물질별로 구분을 쉽게 만들었다. 정리된 데이터는 최종적으로 excel file로 외부에 저장하였다. 이렇게 만들어진 excel file들을 모아서 database를 구성하였다. Database 생성 순서는 Fig. 5의 흐름도와 같이 진행된다. 검색 대상이 되는 화학물질은 PubChem에서 얻은 데이터 중 유기 화학물질을 무작위로 선택했다. 유해물질 관련 데이터는 Occupational Safety and Health Administration (OSHA)의 고 위험군 화학물질 목록을 참고하였다[21].

Database를 기반으로 탐색을 진행하여 목표 물질에 대한 경로를 제안하는 알고리즘은 DFS를 기반으로 재귀적인 탐색을 진행하는 코드에 가지치기(Pruning) 알고리즘을 결합하여 만들었다. DFS만 사용하게 된다면 목표 물질에 대하여 database상의 모든 합성정보에 대해 탐색을 진행한다. 이렇게 되면 시간이 매우 오래 걸리게 된다는 단점이 생긴다. 이 문제를 해결하기 위해 가지치기를 실행하였고, 가지치기를 통해 잘라낸 노드들에 대해서는 더 이상 탐색을 진행하기 않기 때문에 목표 물질과 관련이 없는 데이터가 많을수록 속도는 비약적으로 상승한다. 코드 상 진행되는 알고리즘의 순서는 다음과 같다. 우선 목표 생성 물질 및 초기에 포함되어야 할 반응 물질을 입력 받는다. 그리고 목표 생성 물질을 목표 노드로 설정하고, 1차적으로 가지치기를 실행하는데, 이 때 목표 노드와 관련이 없는 다른 노드들은 배제한다. 이후 남아있는 노드들을 큐에 배정을 하고, 각 노드들에 대해 재귀적으로 하위 노드에 대해서 탐색을 진행한다. 진행하면서 포함되어야 할 반응 물질에 해당되는 노드를 찾으면, 목표 노드부터 해당 노드까지 이어지는 경로를 최종 그래프 상에 나타낼 수 있도록 임시로 저장한다. 큐에 배정된 특정 노드가 제한된 단계까지 재귀적으로 탐색을 진행해도 포함되어야 할 반응 물질을 찾지 못한다면 그 노드는 배제된다. 이후 다른 노드의 하위 노드로 배제된 노드가 나올 경우, 2차적으로 가지치기가 실행되어 하위 노드로 내려가지 않고 다음으로 넘어간다. 탐색이 끝난 노드는 큐에서 제거되며, 큐에 더 이상 남아있는 노드가 없는 경우 종료되며, 임시로 저장해둔 경로들을 모아서 최종 그래프로 표현한다. 최종 그래프는 합성 데이터를 노드로 표현하여 경로를 연결하고, database상 저장된 유해물질이 포함되어 있다면 표시를 한다. 시스템 작동 순서는 Fig. 6과 같이 진행된다.

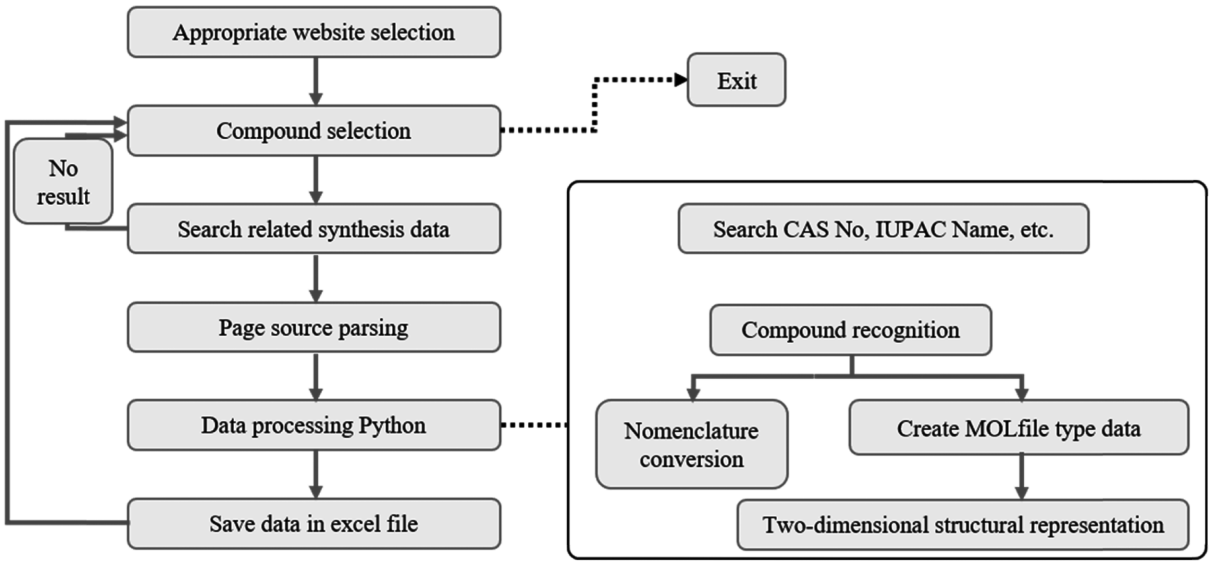

Fig. 5. Database building flowchart. After selecting the appropriate website, the code is run. The data used in the selection of compound were based on the data from PubChem and OSHA. Repeated work is performed to retrieve and store related synthetic data.

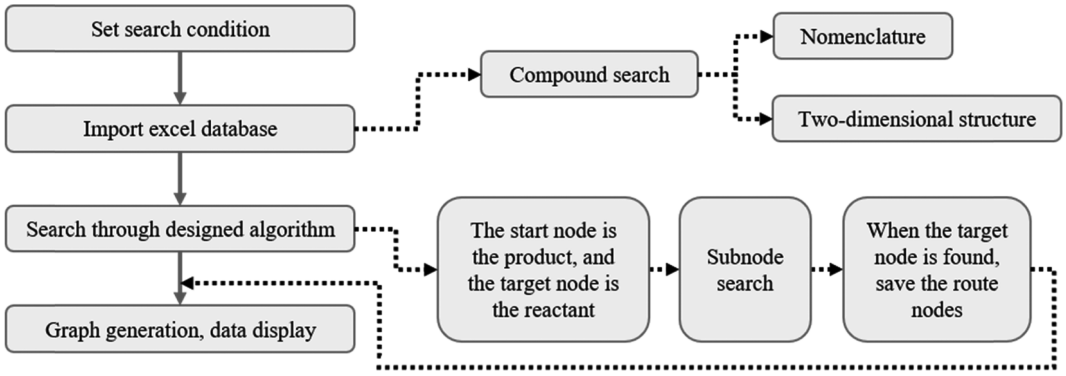

Fig. 6. System running flowchart. The algorithm is built as recursive function and iterates, searching all paths that match the conditions. Compound search is executed through the imported Excel database.

3-2. Graphical User Interface (GUI) 구성

시스템 내부 코드를 작성한 후, 그것을 표현하기 위해 GUI를 사용한다. 시각적인 결과 화면 생성을 위해 주로 PyQt5 패키지 및 Qt Designer가 사용되었다. Qt Designer는 PyQt5의 내장 툴로서 코드를 작성하여 GUI를 구성하는 형식이 아니라, 직관적인 인터페이스를 통해서 쉽게 GUI 구성이 가능하다는 장점을 가진다. 또한 특정한 Python 명령어를 이용하여 구성된 GUI를 Python 언어로 변경하는 것이 가능하다. 때문에 Qt Designer를 사용하여 기본 화면들의 뼈대를 생성하여 Python 언어로 변경하여 세부 코드를 추가하였다. 알고리즘을 통해 목표 물질까지 이어지는 노드들을 연결하여 그래프로 표현하였고, 해당하는 정보를 확인할 수 있다.

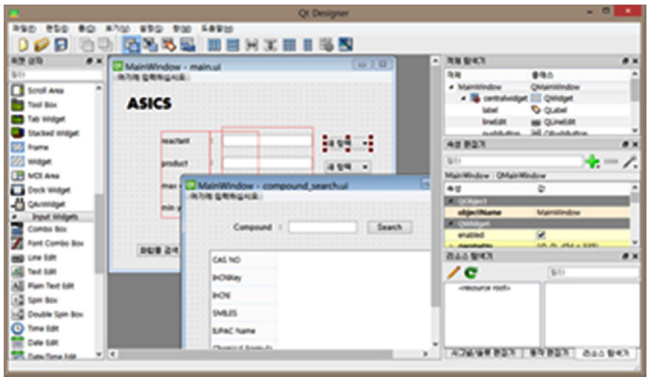

Fig. 7. Qt Designer interface. It is easy to configure the GUI because drag-and-drop is possible.

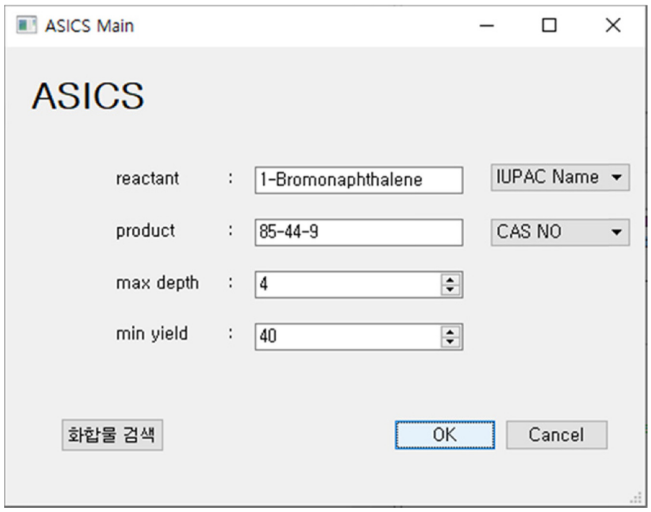

Fig. 8. System main screen. Set reactant, product, maximum depth and minimum yield. Press the OK button to go to the result screen.

시작 화면(Fig. 8)에서 초기 반응물에 포함된 화학물질, 최종 생성물에 포함된 화학물질, 제한할 경로 단계의 수, 최소 수율을 설정한다. 설정 가능한 화학물질 식별자는 IUPAC Name, CAS No, InChI, InChIKey, SMILES 5가지이다. 입력 후 결과 창에서 'start' 버튼을 누르면 검색을 시작하고, 'rxn_route' 버튼을 누르면 검색된 경로 그래프가 생성된다. 'route_data' 버튼을 누르면 반응 경로 데이터를 depth 별로 tree 형식으로 확인이 가능하다. 현재 결과는 화학물질을 CAS No로 표현하며, 화합물 검색을 통해 별도의 식별자나 구조를 확인할 수 있다.

4. 결과 및 토의

4-1. 시스템 테스트 결과

Database 생성을 위해 PubChem의 화학물질 데이터 중 무작위로 선정한 유기 화학물질 약 2000개에 대하여, 합성 관련 정보를 제공하는 사이트에서 Web Scraping을 진행하였다. 그 결과 약 143,000

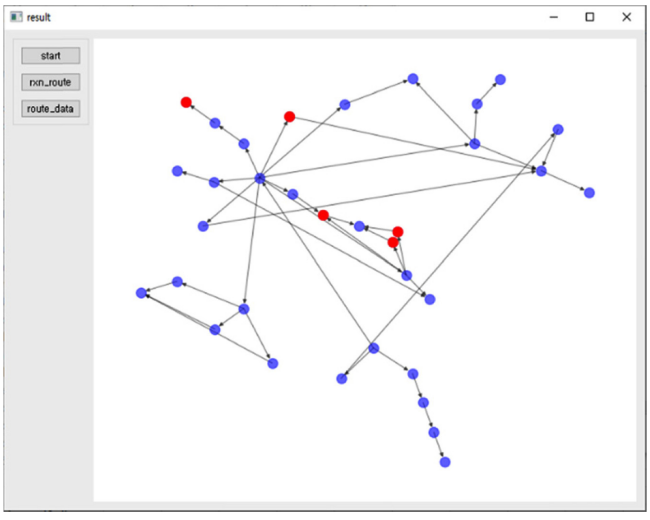

Fig. 9. Graph of system test results. It was set to 75-65-0 (2-Methylpropan-2-ol) for reactant, 78-98-8 (2-Oxopropanal) for product, 5 for depth and 10% for yield. Blue nodes are general nodes, and red nodes are hazardous chemicals.

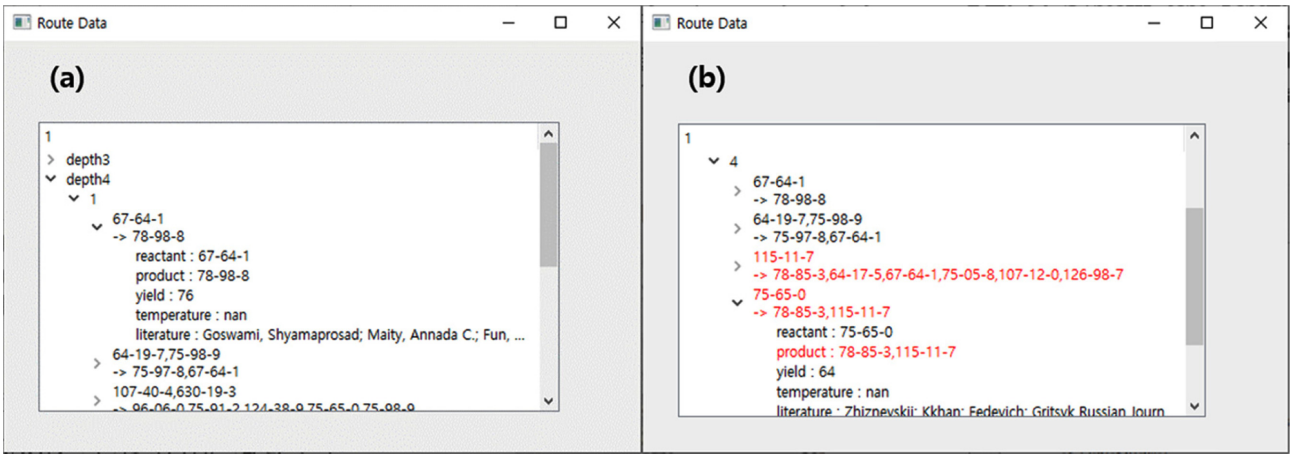

Fig. 10. Data from system test results. (a) Data are displayed in a tree format by depth. It indicates step-by-step synthetic information, which includes reactant, product, yield, temperature and literature. (b) Synthetic information, including hazardous chemicals, is displayed in red text.

개의 합성 데이터를 얻었고, 유해화학물질 데이터는 OSHA의 고 위험군 화학물질 표를 토대로 얻었다. 위 database를 기반으로 테스트를 진행하였으며, 결과는 Fig. 9, Fig. 10과 같이 나타났다.

위 그림들은 목표 생성 물질을 2-Oxopropanal (78-98-8), 포함되어야 하는 반응 물질을 2-Methylpropan-2-ol (75-65-0)로 정하고, 경로의 최대 단계 수를 5로 제한하고 각 합성마다 최소 수율을 10%로 설정하여 진행한 결과이다. 탐색 결과 Fig. 9와 같은 그래프가 나타났고, 생성 물질을 목표 노드로 잡고 진행하였기 때문에, 생성 물질을 기준으로 뻗어 나가서 시작 반응 물질에 도달하는 그래프가 나타났다. 그래프 상의 파란색 노드들은 일반적인 합성 정보를 나타낸다. 그리고 붉은색 노드들은 합성 정보에 유해화학물질이 포함되어 있다는 것을 의미한다. 결과 그래프에 대한 자세한 정보는 Fig. 10의 트리 형식의 경로 데이터에 나타냈다. 데이터는 depth별로 순서대로 정리되게 나타냈으며, 합성 정보에 유해화학물질이 포함되어 있으면 Fig. 10(b)와 같이 붉은색 문자로 나타냈다. 이 붉은색 문자로 나타나는 합성 정보가 그래프의 붉은 노드에 해당한다. Fig. 11은 추가 기능으로, 화학물질 검색 시스템이다. 현재 결과 데이터 상에서 알 수 있는 화학물질 정보는 CAS No를 사용하여 나타내고 있다. 때문에 CAS No의 숫자만 보고는 어떤 화학물질인지 바로 파악하기가

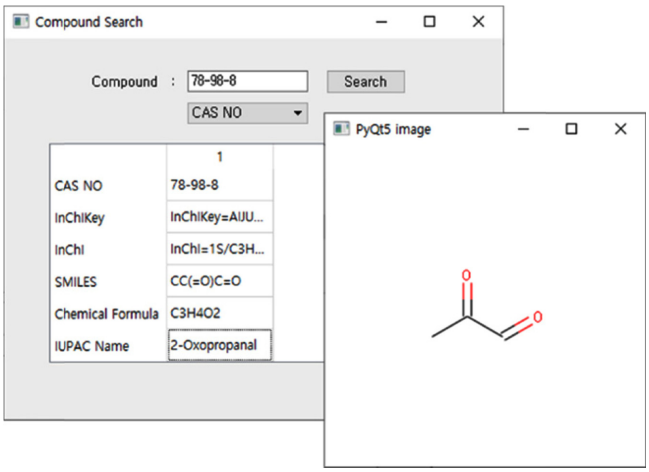

Fig. 11. Additional function in system. It finds other nomenclature of chemicals and draw 2-D structure.

힘들다. 이러한 단점을 보완하기 위해, 위 기능에서 CAS No를 입력하여 검색을 하면 database 상의 화학물질 정보를 빠르게 찾아준다. SMILES, InChI 등 다른 식별자를 선택하여 검색할 수 있으며, 해당 화학물질의 다양한 식별자 및 2차원 구조를 확인할 수 있다.

2-Oxopropanal의 합성 경로에 대한 결과 데이터가 실제로 가능한지 확인하기 위해 CAS의 SciFinder와 SciPlanner를 사용하여 비교했다[23]. SciFinder는 화학물질, 반응 및 문헌을 전문적으로 제공하며(약 9500만 개의 데이터), 여기서 찾은 합성 정보를 SciPlanner에 등록하여 사용자가 수동으로 합성 경로를 연결하는 기능이 있다. 결과 데이터의 3-step 경로와 4-step 경로 각각 하나를 무작위로 선택하였고, 해당하는 합성 경로 데이터를 Fig. 12(a), (b)와 같이 ChemAxon의 MarvinSketch를 사용하여 해당하는 합성 경로를 그려서 나타냈다[24]. Fig. 12(a)의 3-step 경로를 살펴보면, 초기의 반응 물질 2-Methylpropan-2-ol이 포함된 첫 번째 합성, 첫 번째 합성의 주 생성물인 Pivalic acid (75-98-9)를 반응 물질로 가지는 두 번째 합성, 여기서 생성된 Acetone (67-64-1)에 의한 세 번째 합성을 통해서 2-Oxopropanal이 생성된다. Fig. 12(b)의 의 4-step 경로를 살펴보면, 초기 반응 물질 2-Methylpropan-2-ol이 포함된 첫 번째 합성, 첫 번째 합성의 생성물인 Neopentyl alcohol (75-84-3) 을 반응 물질로 가지는 두 번째 합성, 주 생성물인 Pivalic acid를 반응 물질로 가지는 세 번째 합성, 생성된 Acetone을 통한 네 번째 합성으로 2-Oxopropanal이 생성된다. 그리고 SciFinder에 실제로 반응 및 문헌 데이터가 존재하는지 검색을 하였고, 해당하는 반응 데이터를 SciPlanner에 등록하여 합성 경로를 Fig. 12(c), (d)와 같이 직접 연결했다. 3-step 경로는 Fig. 12(a), (c)와 같이 동일한 경로가 있음을 확인했다. 그리고 4-step 경로는 Fig. 12(b), (d)와 같이 반응물, 생성물에 대하여 약간의 차이가 존재하지만, 경로 상의 핵심 물질들(Neopentyl alcohol, Pivalic acid, Acetone)이 일치한다는 것을 확인했다. 이외의 데이터에 대해서도 검색을 진행해본 결과, 핵심 물질들이 일치하는 것을 확인했다.

다른 설정에 대한 결과 데이터를 이용해 위와 같이 검색을 진행하였고, 비슷한 결과를 얻었다. 설정 값은 초기의 반응 물질을 Acrylic acid (79-10-7), 생성물질을 Methacrylic acid (79-41-4), 최대 단계 수를 4로 제한하고 최소 수율을 20%로 했다. 결과 그래프는 Fig. 13과 같이 그려졌고, 2-Oxopropanal 생성 그래프와 같이 다양한 경로를

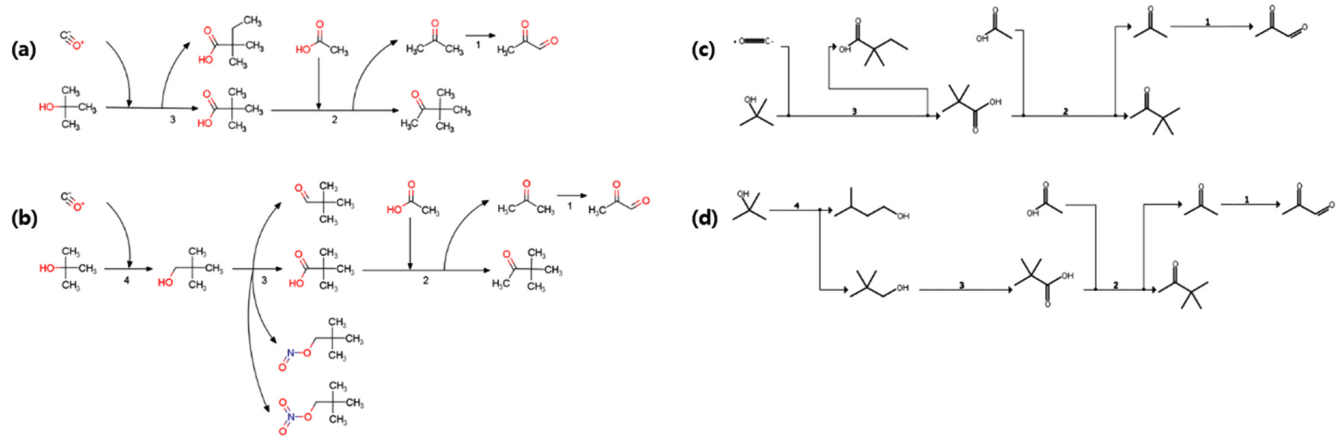

Fig. 12. Synthetic paths for 2-Oxopropanal. (a) The 3-step synthetic path, one of suggested system results, created by MarvinSketch, (b) the 4-step synthetic path, one of suggested system results, created by MarvinSketch, (c) the 3-step synthetic path corresponding to (a) created by SciPlanner in SciFinder, and (d) the 4-step synthetic path corresponding to (b) created by SciPlanner in SciFinder.

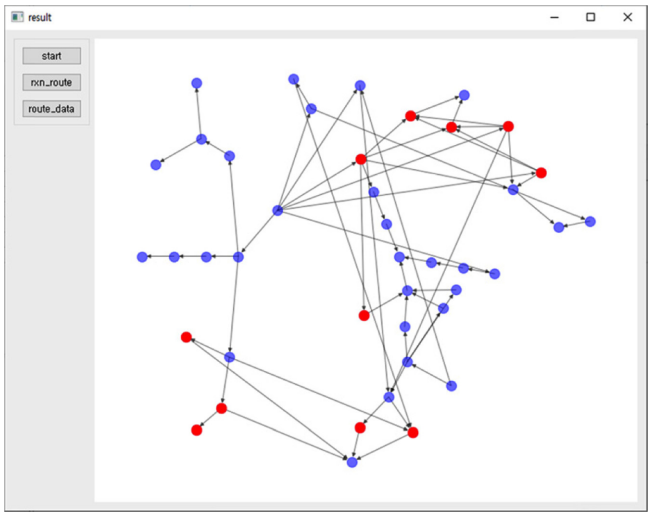

Fig. 13. Graph of system test results for methacrylic acid production.

나타내는 것을 볼 수 있다. 위와 같은 시스템 테스트 결과 데이터를 통해서 목표 생성 물질에 대한 다양한 경로가 있고, 여러 경로 상에 유해화학물질이 포함되어 있으며, 실제 사용되는 SciFinder에서도 합성 정보의 존재를 확인할 수 있는 유의미한 테스트 결과가 나오는 것을 확인하였다.

4-2. 현 시스템의 한계 및 개선방향

본 시스템은 현재 초기 버전으로 테스트를 진행하기 위해 약 143,000개의 데이터를 사용하였지만, 다른 유료 S/W의 database에 비하면 매우 작은 편이다. 때문에 database에 없는 정보에 대해서는 입력을 하지 못하고, 결과를 얻을 수도 없다. 그리고 데이터에 따라서 반응 방법을 상세하게 나타내는 데이터도 있지만, 간단하게 나타내는 데이터가 더 많다. 이 때문에, 반응의 상세 내용을 알기 힘든 점이 있다. 유해성 판정 또한 OSHA 데이터만 기반으로 사용하였기 때문에 단순하다. 향후 발전을 위해서는 Database 확장 및 유해성 판정 모듈의 개선이 필요하다. 따라서 후속 연구를 위해 다음과 같은 개선사항을 제안한다.

4-2-1. Database 개선

Data를 얻는 방법은 여러가지가 있다. 첫째, Web Scraping을 하는 것이다. 웹 자동 조작을 통하여 원하는 데이터만 추출하여 지속적으로 database에 등록하는 것이다. 둘째, 이미지 인식을 활용하는 것이다. 자료 검색 시 화학물질을 2차원 또는 3차원 구조 이미지로 나타내는 경우가 많다. 이러한 경우 구조 이미지를 인식하고, 이에 대한 화합물의 정보를 알아내어 database에 등록하는 것이다. 셋째, 최근 많은 연구가 이루어지고 있는 딥 러닝 기술을 활용하는 것이다. 딥 러닝을 통해 발생 가능한 반응을 예측하고, 이 정보를 database와 함께 사용하는 것이다. 예측된 정보는 검증되지 않은 정보이기 때문에 기존의 database와는 별도로 저장되어야 할 것이고, 검증 또한 추가로 필요할 것이다. 딥 러닝을 활용하는 것은 쉽지는 않겠지만 예측을 통해 자체적으로 데이터를 생성하기 때문에, 정보가 없는 상황에서도 결과를 얻을 수 있다는 장점이 생길 것이다. 본 논문에서는 첫 번째 방법만 사용하여 database를 구축하였으나, 두번째, 세번째 방법을 추가로 연구한다면 database 확보에 큰 도움이 될 것이라 보인다.

4-2-2. 유해성 판정 모듈의 개선

현재 결과 데이터의 표현에서 유해화학물질이 포함된 경로 및 다른 여러가지 경로들을 보여준다. 이 때, 물질의 유해성 판정은 위해 OSHA의 고 위험군 화학물질 표의 데이터를 토대로 진행된다. 이 데이터는 시스템 테스트를 위해 사용된 일정 크기의 데이터로, 합성 정보 데이터와 마찬가지로 상대적으로 작다. 차후 유해성 판정 모듈의 개선을 위하여 화학물질정보시스템(NCIS)의 유독 물질, 제한 물질, 금지 물질 데이터와 물질의 MSDS 정보 그리고 NFPA 704 정보를 등록하여 개선할 수 있다. 이를 통해 물질의 유해성 등급을 세분화하고, 등급별 가중치를 부여한다면 시스템에서 유해성을 판정하는데 도움이 될 것이다.

5. 결 론

웹 상에는 다양한 정보들이 존재하고, 그 정보들 중에는 여러가지 반응에 대한 정보들이 포함되어 있다. 그리고 정보를 다루는 방법이나 알고리즘을 구축하는 이론적인 방법론 또한 많이 제안되어 있다. 본 연구에서는 이러한 기존에 알려진 이론적인 지식을 활용하여 오

픈소스로 반응경로 탐색 지원시스템을 구현하는 것에 초점을 맞추었다. 그 결과, 반응 정보의 불확실성에 기인한 실험실 반응 사고와 위험물 사고 예방을 위해, 관련 화학반응 정보를 찾아주는 시스템을 공개 S/W (ASICS, Advanced System for Intelligent Chemical Synthesis)로 제안하였다. 다양한 테스트를 통해 시스템을 검증하였으며, 제안 시스템의 사용을 통해 다음과 같은 효과가 기대된다.

(1) 실험자의 검색 시간 단축: 웹상에 널리 퍼져있는 합성 정보를 검색하는데 걸리는 시간을 데이터베이스를 활용한 검색 시스템을 이용하여 단축시킨다.

(2) Open source를 활용한 지속적 개발 및 개선: Open source로 공개하여 위험물 기술자들과 공유/확충을 통해 보다 나은 시스템으로 발전 가능하다.

(3) 지속적인 합성 정보 업데이트 및 축적: Database의 지속적인 업데이트가 가능하여, 언제나 최신 기술 발전을 반영한 결과를 얻을 수 있도록 데이터를 지속적으로 자동으로 수집할 수 있다.

(4) 연구자의 customization 및 비공개 반응 검색 지원: 기업과 같은 경우 자신들만의 비공개 위험물 및 반응 연구 데이터를 open source로 공개한 시스템에 추가하여 사용 가능하다.

(5) 유해 물질을 제외한 합성 경로 제안: 실험자가 알고 있는 합성 정보보다 유해성이 낮은 물질을 포함한 합성 경로를 제안하여 실험 안전사고에 대한 위험을 조기에 예방한다.

현재 만들어진 시스템은 초기 버전으로 개선 가능성이 많다, 다만 위험물 사고 원인의 한 축을 차지하고 있는 복잡한 반응정보의 정확한 제공을 검증된 자료를 바탕으로 지원한다는 측면에서 본 연구는 차별성을 갖는다. 또한 연구 시작 전 초기단계에서 목표 생성물질을 향해 진행되는 다양한 경로들을 단기간에 쉽게 확인할 수 있다는 측면에서 스마트 지원시스템으로의 의의를 충족한다. 보다 나은 합성 경로를 선택하면 실험 시 사용되는 유해 물질의 사용량 또한 줄어들게 되고, 생산공정상 혼합과정을 통해 발생가능한 위험반응의 사전 분석도 제공함으로써 결과적으로는 위험물 사고예방 측면에도 크게 도움이 된다.

감 사

본 연구는 국토교통부의 재원으로 햅틱 기반 플랜트 안전훈련시스템 기술개발 프로그램의 지원을 받아 수행된 연구입니다(MPSS-사회-2014-38).

References

1. National Fire Agency, *2018 Dangerous Goods Statistics Data*, 121-139(2018).
2. Korea Occupational Safety and Health Agency, *Accident Investigation on the Explosion During Butadiene Experiment*, 2016-Specialty-424(2016).
3. Szymkuc, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., Bajczyk, M. and Grzybowski, B. A., "Computer-Assisted Synthetic Planning: The End of the Beginning," *Angew. Chem. Int. Ed.*, **55**(20), 5904-5937(2016).
4. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang J. and Bryant, S. H., "PubChem Substance and Compound Data-Bases," *Nucleic Acids Research*, **44**(D1), D1202-D1213(2015).
5. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J. P., "ChEMBL: a Large-scale Bioactivity Database for Drug Discovery," *Nucleic Acids Research*, **40**(1), D1100-D1107(2011).
6. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E. and Berman, H. M., "The RCSB PDB Information Portal for Structural Genomics," *Nucleic Acids Research*, **34**(1), D302-D305(2006).
7. Pence, H. E. and Williams, A., "ChemSpider: An Online Chemical Information Resource," *Journal of Chemical Education*, **87**(11), 1123-1124(2010).
8. MOLBASE homepage, http://www.molbase.com/.
9. ChemSrc homepage, https://www.chemsrc.com/en/.
10. Landrum, G., *RDKit Documentation, Release* (2017).
11. Mitchell, R., Web Scraping with Python: Collecting More Data from the Modern Web, 2nd ed., *O'Reilly Media, Inc.*, Sebastopol (2018).
12. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R., "Open Babel: An Open Chemical Toolbox," *Journal of Cheminformatics*, **3**(1), 33(2011).
13. Homer, R. W., Swanson, J., Jilek, R. J., Hurst, T. and Clark, R. D., "SYBYL Line Notation (SLN): a Single Notation to Represent Chemical Structures, Queries, Reactions, and Virtual Libraries," *Journal of Chemical Information and Modeling*, **48**(12), 2294-2307(2008).
14. DAYLIGHT Chemical Information Systems, *SMARTS – A Language for Describing Molecular Patterns*, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
15. DAYLIGHT Chemical Information Systems, *A Reaction Transform Language*, http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html.
16. Neapolitan, R. E., Foundations of Algorithms, 5th ed., *Jones & Bartlett Learning*, Burlington(2015).
17. Panico, R., Powell, W. H. and Richer, J. C., "A Guide to IUPAC Nomenclature of Organic Compounds," *Blackwell Scientific Publications*, Oxford, (1993).
18. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D., "InChI, The IUPAC International Chemical Identifier," *Journal of Cheminformatics*, **7**(1), 23(2015).
19. Weininger, D., "SMILES, a Chemical Language and Information System I. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences*, **28**(1), 31-36(1988).
20. Dittmar, P. G., Stobaugh, R. E. and Watson, C. E., "The Chemical Abstracts Service Chemical Registry System I. General design," *Journal of Chemical Information and Computer Sciences*, **16**(2), 111-121(1976).
21. OSHA, *List of Highly Hazardous Chemicals, Toxics and Reactives (Mandatory)*, https://www.osha.gov/law-regs.html.
22. NCI/CADD Group, *Chemical Identifier Resolver*, https://cactus.nci.nih.gov/chemical/structure.
23. CAS, *SciFinder*, https://www.cas.org/products/scifinder.
24. ChemAxon, *MarvinSketch*, https://chemaxon.com/products/marvin.