대규모 범죄 수사기록을 활용한 온톨로지 기반 서비스 구현*

- 침입 절도 범죄 분야를 중심으로 -

Implementation of Ontology-based Service by Exploiting Massive Crime Investigation Records: Focusing on Intrusion Theft

고 건 우 (Gun-Woo Ko)**, 김 선 우 (Seon-Wu Kim)***
박 성 진 (Sung-Jin Park)****, 노 윤 주 (Yoon-Joo No)*****
최 성 필 (Sung-Pil Choi)******

목 차

초 록

온톨로지는 특정 분야의 특정 지식과 관련된 용어 및 용어 사이의 관계를 정의하는 복합 구조 사전이다. 국내외로 다양한 온톨로지 구축의 시도가 있었으나 대규모의 범죄 수사기록을 온톨로지로 구축하고 이를 통한 서비스를 구현한 사례는 존재하지 않았다. 따라서 본 논문은 비정형 데이터인 범죄 수사기록 문서 중 침입 절도 분야로부터 추출한 정보를 통해 온톨로지를 구축하고, 온톨로지 기반의 검색 서비스와 범행 장소 추천 서비스를 구현하는 과정을 설명한다. 검색 서비스의 성능을 파악하기 위하여 사건 검색에 대한 정확도 측정 방법 중 하나인 Top-K 방식의 정확도 측정을 실험하였고, 실험 집합에 대하여 최대 93.52%의 정확도를 얻었다. 또한, 범행 장소 추천 서비스의 성능을 파악하기 위한 실험 결과, 실험 데이터셋의 전체에 대해 적합한 단서 필드 조합을 얻어냈으며, F1-measure 76.19%의 성능으로 데이터베이스 내의 범행 장소 필드 정보를 교정할 수 있음을 확인하였다.

ABSTRACT

An ontology is a complex structure dictionary that defines the relationship between terms and terms related to specific knowledge in a particular field. There have been attempts to construct various ontologies in Korea and abroad, but there has not been a case in which a large scale crime investigation record is constructed as an ontology and a service is implemented through the ontology. Therefore, this paper describes the process of constructing an ontology based on information extracted from instrusion theft field of unstructured data, a crime investigation document, and implementing an ontology-based search service and a crime spot recommendation service. In order to understand the performance of the search service, we have tested Top-K accuracy measurement, which is one of the accuracy measurement methods for event search, and obtained a maximum accuracy of 93.52% for the experimental data set. In addition, we have obtained a suitable clue field combination for the entire experimental data set, and we can calibrate the field location information in the database with the performance of F1-measure 76.19% Respectively.

키워드: 범죄 기록물, 비정형 데이터, 온톨로지, 온톨로지 서비스, 트리플
Criminal Record, Unstructured Data, Ontology, Ontology Service, Triple

---

* 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2018R1D1A1B07048839).
** 경기대학교 일반대학원 문헌정보학과 석사과정(zellyshu@kyonggi.ac.kr/ISNI-0000000474716845) (제1저자)
*** 경기대학교 일반대학원 문헌정보학과 석사과정(kimsw@kyonggi.ac.kr) (공동저자)
**** 스펠릭스(Spelix) 이사(sjpark54@spelix.com) (공동저자)
***** 경찰청 사서주사보(ballen3163@police.go.kr) (공동저자)
****** 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr/ISNI-0000000467729269) (교신저자)
논문접수일자: 2019년 1월 16일 최초심사일자: 2019년 1월 16일 게재확정일자: 2019년 2월 20일
한국문헌정보학회지, 53(1): 57-81, 2019. [http://dx.doi.org/10.4275/KSLIS.2019.53.1.057]

1. 서 론

온톨로지는 사람들이 세상에 대하여 보고, 듣고, 느끼고, 생각하는 것에 대하여 서로 간의 토론을 통하여 합의를 이룬 바를 개념적이고 컴퓨터에서 다룰 수 있게 표현한 모델로, 개념의 타입이나 사용상의 제약조건들을 명시적으로 정의한 기술이다(위키백과 2018). 온톨로지는 시맨틱 웹을 구현할 수 있을 뿐만 아니라, 지식 개념을 의미적으로 연결할 수 있는 도구로서 RDF, OWL, SWRL 등의 언어를 이용해 표현하며 기계가 개념의 이해를 넘어 추론까지 가능하다는 장점을 가진다. 최근, 대규모의 텍스트 데이터를 활용하여 새로운 결과를 도출하는 서비스들이 다양하게 등장하고 있고, 서비스를 위한 모델 구축에 온톨로지가 자주 사용되고 있다. 왜냐하면 온톨로지는 특정 분야(도메인)의 특정 지식과 관련된 용어 및 용어 사이의 관계를 정의하는 것이 가능한 복합 구조 사전이고, 대용량의 비 구조화되고 비 조직화된 정보의 통합을 가능하게 하기 때문이다. 또한, 웹의 급속한 발달로 인해 지능화된 정보 검색 시스템 개발과 웹 자원을 효과적으로 관리할 수 있는 정보 검색의 새로운 도구의 필요성에 대한 목소리가 높아지면서, 온톨로지를 활용한 시맨틱 웹 및 시맨틱 웹 서비스를 통해 문제를 해결하고자 하였다(황미녕 외 2012).

경찰청은 축적한 대규모 수사기록 데이터의 분석 결과를 토대로 다양한 서비스를 구축하였고, 이를 내부에서 수사를 보조하기 위한 목적으로 활용하고 있다(중앙일보 2017.12.8). 경찰청이 구축한 서비스는 수사기록을 토대로 용의자의 패턴을 파악하여 범죄를 사전에 방지하는 등 다양한 방면에 활용되고 있다. 범죄 수사기록의 온톨로지 구축도 수사를 보조하기 위한 활용 방법이라는 목적을 가지며, 기존의 하드 매칭이 아닌 소프트 매칭의 서비스를 구현하여 수사를 보조하는 목적의 서비스로 구현하였다.

범용 목적이 아닌 이용자의 목적을 위한 특정 분야의 온톨로지 활용 서비스는 다수의 구축 사례가 존재하지만, 문서 기반 대규모 데이터와 관련 정보를 통해 특정 목적에 부합하는 실제 사용 가능한 서비스를 구축한 과정은 찾아보기 어렵다(고건우 외 2018). 또한, 수사기록과 같은 범죄 기록물을 온톨로지로 구축하여 하드 매칭이 아닌 소프트 매칭이 가지는 장점을 가지는 서비스를 구현한 선행 연구도 존재하지 않았다. 그렇기 때문에, 온톨로지 구축 시도가 없었던 수사기록 데이터를 본 논문이 제시하는 방법론을 통해 온톨로지로 구축하여 결과의 유의미함을 확인하고자 하였다. 더 나아가, 하드 매칭으로는 볼 수 없던 새로운 측면의 확인 여부와 실용 단계의 온톨로지 기반 서비스 제공을 통하여 수사 시 어떻게 활용할 수 있는지 나누어 설명하였다.

수사기록과 같은 사건 데이터는 일정한 틀이 존재하기는 하지만 동일한 형식을 취하지 않고, 작성자마다 표현 방법부터 글의 전개 방식이 미묘하게 다르기 때문에 유기적으로 필드를 연결하여 활용하는 온톨로지에 적합한 데이터라고 할 수 있다. 보이스피싱, 살인, 성폭행, 침입 절도의 4가지 범죄 분야 수사기록 중에서도 침입 절도 분야는 다른 범죄 분야와 비교하여 범행 동기가 다양한 편이며, 가해자와 피해자 간 관계성도 비면식, 면식에 따라 다양하게 등장하는 등 복잡한 관계성과 많은 단서 필드들을

가진다. 이를 통하여 방대한 양의 트리플 생성이 가능하고, 온톨로지 구축뿐 아니라 온톨로지 기반 서비스에 사용하였을 때 수사에 도움이 될 수 있는 결과를 출력 가능한 데이터라 판단하였다. 또한, 온톨로지 구축 및 온톨로지 서비스의 가능성을 통계적인 방법을 통해 확인하고자 하였다.

따라서 본 논문은 경찰청의 대규모 수사기록 데이터 중 침입 절도 분야 데이터에서 추출한 정보를 토대로 트리플을 구성하여 온톨로지 구축 방법론을 통해 실제 온톨로지를 구축하고, 이를 통한 온톨로지 기반 검색 서비스 및 범행 장소 추천 서비스를 소개한다.

2. 관련 연구

국내・외로 온톨로지는 범용적인 목적은 물론 전문 분야 활용을 위한 목적으로도 다양하게 구축되었으며, 국내외적으로 인공지능 분야, 정보 검색 분야, 전자 상거래 분야 등 다양한 분야에서 구축되었다.

국내에서는 한의학 분야에서 사용되는 약재와 병증 등을 온톨로지로 구축하여 약재의 이름을 표준화하는 효과와 더불어 병증으로 예상 처방 약재를 추론하는 시스템을 개발하였다(박경모, 임희숙, 박종현 2003). 또한, 국립 중앙도서관은 OPEN API의 형태로 데이터를 GUI (Graphic User Interface) 환경에서 탐색할 수 있도록 지원하는 데이터 브라우저 서비스와 RDF 트리플이 포함하고 있는 Datatype Property에 대해 검색어와 매칭되는 결과를 결과 값으로 노출하여 주는 검색 서비스에 온톨로지를 활용하고 있으며(국립중앙도서관 2013), KISTI (한국과학기술정보연구원)에서는 대규모의 논문, 연구자 등의 데이터를 기반으로 온톨로지를 구축하여 연구 시 연구자 및 연구 프로젝트의 참여자를 돕는 서비스를 진행하고 있다. 특히 OntoFrame이라는 이름으로 진행되는 서비스는 연구 프로젝트의 진행 시 해당 연구 주제 분야의 최고 권위 연구자 및 관련 연구 기관을 추천해주는 추천 서비스를 비롯하여, 연구 Trend 및 해당 연구의 미래성까지 판단해주는 시각화 서비스까지 가능하기 때문에 온톨로지를 활용한 다양한 서비스를 확인할 수 있다(김평 외 2008). 〈그림 1〉은 OntoFrame의 검색 결과인 연구 Trend 등을 시각화하여 이용자에게 출력하는 결과 화면이다. 온톨로지 기반 서비스의 특징인 시각화된 결과는 보다 직관적으로 내용을 이해할 수 있다는 장점을 가지며, 본 논문의 온톨로지 기반 서비스도 이러한 점을 반영하여 마우스의 클릭만으로 검색이 가능하도록 시각화의 장점을 최대한 살려 구현되었다.

국외에서는 음악 분야에서 'MusicBrainz'가 기존의 각 회사 혹은 각 개인마다 각자의 방식대로 작성하여 정확한 표준이 없던 음악에 대한 메타데이터를 온톨로지를 통해 체계적으로 관리하고자 하였고(Raimond et al. 2007), 온톨로지를 활용한 의학 분야 문헌 검색 에이전트 시스템인 MELISA(MEdical Literature Search Agent)를 통해 기존의 의학 분야 문헌 검색 시스템의 성능을 향상시키고자 한 의학 분야의 사례도 있다(Abasolo and Gomez 2000). 또한, 이탈리아에서 범죄 온톨로지를 구축한 사례가 존재하나 이는 수사에 활용하기 위하여 구축한 것이 아니라 법조계에서 사건이 범죄인지 아닌지를 판단

〈그림 1〉 OntoFrame의 연구 Trend 등을 시각화하여 나타내는 결과 화면

하는 기준으로서 사용되었을 뿐 그 이상의 서비스로는 발전하지 않았다(Asaro et al. 2003).

국내외로 온톨로지를 범용 목적뿐만 아니라 다양한 분야에서 특정 목적을 가지고 구축하였고, 온톨로지의 구축을 위하여 정형화된 데이터를 사용하거나 비정형 데이터를 정형화하는 과정을 거치는 등 데이터의 성격에 맞게 방법론을 적용한 것을 확인할 수 있었다. 특정 분야의 온톨로지 구축이라는 측면에서 살펴 보았을 때, 국내외에서 수사 과정에 도움을 얻고자 범죄 온톨로지를 구축하고 이를 활용한 연구는 이루어진 바가 없다. 또한, 국내에서 비정형 데이터인 수사기록의 메타데이터와 수사기록 내 정보를 통해 온톨로지를 구축하는 연구도 이루어진 바가 없었다. 서비스 측면에서도, 온톨로지를 활용한 서비스는 단독적인 서비스라기보다 기존 서비스의 시각적인 요소 등을 보완하는 측면이 강하였으며, 온톨로지가 주가 되어 수사 과정을 돕는 등의 서비스는 존재하지 않았다. 따라서 온톨로지의 구축에 있어 수사기록이라는 범죄 기록물이 타당한 데이터인지 확인하기 위하여, 범죄 기록물을 통한 데이터 마이닝 등의 어떠한 시도들이 있었는지 알아보았으며 관련 연구는 다음과 같다.

수사기록과 같은 비정형 데이터의 활용에 관한 측면에서, 기존의 정형 데이터만을 이용한 분석이 아닌 텍스트 문서 및 이미지, 동영상, 음성과 같은 비정형 데이터를 추가로 활용하여 전문 수사관 및 프로파일러의 수사를 돕고, 정보의 수집 및 분석이 가능한 시스템을 제안한 연구가 존재한다(김용훈, 정목동 2017). 텍스트 비정형 데이터의 정형화를 위하여 정제 과정을 거쳤는데, 문장 내 출현 정보 및 유사성 등의 통계 정보뿐만 아니라 형태소 분석, 품사 부착 등의 자연어 처리 방법을 사용하였다. 범죄 패턴 분석, 범죄 예측 및 예방, 범죄 유형 분류 등에 범죄 기록물

데이터를 활용한 사례가 다수 존재한다. 범죄 예측에서의 데이터마이닝 적용 가능성을 파악하기 위하여, 절도범죄 데이터로 군집분석을 적용한 뒤 데이터마이닝의 결과와 범죄 환경요소와의 관련성을 확인하고, 유의미한 결과를 도출한 사례가 있었다(방승환, 김태훈, 조현보 2014). 또한 범죄 기록물 빅데이터를 분석 및 활용함에 있어 법적 근거와 책임을 확인하기 위한 연구가 이루어지기도 하였다(권양섭 2017). 범죄 예측과 관련된 연구로는 과거 범죄 기록 데이터를 기반으로 랜덤 포레스트 알고리즘 기반의 범죄 유형 분류모델 및 시스템 인터페이스 디자인을 제시한 연구(박준영, 채명수, 정성관 2016)가 있고, 미래 예측 모델인 마코프 체인을 이용하여 범죄 예측 모델링을 구현하고, 5개 범죄(강도, 살인, 강간, 절도, 폭력)에 대해 실제 발생 건수와 모델을 통한 범죄 발생 예측 건수의 비교 결과가 높은 일치도를 보인 연구가 있었다(정영석, 김진묵, 박구락 2012).

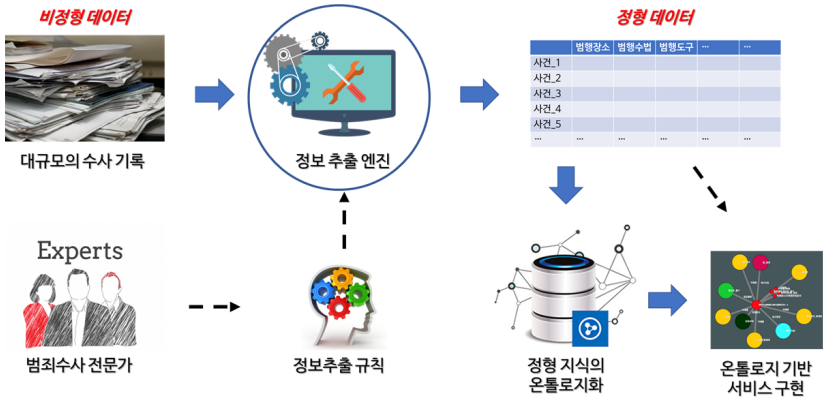
수사기록이라는 범죄 기록물을 온톨로지화 하기에 앞서, 수사기록이 데이터 마이닝 기법을 도입하기에 적합한지 관련 연구를 통해 확인하였고, 범죄 패턴 분석, 범죄 예측, 범죄 유형 분류 등의 분야에서 유의미한 결과를 도출할 수 있음을 알 수 있었다. 또한, 수사기록을 온톨로지로 구축한 연구사례가 존재하지 않으므로 최초의 시도라는 점에서 의의가 있다. 따라서 본 논문은 비정형 데이터의 정형화를 통한 방법론과 특정 분야의 온톨로지 구축 방법론을 통해 경찰청이 기구축한 수사기록 빅데이터를 온톨로지로 구축하였고, 실제 수사관과 같은 이용자에게 도움이 될 수 있는 방향의 온톨로지 기반 서비스를 구현하였다.

3. 온톨로지 기반 분석 서비스

본 논문은 범용적인 목적이 아닌 범죄 분야라는 특수성을 가지는 전문 분야의 온톨로지를 구축하였다. 또한, 비정형 데이터인 수사기록을 통해 온톨로지를 구축하고자 하였으므로, 다음과 같은 절차를 통해 진행하였다. 구축 과정에서 범죄 분야라는 특수성으로 인하여 매주 주제전문가와 형사, 기술연구자들이 회의를 통해 직간접적인 관계추출과 클래스 트리 및 제약조건을 정의하였고, 구축 기간은 약 9개월 정도가 소요되었다. 참여 인력은 온톨로지 구축 경험이 있는 교수급 1인과 석사급 2인, 학부생 1인이며, 경찰청의 범죄 분야 주제 전문가 및 형사들로부터 조언을 받았다. 구축한 온톨로지는 수사관들에게 실제 도움이 될 수 있는 서비스로써 활용될 수 있도록, 필드 검색 서비스와 범행 장소 추천 서비스로 각각 구현하였다. 서비스는 온톨로지 구축이 어느정도 이루어진 후 병행하여 진행하였고, 약 6개월이 소요되었으며, 투입된 인력은 구축 인원과 동일하다. 〈그림 2〉는 일련의 프로세스를 도식화한 것으로, 비정형 데이터를 정형화하여 이를 온톨로지로 구축하고, 온톨로지 기반의 서비스를 구현하는 것을 의미한다.

3.1 비정형 데이터로부터의 온톨로지 구축

전문 분야의 온톨로지 구축과 이를 기반으로 하는 온톨로지 서비스의 구현에 앞서, 문서와 같은 비정형 데이터는 데이터를 분석하고 정형화하는 과정이 선행되어야 한다(조대웅, 최지웅, 김명호 2014). 본 절에서는 수사기록이라는 비

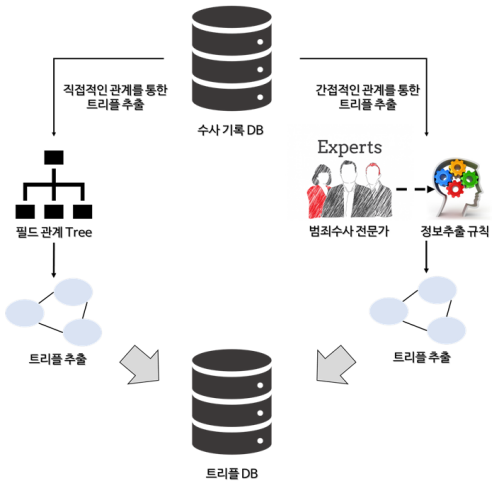


〈그림 2〉 온톨로지 구축 및 기반 서비스 도식도

정형 데이터와 경찰청이 이를 토대로 지속적으로 구축한 정형화된 범죄 빅데이터를 바탕으로 온톨로지의 구축을 위해 트리플을 추출하였다. 트리플이란 RDF에서 사용하는 주어-술어-목적어와 같이 노드와 노드 간을 연결하는 관계로 묶이는 하나의 묶음이다(황미녕 외 2012). 온톨로지의 구축 과정은 다음과 같다.

3.1.1 데이터 정형화를 위한 문서 정보 추출

온톨로지의 구축을 위해 문서로부터 정보를 추출하는 과정이 우선적으로 진행되어야 한다. 문서에서 추출 가능한 정보는 기본적으로 저자, 표제, 서명, 출판일 등이 있다. 기본적인 정보 이외에 문서의 성격에 따라 추출할 수 있는 정보는 매우 다양하며, 본 논문에서 활용하는 수사기록 데이터 같은 경우 자연어에서 피해품, 범행도구, 범행수법 등의 정보를 추출하는 과정을 거쳤다. 또한, 직접적으로 추출할 수 있는 정보 이외에 필드와 필드 간의 관계를 통해 유추할 수 있는 간접적인 정보의 추출도 진행하였다. 이는 직접적인 관계가 보이지는 않지만 추론이 가능한 관계에 대하여 연산을 줄이는 것과 동시에 추론 작업 자체에 활용하기 위한 과정이다. 직접적인 관계와 간접적인 관계의 정보 추출을 통해 대규모의 트리플 집합을 구축할 수 있다. 〈그림 3〉은 수사기록 데이터베이스에서 트리플을 추출하는 과정이다.



〈그림 3〉 트리플 추출 과정

직접적인 관계를 추출하기 위해 필드 간의 관계를 담고 있는 Tree 데이터베이스를 통해 필드와 필드 간을 관계로 연결하여 트리플을 구성

한다. 간접적인 관계의 추출을 위하여 주제 전문가와의 회의를 통해 추론이 가능한 관계의 기준을 정하고, 추가적인 트리플을 구성한다. 구성한 대규모의 트리플 집합은 데이터베이스에 적재하여 온톨로지 구축 및 온톨로지 서비스에 사용된다.

3.1.2 온톨로지 구축

온톨로지의 설계를 위해 보편적인 설계 방법인 Ontology Development 101을 활용하였다. 방법론에서 제시하는 온톨로지의 도메인(Domain)과 범위 결정, 기존 온톨로지의 재사용 고려, 중요 용어 열거, 클래스(Class) 간의 계층 정의, 클래스 속성(Property) 정의, 속성 정제, 인스턴스(Instance) 생성의 7가지 과정을 거쳤다(Noy and McGuinness 2001). 온톨로지의 도메인과 범위 결정 단계는 경찰청이 기 구축한 수사기록 데이터 중 필드와 필드 간의 관계가 가장 다양하고 복잡한 범죄 분야인 침입 절도 분야로 한정하였다. 기존 온톨로지 재사용의 여부에 관련하여, 경찰청이 사용하는 용어가 일반적인 용어와 다른 의미를 가지는 특수성으로 인해 주제 전문가와의 협의 하에 기존 온톨로지는 참고 사항으로만 활용하였다. 이후의 단계는 다음과 같다.

(1) 중요 용어 열거

수사기록 데이터에서 추출한 정보는 같은 의미지만 조사 및 어미가 가지는 한국어의 특징 때문에 형태가 다른 경우가 존재한다. 온톨로지의 구축을 위하여 이를 정제하는 과정을 거쳤고, 추상화 및 정규화 과정을 통해 개념적으로 같은 단어들의 집합인 대표 단어를 뽑아냈다. 대표 단어를 뽑아내는 과정에서 경찰청에서 사용하는 용어의 의미와 일반적인 의미 간 차이가 발생하는 문제가 발생하였고, 경찰청 측 주제 전문가와의 지속적인 협의를 통해 용어의 의미 충돌 문제를 해결하였다. 또한, 추상화 및 정규화 과정을 거친 데이터 간 중복 등의 오류가 없는지 확인하였다.

(2) 클래스 간의 계층 정의

클래스는 개념적으로 같은 단어들의 집합을 표현할 수 있는 대표 용어이다. 이전 단계에서 열거한 용어들의 계층을 상위 클래스와 하위 클래스로 나누었고, 클래스와 클래스 간의 계층적 구조를 설정하였다. 경찰청에서는 수사기록의 범행 장소, 범행도구, 범행수법 등의 정보를 추출하여 정제한 뒤 이를 지속해서 데이터베이스화하였는데, 계층을 정의하는 과정에서 기구축한 데이터베이스와 스키마를 참고하였다. 이러한 과정을 통해, 경찰청에서 사용하고 있는 상・하위 계층과 일반적인 용어의 상・하위 계층이 다른 점을 해결하였고, 추가적으로 발생한 모호한 사항에서는 주제 전문가와 협의를 거쳐 클래스 간의 계층을 정의하였다.

(3) 클래스 속성 정의

클래스의 속성은 두 클래스 간의 관계를 설명하기 위해 사용된다. 앞서 설정한 클래스의 상・하위 계층도 하나의 속성으로 활용하였으며, 경찰청에서 기구축한 데이터베이스의 스키마를 기준으로 하여 클래스의 속성을 정의하였다. 대규모 트리플 집합의 복잡한 관계를 통해 새로운 추론 결과를 얻어내기 위하여, 문서 정보의 직접적인 관계뿐만 아니라 추론 등의 과

정을 통해 간접적으로 보이는 관계들도 정의하였다. 간접적인 관계를 정의할 때도, 주제 전문가와의 협의를 거쳐 수사관이 사용하기에 알맞은 속성인지 확인하였다.

(4) 속성 정제

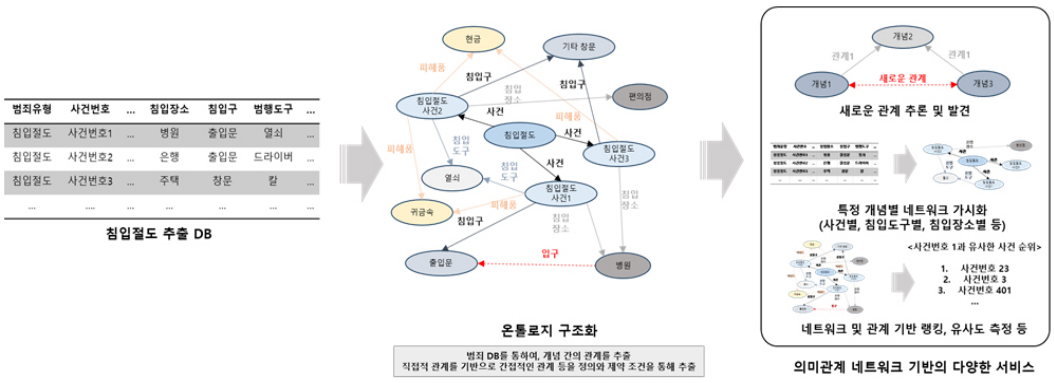
속성은 온톨로지를 구성하고 완성하는 단계까지 발생하는 오류들을 통해 지속적인 정제 과정을 거쳐야 한다. 침입 절도 온톨로지를 구축하는 과정에서 발생하는 논리적 오류 뿐만 아니라, 검색 서비스의 질의 및 추론 단계에서 발생하는 오류 등의 사항을 통해 속성을 지속적으로 정제하는 과정을 거쳤다.

(5) 인스턴스 생성

인스턴스는 클래스에 포함되지 않는 용어를 말한다. 본 논문에서는 수사기록 데이터 필드의 필드 값을 통하여 인스턴스를 생성하였으며, 정제된 값과 정제되지 않은 값을 필드의 특성에 따라 구분하여 생성하였다. 정제된 값과 정제되지 않은 값을 사용한 이유는 온톨로지 기반의 검색 서비스에서 수사관이 목적에 맞게 검색을 용이하게 하기 위해서이다.

온톨로지 기반의 서비스는 대규모 트리플 집합의 복잡한 관계를 통해 새로운 추론 결과를 얻어낼 수 있기 때문에 문서 내 정보의 광범위한 추출은 필수적이다. 트리플 추출에 앞서 문서 정보를 통해 직접적으로 보이는 관계와 직접적으로 보이지는 않으나 추론 등의 과정을 통해 앞서 정의한 간접적으로 보이는 관계들을 통해 트리플을 자동추출하는 모듈을 구성하여 트리플을 추출하였다. 지속적으로 문서 및 문서에서 추출한 정보를 분석하여 새로운 관계를 도출하고 이를 통해 추가적으로 트리플을 추출한다면 더욱 신뢰도와 정확성이 높은 트리플 데이터베이스를 생성할 수 있다.

〈그림 4〉는 수사기록으로부터 추출한 경찰청의 범죄 빅데이터를 바탕으로 온톨로지를 구축하고, 이를 통해 구현 가능한 서비스 방안들을 표현한 그림이다. 입력 조건과 정확하게 일치하는 결과만을 출력하는 검색 방법인 하드 매칭 검색으로 알아낼 수 없었던 정보들을 대규모로 생성된 트리플이 광범위하게 엮여 이를 통한 추론으로 알아낼 수 있다는 것이 온톨로



〈그림 4〉 온톨로지 구축 및 기반 서비스

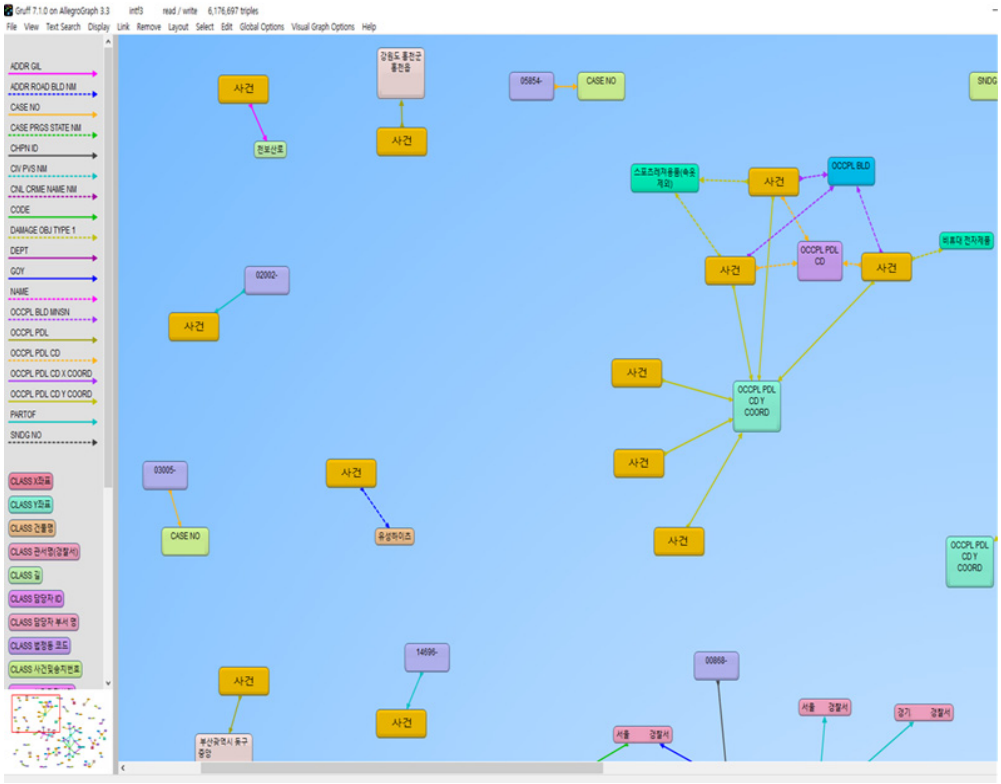
지 서비스의 특징 중 하나이다.

자동화 과정을 거쳐 생성된 트리플 데이터베이스를 Gruff 7.1.0 for AllegroGraph 3.3 (64-bit) (Franz INC., 2013) 프로그램 내부의 Allegro Graph 3.3 Database에 저장한 뒤, 시각화 프로그램중 하나인 Gruff 7.1.0을 활용하여 시각화된 결과를 확인하였다. 〈그림 5〉는 Gruff 7.1.0을 통해 온톨로지를 시각화하여 출력한 화면의 예시이다. 시각화된 결과를 통해 기존 데이터베이스 검색의 단방향적 모델이 아닌, 시맨틱 검색의 양방향적 모델이라는 장점을 가지는 온톨로지만이 가능한 서비스 시나리오의 도출할 수 있다.

3.2 침입 절도 온톨로지 기반의 서비스 구현

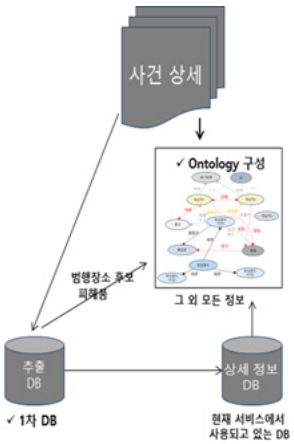
방법론대로 경찰청의 범죄 빅데이터 중 침입 절도 범죄 데이터를 통해 온톨로지를 구현하였고, 온톨로지를 기반으로 하여 실제 활용할 수 있는 서비스 시나리오의 구성과 서비스 구현 작업을 실시하였다. 이 때, 데이터베이스를 세 종류로 나누어 사용하였는데 각각의 데이터베이스는 다음과 같다.

첫 번째는 사건 정보를 담고 있는 범죄기록 데이터베이스로, 실제 자연어로 구성된 범죄기록 정보를 그대로 분석할 수 있도록 하였다. 두 번째는 정보 추출 데이터베이스로, 범죄기록 내



〈그림 5〉 침입 절도 데이터베이스를 기반으로 구축한 온톨로지의 시각화 출력 예시

에서의 1차적인 개체명 추출이 이루어진 데이터베이스를 활용한다. 이를 통하여 상세 피해품이나 범행 장소 후보 필드 등을 구성하여 트리플 정보로 활용한다. 마지막으로 실제 경찰청에서 사용하고 있는 데이터베이스 기반 검색 서비스의 상세 정보 데이터베이스를 사용한다. 이는 정형화된 필드를 활용하여 기본적인 트리플 정보로서 활용된다. 결과적으로, 온톨로지 구성 작업에 활용되는 데이터베이스에 대한 간략한 도식은 〈그림 6〉과 같다.



〈그림 6〉 온톨로지 구성 과정과 사용되는 데이터베이스의 종류

〈표 1〉은 침입 절도 범죄 데이터와 트리플 추출 작업 결과에 대한 통계 정보이며, 데이터 내 필드와 필드 간의 상하 관계 및 다양한 관계를 통해 트리플을 추출한 결과이다. 이를 활용하여 범죄 분야 중 '침입 절도'에 대한 온톨로지 기반의 서비스를 구축하였다. 온톨로지 기반 필드 검색 서비스와 범행 장소 추천 서비스로 나뉘며, 본 논문에서는 이를 나누어 자세히 설명한다.

〈표 1〉 침입 절도 사건 데이터 기반 트리플 추출 작업 결과에 대한 통계 정보

범죄 수	42,306
문서 수	25,517
추출된 트리플 수	2,657,015

3.2.1 침입 절도 온톨로지 기반 필드 검색 서비스

온톨로지 기반의 검색 서비스는 각 필드 간의 연결과 가중치를 정해두기 때문에 각 필드 간의 연결성을 파악하여 검색 결과를 각 필드 단위로 추출할 수 있다는 점이 특징이며, 이는 일반적인 키워드 매칭 검색과 다른 점이라 할 수 있다. 예를 들어, 범죄 사건에 대하여 검색을 진행할 때 일반적인 검색은 사건의 단서가 되는 필드를 통해 사건을 검색하는 형태로 이루어지는 반면, 온톨로지를 적용한 검색은 단서 필드를 통하여 유사한 단서나 관련성이 높은 단서 정보를 검색하는 형태로 이루어진다. 일반적인 검색 방법인 하드 매칭 검색이 아니라, 입력한 검색어와 일치하지 않더라도 관계도가 높은 정보를 출력하는 소프트 매칭 검색을 수행함으로써 필드 간의 연관성을 고려할 수 있다. 또한, 해당 검색어가 정확한 검색어가 아니더라도 주변의 유사한 검색어까지 활용하여 검색 결과를 도출할 수 있다. 앞서 예시로 설명한 범죄 사건의 경우, 단서를 정확히 모를 때 사건 정보 간의 검색 및 단서 간의 패턴 정보 등을 온톨로지 기반의 검색을 통해 파악함으로써 사건 해결의 실마리를 제공하는 등으로 발전할 수 있다.

온톨로지 기반 검색의 형태는 일반적인 검색 형태에 비하여 적은 검색어를 가지고도 유사성과 관계성이 높은 검색어를 추출하면서 좀 더

구체적인 내용으로 다가갈 수 있다는 점과 하드 매칭만으로는 발견하지 못하는 정보와 이용자가 사전에 예상하지 못한 검색어 간의 패턴 등을 발견할 수 있다는 점이 있다. 이를 각각 범죄 분야에 대해 적용하면, 수사 과정에서 가지고 있는 단서가 적어 수사가 어려운 상황이 발생했을 때 유사하면서도 관계성이 높은 다른 단서의 추천을 통하여 추가적인 단서 확보가 가능하기 때문에 사건을 보다 구체화 시키는 과정에 하나의 계기가 될 수 있고, 수사 과정에서 찾은 단서들 간의 예상하지 못한 패턴 정보를 분석하여, 더욱 빠르게 사건을 구체화시키는 것이 가능하다. 예를 들면, 침입 절도 사건 현장에 도착하여 수사를 시작할 때, 범행 장소와 피해품 외의 단서를 찾기 어려운 경우 다른 단서를 추천받아 수사를 진행할 수 있으며, 이를 통해 전혀 연관성이 없는 것처럼 보였던 다른 단서 및 관련된 구체적인 사건을 발견할 수 있다.

추가적으로, 온톨로지를 이용한 검색은 필드만이 아닌 사건을 대상으로도 수행할 수 있다. 그 때문에 필드를 대상으로 한 검색 중간에 가지고 있는 단서들에 기준하여 구체적인 사건을 검색할 수도 있다. 주어진 사건 단서를 기준으로, 찾고자 하는 구체적 사건이 나타나는지를 확인하며 검색을 중단할 것인지, 계속 이어나갈 것인지를 결정할 수 있다. 필드 검색 서비스에서 활용하는 자질 정보는 〈표 2〉와 같다.

〈표 2〉는 필드 검색 서비스에서 활용하는 자질 정보를 나타낸 표로, 자질 정보는 다음과 같이 사용된다. 〈표 2〉의 ①은 해당 사건의 주어진 단서 필드에 대한 온톨로지 내의 동시 출현 범행 장소를 계수한 뒤 가중치화 하여 계산한 정보로, 가장 단순하게 적용할 수 있는 가중치

〈표 2〉 필드 검색 서비스에서 활용하는 자질 정보

	활용 자질 정보
①	단서 필드에 대한 동시 출현 정보
②	단서 필드와의 TF-IDF 가중치 정보
③	단서 필드 간의 다중 TF-IDF 가중치 정보
④	단서 필드와의 통계적 가중치 정보
⑤	단서 필드의 상하 관계 가중치 정보
⑥	범행 장소 후보 필드의 상하 관계 정보

정보이다. 〈표 2〉의 ②는 각 사건 인스턴스를 "문헌"으로, 각 필드 정보를 "키워드"로 두고 문헌빈도(TF, Term Frequency)와 역문헌빈도(IDF, Inverse Document Frequency)를 구하여 둘을 곱하는 가중치 기법이다. $tf_{x,y}$는 문헌(y) 내 키워드(x)의 빈도수를 의미하고, df_x는 키워드(x)가 문헌 내에 등장하는 개수를 의미하며, N은 문헌의 전체 수를 의미한다. 이는 가장 대표적인 용어 가중치 기법으로, 각 단서 필드 마다의 범행 장소에 대한 가중치 계산을 가능하도록 한다. 결과적으로, 다음 (1)과 같은 연산을 수행한다.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

〈표 2〉의 ③은 각 사건 인스턴스를 "문헌"으로, 각 필드 정보간의 조합을 "키워드"로 두고 문헌빈도(TF, Term Frequency)와 역문헌빈도(IDF, Inverse Document Frequency)를 구하여 둘을 곱한다. 〈표 2〉의 ②의 TF-IDF 연산과 같은 방식으로 구성되며, "키워드"에 해당하는 값이 각 필드에서 필드 간의 조합 정보로 바뀐 점이 다르다. 이는 경우의 수로 완전 연결된 필드 간 조합 정보를 통한 패턴 정보 가

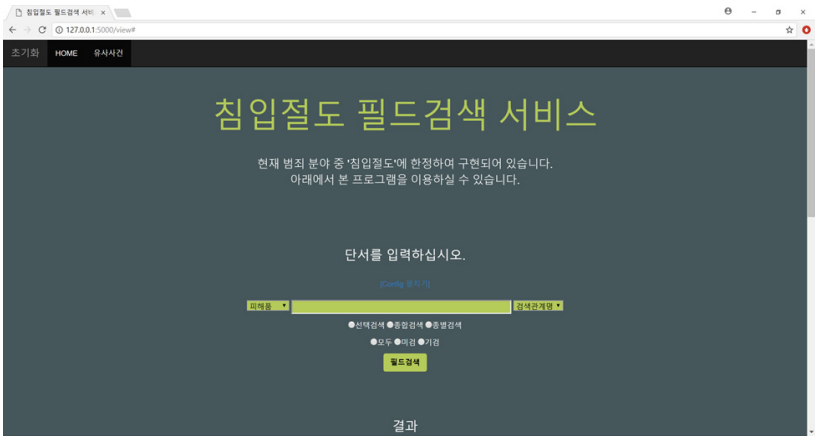
중치를 강화시키기 위한 방안으로 활용되었다. 〈표 2〉의 ④는 각 단서 필드의 정보를 가지고 있는 사건 인스턴스를 임의로 30건씩을 5 세트로 뽑아 그 안에서의 각 범행 장소 필드에 대한 동시 출현 빈도를 계산하여 평균을 낸 가중치 정보로, 〈표 2〉의 ①의 동시 출현 정보에 대해 표준편차를 적용하여 통계적 당위성을 부여한 가중치 σ이다. 임의로 뽑는 사건 인스턴스 수를 n이라 하고, 세트의 횟수를 s라고 하며, a라는 단서가 모든 세트에 포함되면서 b라는 단서가 동시에 출현하는 횟수를 c라고 정의할 때, 분산의 표준편차를 적용하여 연산을 수행하는 가중치 기법이다. 결과적으로, 다음 (2)와 같은 연산을 수행한다.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(c)}{n}} \qquad (2)$$

〈표 2〉의 ⑤는 각 단서 필드의 상하 관계를 파악하여, 동등한 상위 값을 가지고 있는 다른 필드를 상하 유사 필드로 두고, 필드 간 일정 수준 이상의 유사한 패턴이 있을 것이라 가정하여, 이에 대한 〈표 2〉의 ①, ②, ③, ④ 가중치 정보를 합산하여 일정 비율의 가중치로 반환하는 가중치이다. 〈표 2〉의 ⑥은 범행 장소 후보 필드의 상하 관계를 파악하여, 동등한 상윗값을 가지고 있는 다른 범행 장소 필드를 상하 유사 필드로 두고, 이러한 필드 간의 일정 수준 이상의 유사한 패턴이 있을 것이라 가정하여, 이에 대한 〈표 2〉의 ①, ②, ③, ④ 가중치 정보를 합산하여 일정 비율의 가중치로 반환하는 가중치이다.

최초 검색 시에는 침입 절도 범죄 내에서 가장 중요한 단서인 "피해품"을 통하여 검색을 실시한다. 이후, 다른 단서 정보를 추가하면서 필드 조합을 구성해가면서 검색을 수행할 수 있다. 또한, 전반 검색 과정에서 "미검"과 "기검" 여부에 따른 사건을 나누어 검색할 수도 있다. 이러한 전반의 검색 방식은 주제전문가의 의견을 수용하여 수사관들이 실제 사용하기에 용이한 방향으로 설계한 결과이다. 〈그림 7〉은 해당 검색 서비스의 메인 화면 UI이다.

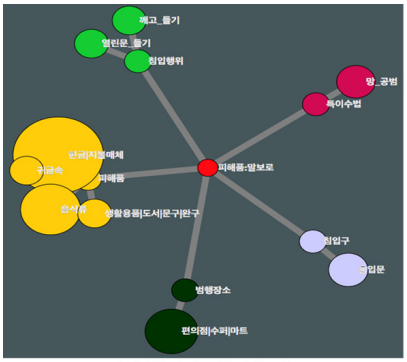
검색의 방식은 총 4가지로, 찾고자 하는 단



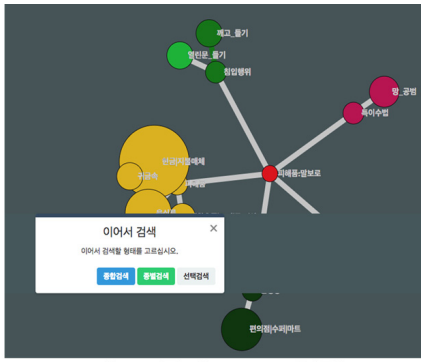
〈그림 7〉 침입 절도 온톨로지 기반 검색 서비스 메인 UI

서 필드의 유형을 정해주는 선택 검색, 각 필드 유형별로 상위 순위 값을 전부 보여주는 종별 검색, 각 필드 유형에 상관없이 전체 필드 중 상위 순위 값을 검색하는 종합 검색, 주어진 단서 필드 조합에 대한 사건을 검색하는 사건 검색이다. 주어진 단서에 대해서 계산된 결과는 필드 간의 네트워크 형식으로 출력하여, 이용자에게 제시된다. 〈그림 8〉과 〈그림 9〉는 UI로 출력된 검색 방식별 결과의 예시이다.

〈그림 8〉의 종합 검색 결과 예시에서 확인할 수 있듯이, 중심의 단서가 된 필드 노드를 기준으로, 검색 결과로 출력된 단서 필드 노드가 연결되어 있는 형태이다. 검색 방식에 따라, 노드 사이에 중간 노드가 존재할 수 있는데, 이는 검색 결과 노드를 설명하는 단서유형 및 필드노드 유형 등으로 구성된다. 제시된 네트워크를 통하여 이용자는 다시 검색을 이어나갈 수 있다. 이어나갈 경우에는 기존 사용된 단서 노드가 그대로 유지된 상태에서, 추가적인 단서 노드가 적용된 검색 결과를 보여준다. 추가 검색 시에도 검색 유형을 재설정할 수 있고, 미검 및 기검 여부와 출력을 원하는 필드도 선택이 가

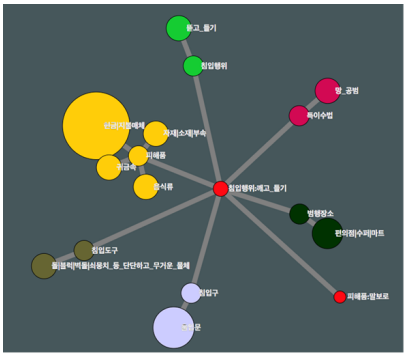


종합 검색 결과

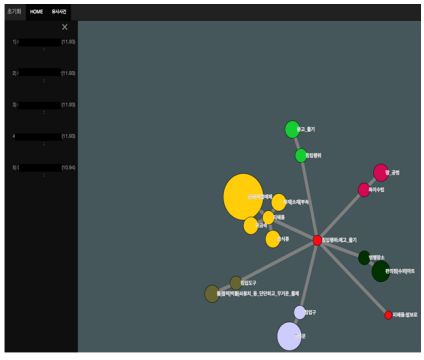


"깨고 들기" 노드 선택

〈그림 8〉 이어서 검색하는 과정과 사건 검색 수행에 대한 예시 1



이어서 검색 결과



검색 중간의 사건 검색

〈그림 9〉 이어서 검색하는 과정과 사건 검색 수행에 대한 예시 2

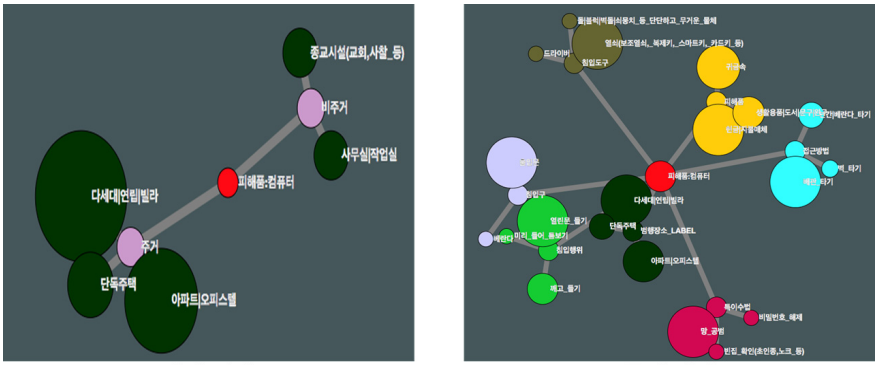
능하다. 검색을 이어가다가 검색 도중에 사건 검색을 시도하는 것도 가능하며, 사건 검색 결과는 다음과 같은 네트워크로 제시되지 않고 상단 좌측의 “유사사건” 버튼을 통하여 확인할 수 있는 순위화된 사건 번호로 확인할 수 있다. 사건 번호를 클릭하면, 해당 사건의 정보가 〈그림 9〉 우측 하단의 예시와 같이 “사건 검색”에 해당하는 네트워크 형식으로 출력된다. 〈그림 10〉과 〈그림 11〉은 이어서 출력 결과인 네트워크를 통하여 이어서 검색을 하는 모습과 사건 검색을 수행하는 일련의 과정이다.

(1) 선택 검색

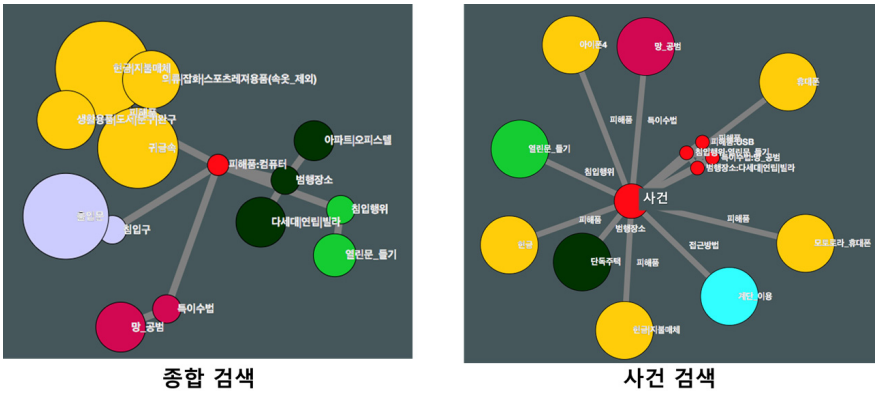
〈그림 10〉의 좌측과 같이, 검색 필드와 유사도가 높은 필드를 이용자가 선택한 관계만 1위부터 5위까지 출력하는 검색 유형이며, 상위 레벨 값으로 묶어(Chunking) 필드를 출력하기 때문에 한눈에 결과를 확인할 수 있다.

(2) 종별 검색

〈그림 10〉의 우측과 같이, 검색한 필드와 유사도가 높은 각 필드별 1위부터 3위까지의 필드들을 출력하는 검색 유형이며, 선택 검색 유



〈그림 10〉 검색 방식별 네트워크 출력 예시 1



〈그림 11〉 검색 방식별 네트워크 출력 예시 2

형과 동일하게 필드를 상위 레벨 값으로 묶어(Chunking) 출력한다.

(3) 종합 검색

〈그림 11〉의 좌측과 같이, 검색한 필드와 유사도가 높은 1위부터 9위까지의 필드를 출력하는 검색 유형이며, 선택 검색 유형과 동일하게 필드를 상위 레벨 값으로 묶어(Chunking) 출력한다.

(4) 사건 검색

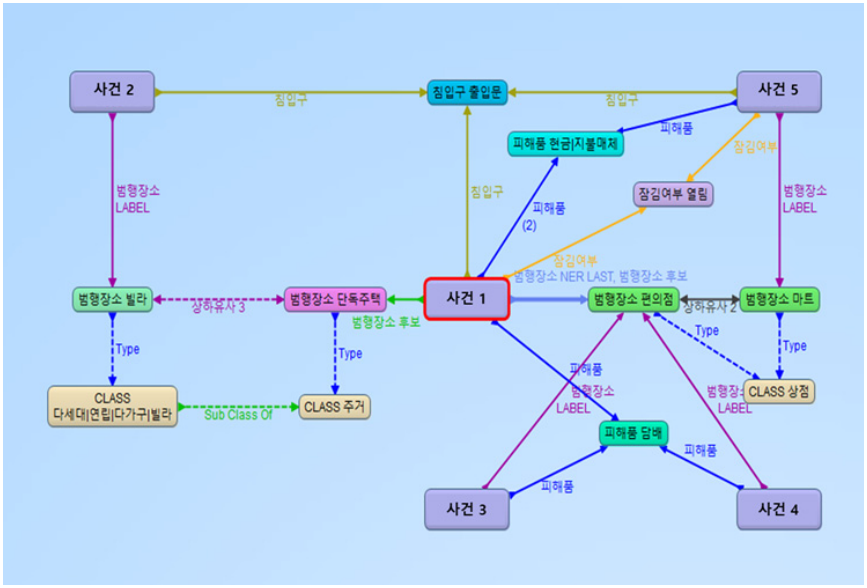
〈그림 11〉의 우측과 같이, 이어서 검색을 계속하여 수행하게 되면 누적된 검색 결과를 토대로 1위부터 5위까지의 유사 사건들이 추천되며, 유사 사건 탭에서 유사 사건과 유사도(절댓값)를 출력한다. 유사 사건 탭은 이용자가 사용할 때만 호출하여 사용 가능하며 숨김이 가능하게 하였다. 유사 사건을 클릭하면 해당 사건 노드를 중심으로 해당 사건의 필드들과 누적된 검색 필드가 전부 시각화되어 화면에 출력된다.

3.2.2 침입 절도 온톨로지 기반 범행 장소 추천 서비스

현재 경찰청 내부에서 진행하는 침입 절도 정보추출 과정의 경우, 범행기록 내에서 규칙 기반의 개체명 추출 과정을 통하여 정보추출이 이루어지는 형태이다. 규칙 기반의 개체명 추출은 정규표현식과 같은 패턴과 개체명 사전을 이용하는 방법으로, 좋은 패턴의 생성 방법과 개체명 사전의 크기가 성능 향상을 위한 요건이 된다(송영길, 정석원, 김학수 2015). 하지만 범행기록 내에서 범행 장소가 하나만 등장하는 경우의 패턴만이 아니라, 범행기록 내에서 범행 장소가 둘 이상 등장할 경우와 같은 새로운 패턴이 등장했을 때 동등한 범행 장소 중에서 어떤 범행 장소가 더 적합한 것인지에 대한 판단은 어려운 문제이다. 해당 문제에 대한 해결 방안으로, 별도의 규칙을 추가하거나 다른 단서 필드에 대한 동시 출현 통계 등으로 판단하는 등의 방법이 활용되고 있다. 본 논문에서는 이러한 문제에 대하여 각 필드 간 상호 관계와 관계별 가중치를 활용하여 적합한 범행 장소를 선별할 수 있으리라 생각하였고, 이에 따라 침입 절도 온톨로지 기반의 적합 범행 장소 추천 서비스를 구현하였다. 본 서비스는 범행기록 내의 다양한 다른 단서 필드를 활용해 필드 간의 관계와 관계별 가중치 등을 합산하여 추출된 후보 범행 장소 중에 해당 사건에 적합한 범행 장소를 선별하여 적용하는 형태로 이루어진다. 〈그림 12〉는 다른 단서 필드를 활용하여 관계성을 기반으로 적합 범행 장소를 선별하는 네트워크 예시를 간략히 표현하고 있다.

적합 범행 장소 추천 서비스는 침입 절도 온톨로지 내의 다양한 자질을 활용한 순위 정보를 통해 이루어지는데, 가장 주요한 자질 정보는 범행 장소 후보 필드이다. 다양한 합산 가중치를 순위 정보로 구성하여 후보 범행 장소 중에서 선택하기 때문에 〈그림 12〉와 같이, 데이터베이스를 활용하여 범행 장소 후보 필드 정보와 개체명 추출 순서 정보 등을 별도로 추출한 온톨로지를 활용하여 트리플 네트워크를 구성하고, 단서 필드를 “피해품”, “침입구”, “침입행위”, “특이 수법”, “접근방법”, “제거 장애물”, “침입 도구”로 총 7가지 선정하여 적용한다.

단서 필드 선정에는 범행 장소에 대하여 그 영향력이 크다고 판단되는 필드를 기반으로 주제 전문가의 선별을 통해 이루어졌다. 각 단서



〈그림 12〉 침입 절도 온톨로지 기반 적합 범행 장소 추천 서비스의 네트워크 예시

필드에 대한 전반적인 계산은 다음에서 설명하는 6종의 가중치 및 자질을 통하여 이루어진다. 각 결괏값은 주어진 변숫값의 비율로 곱하여 합산하고, 이를 기반으로 각 범행 장소 후보의 순위화가 이루어진다. 상위 순위를 기준으로 단서 필드에 적합한 범행 장소를 추천할 수 있으며, 이를 데이터베이스 내부에 적용하여 자동으로 데이터베이스의 범행 장소 필드를 보강하는 방식으로 서비스가 이루어진다. 범행 장소 추천 서비스에서 활용하는 자질 및 관계 정보 중 단서 필드에 대한 동시 출현 정보, 단서 필드와의 통계적 가중치 정보, 단서 필드의 상하 관계 가중치 정보, 범행 장소 후보 필드의 상하 관계 정보는 〈표 3〉과 같다.

〈표 3〉의 ①, ②, ③, ④는 〈표 2〉와 동일한 형태로 활용되는 자질 정보이며, 범행 장소 추천 서비스에서는 별도로 ⑤, ⑥의 자질 정보를 활용한다. 〈표 3〉의 ⑤는 각 사건 인스턴스를

〈표 3〉 범행 장소 추천 서비스에서 활용하는 자질 정보

	활용 자질 정보
①	단서 필드에 대한 동시 출현 정보
②	단서 필드와의 통계적 가중치 정보
③	단서 필드의 상하 관계 가중치 정보
④	범행 장소 후보 필드의 상하 관계 정보
⑤	단서 필드와의 TF-IDF 가중치 정보
⑥	수사기록 내의 개체명 추출 순서 정보

"문헌"으로, 각 필드 정보를 "키워드"로 두고 문헌빈도(TF, Term Frequency)와 역문헌빈도(IDF, Inverse Document Frequency)를 구해 둘을 곱한다. 이는 가장 대표적인 용어 가중치 기법으로, 각 단서 필드 마다의 범행 장소에 대한 가중치 계산을 가능하도록 한다. 이에 대해 다중 단서 필드 정보를 경우의 수에 맞게 매핑하여 활용할 수 있으며, 이를 다중 TF-IDF 가중치로 별도로 계산하여 적용하였다. 〈표 3〉의 ⑥은 자연어로 구성된 수사기록 내의 개체

명 추출 당시의 추출 순서가 첫 번째인지, 마지막인지 등의 등장 순서에 대한 패턴이 있을 것이라 예상하여 추출한 개체명 추출 순서 정보이다. 이에 대해서는 일정 비율의 가중치로 변환하는 작업이 이뤄지며, 주로 첫 번째와 마지막 추출 정보를 추출하였다.

4. 실험 및 환경

침입 절도 온톨로지에 기반한 필드 검색 서비스와 범행 장소 추천 서비스의 성능을 파악하기 위하여 실험을 진행하였고, 실험에 사용된 데이터 및 실험 내용은 다음과 같다.

4.1 실험 데이터

적합한 평가를 위하여 범행 장소 후보 필드 중 실제 범행 장소가 존재하면서 사건기록과 대비하였을 때 실제로 추출된 범행 장소가 일치하는 무결점 인스턴스를 수작업으로 추출하였다. 침입 절도 범죄 분야에 해당하는 전체 사건 42,306건의 약 10% 규모에 해당하는 4,000건을 추출하였으며, 이를 샘플 데이터 셋으로 활용하였다. 샘플 데이터 셋 4,000건 내에서 무작위로 400건의 사건 인스턴스를 구성하여 실험 성능을 비교할 실험 데이터 셋으로 구성하였다.

4.2 성능 최적화 실험

가중치별 합산 비율 변수를 최적화하기 위한 최적화 실험을 수행하였다. 실험 데이터로는 4.1의 4,000건의 샘플 데이터셋을 사용하였으며, 비율에 맞게 적용되는 가중치별로 일정 범위를 지정하고, 변수별 경우의 수를 모두 충족하여 변수를 구성할 수 있도록 루프를 구성하여 최적화 실험을 진행하였다. 〈표 4〉는 통계 분야 주제 전문가의 의견을 반영하여 변수의 범위를 설정한 뒤, 실험을 진행한 내용이다.

〈표 4〉 최적화 실험을 위한 변수 범위

변수	범위		
	시작	끝	증가
TF-IDF 가중치	0.0	1.0	0.001
다중 TF-IDF 가중치	0.0	1.0	0.01
통계 가중치	0.0	1.0	0.01
상하 유사 가중치	0.0	1.0	0.01
상세 피해품 가중치	0.0	2.0	0.05
수사기록 내 피해품 최초 출현 가중치	0.0	1.0	0.05

〈표 4〉의 가중치는 앞서 〈표 2〉와 〈표 3〉에서 활용한 자질 정보의 내용적인 개념과 같다. 최적화된 가중치 변수를 찾고자 실험을 통해 경험적으로 이를 선별하였고, 필드 검색 서비스와 범행 장소 추천 서비스에 최적화된 가중치 변수를 넣어 최적의 성능을 얻고자 하였다. 각각의 가중치에 대해 일정한 값으로 증가시켜 그때마다 성능을 평가하였고, 최적화된 값을 얻어내었다.

4.3 침입 절도 온톨로지 기반 필드 검색 서비스 실험

침입 절도 온톨로지 기반 검색 서비스의 검색 성능을 파악하기 위하여 사건 검색에 대한 정확도 측정 방법의 하나인 Top-K 방식의 정확도 측정 실험(Oh 2017)을 수행하였다. Top-K 정

확도 측정 방식은 확률이 높은 순서대로 K개 안에 정답이 있으면 그 데이터에 대해서는 옳은 데이터라고 세는 방식이다. 필드 간의 관계에 대한 정확한 지표나 성능 척도가 없는 관계로 필드 간의 유사도 등을 파악할 수 있는 실험은 별도로 수행하지 않았으나, 샘플 데이터를 구성하여 유사 사건 검색 결과를 통한 정확도를 측정하는 방식으로 실험을 수행하였다. 동일한 가중치를 기반으로 검색을 수행하기 때문에, 이에 대한 성능 평가로 대체할 수 있으리라 판단하였다. 실험의 방식은 주제 전문가가 선별한 단서 필드 유형 내에서도 침입 절도 내에서 중요한 단서 필드 유형으로 판단된 “피해품”, “범행 장소”, “침입행위”, “특이 수법”, “침입구” 등 5가지의 유형을 통하여 실험을 시행하였다. 이에 대해서는 필드 유형 간의 순서를 섞어 각 조합에 대한 성능을 확인하였다. 또한, 소프트 매칭 성능을 파악하기 위하여 실제 필드의 유사한 필드 정보인 상하 유사 관계를 가진 필드를 활용한 검색을 실험하였다. 해당 실험의 경우, 상하 유사 관계가 확실하게 존재하는 “범행 장소”, “침입행위”를 기준으로 하여 유사 필드를 추출하여 적용하여 실험하였다.

4.4 침입 절도 온톨로지 기반 범행 장소 추천 서비스 실험

침입 절도 온톨로지 기반 적합 범행 장소 추천 서비스의 우수성 입증을 위한 실험을 크게 2가지로 진행하였으며 실험 데이터셋은 4.1의 샘플 데이터 셋을 활용하였다. 실험 성능 측정에는 실제 범행 장소에 해당하는 정답 중에서 예측한 값에 해당하는 비율인 재현율(Recall), 예측한 값 중에서 실제 정답의 비율에 해당하는 정확률(Precision), 재현율과 정확률의 평균 분포에 해당하는 F1-Measure로 나누어 평가하였으며, 이 중에서도 F1-Measure를 핵심적인 평가 척도로 삼았다. F1-Measure의 연산은 (3)과 같다.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

5. 실험 및 결과 분석

온톨로지 자체에 대한 성능의 평가를 시도한 연구는 존재하지 않아 침입 절도 온톨로지를 활용하여 구현한 필드 검색 서비스와 범행 장소 추천 서비스의 성능을 파악하기 위한 실험을 진행하였다. 실험을 통해 본 논문에서 제시하는 서비스의 성능 평가와 결과를 분석하여 객관적인 지표와 시스템의 우수성을 입증하고자 하였다. 실험 방법과 실험에 대한 결과 분석은 다음과 같다.

5.1 성능 최적화 실험 및 결과 분석

각 가중치의 최적화된 가중치 변숫값을 찾기 위하여 4.2의 성능 최적화 실험 절차에 따라 실험을 진행하였다. 실험 결과는 〈표 5〉와 같다.

수사기록 내 피해품 최초 출현 가중치의 경우, 필드 검색 서비스에서만 사용된다. 왜냐하면, 범행 장소 추천 서비스의 경우 상세 피해품의 최초 출현을 다루는 가중치이므로, 전체 필드에 대한 조합을 중요시하는 서비스에는 적합한 가중치가 아니기 때문이다. 서비스의 성격

을 고려하여 〈표 5〉의 가중치 변숫값을 토대로, 필드 검색 서비스와 범행 장소 추천 서비스의 자질에 적용하여 서비스에 반영하였다.

〈표 5〉 필드 검색 서비스에 사용되는 최적화된 가중치 변수

변수	값
TF-IDF 가중치	0.854
다중 TF-IDF 가중치	0.99
통계 가중치	0.31
상하 유사 가중치	0.99
상세 피해품 가중치	1.5
수사기록 내 피해품 최초 출현 가중치	0

5.2 침입 절도 온톨로지 기반 필드 검색 서비스 실험 및 결과 분석

필드 검색 서비스의 경우, 순위화 당시에 동일한 가중치를 가질 때 무작위로 그 값을 배열하도록 설정하였기 때문에 실험마다 성능이 달라질 수 있다. 이 점을 고려하여, 5회의 실험을 하고 그 성능들의 평균을 내어 이를 보완하고자 하였다. 하드 매칭과 소프트 매칭을 각각 나누어 성능을 파악하고자 하였고, 완전히 일치하는 경우와 유사 필드 2종을 활용한 경우로 나누어 실험을 진행하였다.

검색 시스템의 필드 완전 일치 Top-5 실험 결과는 〈표 6〉과 같다. 완전 일치 실험 결과, 가장 성능이 높게 나온 단서 필드 유형 조합은 "피해품", "범행 장소", "침입행위", "특이 수법"의 조합으로 93.52%의 정확도를 보였다. 해당 성능은 "침입구"까지 활용한 검색 성능(92.89%)보다 0.63% 더 높은 성능이다. 이는 "침입구" 필드의 경우, 그 필드 값이 "출입문", "베란다", "구멍(환기구)" 등으로 3가지 밖에 존재하지 않으면서, "출입문"에 해당하는 값에 대부분 편중되어 있기 때문에 단서로써 활용할 때 오히려 검색 성능을 낮춘 것으로 판단된다.

한편, 유사 필드 2종을 활용한 Top-5 실험에서 필드 조합 활용 별로 가장 높은 성능을 보였던 결과는 〈표 7〉과 같다. 유사 필드 활용 실험 결과, 성능이 가장 높게 나온 단서 필드 유형 조합은 "피해품", "특이 수법", "침입행위", "범행

〈표 6〉 검색 시스템의 필드 완전 일치 Top-5 실험 결과

필드 활용	1	2	3	4	5	ACC
3	피해품	범행 장소	침입행위			90.00
4	피해품	범행 장소	침입행위	특이 수법		93.52
5	피해품	범행 장소	침입행위	특이 수법	침입구	92.89

〈표 7〉 검색 시스템의 유사 필드 2종 활용 Top-5 실험 결과

필드 활용	1	2	3	4	5	ACC
3	피해품	특이 수법	침입행위			82.75
4	피해품	특이 수법	침입행위	범행 장소		88.91
5	피해품	특이 수법	침입행위	범행 장소	침입구	88.83

장소" 조합이다. 본 실험에서는 순서가 검색 성능에 영향을 미치지 않기 때문에 사실상 완전 일치와 유사한 결과로 판단된다. 하지만 필드 활용을 3종으로 했을 경우에는 차이를 보인다. 이는 "범행 장소"와 "침입행위"가 유사 필드로써 활용되었기 때문에 특이 수법이 상단에 위치하였으며, "침입행위"가 "범행 장소"보다 유사 필드 사용시 더욱 좋은 성능을 보였기 때문으로 이해할 수 있다. 또한, "침입구" 필드 유형이 유사필드를 사용했을 때보다 성능 저하에 영향을 미치기 때문에 일치 필드임에도 3종에 들지 못하였다고 판단된다.

결과적으로, 본 논문에서 제시하는 온톨로지 기반의 검색 서비스는 완전 일치 필드를 활용할 경우 93.52%의 성능으로 해당 사건을 검색할 수 있다. 또한, 완전히 일치하지 않은 단서의 경우 유사한 필드를 2종까지만 활용하여도 88.91%의 성능으로 해당 사건을 검색할 수 있다. 향후에는 더욱 다양한 단서 필드와 가중치 적용 방법론을 고민하고, 규칙 기반의 방법론 등 다른 방법론과의 복합 적용을 고려하여 온톨로지 기반 검색의 성능 향상을 꾀할 것이며, 이러한 추가적인 방법론에 대해서는 옵션으로서 사용자가 선택하여 활용할 수 있도록 API를 구현할 것이다. 또한, 이용자 입장에서의 편의성을 고려하여, 다양한 방향으로 GUI를 수정・보완해가며 이용자의 만족도를 높일 예정이다.

5.3 침입 절도 온톨로지 기반 범행 장소 추천 서비스 실험 및 결과 분석

5.1의 성능 최적화 실험 이후, 각 범행 장소별 유효한 단서 필드 조합의 추출을 위해 7종의 단서 필드 간 모든 경우의 수를 고려한 단서 필드 조합을 구성하여 실험을 수행하였다. 7종의 단서 필드는 접근방법, 제거 장애물, 침입구, 침입 도구, 침입행위, 특이 수법, 피해품이며 F1-measure 성능에 중점을 두었다. 실험 결과, 각 범행 장소별로 가장 우수한 성능을 얻은 단서 필드 조합과 그 성능은 〈표 8〉과 같다.

〈표 8〉을 분석해 보았을 때, 범행 장소별로 적합한 단서 필드 조합이 매우 다양하게 나타난 것을 확인할 수 있다. 분석 결과, 적은 단서 필드 조합으로도 높은 성능을 보이는 "금은방", "스포츠시설"과 같은 경우는 범행 장소에 대한 단서 필드의 패턴이 비교적 명확한 것으로 분석되었다. 한편, "다세대/연립/빌라"와 같은 경우는 유사한 필드(상하 관계상, "주거"에 해당하는 다양한 필드가 존재)가 많거나, 사건의 건수가 매우 많고 다양한 패턴을 가지고 있는 범행 장소라고 판단되었다. 실제로 "다세대/연립/빌라"의 사건 건수는 해당 데이터셋 내에서 가장 높은 비율을 차지하고 있는 범행 장소이다.

최종적인 실험 결과, 본 서비스는 실험 데이터셋의 전체에 대해 적합한 단서 필드 조합은 "침입 도구", "피해품", "특이 수법", "침입구", "접근방법", "침입행위"의 6가지 필드를 활용하는 것이며, F1-measure 76.19%의 성능으로 데이터베이스 내의 범행 장소 필드 정보를 교정할 수 있음을 확인하였다. 서비스 제공을 위해 Python API 형식의 폼을 통해 입력된 데이터베이스의 범행 장소 후보 중, 적합한 범행 장소를 추천받아 자동 교정할 수 있도록 구현하였다. 향후에는 더욱 다양한 단서 필드와 가중치 적용 방법론을 고민하고, 규칙 기반의 방법론 등의

〈표 8〉 침입 절도 온톨로지 기반 적합 범행 장소 추천 서비스의 실험 결과

장소유형	단서 필드	Recall	Precision	F1
PC방/게임장/오락실	침입구, 침입행위, 피해품	66.67	**100.00**	80.00
공장/공사현장	접근방법, 침입행위, 특이 수법, 피해품	70.97	84.62	77.19
금은방	**침입행위, 피해품**	**90.00**	**100.00**	**94.74**
다세대/연립/빌라	접근방법, 제거 장애물, 침입구, 침입 도구, 침입행위, 특이 수법, 피해품	91.06	75.41	82.50
단독주택	접근방법, 침입구, 침입 도구, 특이 수법, 피해품	56.85	60.05	58.41
병원/의원	접근방법, 침입구, 특이 수법, 피해품	34.48	**100.00**	51.28
사무실/작업실	침입 도구, 침입행위, 특이 수법, 피해품	58.24	54.08	56.08
사우나/찜질방	접근방법, 특이 수법, 피해품	53.85	87.50	66.67
숙박업소	접근방법, 침입구, 침입 도구, 침입행위, 피해품	66.67	**100.00**	80.00
스포츠 시설	침입 도구, 특이 수법	37.50	**100.00**	54.55
식당/카페/유흥/풍속	접근방법, 침입구, 침입 도구, 침입행위, 특이 수법, 피해품	87.65	83.24	85.39
아파트/오피스텔	제거 장애물, 침입구, 특이 수법, 피해품	**95.42**	80.52	87.34
업소(매장) 창고	침입구, 침입행위, 특이 수법, 피해품	57.14	95.92	78.99
전용창고(공장, 물류 등)	접근방법, 침입 도구, 침입행위, 피해품	86.54	73.77	79.65
종교시설	접근방법, 침입행위, 특이 수법, 피해품	76.11	98.85	86.00
주거창고	**접근방법, 침입구, 특이 수법, 피해품**	**2.94**	**100.00**	**5.71**
편의점/수퍼/마트	침입행위, 특이 수법, 피해품	89.80	88.44	89.80
학교/학원/도서관	접근방법, 침입구, 특이 수법, 피해품	61.30	97.73	75.44
전체	접근방법, 침입구, 침입 도구, 침입행위, 특이 수법, 피해품	76.58	78.46	76.19

다른 방법론과의 복합 적용을 고려하여 적합 범행 장소 추천 성능 향상을 꾀할 것이다.

6. 결론 및 향후 연구

온톨로지는 시맨틱 웹을 구현할 수 있는 도구로서 여러 지식 개념들을 의미적으로 서로 연결할 수 있는 도구라는 장점 때문에 인공지능, 정보검색 분야 등 다양한 분야에서 구축 및 활용되고 있다. 본 논문은 경찰청의 대규모 범죄 수사기록 문서 데이터 중 침입 절도 분야를 중심으로 정형화한 정보를 통해 대규모의 트리플을 구성하고, 이를 바탕으로 특정 분야의 온톨로지를 구축하였으며, 수사 시 도움을 주기 위한 목적의 온톨로지 기반 검색 및 범행 장소 추천 서비스를 구현하였다. 수사 기록이라는 데이터를 활용하여 온톨로지를 구축한 선행 연구는 물론, 온톨로지 기반의 서비스를 구현한 관련 연구가 부족한 실정이나, 범죄 기록물에 대하여 다양한 데이터 마이닝 기법이 도입되고 있는 시기에 온톨로지라는 새로운 측면의 시도를 했다는 데 의의가 있으며, 침입 절도 기반 범죄 온톨로지의 구축뿐 아니라 실제 온톨로지에 기반한 서비스를 구현하여 유의미한 성능을 입증하였다.

온톨로지 기반 검색 서비스의 유용성에 대한 성능을 파악하기 위하여 사건 검색에 대한 정확도 측정 방법의 하나인 Top-K 방식의 정확도 측정 실험을 수행하였다. 실험은 필드가 완전히 일치하는 경우인 완전 일치 실험과 실제 필드와 유사 관계를 가지는 경우인 유사 필드

활용 실험의 2가지 형태로 진행하였으며, 실험 결과 완전 일치 실험에서는 약 93.52%, 유사 필드 활용 실험에서는 약 88.91%라는 유의미한 결과를 얻어낼 수 있었다.

온톨로지 기반 적합 범행 장소 추천 서비스는 우수성 입증을 위한 실험을 크게 2가지로 진행하였고, 실험 데이터셋의 전체에 대해 적합한 단서 필드 조합은 "침입 도구", "피해품", "특이 수법", "침입구", "접근방법", "침입행위"의 6가지 필드를 활용하는 것이며, F1-measure 76.19%의 성능으로 데이터베이스 내의 범행 장소 필드 정보를 교정할 수 있음을 확인하였다.

향후 연구에서는 수사기록 데이터에서 자연어 처리를 통한 개체명 인식, 관계 추출 등의 정보 추출을 추가로 진행한 뒤, 추출한 정보와 관계를 기반으로 온톨로지의 확장을 통해 검색 품질을 높일 것이다. 또한, 온톨로지 기반 서비스에 대하여 이용자 만족도 조사를 시행한 뒤, 만족도 조사 결과를 반영한 이용자 중심의 GUI를 통해 이용자의 만족도를 높일 것이며, 가중치 및 관계성에 대하여 최근 심층학습 등의 기계학습 기법에서 주로 활용되는 키워드 가중치 기법인 워드 임베딩(Word Embedding) 벡터를 구성하여 적용하는 등 추가적인 방법론을 통해 성능을 향상할 것이다.

참 고 문 헌

[1] [단독] 52억짜리 AI 수사관 '클루'가 '살인의 추억' 재발 막는다. 2017『중앙일보』. 12월 8일.

[2] 고건우 외. 2018. 대규모 범죄 수사 기록을 활용한 온톨로지 기반 서비스 구현.『한글 및 한국어 정보처리 학술대회 논문집』. 2018년 10월 13일, 서울: 고려대학교 현대자동차 경영관: 477-481.

[3] 국립중앙도서관. 2013. 국립중앙도서관 국가서지 Linkd Open Data 서비스. [online] [cited 2018. 9. 28.] 〈https://lod.nl.go.kr/home/about/introduction.jsp〉

[4] 권양섭. 2017. 범죄예방과 수사에 있어서 빅데이터 활용과 한계에 관한 연구.『법학연구』, 17(1): 179-198.

[5] 김용훈, 정목동. 2017. LSA를 이용한 정형·비정형데이터 분석과 범죄 프로파일링 시스템 구현.『멀티미디어학회논문지』, 20(1): 66-73.

[6] 김평 외. 2008. OntoFrame 기반 학술정보 분석 서비스.『정보과학회논문지: 소프트웨어 및 응용』, 35(7): 431-441.

[7] 박경모, 임희숙, 박종현. 2003. Protege를 이용한 한의학의 구조화된 증상 입력을 위한 온톨로지 개발.『동의생리병리학회지』, 17(5): 1151-1156.

[8] 박준영, 채명수, 정성관. 2016. 실시간 범죄 예측을 위한 랜덤포레스트 알고리즘 기반의 범죄 유형 분류모델 및 모니터링 인터페이스 디자인 요소 제안.『정보과학회 컴퓨팅의 실제 논문지』, 22(9):

455-460.

[9] 방승환, 김태훈, 조현보. 2014. 범죄예측에서의 데이터마이닝 적용 가능성 연구: 절도범죄를 중심으로.『한국컴퓨터정보학회논문지』, 19(12): 309-317.

[10] 송영길, 정석원, 김학수. 2015. 위키피디아 기반 개체명 사전 반자동 구축 방법.『정보과학회논문지』, 42(11): 1397-1403.

[11] 위키백과. 2018. "온톨로지." [online] [cited 2018. 6. 30.] 〈https://ko.wikipedia.org/wiki/%EC%98%A8%ED%86%A8%EB%A1%9C%EC%A7%80〉

[12] 정영석, 김진묵, 박구락. 2012. 범죄유형별 범죄발생 예측확률을 높일 수 있는 방법에 관한 연구.『한국컴퓨터정보학회논문지』, 17(4): 163-172.

[13] 조대웅, 최지웅, 김명호. 2014. 비정형 문서의 정보추출을 통한 OWL 온톨로지 구축 시스템의 설계 및 구현.『한국컴퓨터정보학회논문지』, 19(10): 23-33.

[14] 한국정보화진흥원 지식자원활용부. 2014.『2014 링크드 오픈 데이터 국내 구축 사례집』. 서울: 한국정보화진흥원 지식자원활용부.

[15] 황미녕 외. 2012. 연구 개발 트렌드 분석을 위한 기술 지식 온톨로지 구축.『한국콘텐츠학회논문지』, 12(12): 35-45.

[16] Abasolo, J. M. and Gomez, M. 2000. "MELISA: An ontology-based agent for information retrieval in medicine." In *Proceedings of the first international workshop on the semantic web (SemWeb2000)*, 73-82.

[17] Asaro, C. et al. 2003. "A domain ontology: Italian crime ontology." In *Proceedings of the ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management.*

[18] Franz INC. Gruff. [online] [cited 2018. 9. 28.] 〈https://franz.com/agraph/gruff/〉

[19] Noy, N. F. and McGuinness, D. L. 2001. "Ontology development 101: A guide to creating your first ontology." *Stanford Knowledge Systems Laboratoryand Technical Report KSL-01-05 and Stanford Medical Informaticsand Technical Report SMI-2001-0880.*

[20] Oh, S. 2017. "Top-k Hierarchical Classification." In *AAAI*, 2450-2456.

[21] Raimond, Y. et al. F. 2007. "The Music Ontology." In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, September 23-27, 2007, Vienna, Austria.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] [Exclusive] 5.2 Billion AI Investigator 'Clue' Blocks 'Memories of Murder' Recurrence. 2017.

Korea Joongang Daily. December 8.

[2] Ko, Gun-Woo et al. 2018. "Implementation of Ontology-based Analytics Service by Exploiting Massive Crime Investigation Records." *The 30th Annual Conference on Human & Cognitive Language Technology*, October 13, 2018, Seoul: Korea University Business School.

[3] National Library of Korea. 2013. National Library of Korea National Survey Linked Open Data Service. [online] [cited 2018. 9. 28.] 〈https://lod.nl.go.kr/home/about/introduction.jsp〉

[4] Kwon, Yang-Sub. 2017. "Study on the Application and Legal Limits of Big Data for Crime Prevention and Investigation." *Law Review*, 17(1): 179-198.

[5] Kim, Yong-Hoon and Chung, Mok-Dong. 2017. "Analysis of Structured and Unstructured Data and Construction of Criminal Profiling System using LSA." *Journal of Korea Multimedia Society*, 20(1): 66-73.

[6] Kim, Pyung et al. 2008. "The Academic Information Analysis Service using OntoFrame-Recommendation of Reviewers and Analysis of Researchers' Accomplishments." *Journal of KIISE: Software and Applications*, 35(7): 431-441.

[7] Park, Kyung-Mo, Lim, Hee-Sook and Park, Jong-Hyun. 2003. "Building an Ontology for Structured Data Entry of Signs and Symptoms in Oriental Medicine." *Journal of Physiology & Pathology in Korean Medicine*, 17(5): 1151-1156.

[8] Park, Joon-Young, Chae, Myung-Su and Jung, Sung-Kwan. 2016. "Classification Model of Types of Crime based on Random-Forest Algorithms and Monitoring Interface Design Factors for Real-time Crime Prediction." *KIISE Transactions on Computing Practices*, 22(9): 455-460.

[9] Bang, Seung-Hwan, Kim, Tae-Hun and Cho, Hyun-Bo. 2014. "A Study on the Applicability of Data Mining for Crime Prediction: Focusing on Burglary." *Journal of the Korea Society of Computer and Information*, 19(12): 309-317.

[10] Song, Yeong-Kil, Jeong, Seok-Won and Kim, Hark-Soo. 2015 "A Semi-automatic Construction method ofa Named Entity Dictionary Based on Wikipedia." *Journal of KIISE*, 42(11): 1397-1403.

[11] Wikipedia. 2018. "Ontology." [online] [cited 2018. 6. 30.] 〈https://ko.wikipedia.org/wiki/%EC%98%A8%ED%86%A8%EB%A1%9C%EC%A7%80〉

[12] Chung, Young-Suk, Kim, Jin-Mook and Park, Koo-Rack. 2012. "A study of improved ways of the predicted probability to criminal types." *Journal of the Korea Society of Computer and Information*, 17(4): 163-172.

[13] Jo, Dae-Woong, Choi, Ji-Woong and Kim, Myung-Ho. 2014. "The Design and Implementation of OWL Ontology Construction System through Information Extraction of Unstructured Document." *Journal of The Korea Society of Computer and Information*, 19(10): 23-33.

[14] Korea Information Science Agency. 2014. 2014 Linked Open Data Domestic Casebook, Seoul.
[15] Hwang, Mi-Nyeong et al. 2012. "Ontology Construction of Technological Knowledge for R&D Trend Analysis." *The Journal of the Korea Contents Association*, 12(12): 35-45.