기술문서 정의문 패턴을 이용한 전문용어사전 자동추출 및 활용방안*

Automatic Extraction and Usage of Terminology Dictionary Based on Definitional Sentences Patterns in Technical Documents

한희정 (Hui-Jeong Han)**, 김태영 (Tae-Young Kim)***
두효철 (Hyo-Chul Doo)****, 오효정 (Hyo-Jung Oh)*****

초 록

기술문서는 지식정보사회에서 생성되는 중요 연구 성과물로, 이를 제대로 활용하기 위해서는 정보 요약 및 정보추출과 같은 개선된 정보 처리 방법을 토대로 기술문서 활용의 편의성을 높여줄 필요가 있다. 이에 본 연구는 기술문서의 핵심 정보를 추출하기 위한 방안으로, 기술문서의 구조와 정의문 패턴을 기반으로 전문용어 및 정의문을 자동 추출하고, 이를 기반으로 전문용어사전을 구축할 수 있는 시스템을 제안하였다. 나아가 전문용어사전을 지식메모리로서 보다 다양하게 활용할 수 있도록 전문용어사전에 기반한 개인화서비스 제공방안을 제안하였다. 이처럼 전문용어 및 정의문 자동추출을 기반으로 전문용어사전을 구축하게 되면 새롭게 등장하는 전문용어를 빠르게 수용할 수 있어 이용자들이 최신정보를 보다 손쉽게 찾을 수 있다. 더불어 개인화된 전문용어사전을 이용자에게 제공한다면 전문용어사전의 가치와 활용성, 검색의 효율성을 극대화할 수 있다.

ABSTRACT

Technical documents are important research outputs generated by knowledge and information society. In order to properly use the technical documents properly, it is necessary to utilize advanced information processing techniques, such as summarization and information extraction. In this paper, to extract core information, we automatically extracted the terminologies and their definition based on definitional sentences patterns and the structure of technical documents. Based on this, we proposed the system to build a specialized terminology dictionary. And further we suggested the personalized services so that users can utilize the terminology dictionary in various ways as an knowledge memory. The results of this study will allow users to find up-to-date information faster and easier. In addition, providing a personalized terminology dictionary to users can maximize the value, usability, and retrieval efficiency of the dictionary.

키워드: 기술문서, 전문용어, 정의문, 용어사전, 자동추출
technical documents, terminology, definitional sentences, dictionary, automatic extraction

* 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068186).
이 논문은 한국과학기술정보원(KISTI)이 주관한 'NTIS 경진대회'에서 제공한 데이터를 활용하였음.
** 전북대학교 문화융복합아카이빙 연구소 전임연구원(freebirdhhj@naver.com) (제1저자)
*** 전북대학교 일반대학원 기록관리학과 박사과정(fnty127@hanmail.net) (공동저자)
**** 전북대학교 일반대학원 기록관리학과 석사과정(encl2@naver.com) (공동저자)
***** 전북대학교 기록관리학과 조교수, 문화융복합 아카이빙연구소 연구원(ohj@jbnu.ac.kr) (교신저자)
■ 논문접수일자 : 2017년 11월 18일 ■ 최초심사일자 : 2017년 12월 7일 ■ 게재확정일자 : 2017년 12월 12일
■ 정보관리학회지, 34(4), 81-99, 2017. 〔http://dx.doi.org/10.3743/KOSIM.2017.34.4.081〕

1. 서 론

1.1 연구배경 및 필요성

2015년 우리나라 총 연구개발비는 65조 9,594억 원으로 GDP 대비 연구개발비 비중으로 볼 때 상당히 높은 수준을 나타내고 있다(미래창조과학부, 한국과학기술기획평가원, 2017). 정책적으로도 연구개발성과 분야별로 전담기관을 지정하여 연구 성과 관리・유통 제도를 마련함으로써 국가연구개발사업의 관리 등을 지원하고 있다(「국가연구개발사업의관리 등에 관한 규정」 제25조(연구개발정보의 관리) 13항). 그러나 이러한 국가 차원의 R&D 사업 및 성과물의 양적 증가와 정책적 지원에도 불구하고 현재까지 연구 성과 활용 실적은 여전히 낮은 편이다. 연구 성과물은 연구개발을 통해 창출되는 과학기술적 성과이자 유・무형의 경제・사회・문화적 성과로서 국가 차원의 R&D의 효율성을 제고하기 위해서는 공공부문의 연구 성과물에 대한 활용 및 확산 노력이 필요하다(한국과학기술정보원, 2014).

우리나라의 경우 국가 차원에서 수행한 R&D 보고서를 디지털화하고 이를 체계적으로 데이터베이스화하여 연구자들에게 제공함으로써, 연구 성과의 확산과 성과 결과를 활용한 후속 연구의 활성화를 유도하고 있다. 그러나 다양한 정책에도 불구하고 고부가가치 서비스의 창출로 연결되지는 못하고 있으며, 특히 공공연구기관의 R&D 성과가 민간 부문에 효과적으로 이전・확산되지 못하고 있는 것이 현실이다. 따라서 국가 R&D보고서에 쉽게 접근하고 효율적으로 활용할 수 있는 계기 마련이 필요하다. 한국과학기술정보원이 실제 국가 R&D 보고서 원문의 유용성을 조사・분석한 결과를 보면 이용자들은 국가 R&D 보고서원문을 상당히 유용하게 생각하고 있으며, 국가 R&D 보고서원문을 상당히 높은 수준에서 계속 이용할 의향이 있으나 관련 서비스들이 미진하여 연구 성과물 활용이 제대로 이루어지지 않고 있다(한국과학기술정보원, 2014). 따라서 연구보고서를 비롯한 다양한 기술문서의 활용을 위해서는 이용자들을 위한 서비스를 개발하여 연구자나 일반 이용자들이 쉽게 관련정보를 이용할 수 있는 기반을 마련할 필요가 있다.

기술문서(TD: Technical Documents)는 빠르게 변화 발전하는 지식정보사회에서 생성되는 방대한 양의 지식들과 사회적 변화를 담아내는 중요 연구 성과물이다. 「연구 성과 관리・유통 전담기관 지정고시」에 따르면 연구 성과물에는 논문, 특허, 보고서원문(전자원문 포함), 연구시설 장비, 기술요약정보, 화합물, 생명자원, 소프트웨어가 있으며, 기술 문서는 이러한 연구 성과물 중 문서 유형에 해당한다. 기술문서에 담긴 전문지식의 수준과 범위 그리고 정보량은 매우 방대할 수밖에 없으며, 기술문서가 제대로 활용되기 위해서는 관련 전문지식에 대한 이해도가 높아야 한다. 관련 전문지식에 대한 개인의 이해도에 따라 기술문서의 이용만족도 및 활용도의 차이가 발생할 수밖에 없으며, 따라서 기존의 정보검색이나 정보 분류를 뛰어넘어 정보에 대한 요약 및 핵심 정보 추출과 같은 좀 더 세밀한 정보의 가공을 통해 이용자의 연구보고서 활용의 편의를 높여줄 필요가 있다.

이에 본 연구는 기술문서의 핵심 정보 추출을 위한 방안으로서 먼저 기술문서의 전문용어

(Terminology) 자동추출을 통한 전문용어 사전 구축을 토대로 이용자들에게 전문용어에 대한 접근성과 검색 편의성을 높일 수 있는 방안을 제공하고자 한다. 기술문서 전문용어의 자동추출을 통한 전문용어 사전 구축의 필요성은 다음과 같다.

첫째, 과학기술이 발전하면서 새롭게 등장하는 전문용어를 빠르게 수용할 수 있는 전문용어 사전 구축 방안을 마련할 필요가 있다. 4차 산업혁명 시대를 맞이하여 지식정보화 사회에서 지능정보화 사회로 변화함에 따라 기존의 용어집에 수록되지 않은 전문용어의 수가 폭발적으로 증가하고 있다. 그러나 기존 수작업 기반의 전문용어사전은 이러한 시대적 흐름에 발맞추어 새로운 전문용어를 포괄하기에는 한계가 있다. 따라서 새로운 전문용어를 자동으로 탐색하여 전문용어사전에 포함시킬 수 있는 자동구축 방안을 마련해야 한다.

둘째, 전문용어사전 구축을 통해 보고서 내용 검색의 한계점을 보완할 필요가 있다. 전문용어는 주로 최신 기술문서에 많이 등장하지만, 방대한 양의 기술문서에서 원하는 용어를 찾기 위해서는 검색을 해야 한다. 그러나 이러한 키워드 입력의 검색 방식은 자신이 최신 전문용어를 정확하게 알고 있을 때는 유용하지만 그렇지 않은 경우에는 원하는 용어를 찾기는 어렵다. 특히, 기술문서의 내용이 방대하기 때문에 이들 보고서에 수록된 전문용어를 자동으로 추출하여 사전을 구축한다면 이용자가 전문용어를 찾는데 드는 시간과 노력을 절감할 수 있다. 따라서 기술문서에 담겨있는 최신 전문용어를 자동으로 추출해서 전문용어 사전을 구축하여 이용자에게 제공함으로써 이용자가 수많은 기술문서를 일일이 검색할 필요 없이 전문용어사전을 이용하여 자신이 원하는 정보를 손쉽게 찾을 수 있도록 해야 한다.

셋째, 이용자들이 전문용어사전을 보다 다양하게 활용하여 전문용어사전의 가치를 증대시킬 수 있는 방안을 마련할 필요가 있다. 즉, 전문용어사전을 이용하여 개별 이용자들의 요구에 잘 부합되는 결과물을 제공함으로써 전문용어사전의 가치를 증대시킬 수 있는 방안을 마련할 필요가 있다. 또한 전문용어는 가장 기본적인 지식정보로서 이들 용어의 정의뿐만 아니라 전문용어가 포함된 문서 내 위치정보 역시 중요하다. 따라서 전문용어 사전 구축 시 전문용어의 정의뿐만 아니라 문서 내 전문용어가 위치한 곳의 정보까지 연결해서 제공한다면 관련 맥락정보와 핵심정보를 확장해서 파악할 수 있어 이용자가 원하는 지식정보를 보다 빠르고 풍부하게 얻을 수 있다.

1.2 연구목적 및 방법

본 연구의 최종목적은 방대한 양의 기술문서에서 이용자가 원하는 정보를 자동으로 추출하여 이용자에게 제공할 수 있는 개인화 서비스를 개발하는 데 있다. 이를 위해 먼저 기술문서에 수록된 전문용어 및 정의문(definition)을 자동으로 추출하여 전문용어사전을 구축할 수 있는 방안을 제안하고자 한다. 전문용어사전 구축 방법은 기술문서의 구조(structure)와 정의문 특성패턴(definitional sentences patterns) 및 규칙으로부터 정의문을 추출하고, 이를 기반으로 전문용어를 자동으로 추출하는 방식이다. 기술문서의 구조와 정의문 특성패턴 및 규칙은 한국

과학기술정보연구원(KISTI)에서 제공한 기술문서 중 연구보고서 XML 파일의 문장을 분석하여 도출하였다. 이후 기술문서 전문용어사전 구축에 필요한 정의문 특징패턴과 용례들을 토대로 정의문을 추출한 후 정의문 패턴의 순위와 출현 비율 등을 분석하였으며, 이상의 방식을 통해 구축된 전문용어사전에 기반한 개인화서비스 제공방안에 대해 제안하였다.

1.3 선행연구

전문용어 관련 연구는 정의문 자동 추출에 관한 연구와 전문용어 자동 추출에 관한 연구로 구분된다. 먼저, 정의문 자동 추출에 관한 연구는 신효식, 김재호, 이해윤, 최기선(2002)의 텍스트 코퍼스로부터 용어 정의문을 자동 추출하는 방법을 제안한 연구와 김재호, 배선미, 신효식, 최기선(2005)의 의학분야 코퍼스에서 주어진 전문용어에 대한 정의문을 자동으로 추출하는 방법을 제안한 연구가 있다. 두 논문 모두 용어의 정의문을 자동으로 추출하기 위하여 텍스트 코퍼스로부터 용어 정의문에 해당되는 정보를 자동으로 추출하는 방법을 제시하여 그 성능을 평가하였다.

다음으로 전문용어 자동 추출에 관한 연구로는 먼저, 박정오와 황도삼(2000)이 전문용어의 어절 패턴을 이용하여 후보 전문용어를 추출한 후, 전문용어를 구성할 수 있는 단어의 위치정보를 이용한 전문용어 추출 방법을 제안한 연구가 있다. 이 외에도 기계가독형 전문분야 사전들을 이용하여 사전 간의 계층관계를 구축하고 이를 이용하여 전문용어를 추출하는 방법을 제시한 연구와 EM 알고리즘을 이용한 전문용어 추출기법을 제안한 연구, 전문분야 사전과 문서를 이용해 전문용어의 특성을 파악하고 이를 이용하여 전문용어를 추출하는 방법을 제안한 연구가 있다(오종훈, 김재호, 최기선, 2003; 오종훈, 이경순, 최기선, 2002; 오종훈, 최기선, 2004). 이들 연구들은 전문용어 및 정의문을 자동으로 추출하는 방법을 제안하고, 이에 대한 성능을 평가하는 데 그쳤다.

상기 나열한 선행연구와는 달리 본 논문은 기술문서의 구조와 정의문의 다양한 문장 패턴을 기반으로 정의문을 인식하고, 이로부터 전문용어를 자동으로 추출하여 사전을 구축할 수 있는 전문용어 자동추출 시스템을 제안하였다. 즉, 기존 선행연구는 문장에 나타난 어휘적 특성만을 패턴으로 인식한 반면, 본 연구는 기술문서의 특성을 활용해 구문특징, 형식특징, 항목특징 등을 패턴화 하였다. 또한 개인 용어사전을 관리하여 새로운 정의가 추출되는 경우 자가학습을 통해 확장하는 방안을 제안함으로써 이용자들이 해당 용어와 정의문을 지식메모리로서 보다 다양하게 활용할 수 있도록 전문용어사전에 기반한 개인화서비스 제공방안을 제안하였다.

2. 이론적 배경

2.1 전문용어

전문어는 일반인들이 일상생활 속에서 사용하는 일반어와 대립되는 것으로 전문인들이 자신의 활동 영역 안에서 특수한 목적을 가지고 사용하는 의사소통 언어이다. 전문용어는 특정

한 전문 분야에서 주로 사용하는 용어로서 전문적 개념을 지칭하는 어휘 또는 어휘의 집합이다. 일반 단어는 다의성을 가지지만 전문용어는 그가 속하는 전문 영역 안에서 하나의 의미만을 갖는다. 즉 개념과의 일대일 대응성(일시성, 일의성)은 전문용어만의 특징이다. 전문용어가 여러 개의 의미로 쓰이는 경우 용어 체계 전체에 혼돈을 주게 되므로 전문용어는 표준화의 대상이며, 용어의 명칭과 개념 사이의 항상성, 안정성이 전제되어야만 하는 도구이다(한국학술단체총연합회, 2006).

한편, 전문용어의 개념과 범위는 학술적 전문용어와 공공언어로서의 전문용어로 구분할 수 있으며, 공공언어로서의 전문용어는 국민들의 일상생활과 관련된 용어로 사용 범위가 확장되면서 그 개념 또한 변화하고 있으며, 이와 관련하여 강현화(2009)는 사용 영역(주체)별로 전문용어의 개념 및 범위를 〈표 1〉과 같이 규정하였다.

전문 분야마다 사용하는 전문용어는 계속적으로 생성되는 특성이 있기 때문에 최근 전문용어 추출에 대한 연구가 활발히 진행되고 있다. 전문용어 추출은 특정 전문영역에서 사용하는 전문용어를 해당 분야 문서에서 파악하는 작업을 의미한다. 전문용어후보 추출 과정은 주로 구문 규칙이나 용어의 형성 패턴 등의 언어적 지식을 이용한 '언어적 필터링(Linguistic Filtering)' 과정 혹은 문서 내에서의 전문용어후보의 빈도수, 문맥정보, 의미정보 등과 같은 통계적 지식을 이용한 '통계적 필터링(Statistical Filtering)' 과정을 통해 수행된다(오종훈, 최기선, 2004). 새로운 전문용어의 추출은 전문용어 정비를 위한 기초자료로 활용될 수 있다. 추출된 전문용어의 언어학적 연구를 병행하게 되면 해당 전문분야에 대한 언어공학적 처리를 위한 연구에 활용될 수 있으며, 용어 데이터베이스 구축을 위한 연구에도 활용될 수 있다.

2.2 정의문 기술형식

용어의 정의문은 단어의 개념을 정확하게 표현하면서 단어의 개념을 간결하게 기술하고, 단어에 대한 핵심적인 정보를 포함하고 있다(최선화, 2006). 이러한 정의문을 형식적으로 규정하는 일은 쉬운 일은 아니다. 정의문 유형의 형식적 기준에 대한 논의는 Trimble(1985)과 Flowerdew(1992)가 진행한 바 있으며, 이를 바탕으로 국내에서는 남길임(2016)이 과학

〈표 1〉 전문용어의 개념 및 범위

구분	개념	사용영역(주체)
학술적 전문용어	학술전문가 집단에게만 사용되는 전문용어로 해당 집단 내의 의사소통에 사용됨을 목적으로 하는 용어	학술적 전문용어
공공언어로서의 전문용어	학술적 전문용어에서 출발하였지만 일반 언중에게 고시되거나 사용될 가능성이 있는 전문용어로 대중성을 필요로 하는 전문용어	행정용 전문용어 언론용 전문용어 교육용 전문용어 산업용 전문용어

(출처: 강현화, 2009)

텍스트를 대상으로 정의문을 Trimble(1985)과 Flowerdew(1992)가 제시한 형식적, 기능적 유형을 참고하여 정의문의 유형과 뜻풀이 패턴을 분석하였다. 남길임(2016)은 『표준국어대사전』의 〈물리〉 분야 전문용어 총 7,194개 표제어를 분석하였으며, 총 빈도 10 이상을 보이는 패턴 748개의 3-gram, 4-gram을 중심으로 각 용례를 확인하여 패턴의 기능을 분석하여 분류하였고, 그 결과는 〈표 2〉와 같다.

김재호 외(2004)는 정의문 기술방식을 "용어=def의미특성소+상위개념"으로 채택한 후, 표준의학사전의 용어 정의문 중 55,791개를 훈련 정의문으로 사용하여 정의문 패턴을 추출하였다. 그 결과는 〈표 3〉과 같다.

〈표 2〉와 〈표 3〉은 특정 분야의 전문용어사전에 등장한 정의문을 분석한 결과로서 기술문서에 일부 적용될 수 있으나 전체를 아우르기에는 한계가 있다. 기술문서의 경우 전 주제 영역을 다루기 때문에 정의문의 기술 형식 또한 보다 다양한 형태로 표현될 것으로 예측된다. 따라서 본 논문에서는 상기 나열한 패턴 외 기술문서에 등장하는 다양한 정의문의 패턴을 추출하고자 하였다.

〈표 2〉 전문용어 정의문의 형식적 특성에 의한 분류

구분	내용
공식적 정의패턴	• 유개념과 종차를 포함하는 뜻풀이 패턴 • ~를 띠는 입자, ~를 띠는 물체, ~를 재는 장치, ~을 나타내는 단위, 역학의 한 분야, 물리학의 한 분야, 소립자의 하나, ~의 한 유형, 이르는 말, ~를 이르다, 통틀어 이르는 말, -ㄴ 법칙이다, ~를 나타내다, -는 말이다, 아울러 이르는 말, ~라(고) 한다 등
준공식적 정의패턴	• 유개념과 종차 중 유개념이 생략된 채로 종차를 위주로 제시되는 패턴 • ~에 쓰다, ~에 이용하다, ~에 속하다, ~로 만들다, ~어서 만들다, ~로 이루어져 있다, ~로 되어있다, ~가 발견하다, ~로 나타내다, ~로 표시하다, ~가 제창하다, ~때에 생기는, ~에 일어난다, ~를 나타내다, ~를 이르다 등

(출처: 남길임, 2016)

〈표 3〉 훈련 정의문에서 추출한 상위개념의 구문적 패턴

규칙	패턴
0	~ N_{sup} (이다) $
1	~ N_{sup} 로(,) / ~ N_{sup} 으로(,)
	~ N_{sup} 로(,) / ~ N_{sup} 으로서(,)
2	~ N_{sup} 의 PART + J
3	~ N_{sup} 을 〔말하- | 총칭하 | 지칭하- | 가리키-〕
4	~ N_{sup} 나, / ~ N_{sup} 이나,
	~ N_{sup} 며, / ~ N_{sup} 이며(,)

• PART는 휴리스틱으로 추출한 '한형', '하나', '분야', '총칭', '일종' 등의 30개의 어휘
• J는 규칙 0, 1, 2, 3, 4에서 N_{sup} 다음에 기술된 조사 혹은 어미
(출처: 김재호 외, 2004)

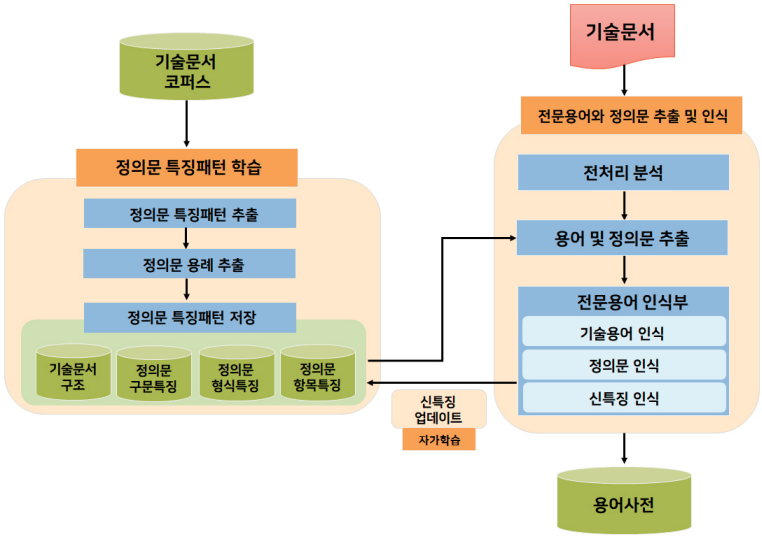
3. 기술문서 전문용어 및 정의문 자동추출

3.1 시스템 구조도

본 논문에서 제안하는 기술문서 전문용어 및 정의문 추출의 전체과정은 〈그림 1〉과 같다. 기술문서 전문용어 및 정의문 자동추출 시스템은 기술문서의 구조와 정의문의 구문적·형식적 규칙 및 패턴으로부터 정의문을 추출하고, 이로부터 전문용어를 자동으로 도출하여 전문용어 사전을 구축한다.

구체적으로 본 시스템은 기술문서로부터 정의문의 특징패턴을 학습하는 과정과 이를 기반으로 전문용어와 정의문을 추출 및 인식하는 과정으로 구성된다. 먼저 정의문의 특징패턴은 기술문서에서 정의문을 추출하는 기준으로서 크게 구문특징패턴, 형식특징패턴, 항목특징패턴 등으로 구분할 수 있다. 정의문을 추출하는 과정 중 하나 이상의 특징패턴에 대응하는 정의문이 추출되며, 기존에 저장되어 있지 않은 새로운 패턴(unseen patterns)이 발견될 경우 정의문 특징패턴에 새롭게 저장되어 학습된다. 이러한 방식으로 정의문 특징패턴의 지속적인 자가학습을 통해 패턴을 확장해감으로써 전문용어 및 정의문의 추출 범위와 정확도를 높일 수 있게 한다. 각 특징패턴에 대한 자세한 내용은 다음 장에서 설명하도록 하겠다.

정의문의 특징패턴을 기반으로 전문용어와 정의문을 추출 및 인식하는 과정에서는 우선 수집된 기술문서에 대한 다양한 언어처리 과정을 거쳐 구문분석을 수행하고, 구문분석 결과로부터 정의문을 추출한다. 앞서 설명했듯이 정의문은 기술문서 코퍼스에서 학습된 정의문 특징패턴을 통해 추출된다. 전문용어는 기본적으로 추출된 정의문의 주어에 해당하는 용어가 추출되며, 빈도수에 기반한 전문용어 추출방식이나 빈도수 및 명사구간의 내포관계에 기반한 전문



〈그림 1〉 기술문서 전문용어 및 정의문 자동추출 시스템

용어 추출, 사전계층관계에 기반한 분야 간 유사도와 통계기법을 이용한 전문용어 자동추출 기법 등 다양한 방식이 활용될 수 있다. 이들 방식을 활용하여 정의문으로부터 전문용어를 자동 추출하여 전문용어사전을 구축한다.

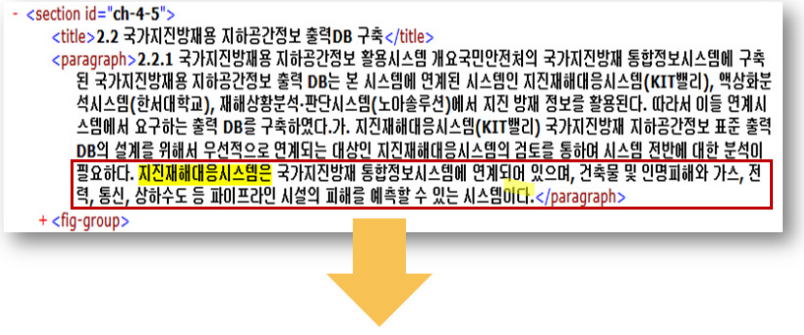
3.2 정의문 패턴 추출과정

3.2.1 정의문 특징패턴 추출

기술문서에 등장하는 전문용어의 정의문의 문장 특징패턴을 추출하기 위한 분석 대상 문서집합으로 먼저 연구보고서 총 28건을 살펴보았으며, 이들 문서의 XML 파일을 분석하여 정의문 패턴을 〈그림 2〉와 같이 추출하였다. 이와 같은 방식으로 기술문서에 등장하는 전문용어의 정의문 특징을 분석한 결과 〈표 4〉와 같이 구문특징패턴, 형식특징패턴, 항목특징패턴이 추출되었다.

구문특징패턴은 특정 개별 어휘(중심어)와 이와 공기하는 일련의 성분들을 규칙에 따라 형식화시킨 것으로 정의문 패턴 중 가장 많은 비중을 차지하고 있다. 본 논문에서 정의문을 추출한 결과 조사는 '은/는' 계열과 '라/란' 계열로 구분할 수 있었으며, 서술어는 완전한 정의문 형태인 '이다' 외에도 '~으로', '~말하-' 등과 같이 불완전한 정의문 형태를 보이는 경우도 많았다. 형식특징패턴은 개별 어휘(중심어)에 특정 기호를 사용하여 형식화시킨 것으로 본 논문에서는 전문용어의 앞 또는 뒤에 ': ', ' - '와 같은 기호를 붙여서 정의문임을 나타내는 형식특징패턴을 추출하였다. 항목특징패턴은 개별어휘(중심어)에 개념을 나타내는 특징용어를 사용하여 형식화시킨 것으로 본 논문에서 추출한 항목특징은 '용어(의) 정의'를 사용한 패턴이었다.

본 논문에서 추출한 정의문 특징패턴 중 가장 많은 비중을 차지하는 구문특징패턴을 구체적으로 분석한 결과 먼저 '은/는' 계열은 '~이다', '~으로(서/써)', ~'말하-', '~로(서/써)', '~ 정의/총칭/지칭/용어', '~이고/이며', '~을/를 의미하-', '~ㅁ', '~라(고) 할 수 있-', '~을/를 뜻하-'와 같은 비교적 다양한 서술어 패턴을 보이고 있었으며, '라/란' 계열은 '은/는' 계열보다 다소 적지만 거의 유사한 서술어 패턴을 보이고 있었다(〈표 5〉 참조).



〈그림 2〉 기술문서 정의문 패턴 추출과정

〈표 4〉 추출된 기술문서 정의문 패턴

구문특징패턴	형식특징패턴	항목특징패턴
N 은/는 ~이다.	N :	용어(의) 정의
N 은/는 ~으로(서/써)	- N	
N 은/는 ~말하-	* N :	
N 은/는 ~로(서/써)	N -	
N 은/는 ~ 정의/총칭/지칭/용어		
N 라/란 ~을/를 말하-		
N 은/는 ~이고/이며		
N 은/는 ~을/를 의미하-		
N 은/는 ~ㅁ		
N 은/는 ~라(고) 할 수 있-		
N 라/란 ~으로(서/써)		
~을/를 N 라 하-		
N 라/란 ~이다.		
N 라/란 ~용어		
N 은/는 ~을/를 뜻하-		
N 라/란 ~을/를 의미하-		
N 라/란 ~라(고) 하-		

〈표 5〉 기술문서 정의문 구문특징패턴 계열 비교

구문특징패턴
N 은/는 ~이다.
N 은/는 ~으로(서/써)
N 은/는 ~말하-
N 은/는 ~로(서/써)
N 은/는 ~ 정의/총칭/지칭/용어
N 라/란 ~을/를 말하-
N 은/는 ~이고/이며
N 은/는 ~을/를 의미하-
N 은/는 ~ㅁ
N 은/는 ~라(고) 할 수 있-
N 라/란 ~으로(서/써)
~을/를 N 라 하-
N 라/란 ~이다.
N 라/란 ~용어
N 은/는 ~을/를 뜻하-
N 라/란 ~을/를 의미하-
N 라/란 ~라(고) 하-

→

은/는 계열	라/란 계열
N 은/는 ~이다.	N 라/란 ~이다.
N 은/는 ~으로(서/써)	N 라/란 ~으로(서/써)
N 은/는 ~말하-	N 라/란 ~을/를 말하-
N 은/는 ~로(서/써)	
N 은/는 ~ 정의/총칭/지칭/용어	N 라/란 ~용어
N 은/는 ~이고/이며	
N 은/는 ~을/를 의미하-	N 라/란 ~을/를 의미하-
N 은/는 ~ㅁ	
N 은/는 ~라(고) 할 수 있-	N 라/란 ~라(고) 하-
N 은/는 ~을/를 뜻하-	
~을/를 N 라 하-	

3.2.2 정의문 용례 추출

정의문 특질패턴 분석내용을 기반으로 기술문서의 정의문 용례와 전문용어를 추출해보았다. 앞서 언급된 정의문 패턴 추출 방식과 동일한 방식으로 기술문서 64건으로부터 정의문 용례 251문장을 추출하였다. 전문용어는 추출된 정의문 패턴 용례로부터 〈그림 2〉와 같은 방식으로 도출하였으며, 이들을 구문특징패턴과 형식특징패턴, 항목특징패턴으로 분류하였다(〈표 6〉~〈표 8〉 참조). 예를 들어, "지진재해대응시스템은 국가지진방재 통합정보시스템에 연계되어 있으며, 건축물 및 인명피해와 가스, 전력,

〈표 6〉 기술문서 정의문 패턴 용례: 구문특징패턴

구문특징패턴	용례
N 은/는 ~이다.	지진재해대응시스템은 국가지진방재 통합정보시스템에 연계되어 있으며, 건축물 및 인명피해와 가스, 전력, 통신, 상하수도 등 파이프라인 시설의 피해를 예측할 수 있는 시스템이다.
N 은/는 ~으로(서/써)	기성테라조는 인조석의 일종으로 화강암과 같은 돌 조각에 백색시멘트를 넣어 견고하게 굳힌 후 미려한 무늬와 광택이 나도록 연마해서 제작된 것으로
N 은/는 ~말하-	에너지 Harvesting은 빛, 풍력, 열전, 진동, 태양열 등 주변에서 미 활용되고 있는 에너지를 전기에너지로 전환 시키는 기술을 말함
N 은/는 ~로(서/써)	페트로코크스 탈황석고는 노내탈황을 하는 유동층 보일러를 운전하는 발전소에서 페트로 코크스+석회석을 혼합 연소하고 남은 산업부산물로
N 은/는 ~ 정의/총칭/지칭/용어	PV 시스템(Photovoltaic system)으로 불리는 태양광 발전 시스템은 태양광을 전기에너지로 변환하는 장치 전반을 총칭한다.
N 라/란 ~을/를 말하-	"국가지진방재 통합정보시스템"이란 국민안전처장이 지진재해대응체계의 구축 등 국가지진방재체계의 확립에 활용하기 위해 각종 국가지진방재 관련 정보를 통합하여 수집·가공·분석 등을 수행하는 시스템을 말한다.
N 은/는 ~이고/이며	폐석회(Waste Lime)는 소다공장을 비롯한 화학공장, 폐수처리장, 제철소 등에서 부산물로 발생되는 무기성슬러지로서pH가 약 12.0~12.2의 알칼리성 물질이며
N 은/는 ~을/를 의미하-	"급경사지정보"는 국민안전처 관리 대상의 급경사지의 상태와 급경사지 주변의 지질, 지반, 지하수 조건을 파악할 수 있는 정보를 의미하며
N 은/는 ~ㅁ	웜홀 파이프라인 공정 관리 시스템은 3차원 그래픽 디지털자산 제작 공정을 관리하고 웹 환경에서 국내외 업체 간 공유 및 협업 가능하도록 개발된 공정 관리 시스템임
N 은/는 ~라(고) 할 수 있-	냉방 부하는 냉방 기준 온도와 외기와 의 온도 차이에 의하여 구조체 및 환기에 의하여 획득열과 일사에 의한 획득열, 인체 발생열, 조명 기구 발생열, 실내 기기 발생열과 같은 내부 발생열의 총합으로서 냉방 기준 온도를 유지하기 위하여 제거해야 할 에너지양이라 할 수 있다.
N 라/란 ~으로(서/써)	GeoHash코드란 포인트/폴리곤을 32진수의 고유한 아이디 체계로 표현하는 시스템으로써
~을/를 N 라 하-	대기층과 구름에서 확산, 투과, 반사되어 지표면에 도달하는 빛을 천공광이라 한다.
N 라/란 ~이다.	ABC 수송체란, ATP 결합상자 수송체라는 단백질 슈퍼패밀리의 한 구성요소로 원핵생물부터 인간에 이르기 까지 널리 퍼져있는 가장 오래된 패밀리이다.
N 라/란 ~용어	사전적으로 외피란 '내부 볼륨을 감싸는 건물의 외부 부분'을 일컫는 용어이다.
N 은/는 ~을/를 뜻하-	디지털 정보는 콘텐트를 가리키며, 더 구체적으로 디지털 콘텐트를 뜻함.
N 라/란 ~을/를 의미하-	실감미디어(immersive media)란 가상의 환경에서 공간과 시간의 제약을 극복하면서 실재감(presence)과 몰입감(immersion)을 제공할 수 있는 다양한 형태의 요소 미디어 정보들의 통합된 표현(representation)을 의미한다
N 라/란 ~라(고) 하-	가시광이란 전자파 중에서 인간의 눈으로 검지할 수 있는 파장 영역을 가시광 혹은 가시광선이라고 하며

〈표 7〉 기술문서 정의문 패턴 용례: 형식특징패턴

형식특징패턴	용례
N :	Break-Wire system (단선감지시스템) : 스크린이나 +자 형태로 배열된 전선을 이용하여 단선에 의해 작동하는 시스템
- N	- 화이트노이즈(White Noise) 모든 소리를 혼합하면 주파수, 진폭, 위상이 균일하게 끊임없이 변하는 완전 랜덤파형을 형성하며 이를 화이트 노이즈라 한다.
* N :	* K-factor : Harmonic loss factor 라고도 하며, 국제규격 IEEE, IEC에서 고조파 전류가 변압기에 유입될 때 변압기의 증가하는 손실을 반영하기 위해서 eddy-current losses 및 stray losses에 곱해지는 factor임.
N -	① 소음통계레벨(LN, Percentile Noise Level) - 전체 측정기간 중 그 소음레벨을 초과하는 시간의 총합이 N%가 되는 소음레벨을 말하며,

〈표 8〉 기술문서 정의문 패턴 용례: 항목특징패턴

항목특징패턴	용례
용어(의) 정의	참고: 용어의 정의 o 탄소중립(Carbon Neutral)은 CO2 발생 저감 및 발생한 CO2의 흡수, 전환, 해소를 통하여 CO2 발생효과가 '0'인 상태 o 탄소중립 도로(Carbon Neutralized Road)는 도로의 계획/설계/시공/운영/유지관리 등 전생애주기 동안 CO2 발생을 최소화하고, 발생한 CO2를 흡수, 전환, 해소하여 궁극적으로 CO2 발생효과가 '0' 상태인 도로

통신, 상하수도 등 파이프라인 시설의 피해를 예측할 수 있는 시스템이다."라는 문장은 'N 은/는 ~이다.'라는 구문특징패턴으로 분류되며, "지진재해대응시스템"이란 전문용어가 추출된다. 또한 "Break-Wire system(단선감지시스템) : 스크린이나 +자 형태로 배열된 전선을 이용하여 단선에 의해 작동하는 시스템"이라는 문장은 'N : '라는 형식특징패턴으로 분류되며, "Break-Wire system(단선감지시스템)"이란 전문용어가 추출된다. 이와 같은 방식으로 분류한 결과와 그 용례는 〈표 6〉, 〈표 7〉, 〈표 8〉과 같다.

기술문서에 등장한 전문용어 정의문의 구문특징패턴은 '피정의항(종개념) = 정의항(종차 + 유개념)'과 같은 정의문의 전형적인 형식으로 구성되어 있는 경우도 있지만, 완전한 정의문으로 구성되어 있지 않은 경우도 많다. 일반적으로 '유개념'은 종개념을 포함하는 상위개념이며, '종차'는 (같은 유개념에 속한) 종개념이 다른 종개념들과 변별되는 차이점을 의미한다. 이와 관련하여 본 논문에서는 Trimble(1985), Flowerdew(1992)가 구분한 정의에 의거하여 정의문을 형식적 완결성에 따라 공식적인 정의, 준공식적 정의, 비공식적 정의로 구분하였다.

예컨대, "지진재해대응시스템은 국가지진방재 통합정보시스템에 연계되어 있으며, 건축물 및 인명피해와 가스, 전력, 통신, 상하수도 등 파이프라인 시설의 피해를 예측할 수 있는 시스템이다."라는 용례에서 '지진재해대응시스템'은 피정의항(종개념)이며, '시스템'은 지진재해대응시스템의 상위개념으로서 유개념이라 할 수 있다. 그리고 '국가지진방재 통합정보시스템에

연계되어 있으며, 건축물 및 인명피해와 가스, 전력, 통신, 상하수도 등 파이프라인 시설의 피해를 예측할 수 있는'은 지진재해대응시스템이 다른 시스템과 변별되는 점을 설명하는 종차라 할 수 있다. 즉, 위의 예제는 '피정의항, 유개념, 종차'가 모두 포함된 전형적인 정의로서 공식적 정의에 해당한다.

이에 반해 준공식적 정의의 경우 유개념이 빠진 두 가지 요소(피정의항, 종차)만을 포함하는 정의이다. 예를 들어 "EMT는 양성종양이 악성 종양으로 변화하는데 필수적인 단계임"에서 'EMT'는 피정의항이며, '양성종양이 악성 종양으로 변화하는데 필수적인 단계임'은 EMT를 설명하는 종차의 일종이라 할 수 있으므로 준공식적 정의라 할 수 있다. 마지막으로 비공식적 정의는 "디지털 정보는 콘텐트를 가리키며, 더 구체적으로 디지털 콘텐트를 뜻함."에서와 같이 피정의항의 대치 가능한 명칭을 보여주는 경우이다.

기술문서의 전문용어는 특정 주제 영역에서만 한정적으로 사용되는 전문용어이거나, 일반적인 의미가 아닌 해당 영역에서 특정한 의미로 사용되는 경우도 많다. 또한 최신용어인 경우 전형적인 정의문의 형식 보다는 의미를 설명하는 방식으로 정의문을 표현하기도 한다. 따라서 기술문서에 등장하는 정의문의 유형[1]인 공식적 정의, 준공식적 정의, 비공식적 정의를 모두 고려하여 추출할 필요가 있다. 실제로 기술문서에서 추출한 전문용어 정의문의 유형의 대표적인 용례는 〈표 9〉와 같다.

3.2.3 정의문 추출 및 순위화

본 논문에서 제안한 방법을 통해 분석한 결과를 정리해보면, 기술문서 총 64건 중 정의문이 등장한 기술문서는 42건(65%)이며, 이로부터 정의문 251개, 전문용어 253개(중복 제거)가 추출되었다. 이들의 정의문 패턴의 출현빈도를 비교 분석한 결과는 〈표 10〉과 같다. 정의문의 패턴 중 '항목특징' 패턴은 하나의 정의문에 여러 개의 전문용어가 추출되었기 때문에 추출된 정

〈표 9〉 기술문서 정의문의 유형과 용례

유형	용례
공식적 정의	**지진재해대응시스템**은 국가지진방재 통합정보시스템에 연계되어 있으며, 건축물 및 인명피해와 가스, 전력, 통신, 상하수도 등 파이프라인 시설의 피해를 예측할 수 있는 시스템이다.
준공식적 정의	**EMT**는 양성종양이 악성 종양으로 변화하는데 필수적인 단계임
비공식적 정의	**디지털 정보**는 콘텐트를 가리키며, 더 구체적으로 디지털 콘텐트를 뜻함.

1) 공식적 정의는 '피정의항, 유개념, 종차'의 세 가지 요소를 모두 포함하는 전형적인 정의이며, 준공식적 정의는 이 세 가지 요소 중 유개념이 빠진 두 가지 요소만을 포함하는 정의이다. 그리고 비공식적 정의는 피정의항과 더불어 대치 가능한 명칭이나 두드러진 특징을 보여주는 단어를 정의항에 제시하는 정의이다. 구체적인 용례는 아래와 같다(남길임, 2016, p. 120).
〈정의문 용례〉
- 공식적 정의: 이와 같이 두 물체의 접촉면에서 미끄러짐을 방해하는 힘을 마찰력이라고 한다.
- 준공식적 정의: 마찰력과 탄성력은 두 물체가 접촉할 때 작용한다.
- 비공식적 정의: 블랙홀은 현대물리학의 주요 연구 주제 중의 하나이다.

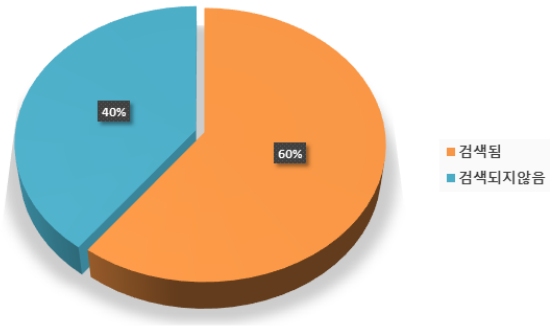
〈표 10〉 기술문서 정의문 패턴 순위화

Ranking	구문특징패턴	출현빈도	형식특징패턴	출현빈도	항목특징패턴	출현빈도
1	N 은/는 ~이다.	64	N :	30	용어(의) 정의	3
2	N 은/는 ~ㅁ	30	- N	6		
3	N 은/는 ~로(서/써)	29	* N :	6		
4	N 은/는 ~으로(서/써)	14	N -	1		
5	N 은/는 ~말하-	12				
6	N 은/는 ~이고/이며	12				
7	N 은/는 ~을/를 의미하-	10				
8	N 은/는 ~ 정의/총칭/지칭/용어	8				
9	N 라/란 ~을/를 말하-	7				
10	~을/를 N 라 하-	6				
11	N 은/는 ~라(고) 할 수 있-	4				
12	N 라/란 ~으로(서/써)	3				
13	N 라/란 ~이다.	2				
14	N 라/란 ~용어	1				
15	N 은/는 ~을/를 뜻하-	1				
16	N 라/란 ~을/를 의미하-	1				
17	N 라/란 ~라(고) 하-	1				
합계		205		43		3

의문의 수와 전문용어의 수 사이에 약간의 차이가 발생하였다(〈표 8〉 참조). 먼저 정의문의 특징패턴 유형으로 보면 전문용어 정의문 총 251개 중 구문특징패턴이 205개로 가장 많은 비중을 차지하고 있으며, 이 중 완전한 형태의 정의문인 'N 은/는 ~ 이다.'가 총 64건으로 가장 높은 빈도를 보이고 있다. 그 다음으로 'N 은/는 ~ㅁ', 'N 은/는 ~로(서/써)', 'N 은/는 ~으로(서/써)' 순으로 나타났다. 그리고 정의문 구문특징패턴의 조사는 '은/는' 계열이 '라/란' 계열보다 압도적으로 높게 나타났다. 한편, 전체적으로 비교하면 구문특징패턴인 'N 은/는 ~ 이다.'가 가장 많은 비중을 차지하고 있지만, 그 다음으로 형식특징패턴인 'N :'가 30건으로 다른 구문특징패턴에 비해 상당히 높은 비중을 차지하고 있음을 알 수 있다. 반면 항목특징패턴은 다른 패턴에 비해 빈도수가 낮은 편이다.

3.3 추출된 전문용어사전의 효과

최신 전문용어에 대한 정의를 시의적절하게 추출하여 제공하는 것이 본 논문에서 제안한 전문용어사전 구축 방법의 목표이다. 이러한 효과를 파악하기 위해 앞서 기술문서의 용례로부터 도출된 정의문 251개에서 추출된 253개(중복제거)의 전문용어가 온라인에 검색되는지 살펴보았다(〈그림 3〉 참조). 특히 기존의 일반사전에서 기술문서의 전문용어가 검색되는지를 확인하기 위해 국내 최대 포털인 네이버에서 제공하는 지식백과(terms.naver.com) 사이트에서 추출된 전문용어를 검색한 결과, 총 253개 중 151개(약 60%)가 검색되었고, 나머지 102개(약 40%)는 검색되지 않았다(〈표 11〉 참조). 검색 시 지식백과에서 해당 용어가 검색되더라도 추출된 용어에 대응하는 정의 설명문이 아닌



〈그림 3〉 기술문서 전문용어 온라인 검색결과

〈표 11〉 추출된 전문용어의 검색결과 일부 예시

(2017년 10월 기준)

검색된 전문용어	검색되지 않은 전문용어
지질지반조사, 감리, 폐석회, 증기양생, 심층혼합처리공법, DES(Data Encryption Standard), Soil cement, SEED 알고리즘, 지오 펜스, 탄소중립(Carbon Neutral), 녹색도로(Green Highway), 입력기/입력 방식 편집기(input method editor, IME), 하스스톤, HMM, RoomAlive, Parametric Surface, 카툰 렌더링, GeoHash 코드, 폴리올, 레이저 다이오드(LD), 아두이노(Arduino), 공간형 콘텐츠, 천공광 등	액상화분석시스템, KISCON, 지반고화, 페트로코크스, 페트롤 코크스 탈황석고, 기성테라조, 표층고화처리공법, 사면정재용 결합재, 탄소중립 도로(Carbon Neutralized Road), 원홀 파이프라인 공정 관리 시스템, 마이마스터즈(MyMasters), Geo-Calender, 제본용 PUR 핫멜트 접착제, DMA, TPS, 앵커리지(Anchorage), 콜체미드, Autotaxin(ATX), 공칭케이블강도(Nominal Cable Strength) 등

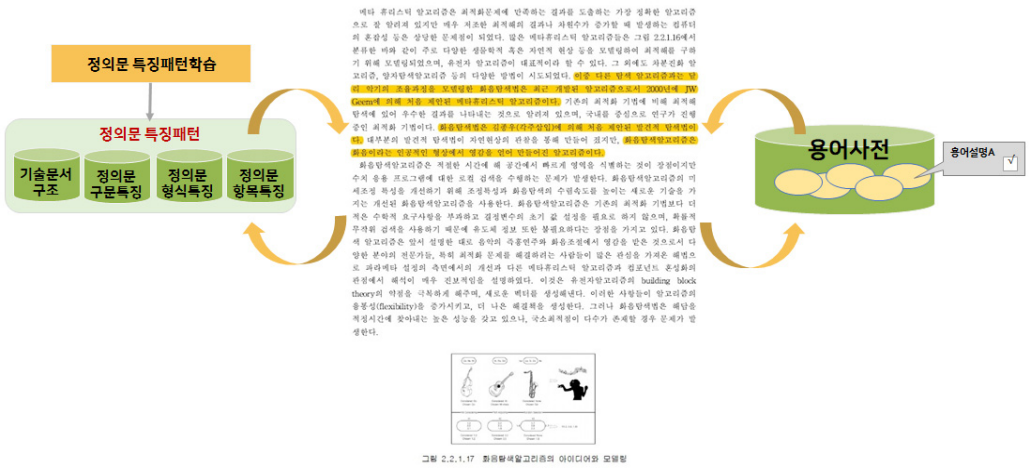
경우 실패한 것으로 평가하였다. 이를 통해 고도의 전문지식을 다루는 기술문서에 나타난 전문용어가 일반사전에서 의미 검색이 불가능한 경우가 있음을 확인하였다. 이는 기술문서에서 발견되는 전문용어를 추출 및 인식하여 사전을 구축해야 하는 타당성을 부여한다.

4. 활용방안

4.1 자가학습을 통한 전문용어사전 구축

본 논문에서는 기술문서 구조와 구문적・형식적 규칙 및 패턴으로부터 정의문들을 도출하고, 이를 기반으로 전문용어를 자동 추출하여 사전을 구축할 수 있는 전문용어 및 정의문 자동추출 시스템을 제안하였다. 기술문서에 나타난 특징 패턴을 분석하여 핵심 전문용어를 추출하여 사전을 구축한 후, 이를 토대로 전문용어에 대한 정의 및 설명을 이용자에게 제공한다면, 이용자들은 보다 풍부한 전문지식과 정보들을 탐색할 수 있다. 또한 전문용어사전의 구축을 통해 일반사전에 등장하는 용어들과 차별화되는 전문용어의 특징을 밝힘으로써 전문가들이 자신의 활동영역 안에서 보다 분명하고 기능적인 의사소통을 하는 데 기여할 수 있다.

〈그림 4〉는 자가학습을 통해 전문용어사전을 구축하는 모습을 도식화 한 그림이다. 기존 용어사전에 있는 전문용어 정의문과 관련하여 새로운 패턴이 인식되는 경우, 이를 추출하여 저장



〈그림 4〉 자가학습을 통한 전문용어사전 구축

함으로써 전문용어사전을 확장해나갈 수 있다. 이와 같은 자가학습 기반 통한 전문용어사전 구축은 기술문서를 보다 다양하고 폭넓게 활용하는 것을 가능하게 한다. 먼저, 전문용어사전은 개체명 인식, 신조어 분석 등 고차원 언어분석을 위한 코퍼스로서 활용할 수 있다. 이를 통해 기술문서 검색의 복잡도를 감소시키고, 검색의 정확도 및 분석 속도를 향상시키는 등 기술문서의 활용도를 높일 수 있다. 또한 구축된 전문용어사전은 개인화서비스를 제공하기 위한 지식메모리로서 활용이 가능하다.

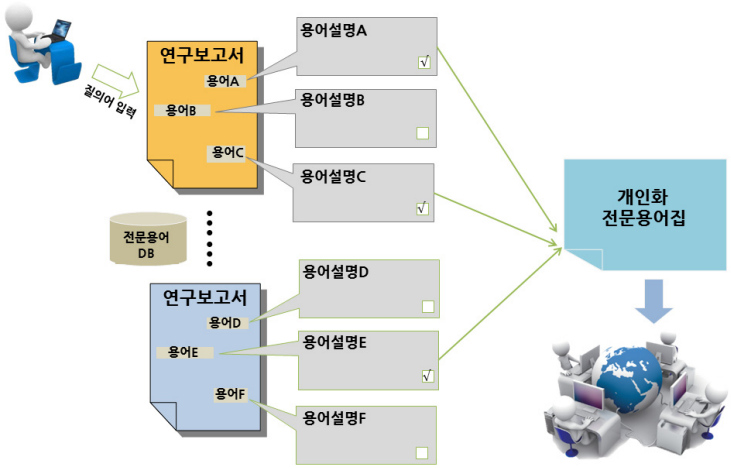
4.2 전문용어사전에 기반한 개인화서비스 제공방안

전문용어사전을 지식메모리로서 보다 다양하게 활용하기 위해서는 개별 이용자들의 선호에 보다 잘 부합되는 결과물을 제공함으로써 서비스의 가치를 증대시킬 수 있는 개인화서비스를 제공할 필요가 있다. 급증하고 있는 정보와 다양한 서비스들 사이에서 이용자들은 자신이 원하는 정보를 검색하는 것이 점점 더 어려워지고 있다. 따라서 이용자의 직접적인 정보요구가 없을지라도 관심이 있을만한 정보를 적극적으로 제공하여 이용자의 잠재적 욕구를 충족시킬 수 있는 개인화서비스를 개발할 필요가 있다. 이와 관련하여 본 연구에서 제안하는 방안은 다음과 같다.

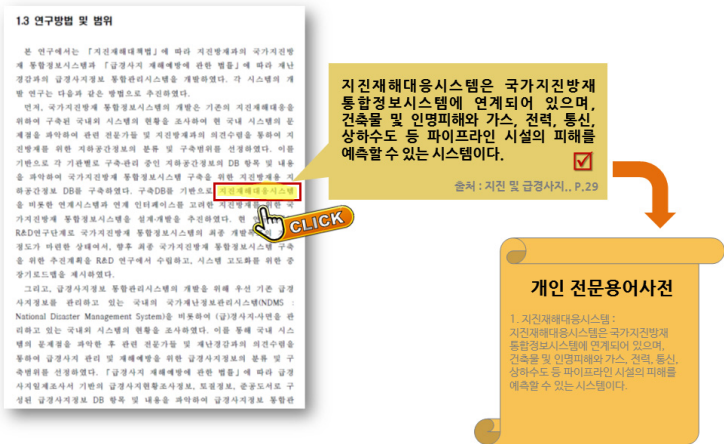
먼저, 〈그림 4〉와 같이 자가학습을 통해 자동적으로 구축된 전문용어사전을 활용 및 공유할 수 있는 대규모 지식메모리로 구축한 다음, 이를 기반으로 기 구축된 지식메모리로부터 입력된 질의어에 상응하는 전문용어를 매핑하여 다양한 기술문서에 담긴 지식정보를 복합적이고 심층적으로 이용자에게 제공하도록 한다. 기존에는 이용자가 원하는 정보가 담긴 기술문서를 개별적으로 검색하여 찾은 후, 각각의 기술문서에 질의어를 하나하나 입력하여 검색해야 했기 때문에 시간과 노력이 많이 소요되었다. 그러나 본 논문에서 제안하는 지식메모리에는 이

미 기술문서의 전문용어와 정의문들이 추출되어 저장되어 있기 때문에 이용자가 질의어를 입력하면 해당 전문용어들과 정의문을 하나의 문서로 통합하여 개인화 전문용어집의 형태로 이용자에게 제공할 수 있게 된다. 따라서 이용자가 질의어를 입력하면 관련 내용만 추출하여 통합적으로 이용자에게 제공할 수 있기 때문에 이용자의 검색효율성과 편의성을 높일 수 있다(〈그림 5〉 참조).

둘째, 이용자가 전문용어에 대한 정의뿐만 아니라 관련 정보들을 보다 손쉽게 찾을 수 있도록 용어사전에 있는 용어들은 기술문서에 하이라이팅(highlighting)으로 표시하여 이용자에게 제공하도록 한다. 즉, 자가학습을 통해 구축된 용어사전에 있는 용어를 〈그림 6〉과 같이 기술문서 내에 하이라이팅으로 표시한 다음, 이용



〈그림 5〉 전문용어 자동추출기반 개인화서비스



〈그림 6〉 전문용어 자동추출기반 개인화서비스 예시

자가 전문용어를 클릭하면 해당 정의문을 쉽게 볼 수 있도록 서비스를 제공하도록 한다. 나아가 이용자가 원하는 정보일 경우 해당 정의문을 체크하면 개인 전문용어사전에 자동으로 저장되게 함으로써 반복적인 검색을 피하고, 자신만의 개인 전문용어사전을 구축하여 활용할 수 있도록 지원하는 방안을 모색해볼 수 있다.

셋째, 검색엔진을 이용하여 핵심 전문용어가 포함된 기술문서가 자동으로 검색되어 높은 가중치의 기술문서를 우선적으로 이용자에게 제공하는 한편, 이용자의 검색의도를 분석하여 다른 이용자들이 정제한 전문용어집을 추천하여 공유 및 제공한다면 이용자 만족도를 더욱 향상시킬 수 있을 것으로 본다. 이는 이용자가 미처 파악하지 못한 정보요구를 스스로 인식할 수 있게 할 뿐만 아니라, 기술문서에 대한 심도 깊은 이해를 바탕으로 본인의 정보요구에 적합한 기술문서 탐색을 보다 용이하게 하는데 도움을 준다.

5. 결론 및 향후 과제

IT 산업의 발전으로 우리 사회는 수많은 지식정보가 넘쳐나고 있으며, 전문성이 요구되는 정보의 양 또한 폭발적으로 증가해 일반인이 이해하기 어려운 수준까지 도달했다. 특히 기술문서는 빠른 과학기술의 변화를 반영한 연구 성과물로서 전문지식의 수준과 그 양 또한 방대하다. 따라서 관련 전문지식에 대한 개인의 이해도에 따라 기술문서의 이용만족도 및 활용도의 차이가 발생할 수밖에 없다. 따라서 기술문서가 제대로 활용되기 위해서는 기존의 정보검색이나 정보 분류를 뛰어넘어 정보에 대한 요약 및 핵심 정보추출과 같은 좀 더 세밀한 정보의 가공을 통해 이용자의 기술문서 활용의 편의를 높여줄 필요가 있다.

이에 본 연구에서는 기술문서의 특징을 기반으로 전문용어 및 정의문을 자동 추출하고, 이를 기반으로 전문용어사전을 구축할 수 있는 시스템을 제안하였다. 실제 기술문서 중 연구보고서에 등장하는 전문용어와 정의문을 분석하여 패턴을 추출하였으며, 이를 기반으로 전문용어 자동추출기반의 개인화서비스를 제공방안에 대해 제안하였다. 전문용어 및 정의문 자동추출을 기반으로 전문용어사전을 구축하게 되면 새롭게 등장하는 전문용어를 빠르게 수용할 수 있어 이용자들이 최신정보를 보다 빠르고 손쉽게 찾을 수 있다. 나아가 개인화된 전문용어사전을 이용자에게 제공한다면 전문용어사전의 가치와 활용성, 검색의 효율성을 극대화할 수 있다.

한편, 본 논문은 기술문서에 등장하는 정의문의 패턴을 수작업으로 추출하였기 때문에 정의문 패턴의 질적 정확성은 높였으나, 그 수가 적다는 한계점이 있다. 따라서 자가학습을 통한 전문용어사전 구축을 위해 전문용어 정의문 패턴을 더 많이 확보할 필요가 있으며, 이에 대한 후속 연구가 필요하다.

참 고 문 헌

강현화 (2009). 전문용어 표준화 제도 장비를 위한 정책 연구. 서울: 국립국어원.

김재호, 배선미, 신효식, 최기선 (2004). 의학 전문용어의 정의문 자동 추출. 한국정보과학회 학술발표논문집, 31(1B), 922-924.

남길임 (2016). 과학텍스트 정의문의 유형분석. 한국어 의미학, 52, 111-138. https://doi.org/10.19033/sks.2016.06.52.111

미래창조과학부, 한국과학기술기획평가원 (2017). 2016 과학기술통계백서. 서울: 휴먼컬처아리랑.

박정오, 황도삼 (2000). 전문용어 추출시스템. 한국정보과학회 학술발표논문집, 27(1B), 381-383.

신효식, 김재호, 이해윤, 최기선 (2002). 텍스트로부터 용어 정의문의 자동 추출 방법. 한국정보과학회 언어공학연구회 학술발표 논문집, 292-299.

오종훈, 김재호, 최기선 (2003). EM 알고리즘을 이용한 전문용어의 자동추출. 한국정보과학회 학술발표논문집, 30(2), 487-489.

오종훈, 이경순, 최기선 (2002). 분야간 유사도와 통계기법을 이용한 전문용어의 자동추출. 정보과학회 논문지, 29(4), 258-269.

오종훈, 최기선 (2004). 정보통합을 통한 생물/의학 분야 전문용어의 자동 추출. 한국정보과학회 학술발표논문집, 31(2), 775-777.

최선화 (2006). 사전 정의문의 구문특징패턴에 기반한 상위어 판별규칙 학습. 박사학위논문, 전남대학교 대학원, 전산학과.

한국과학기술정보연구원 (2014). 국가R&D 보고서원문 성과 활용 분석 및 경제적 기여도 분석 연구보고서. 서울: 한국과학기술정보연구원.

한국학술단체총연합회 (2006). 전문용어정리방법론 개발 연구보고서. 서울: 한국학술단체총연합회.

Flowerdew, J. (1992). Definitions in science lectures. Applied Linguistics, 13(2), 202-221. https://doi.org/10.1093/applin/13.2.202

Trimble, L. (1985). English for science and technology: A discourse approach. Cambridge: Cambridge University Press.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Choi, Seon-Hwa (2006). Learning of hypernym identification rules based on syntactic patterns in definition sentences of dictionaries. Ph.D. dissertation, Graduate School of Chonnam

National University, Department of Computer Science.

Kang, Hyeon-Hwa (2009). Policy study for standardization of terminology standardization system. National Institute of Korean Language. Seoul: National Institute of the Korean Language.

Kim, Jae-Ho, Bae, Sun-Mee, Shin, Hyo-Shik, & Choi, Key-Sun (2004). Automatic extraction of medical term definition from texts. Proceedings of the Korea Information Science Society, 31(1B), 922-924.

Korea Institute of Science and Technology Information (2014). Outcome measurement and degree of economic contribution for the national R&D reports. Seoul: Korea Institute of Science and Technology Information.

Korean Association of Academic Societies (2006). Development of terminology methodology. Seoul: Korean Association of Academic Societies.

Ministry of Science and ICT, & Future Planning & Korea Institute of S&T Evaluation and Planning (2017). 2016 White paper of science and technology statistics. Seoul: Human, Culture, Arirang.

Nam, Kil-Im (2016). A study on types of defining sentences in science text. Korean Semantics, 52, 111-138. https://doi.org/10.19033/sks.2016.06.52.111

Oh, Jong-Hoon, & Choi, Key-Sun (2004). Recognizing biomedical terminologies through integration of heterogeneous information. Proceedings of the Korea Information Science Society, 31(2), 775-777.

Oh, Jong-Hoon, Kim, Jae-Ho, & Choi, Key-Sun (2003). Automatic term recognition through EM algorithm. Proceedings of the Korea Information Science Society, 30(2), 487-489.

Oh, Jong-Hoon, Lee, Kyung-Soon, & Choi, Key-Sun (2002). Automatic term recognition using domain similarity and statistical methods. Korea Information Science Society, 29(4), 258-269.

Park, Jung-Oh, & Hwang, Do-Sam (2000). A terminology extraction system. Proceedings of the Korea Information Science Society, 27(1B), 381-383.

Shin, Hyo-Shik, Kim, Jae-Ho, Lee, Hae-Yun, & Choi, Key-Sun (2002). A method for automatic extraction of term definition from text. Proceedings of the Korea Information Science Society Language Engineering Research Society, 292-299.