Korean Journal of Sport Science
2020, Vol. 31. No. 4, 650-661
https://doi.org/10.24985/kjss.2020.31.4.650

체육과학연구

The study of ontology model for soccer player's social media contents analysis

Joo-Hak Kim, Sun-Mi Cho*, Ji-Yeon Kang

Myongji Univ.

[Purpose] Soccer-related social media BigData includes complex information related to soccer players and is continuously and instantly generated. Text mining research is actively carried out for this kind of social media contents analysis, but it tends to be analyzed with limited linguistic characteristics such as understanding of language complexity and context, ambiguous terms, rhetoric, and new terms. This can be attributed to the fact that the tools commonly used for text mining use universal terminology dictionaries and packages that exclude the peculiarities of the analysis themes. The purpose of this study is to develop an Ontology model, which are representative tools for defining semantic ambiguity and relationships and systems between terms of text data. **[Methods]** In order to achieve the research objectives, we applied the 7-step development method of 'Ontology Development 101: A Guide to Creating Your First Ontology', which is useful for ontology development. Each step includes 1) Determine the domain and scope of the ontology 2) Consider reusing existing ontology 3) Enumerate important terms in the ontology 4) Define the classes and the class hierarchy 5) Define the properties of classes-slots 6) Define the facts of the slots 7) Create instances. In particular, the 3rd-step of this study, the glossary stage, is to extract core terms that make up he ontology, but since the goal of this study is to develop the ontology that can be used in social media contents analysis of soccer players, we conducted a social media text analysis related to actual soccer players and selected 484 core terms. **[Results]** The ontology which was developed in this research for social media contents analysis of soccer players consisted largely of four parts(General terms, performance results terms, common terms, and Characteristic term) and classified according to the content characteristics of the term. **[Conclusion]** Developed ontology in this study is object-oriented that defining classes and objects to define divisions and relationships between terms and also means a social media contents knowledge system of soccer players. In addition, it performs a function as a secondary tool which can be utilized for atypical data analysis.

Key words: ontology, BigData, text mining, object orientation, social media

서 론

최근 사회적 변화와 요구 속에서 빅데이터 분석 기술의 적용은 많은 분야에 걸쳐서 증가하고 있다(Choi, 2020). 3Vs로 대변되는 현재의 빅데이터 환경은 데이터의 생산과 처리속도, 다양성에 혁신을 가져오며 텍스트나

논문 투고일 : 2020. 08. 18.
논문 수정일 : 2020. 11. 11.
게재 확정일 : 2020. 12. 03.
* 교신저자 : 조선미(liff99@gmail.com).
* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A01025931).

이미지, 그림과 같은 비정형 데이터를 데이터 사이언스의 영역으로 견인하였다. 즉, 과거의 데이터 사이언스가 수량화, 계량화에 근거한 정량적 분석에 초점이 맞춰져 있었다면 최근에는 대용량의 비정형 데이터를 활용한 정성적 분석으로 연구 패러다임의 확장이 이루어지고 있다(Kim, Kang & Cho, 2020).

스포츠 영역에서의 빅데이터 연구는 발전 방향 및 활용 방안과 같은 개괄적인 연구에서(Cho, 2018; Park & Lee, 2017) 소셜 빅데이터 분석, 텍스트 마이닝 등과 같은 빅데이터 분석 기법을 활용한 연구(Choi & Yun-Soo, 2020; Lee & Kim, 2019; Yun, 2018)로 다각적인 연구가 진행되고 있다. 뿐만아니라 스포츠 산업에서 경기력 분석까지 분석 주제 범위의 점진적인 확대가 이루어지고 있다.

축구는 범세계적인 인기 스포츠로써 경기기록 이외에도 다수의 기록이 다양한 형태로 존재한다. 특히, IT 기술, 인터넷, 모바일 등의 발전으로 축구 정보의 접근성이 높아짐에 따라 축구 경기로 인한 빅데이터의 생산 속도와 양은 기하급수적으로 증가하고 있다. 이러한 데이터의 증가 경향은 2020년 한국프로스포츠협회가 발표한 '프로스포츠 x 소셜빅데이터' 보고서의 K리그 소셜빅데이터 증가율에서도 확인할 수 있다. 보고서에 따르면 소셜 미디어에서 K리그 관련어의 등장빈도가 2018년 대비 2019년 62.1% 증가하였음을 명시하고 있다.

빅데이터의 관점에서 축구 데이터는 경기기록, 기사, 블로그, SNS 등과 같은 소셜 미디어의 모든 비정형 데이터를 포함한다(Kim et al., 2020). 소셜 미디어 데이터는 축구 경기와 관련한 다각적 정보를 포함하며, 빅데이터 분석 환경에서 중요한 정보원의 가치를 가진다. 소셜 미디어의 비정형 데이터는 작성자의 의견, 느낌, 관심, 선호, 평판 등을 직·간접적으로 반영하고 있으며 다양한 의사결정의 도구로 활용되고 있다. 이러한 비정형 데이터를 분석하는 방법으로 자연어 처리 기반 텍스트 마이닝을 활용할 수 있다.

과학기술정보통신부의 국립중앙과학관은 자연어 처리 기반 텍스트 마이닝을 '언어학, 통계학, 기계 학습 등을 기반으로 한 자연어 처리 기술을 활용하여 반정형/비정형 텍스트 데이터를 정형화하고, 특징을 추출하기 위한 기술과 추출된 특징으로부터 의미 있는 정보를 발견할 수 있도록 하는 텍스트 마이닝 기술'이라고 정의하고 있다. 즉, 텍스트로부터 유의미한 정보를 추출하는 과정이 자연어 처리의 주요 기술이라 할 수 있으며, 자연어 처리 기반 텍스트 마이닝은 비정형 데이터를 정형 데이터와 같이 양적인 분석을 가능하게 하는 일련의 과정을 포함한다.

그러나 언어가 가지고 있는 복잡성과 문맥에 대한 이해, 중의어, 수사어, 신조어 등 언어적 특성으로 인해 텍스트 데이터의 완전한 자동 분석에 난점이 존재하는 경우가 있다. 즉, 빅데이터 분석 기법을 활용하여 텍스트 마이닝을 진행할 때, 시소러스, 형태소분석기, 텍사노미 등의 기본적인 언어 사전으로 단순한 의미 분석과 학습은 가능하지만, 단어 간 관계 또는 특정 주제의 영역에서의 사회·문화에 따른 구분을 모두 포함하기에는 한계(Kim & Jeon, 2019)가 있다는 것이다. 특히, 스포츠와 같이 특정 영역의 소셜 미디어가 연속적으로 빠르게 생산되는 환경에서는 복합적 차원의 소셜 미디어의 텍스트를 명확히 해석하기에 다소 제한적일 수밖에 없다.

스포츠 분야의 텍스트 마이닝의 특성에 따른 난점에 대한 논의는 선행연구에서도 확인할 수 있는데, Oh, Han & Kim(2019)은 수상 레저스포츠의 소셜 빅데이터 분석에서 수집 키워드의 정제와 특정 이슈와 트렌드의 반영에 대한 제한점을 논의한 바 있다. 또한, Lee & Kim(2019)은 텍스트 마이닝, TF-IDF, 의미연결망 등의 빅데이터 분석 기법을 활용하여 스포테인먼트 마케팅의 현황과 스포츠 키워드를 분석하여 마케팅의 방향과 전략을 제시하는 연구를 진행하였는데, 이 연구에서 빅데이터 분석을 통해 도출된 키워드의 다양한 잠재변수와 명확한 해석의 한계를 논의하며 후속연구를 통해 개선되어야 함을 개진하였다.

예컨대, 텍스트에서 추출한 용어, 문장 등의 개체 간 관계를 인식하고 학습할 때, 기본적인 개념을 정의해야 하는데 기본적으로 빅데이터 분석 환경에서 주로 사용되고 있는 감성사전 및 용어사전은 일반적 사전의 정의를 내포하고 있어 스포츠 선수의 별명, 수사어, 지칭어에 대한 분석은 배제되는 경우가 발생하기도 한다. 특히, 축구와 같이 거대한 팬덤을 형성하고 있는 스포츠 종목의 경우 용어가 가지는 속성에 대한 의미와 관계를 파악하는 용어 체계의 개발이 선행되어야 한다. 텍스트 데이터의 의미적 모호성과 용어 간 관계 및 체계를 정의하는 대표

적인 도구로는 온톨로지(Ontology)를 들 수 있다.

Jung(2018)과 Han, Kim, Song & Song(2019)은 비정형 소셜 데이터의 정교한 의미 분석을 위해 온톨로지(ontology)가 필요하며, 특정 분야의 온톨로지는 소셜 데이터와 같은 빅데이터를 수집하고 분석하는 데 필요한 분석틀(framework)로써 활용됨을 주장하였다.

온톨로지는 용어 간 개념을 정의하고 구조와 관계를 표현하는 대표적인 도구로 온톨로지의 객체 지향적 특성은 해당 분야의 지식체계를 표현하는 데 효과적이다. 존재론에 대한 철학적 관점에서 온톨로지는 '존재를 설명하는 특징들과 그 구조(Guarino, Oberle & Staab, 2009)'로 정의할 수 있다.

한국정보통신기술협회(Telecommunications Technology Association;TTA)에서 발행한 '정보통신용어사전'에 따르면 온톨로지는 어떤 일정 범위에서 사용되는 단어들의 개념, 특성, 연관 관계 등을 표현하여 단어에 대한 일반적 지식이 명시적으로 드러나고, 단어 간 관계 정의를 통해 문장의 의미를 파악할 수 있는 도구로써의 기능을 가진다고 하였다. 또한, 온톨로지의 활용 영역에 대하여 인공지능, 시맨틱 웹, 자연어 처리, 문헌정보학 등 여러 분야에서 지식 처리, 공유, 재사용 등에 활용됨을 명시하였다.

최근 정보학 분야에서 온톨로지의 활용이 확산되고 있는 가운데, 온톨로지의 개념은 '서로 공유하는 개념적 이해에 대한 형식적이고 명시적인 구체화 과정의 결과물'로 재정립되어 사용되고 있다. 온톨로지 구축의 목적은 비구조화된 데이터로부터 정보의 체계적인 수집과 분석을 가능하게 하기 위함이며, 온톨로지 구축 과정은 특정 주제와 관련하여 그동안 입증되어 온 이론들을 바탕으로 위계적인 체계를 정리하는 것으로 볼 수 있다(Han et al., 2019).

온톨로지는 클래스(class, 또는 개념), 인스턴스(instance), 속성(property), 관계(relation) 등의 구성요소로 표현된다. 클래스는 사물의 개념(concept), 즉 범주(category)를, 인스턴스는 개별 요소인 실체(entity)를 뜻한다. 속성은 클래스와 인스턴스의 특성(feature)을 나타내며, 관계는 클래스 및 인스턴스 간의 관계성을 표현한다. 예를 들어, '최강희' 인스턴스는 '감독' 또는 '인물'이라는 속성으로 'K리그' 또는 '전북현대모터스FC' 클래스, 속성과 관계를 맺는다.

즉, 온톨로지는 텍스트의 개념과 개념 사이의 관계와 체계, 속성을 기술하는 정형어휘의 집합이라 할 수 있다. 온톨로지의 개발과 적용을 통해 소셜 미디어에서 나타난 내용 분석을 위한 자연어 처리 기반 텍스트 마이닝 수행 시 언어적 특성에 따른 필연적 시행착오를 줄이고 효율적인 분석을 가능하게 할 것이다. 특히, 스포츠와 같이 복잡하고 다양한 기록이 존재하는 분야의 소셜 미디어 분석에서의 온톨로지는 핵심 용어를 추출하고, 핵심개념을 명시할 뿐 아니라, 해당 개념들 간의 관계를 나타내주는 역할을 수행할 것으로 기대된다.

이 연구는 축구 선수의 소셜 미디어에 나타난 내용 분석을 위한 온톨로지 모형을 개발 제안하는 데 목적이 있으며 개발한 온톨로지는 축구 선수와 관련된 빅데이터를 분석하는 도구로 활용할 수 있다.

연구방법

개발 방법

온톨로지 개발을 위해 Noy & McGuinness(2001)가 고안한 'Ontology Development 101: A Guide to Creating Your First Ontology'의 방법을 적용하였다. 'Ontology Development 101'은 현재 가장 대중적이고 실용적으로 쓰이고 있는 온톨로지 개발 방법 중 하나로 광의의 개념과 협의의 개념이 모호한 어휘 간의 관계를 가장 효율적으로 제시한다는 특징이 있다.

또한, 온톨로지는 1차 개발된 후 사회 · 문화적 특성을 반영하여 지속적으로 보완되어야 하는데, 'Ontology Development 101'은 기존 온톨로지가 존재하지 않거나 온톨로지 개발 초기 단계에서 기초를 세우기 위한 가장 합리적인 지침서의 기능을 한다.

'Ontology Development 101'은 총 7단계로 구성되어 있으며 〈Table 1〉은 각 단계별 세부 내용과 이 연구에서의 수행과정을 나타낸다.

Table 1. 'Ontology Development 101' Process

step	contents	function/task process
1	Determine the domain and scope of the ontology	case study / exploratory analysis
2	Consider reusing existing ontologies	prior-ontologies research
3	Enumerate important terms in the ontology	morpheme analysis / soccer player-related terms (nouns/verbs) extraction
4	Define the classes and the class hierarchy	classification criteria definition of terms and leveling
5	Define the properties of classes—slots	class/term relationship and structure description
6	Define the facets of the slots	facets definition
7	Create instances	object description

단계별 온톨로지 개발 과정

1단계 : 온톨로지 대상 분야의 범위 규정

'Ontology Development 101'에서 제시하는 온톨로지 개발의 첫 번째 단계는 대상 분야의 범위를 규정하는 것이다. 온톨로지는 시스템적 측면, 구축범위, 구축대상에 따라 다양한 형태로 존재하며, 구축 목적에 따라 인공지능, 빅데이터 분석, 정보검색, 유비쿼터스, 전자상거래 등 다양한 분야에 적용된다. 온톨로지 개발 범위와 대상 분야를 규정하기 위해 연구결과의 활용 목적, 온톨로지의 사용 대상, 활용 방법 등의 핵심내용을 함축하는 역량 질문 체크리스트를 개발하여 대상 분야의 범위를 규정하였다.

Table 2. Ontology Development Competency questionnaire

subject	competency question
object	What is the purpose development ontology? Who are the main users of development Ontology? What is the type of development ontology?
range	What is the scale of development ontology? What is the scope of development ontology?
contents	What is the detailed process of development?

〈Table 2〉는 'Ontology Development 101'에서 제안하는 역량질문의 구성내용이다. 역량질문에 따라 연구의 범위를 설정한 결과 이 연구에서 개발하는 온톨로지는 축구 선수의 빅데이터 분석을 위한 참조자료의 특성이 있는 온톨로지를 개발하는 것으로 범위를 규정하였다. 또한 축구 영역의 데이터와 지식을 결합하여 '축구 선수의 소셜 미디어'라는 특정 영역을 설명하는 영역 온톨로지를 개발함을 목표로 하였다.

2단계 : 선행 온톨로지 검토

온톨로지 개발의 두 번째 단계는 기존 온톨로지를 검토하여 활용 및 가공하는 단계인데, 이 연구에서 목표로 하는 온톨로지는 현재 기개발 된 사항이 없다. 또한, 유사한 온톨로지의 응용 여부를 확인한 결과 적용 가능한 온톨로지는 제한적이기 때문에 축구 선수의 소셜 미디어 분석을 위한 새로운 온톨로지 모형 개발을 이 연구의 목표로 하였다.

3단계 : 용어나열

세 번째 단계는 용어나열의 단계로, 이는 온톨로지의 핵심 용어를 추출하는 과정을 의미한다. 보편적으로, 핵심용어를 추출하는 과정은 관련 주제의 보고서나 학술연구 등의 텍스트에서 용어를 추출하는 것이 일반적이다. 그러나 이 연구에서 개발하는 온톨로지는 축구 선수의 소셜 미디어 분석을 위한 2차 자료이기 때문에 실제 소셜 미디어를 포함하는 용어를 추출해야 한다. 따라서 이 연구에서는 축구와 관련한 소셜 미디어를 대상으로 파일럿 테스트(텍스트 마이닝)를 진행하여 용어를 추출하였다.

텍스트 마이닝의 용어 수집 대상은 2018시즌(2018.02.28.부터 2018.12.03까지)의 기사, 댓글, SNS(페이스북)의 비정형 데이터이며, 검색 대상 키워드는 전북현대모터스FC의 '김신욱, 김진수, 로페즈, 이동국, 이용, 전북현대, 최강희'다. 페이스북 페이지 선정 기준은 대한민국 축구 국가대표팀, K리그, 전북현대모터스FC의 공식 운영 페이지와 K리그 키워드로 검색된 개인 운영 페이지 중 〔좋아요〕의 빈도가 2만 이상인 것을 대상으로 선정하였다. 〈Table 3〉은 데이터 수집의 대상을 나타낸다.

데이터 수집을 위해 파이썬(Python) 언어를 활용하여

Table 3. Data collecting

		contents	page	like
NEWS Comment		NAVER SPORTS NEWS / Comment		
		Jeonbuk Hyundai Motors FC official website article		
SNS	official Facebook	Korea Football Team	KoreaFootballTeam	5800K
		K-League	withKLEAGUE	870K
		Jeonbuk Hyundai Motors FC	jeonbuk1994	700K
SNS	unofficial Facebook	ONLY K-League	onlykleague	260K
		Today K-League	todaykleague	220K
		Our K-League	ourkleague	210K
		Love K-League	lovekleague	470K

크롤링 툴을 개발하였으며 웹 사이트에서 대량으로 데이터를 추출할 수 있도록 하는 Scrapy, 텍스트 · 이미지 및 기타요소를 파싱하기 위한 Newspaper, HTML 및 XML 문서의 구문 분석을 위한 BeautifulSoup 패키지를 사용하였다.

데이터 수집 시 중의어의 오류를 최소화하기 위해 불리언 연산자를 활용하여 키워드를 통제하였다. 가령 검색 키워드로 '이용'을 입력하였을 경우 '가수 이용' 과 '사용하다'라는 의미의 '이용'이 함께 검색되므로 이러한 경우 '축구 and 이용' 등의 키워드로 검색하여 부적합 정보의 수집을 최소화하였다. 수집 변수는 인덱스(index), 날짜(date), 기사제목(title), 페이스북 게시페이지 아이디(source_name), 내용(text), 기사내용(contents), 검색 키워드1/2(keyword1/2), 빈도(frequency), 댓글(comment), 페이스북 게시 페이지명(source), 링크(url), 조회수(reactions), 좋아요(likes)등 이다.

수집한 데이터의 자연어 처리를 위해 형태소 분석을 실시하였으며, 파이썬(Python) 패키지인 KoNLPy를 사용하였다. 형태소 분석은 정규화, 토큰화, 어근화, 어구 추출 등 총 4단계로 진행하였다.

데이터 전처리를 통해 수집된 용어 중, 1차적으로 상위 3%의 빈도를 보인 용어를 온톨로지의 개발 대상으로 결정하였다. 1차적으로 선정된 1,100여개의 단어를 대상으로 동의어, 수사어, 용언 등을 정제하여 최종적으로 484개의 용어를 온톨로지 개발 대상으로 최종 선정하였다.

4단계 : 클래스 및 계층 정의

이 연구에서 개발한 온톨로지 개발 단계의 핵심은 4단계와 5단계로 어휘 간 클래스의 계층과 속성을 정의하는 것이다. 텍스트 마이닝으로 추출된 단어를 2016년 한국언론진흥재단에서 발행한 '2016년 뉴스빅데이터 시소러스'를 참조하여 사전적 의미로 분류한 뒤 축구 경기에서 사용되는 용어 및 별명, 수어 등 사회 · 문화적 특성을 반영하여 속성과 계층을 정의하였다. 계층 분류 기준은 'Ontology Development 101' 의 세 번째 용어 나열 단계로 부터 도출된 용어를 대상으로 역량질문의 내용 반영 여부와 내용 기준에 따라 분류하였고, 이에 대하여 연구자 간 검수를 3회 진행하였다.

보편적으로 용어의 클래스를 유목화하는 과정은 하향식(top-down), 상향식(bottom-up), 복합식(combination)의 세 가지 방법으로 구분된다. 이 연구에서는 하향식과 상향식을 모두 사용한 복합식의 방법으로 클래스의 계층을 유목화하였다.

먼저 온톨로지 개발 대상의 용어의 의미에 따라 내용분석 한 뒤 가장 큰 단위의 클래스 기준을 선정하였다. 온톨로지의 클래스의 속성 정의는 해당분야의 지식 구조를 직관적으로 사변하기 때문에 총 4단계로 세분화 된 구조로 유목화하였다. 각 단계의 class 기준은 다음 〈Table 4〉와 같다.

하향식 방법으로 축구 선수의 소셜 미디어 분석 용어로 대분류된 온톨로지는 두 번째 이후 클래스에서 상향식 방법으로 중분류 및 소분류 하였다.

유목화 과정은 계층화 과정이라고 할 수 있으며, 1)우

Table 4. Classification Level Structure

Social Media record Classification Level	
class1	soccer's social media contents terms
class2	subject classification
class3	concept / vocabulary classification
class4	

선어 선정, 2)대응수준 N:N, 3)상위계층은 하위계층을 포함, 4)기본 텍사노미 어휘 규준 등의 규칙에 의거하였다. 우선어는 2016년 한국언론진흥재단에서 발행한 '2016년 뉴스빅데이터 시소러스'를 통제도구로 활용하였다.

5단계 : 클래스의 속성 정의

'Ontology Development 101'에서 제시하는 온톨로지 개발의 다섯 번째 단계는 클래스 별로 유목화된 온톨로지의 클래스 속성 정의 단계이다. 이 단계에서는 클래스의 특징을 반영하는 속성을 파악하며 계층 간 용어 및 어휘 간 관계를 기술하였다.

어휘 분류의 기준은 Cho(2011), Choi(2001) 및 메타데이터의 국제표준인 ISO 15836(2017)에서 제시한 어휘 분류 기준을 부분 선택하여 적용하였다〈Table 5〉.

6단계 : 슬롯의 패싯 정의

여섯 번째 단계는 슬롯의 패싯 정의 즉, 클래스의 속성값을 정의하는 단계이다. 이 연구에서 개발한 온톨로지는 클래스의 속성값은 여러 속성값을 갖는 다중형의 특성을 가지며, 특정 클래스의 개체를 따른 개체와 숫자, 문자 등으로 구성되었다.

7단계 : 개별 사례 생성

'Ontology Development 101'에서 제시하는 온톨로지 단계의 마지막 일곱 번째 단계로 개별 사례를 생성하여 개체를 부여하는 단계이다. 연구과정에서 추출된 축구 선수와 관련된 용어 484개를 대상으로 각각 해당 개체를 모두 정의하고, 속성값과 관계를 부여하였다. 결과적으로 두 번째 단계 클래스에서 구분된 4가지 주제 영역의 온톨로지가 개발되었다.

Table 5. Vocabulary and relationship classification

Vocabulary and relationship classification				
Lexical Meaning class				
category 1	category 2		category 3	
lm-category1	lm-category2		lm-category3	
number	animate noun	AN	organization	ON
entity	inanimate noun	IN	human	HN
set	phenomenon noun	PN	zoo	ZN
event	abstraction noun	AN	plant	PN
volume			state	SN
container			information	IN
size			etc	EN
			location	LN
			time	TN
relationship type (between sub-class and specific sub-class)				
subordinate relationship	up	BT	down	NT
horizontal relationship		CT		
processes relationship		PT		
causal relationship	cause	CT	effect	EF

연구결과

소셜 미디어 내용 분석을 위한 축구 선수 온톨로지 분류 체계

'Ontology Development 101'에서 제시한 과정에 따라 최종 개발한 축구 선수의 소셜 미디어 분석을 위한 온톨로지는 크게 4가지의 영역으로 구성되었다. 각 영역은 용어에 따라 3단계 또는 4단계로 하위분류 되었다.

4가지 영역의 구성 기준은 선별된 484개 용어에 대해 연구 방법(4step:클래스 및 계층정의)에서 기술한 클래스 유목화 기준을 적용하여 구성되었으며, 용어의 내용적 특성에 따라 분류되었다. 분류과정은 총 3단계로 진행되었으며, 1차 사전적 분류 – 2차 미분류 용어 구분 – 3차 클래스 계층 정의의 과정으로 진행되었다. 선별 된 용어를 '2016년 뉴스빅데이터 시소러스'를 참조하여 1차 분류한 결과 몇몇 용어들은 미분류 되거나 일반적인 용어이나

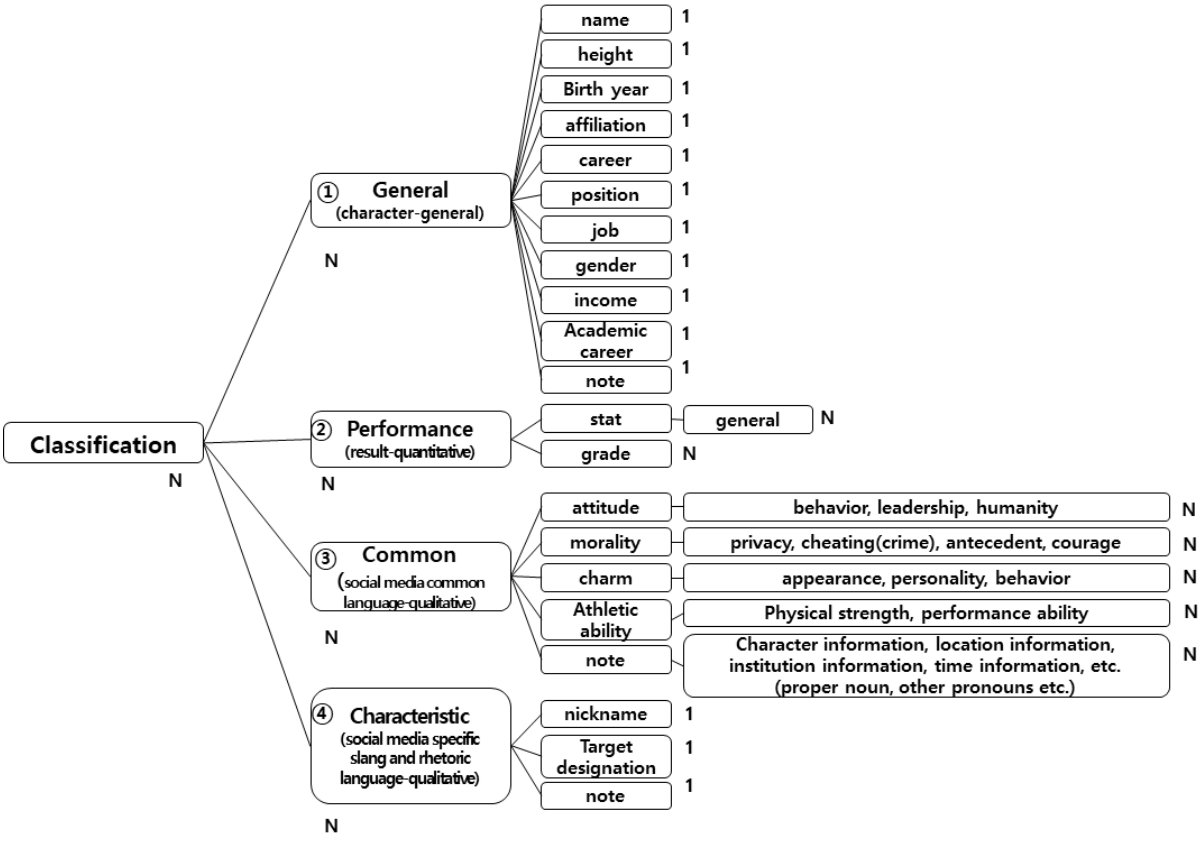

Fig. 1. Soccer player's ontology classification system

특정 선수를 지칭하는 의미의 용어가 나타났다. 예를 들어 '한퀴아오', '일단은', '강희대제'. '봉동이장'과 같은 단어이다. 이들은 선수나 해당 팀을 지칭하는 신조어, 수사어의 일종으로 사전적 정의나 분류를 할 수 없으므로 클래스의 계층을 구분할 때 ④번 소셜 미디어 특정 은어 및 수사어(Characteristic)특성 집단으로 단어 군집 영역을 설정하였다. 〈Figure 1〉은 소셜 미디어 분석을 위한 축구 선수 온톨로지 분류 체계를 도식화한 것이다.

개발 온톨로지의 영역은 크게 ①인물-일반(General), ②수행결과(Performance), ③소셜 미디어 보통어(Common), ④소셜 미디어 특정 은어 및 수사어(Characteristic) 영역 등으로 구성하였다. 온톨로지를 4개의 영역으로 구분하여 개발한 이유는 4가지 영역을 구성하는 용어의 속성 또는 메타데이터의 구성에 차이가 있기 때문이다. 이러한 구성의 차이는 용어의 의미 또는 내용 분석을 반영한 결과로 각각의 분류에 속한 용어는 공통된 속성 군집으로 이해된다.

이 연구에서 개발한 온톨로지의 대표적인 특징은 특정 영역의 지식 구조를 직관할 수 있는 영역 온톨로지라는 점이다. ①번 인물-일반(General)영역과 ②수행결과(Performance) 영역의 경우 일반적인 언어 사전을 적용하여 텍스트 분석을 할 수 있으나 ③번 소셜 미디어 보통어(Common)와 ④번 소셜 미디어 특정 은어 및 수사어(Characteristic)영역의 경우 이 연구에서 개발한 소셜 미디어 분석을 위한 특수한 용어집이라 할 수 있다. 특히, ④번 영역 온톨로지에서 구성된 특정어는 선수의 별명, 대상 지칭어 및 수사어를 포함하며, 이 연구에서 개발한 온톨로지의 핵심 부분이라 할 수 있다. 또한, 개발한 온톨로지는 인물과 관련한 텍스트 정보를 연구 범위로 하고 있기 때문에, 모든 용어에 식별코드를 부여하여 해당 축

구 선수와 관련된 용어의 경우 관계를 N:N, N:1, 1:1로 설정하였다. 예컨대, 개발 온톨로지의 첫 번째 영역의 '이동국' 선수에 관한 인물정보는 네 번째 특정어 영역의 '라이온킹'이라는 단어와 등위의 관계로 설정하였음을 확인할 수 있다.

아울러 선수의 소속이나, 경력, 포지션, 소득 등의 정보는 첫 번째 인물-일반(General) 영역 온톨로지에서 분류되는데, 이러한 정보의 경우 유동적일 수 있기 때문에 선수에 대한 전거 정보 파일을 따로 구성하였다.

소셜 미디어 내용 분석을 위한 축구 선수 온톨로지 구조

〈Table 6〉은 개발한 축구 선수 온톨로지의 구조와 속성를 나타낸다. 개발한 온톨로지 484개의 용어에 대하여 관계 및 정의, 속성값을 기술하였다. 즉, 선별된 모든 용어에 고유의 식별 기호를 부여하고, 해당 내용에 대하여 값(국문/영문), 우선어 선정, 설명 및 정의되었으며 〈Table 5〉에 명시한 어휘 분류 기준의 세 가지 속성과 네 가지 관계를 구성하였다. '우선어'라는 것은 대표적인 용어를 의미한다고 할 수 있는데 이 연구에서 선별된 용어들 중 '강희대제', '봉동이장', '한쿼아오', '전주성' 등의 단

Table 6. Ontology structure for soccer Player's social media contents analysis

class1	class2	class3	property	value (data type)	note1	note2
	medium scale classification	small scale classification		property value	note	authority file
Large scale classification: soccer's social media contents terms	General (character-general) n=17	name	id-code, value(kor/eng), vocabulary-category classification (3 type), relationship (4 type), preferred term	word		
		height		number		
		birth year		number		
		affiliation		word		○
		career		number		○
		position		word		○
		job		word		
		gender		word		○
		income		number		
		academic career		word		○
		the other info.		word/number		○
	Performance (result-quantitative) n=53	stat	id-code, value(kor/eng), description, formula, classification (3 type)	word/number		
		grade		word/number	award info. included	○
	Common (social media common language-qualitative) n=375	attitude	id-code, value(kor/eng), description, definition, vocabulary-category classification (3 type), relationship (4 type), preferred term	word	positive/negative	
		morality		word	positive/negative	
		charm		word	positive/negative	
		athletic ability		word		
		the other info.		word	positive/negative	
	Characteristic (social media specific slang and-qualitative) n=39	nickname		word	common thesaurus unapplicable	comments
		target designation		word		
		the other info.		word		

어는 전북현대모터스FC의 '최강희', '한교원', '전주월드컵경기장'을 지칭하는 단어이다. 따라서 우선어를 선정하여 속성값으로 명시하거나 관련 인물에 관계를 설정하였다.

온톨로지의 구조와 속성값을 기술하는 과정에서 두 번째 영역인 ②수행결과(Performance) 영역의 용어는 스탯을 나타내기 때문에 공식 속성이 추가되었다. 용어의 설명 및 정의에 대한 속성에서는 네 번째 영역인 ④소셜 미디어 특정 은어 및 수사어(Characteristic) 영역을 제외하고는 모두 사전적 정의 및 기술이 수행되었으며, 네 번째 영역은 사전에 있지 않은 용어이므로 문헌고찰, 사례조사를 통한 연구자의 함의를 기준으로 정의 및 기술되었다.

개발 온톨로지는 클래스와 객체를 정의하여 용어 간 구분 및 관계를 정의한 객체 지향적 온톨로지 모형이다. 개발한 온톨로지의 클래스의 관계는 중분류 간, 특정 소분류 간 관계로 정의하였으며, 상하 관계는 클래스의 계층이 '상위 계층은 하위 계층을 포함 한다'는 명제를 전제로 하고 있기 때문에 따로 기술하지는 않았다. 개발 온톨로지의 클래스의 속성 관계는 대표적으로 종속, 수평, 기타-과정, 기타-인과 관계로 구성하였으며, 일반 사전에서 제시하는 어휘 분류를 동시에 진행하였다. 일반적인 사전 및 여러 기준에서 제시하는 어휘 분류를 동시에 진행한 이유는 소셜 미디어에 등장하는 텍스트가 개발 온톨로지의 영역의 네 번째 영역인 별명, 지칭어, 수사어 등 특정 용어만 존재하는 것이 아니라 보편적인 내용의 단어를 포함하기 때문이다.

이 연구에서 개발한 축구 선수의 소셜 미디어 내용 분석을 위한 온톨로지는 포괄적으로 텍스트 문서와 용어에 대한 메타데이터의 의미를 내포하며 다음과 같은 핵심적 기능을 수행한다.

① 특정 분야를 기술하는 데이터 모델로서 특정한 분야(domain)에 속하는 개념과 개념 사이의 관계를 기술하는 정형(formal) 어휘의 집합
② 시맨틱 웹을 구현할 수 있는 도구로서 여러 지식 개념들을 의미적으로 서로 연결할 수 있는 도구/기술
③ 어떤 특정 분야를 개념화하기 위해 이미 형식화(존재)되거나 암묵적 또는 추상적인 것들을 명시적으로 정형화한 정의서이며, 나아가 컴퓨터로 처리할 수 있도록 나타낸 용어들의 논리적 집합
④ 시소러스(통제어휘집)와 차이 : 어휘, 개념과 같은 지식의 관계가 표현된 일종의 사전
⑤ 문맥 분석에 필요한 기초 구조 - 관계 정의
⑥ 효율적이고 합리적인 텍스트 마이닝을 위해 필요한 도구, 기술

즉, 이 연구에서 개발한 온톨로지 모형은 축구 선수의 소셜 미디어 내용 분석 영역에 특화된 온톨로지로써 비정형 데이터 분석에 활용될 수 있는 2차 도구로써의 기능을 가진다고 할 수 있다. 서두에도 개진한 바와 같이 비정형 데이터 분석 시 언어의 복잡성 및 중의성 등의 특성으로 배제되는 용어를 분석할 수 있도록 용어의 체계와 속성, 개념과 구조를 정의하였다.

예를 들어 2018년 수집된 어떤 기사의 제목이 「봉동이장 전북 최강희 감독, 中 텐진으로…14년 닥공은 전주성의 명예로」 였는데, 이 기사를 텍스트 마이닝 하였을 때, '전북', '텐진', '전주성'은 〔지명〕으로, '봉동이장', '닥공'은 〔이형어〕로 분류되는 경향을 보였다. 그러나 이러한 단어는 전북현대모터스FC 팀과 최강희 감독,홈구장을 의미하는 단어로 소셜 미디어에도 빈번히 등장하는 단어이다. 이 연구에서 개발한 온톨로지를 적용하면 이러한 특정 용어의 의미를 명확하게 분석할 수 있다. 즉, 이 연구에서 개발한 온톨로지를 적용하여 의미가 있는 단어임에도 불구하고, 〔이형어〕 또는 〔중의어〕, 〔지명〕등으로 구분되는 단어를 실제 선수와 연결하여 관계를 정의하고, 구조를 개념화함으로써, 텍스트 마이닝에서 의미 있는 정보가 누락되는 문제를 최소화 하였다.

물론, 텍스트 마이닝에서 누락되는 몇몇 용어는 연구자의 주재로 재정의 될 수 있으나, 빅데이터 분석환경에서 모든 용어를 일일이 연구자가 주재하는 것은 시간과 재원의 한계가 존재한다. 아울러 빅데이터 분석 목표가 인공지능(AI)의 자동 분석을 지향한다는 관점에서, 분석의 적합도 및 정확도 향상을 위해 이 연구에서 개발한 온톨로지 모형과 같은 특정 주제의 온톨로지는 매우 효과적인 기능을 수행한다고 할 수 있다.

논의 및 결론

온톨로지는 개발 목적에 따라 시스템적 측면, 구축범위, 구축대상에 따라 다양하게 구성되며, 이 연구에서는 구축범위에 따른 종류인 '영역 온톨로지'를 개발하였다. 보편적으로 '영역 온톨로지'는 특정 영역의 유효한 지식과 대상을 정의하고 관련 지식의 체계를 구축할 때 효과적이다.

이 연구를 통해 개발한 온톨로지는 데이터 지식의 체계를 구성하는 유일한 도구는 아니나 비교적 효율적이고 유용한 방법이다. 개발한 온톨로지는 용어 간 관계, 계층 및 긍정과 부정의 영역에서 각 어휘의 사회 · 문화의 특성을 반영하였기 때문에 축구 선수의 소셜 미디어 내용 분석에 유용한 도구로 활용될 것이다.

소셜 미디어 빅데이터는 거대한 양의 정보가 빠른 속도로 생성되고 있고, 연속적인 데이터의 특성이 있기 때문에 축구 선수의 가치 평가에 적절한 복합적 차원을 모두 포함하는 데이터라고 할 수 있다. 그러나 스포츠 선수의 빅데이터 분석 및 가치 평가에 대한 선행연구를 검토하였을 때, 경기력 분석이 주를 이루거나 소셜 미디어 내용 분석 시 스포츠 영역에 특성화된 감성사전 및 기타 언어사전 등의 부제로, 일반적인 용어사전이 분석 도구로 사용된다. 이러한 이유로 간혹 스포츠 선수의 소셜 미디어 분석에서 유의미한 수사어, 지칭어, 신조어 등의 특성을 반영하지 못하는 다소 제한적인 연구가 진행되는 경향이 있다. 이와 관련하여, Lee & Kim(2019)은 스포츠와 관련한 빅데이터 분석에서 키워드가 가지고 있는 명확한 해석과 잠재변수에 대한 한계를 텍스트 분석의 제한점으로 주장한 바 있다. 이 연구에서 개발한 온톨로지를 통해 이러한 텍스트 마이닝에서의 시행착오 및 제한을 일부 해결할 수 있을 것으로 기대된다. 온톨로지는 개념들을 정의하고, 그 개념들 간의 관계와 계층구조를 논리적이고 명시적으로 정의하여, 컴퓨터가 스스로 정보처리를 할 수 있도록 돕는 2차 자료(Han et al., 2019)라 할 수 있기 때문이다.

특히, 이 연구에서 개발한 온톨로지는 축구 선수의 소셜 미디어 내용 분석을 목표로 개발되었으며, 영역의 범위가 넓지는 않지만 특정 영역의 핵심적 지식을 포함하는 특성이 있다. 이는 축구 선수의 소셜 미디어 영역에서 나타나는 선수의 가치를 평가하는 기초적인 기준을 제시한다고 할 수 있다. 선수의 가치를 직관적으로 사변하였을 때 개인 선수의 경기력이 가장 중요한 요인이 되어야 하는 것은 주지의 사실이나 선수의 가치를 평가하는 유일한 요인이라 할 수 없을 것(Kim et al., 2020)이다. 가치평가는 질적인 속성을 포함하며, 수적으로 비교하거나 양적으로 범위로 판단할 수 없고 상대적인 의미와 해석적 인식이 포함되어야 하기 때문이다. 특히, 연속적으로 빠르게 생성되는 축구 선수의 소셜 미디어의 텍스트는 경기력 이외의 선수의 이미지나 트렌드와 같은 복합적 정보를 포함하기 때문에 경기력 외적인 부분을 해석하기에 좋은 정보원이 되며, 이 연구에서 개발한 온톨로지 모형은 축구 선수와 관련한 용어의 관계와 체계를 구축한 유일한 용어체계라 할 수 있다.

이 연구에서 개발한 온톨로지와 같이 계층적으로 구성된 온톨로지는 메타데이터의 속성을 포함하기 때문에 선수를 평가하는 프로파일링 시스템 기초자료로 활용할 수 있다. 이는 온톨로지 구축의 목적이 비구조화된 데이터로부터 정보의 체계적인 수집과 분석을 가능하게 하기 위함이라 논의한 Han et al.(2019)의 의견과 같은 맥락에서 이해할 수 있다. 아울러 Choi & Lee(2020)는 스포츠 빅데이터 분석을 활용한 데이터의 재해석과 지식의 재창출에 대한 논의를 개진한 바 있는데, 이 연구에서 개발한 온톨로지는 지식의 재해석과 재창출에 대한 논리적 구조를 함의한다고 할 수 있다.

Choi(2020)는 국내 스포츠 빅데이터 분석의 현황을 논의하면서, 지속적인 방법론의 발전과 텍스트 마이닝 분석 방법의 세분화 및 의미 분석의 다양한 연구 흐름을 논의하였다. 이 연구는 이러한 측면에서 시대의 거대담론에 걸맞은 시의적인 연구라고 할 수 있으며, 미시적 차원에서 스포츠 경기분석 및 스포츠 데이터 과학의 방향을 제시하는 연구라 할 수 있다.

아울러 온톨로지 개발 연구의 측면에서 살펴보았을 때, 초창기 온톨로지 연구는 정보학 분야에서 정보검색과 관련한 연구가 주를 이뤘다. 그러나 최근 빅데이터 기술을 활용한 분석이 활발히 진행되면서 소셜 미디어 내용분석, 텍스트 마이닝 등에 활용되는 2차적 도구로써의 온톨로지의 개념이 확산되고 특정 분야의 빅데이터 분석을 위한 온톨로지 모형 개발 연구가 점진적으로 진행되고 있다. 이는 온톨로지의 개념화, 명시화, 체계화, 관계화의

특징에 기인한 결과라고 볼 수 있으며, 이 연구는 온톨로지 연구의 측면에서도 시의적인 연구라 할 수 있다.

다만, 연구과정에서 텍스트 마이닝으로 추출된 용어를 1차 사전적 기준으로 분류하고, 미분류 된 용어에 대하여 2차 연구자 간 함의와 문헌고찰에 의해 분류 및 기술하였으며 온톨로지의 영역 계층을 구분하였기 때문에 Oh et al.(2019), Lee & Kim(2019) 등이 논의한 연구자의 주관적 관점의 배제는 이 연구에서도 제한적으로 이루어졌다. 이는 이 연구가 초기단계의 온톨로지 연구 단계이고, 기존의 비슷한 온톨로지의 재가공이 불가하였기 때문이라 판단한다. *Nahm*(2016)은 연구자의 해석적 관점에서의 분류가 어느 정도의 내적 타당도를 확보한다고 주장한 바 있지만, 이 연구에서 개발한 온톨로지 모형의 평가에 대한 후속 연구가 필요하다. 온톨로지 모형의 평가는 Jung(2018)이 제시한 평가방법과 같이 평가도구를 개발하여 평가하거나 온톨로지 모형을 확장하여 사전의 형태로 개발한 후 적용하거나 전문가 집단을 구성하여 의견을 수렴하는 방법 등이 있다.

일반적인 모든 지식의 영역을 온톨로지로 만드는 것은 불가능한 일이다. 이 연구에서 개발한 온톨로지는 스포츠 분야 중에서도 축구 영역의 기초적인 온톨로지이며 향후 확장 · 발전해 나가야 한다. 이는 스포츠 영역의 빅데이터 분석의 타당성과 정확성, 적합성을 위한 일련의 과정이 될 것이며 이 연구는 이를 위한 시발점적인 연구가 될 것이다.

참고문헌

Cho, E. H. (2018). The science of recording and measuring sports field. *Korean Journal of Sports Science, 142,* 20-29.

Cho, Y. J. (2011). *A Study on the Image Reception of Foreign Athletes in Professional Baseball League.* Unpublished masters dissertation, The Graduate School of Education, Hanyang.

Choi, H. (2020). The current status of sports big data analysis researches in korea. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science. 22*(2), 63-69.

Choi, H., Lee, Y. S. (2020). The big data analytics for performance analysis of tennis. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, 22*(1). 57-68.

Choi, K. (2001). Semantic classification of vocabulary for the construction of the knowledge base - with special reference to the classification of nouns. *Discourse and Cognition, 8*(2), 275-303.

Guarino, N., Oberle, D. & Staab, S. (2009). *What is an ontology?.* In Handbook on ontologies, *Springer,* 1-17.

Han, Y. S., Kim, H. Y., Song, J. Y. & Song, T. M. (2019). Ontology development of school bullying for social big data collection and analysis. *Journal of The Korea Contents Association, 19*(6), 10-23.

ISO 15836(2017). *Information and Documentation – the Dublin Core Metadata Element Set*, ISO-International Organization for Standardization.

Jung, H. (2018). *Development and Evaluation of Youth Depression Ontology for Analyzing Social Data.* Unpublished masters dissertation, The Graduate School of Seoul National University.

Kim, J. H., Kang, J. Y. & Cho, S. M. (2020). Fan-centered soccer player attribute model development using machine learning and kano model. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science, 22*(3), 91-104.

Kim, M. S., Jeon, S. W. (2019). Trends of sports policy through the analysis of big data text-mining : with a focus on the inauguration of the mcst minister. *The Korean Journal of Sport, 17*(2), 519-529.

Korea Press Foundation(2017). *News Big Data Thesaurus and Teksanomi Dictionary.* Seoul : KPF.

Korea Professional Sports Association(2019). *Pro Sports X Social Big data 2019,* Korea Professional Sports Association.

Lee, J. M., Kim, J. H. (2019). A study on the current situation and strategy analysis of sportainment using big data analysis. *The Korean Journal of Sport, 17*(2), 1-13.

Nahm, C. H. (2016). An illustrative application of topic modeling method to a farmer's diary. *Institute of Cultural Studies, 22*(1), 89-135.

Noy, N. F., McGuinness, D. L. (2001). Ontology development 101: a guide to creating your first ontology. *Stanford Knowledge Systems Laboratory*

Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report, 49-69.

Oh, S. W., Han, J. W. & Kim, M. S. (2019). A social big data analysis on perception about water leisure sport tourism. *Korean Journal of Sport Management, 24*(4), 83-95.

Park, S. J., Lee, J. W. (2017). Structure and use knowledge map of sports field through text mining and social network analysis. *Korean Journal of Sports Science 26*(4), 639-653.

Telecommunications Technology Association(2016). *Information Communication Terminology Dictionary:ICT word 101 2016,* Telecommunications Technology Association.

Yun, H. J. (2018). *A Real-time Players Evaluation Model Development based on Social Big Data in Korea Professional Baseball: Sentiment Analysis Using Machine Learning.* Unpublished doctoral dissertation, Graduate School of Korea National Sport University.

축구 선수의 소셜 미디어 내용 분석을 위한 온톨로지 모형 연구*

김주학[1], 조선미[2], 강지연[2]
[1]명지대학교 교수
[2]명지대학교 박사과정

〔목적〕 축구와 관련한 소셜 미디어 빅데이터는 축구 선수와 관련된 복합적 차원의 정보를 내포하며 연속적으로 빠르게 생성되고 있다. 이러한 소셜 미디어 내용 분석을 위해 텍스트 마이닝 연구가 활발히 진행되고 있으나 언어의 복잡성과 문맥에 대한 이해, 중의어, 수사어, 신조어 등 언어적 특성으로 다소 제한적으로 분석되는 경향이 있다. 이는 일반적으로 텍스트 마이닝에 사용되는 도구가 분석 주제의 특수성을 배제한 보편적인 용어 사전이나 패키지를 사용하기 때문이라 볼 수 있다. 이 연구는 텍스트 데이터의 의미적 모호성과 용어 간 관계 및 체계를 정의하는 대표적인 도구인 온톨로지(Ontology) 모형을 개발하는 데 그 목적이 있다. **〔방법〕** 연구의 목적 달성을 위해, 초기 온톨로지 개발에 유용한 'Ontology Development 101: A Guide to Creating Your First Ontology'의 7단계 개발방법을 적용하였다. 각 7단계는 1)온톨로지 대상 분야와 범위 규정, 2)선행 온톨로지 검토, 3)용어 나열, 4)클래스 정의 및 계층 정의, 5)클래스의 속성 정의, 6)슬롯의 패싯 정의, 7)개별 사례 생성의 단계를 포함한다. 특히, 이 연구의 세 번째 단계인 용어 나열 단계는 온톨로지를 구성하는 핵심 용어를 추출하는 단계인데, 이 연구의 목표가 축구 선수의 소셜 미디어 내용 분석에 활용되는 온톨로지를 개발하는 것이기 때문에 실제 축구 선수와 관련한 소셜 미디어에 등장한 텍스트 분석을 진행하여 484개의 핵심용어를 선정하였다. **〔결과〕** 개발한 축구 선수의 소셜 미디어 내용 분석을 위한 온톨로지는 크게 인물, 수행결과, 공통용어, 특정어의 4가지의 영역으로 구성되며 용어의 내용적 특성에 따라 분류되었다. 각 영역에 구성된 484개의 용어에 대하여 관계 및 정의, 속성값을 기술하였다. **〔결론〕**개발 온톨로지는 클래스와 객체를 정의하여 용어 간 구분 및 관계를 정의한 객체 지향적 온톨로지 모형이며 축구 선수의 소셜 미디어에서 나타난 지식체계를 대변한다. 또한, 비정형 데이터 분석에 활용될 수 있는 2차 도구로써의 기능을 수행한다.

주요어: 온톨로지, 빅데이터, 텍스트마이닝, 객체지향, 소셜 미디어