IPTV의 VOD 어노테이션을 위한 반자동 온톨로지 모델링

(Semi-automatic Ontology Modeling for VOD Annotation for IPTV)

최 정 화 † 허 길 † 박 영 택 ††
(Jung-Hwa Choi) (Gil Heo) (Young-Tack Park)

요 약 본 연구는 IPTV의 지능형 검색을 가능하게 하는 VOD 어노테이션을 위해 효율적인 반자동 온톨로지 모델링 기법을 제안한다. 제안하는 방법은 워드넷(WordNet)으로 부터 특정 도메인(또는 장르)을 대표하는 콘텐츠에 관련된 키워드의 상·하위어와 동의어에 해당하는 부분 트리를 추출하고, 워드넷에 없는 외래어, 한자어 등은 확장하여 콘텐츠 온톨로지를 구축한다. 이 온톨로지는 보편적 계층구조와 특정 계층구조를 생성한다. 전자는 콘텐츠와 관련 키워드를 제약 기술(description)을 포함하는 클래스로 정의한 어휘의 의미 모델이다. 후자는 생성된 모델에 함의관계(subsumption) 추론 기술을 적용하여 키워드를 관련 있는 콘텐츠로 추론한 모델이다. 어노테이션은 이 온톨로지를 기반으로 VOD에 콘텐츠와 장르의 메타데이터를 의미 기반으로 생성한다. 보편적 계층구조는 서비스 도메인에 독립적으로 재사용이 가능하며, 특정 계층구조는 서비스 목적에 맞는 완전하고 함축적인 모델을 생성한다. 제안하는 방법은 서비스 도메인에 상관없이 적용 가능한 알고리즘이며, 2,400건의 테스트 데이터로 어노테이션 결과를 평가하여 82%의 정확도를 보였다.

키워드 : 온톨로지, 워드넷, 코렉스, 반자동 어노테이션, 메타데이터, IPTV

Abstract In this paper, we propose a semi-automatic modeling approach of ontology to annotate VOD to realize the IPTV's intelligent searching. The ontology is made by combining partial tree that extracts hypernym, hyponym, and synonym of keywords related to a service domain from WordNet. Further, we add to the partial tree new keywords that are undefined in WordNet, such as foreign words and words written in Chinese characters. The ontology consists of two parts: generic hierarchy and specific hierarchy. The former is the semantic model of vocabularies such as keywords and contents of keywords. They are defined as classes including property restrictions in the ontology. The latter is generated using the reasoning technique by inferring contents of keywords based on the generic hierarchy. An annotation generates metadata (i.e., contents and genre) of VOD based on the specific hierarchy. The generic hierarchy can be applied to other domains, and the specific hierarchy helps modeling the ontology to fit the service domain. This approach is proved as good to generate metadata independent of any specific domain. As a result, the proposed method produced around 82% precision with 2,400 VOD annotation test data.

Key words : Ontology, WordNet, KorLex, Semi-automatic Annotation, Metadata, IPTV

· 본 연구는 숭실대학교 교내 연구비 지원으로 이루어졌습니다.

† 학생회원 : 숭실대학교 컴퓨터학과
cjh7963@hotmail.com
heogil@gmail.com
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
park@ssu.ac.kr

논문접수 : 2009년 11월 23일
심사완료 : 2010년 4월 20일

정보과학회논문지: 소프트웨어 및 응용 제37권 제7호(2010.7)

1. 서 론

IPTV는 기존의 TV를 통해 과거 또는 최신의 TV방송 프로그램, 영화, 뮤직 비디오 등의 VOD(Video On Demand)를 신청해 보는 '주문형' 서비스가 가능하다. 주문형 서비스의 장점 중 하나는 사용자 개개인에 따라 개인화가 용이하다는 것이다. 개인화를 실현하기 위해서는 시청자들이 입력하는 다양한 요구의 어휘를 처리하는 기술이 필요하다. 하지만 사람이 사용하는 어휘는 너무 방대하므로, 기계 또는 소프트웨어 에이전트가 처리하기 위해서는 텍스트에서 어휘를 추출하여 의미를 부여한 어휘의미망을 구축하는 기술이 필요하다[1-3]. 워드넷(WordNet)[4]은 영어 어휘의 의미관계를 계층망 형태로 설정한 어휘사전으로 널리 활용된다. 하지만 워드넷은 너무 방대하고 어휘의 분류체계(taxonomy)만을 제공하여 서비스 도메인에 따른 필터링과 어휘의 의미를 고려한 관련 어휘의 추론이 필요하다.

본 연구는 IPTV의 지능형 검색을 가능하게 하는 VOD 어노테이션(annotation)을 위해 콘텐츠(contents) 온톨로지(ontology)의 반자동 모델링 방법을 제안한다. 제안하는 방법은 특정 도메인을 대표하는 콘텐츠에 관련된 방대한 어휘를 의미기반으로 온톨로지에 반자동 모델링하는 방법과 이 온톨로지 기반의 VOD 어노테이션 자동화를 위한 방법이다. 온톨로지는 메타데이터를 표현하기 위한 해당 분야의 개념 모델이다. 어노테이션은 문장이나 문서에 추가적인 정보를 기입하는 것을 말하며, 이 정보를 메타데이터라고 한다. 예를 들어 '범죄' 관련 영화 VOD는 제목과 줄거리 등의 텍스트를 포함하고, 소프트웨어 에이전트는 '범죄' 관련 용어를 알고 있을 때 그 VOD를 범죄 영화로 추천할 수 있다. 본 논문에서 제안하는 방법의 원리를 살펴보면, 영화는 여러 개의 장르(범죄, 코미디, 액션 등)로 구분되고, 장르를 대표하는 여러 개의 콘텐츠(범죄-형사, 범죄조직 등, 코미디-로맨틱, 가족 등)로 세분화 할 수 있다. 이를 기반으로 콘텐츠에 관련된 키워드(범죄-형사: 범인, 수사 등, 범죄-범죄조직: 보스, 킬러 등)의 어휘 개념이 온톨로지에 정의되어 있을 때, 에이전트는 VOD에 메타데이터를 생성하고, 메타데이터를 기반으로 VOD를 추론 및 추천할 수 있다.

제안하는 온톨로지는 보편적 계층구조(hierarchy)와 특정 계층구조를 생성한다. 보편적 계층구조는 서비스 도메인에 독립적인 어휘 모델링 방법이다. 이 방법은 콘텐츠에 관련된 키워드를 워드넷에 정의된 어휘로 매핑하고, 그 어휘의 상·하위어와 동의어의 부분 트리(partial tree)를 추출하고 온톨로지 설계자가 확인하는 것을 반복하여 온톨로지를 반자동으로 생성한다. 반자동 모델링은 사용자의 검증 단계를 추가하여 자동화로부터 발생할 수 있는 오류를 방지할 수 있다. 제안하는 방법은 또한 어휘에 의미를 부여하기 위해 온톨로지 모델링 시에 콘텐츠는 도메인 제약을 기술(description)한 클래스로 정의하고, 콘텐츠 키워드는 존재 양화사(existential quantifier, '∃') 제약을 가지는 클래스로 기술한다. 하지만 워드넷은 영어 어휘의 의미망이므로 한국어로 된 외래어나 한자어 등은 포함하지 않는다. 따라서 워드넷에 정의되지 않은 단어에 대해서는 대체 어휘로 앵커링(anchoring)하여 어휘의 부분트리를 확장한다. 다음으로 특정 계층구조는 생성된 온톨로지에 함의관계(subsumption) 추론 방법을 적용하여 도메인에 종속적인 어휘 모델을 자동 생성한다. 함의관계 추론은 도메인 제약과 존재 양화사 기술의 상관관계를 이용하여 온톨로지에 정의된 키워드들이 관련된 특정 콘텐츠로 추론되도록 한다. 본 연구에서는 이렇게 구축된 온톨로지를 기반으로 VOD 어노테이션 정보를 사용해서 VOD를 관련 콘텐츠로 자동분류한 후, 온톨로지를 사용해서 IPTV의 지능형 검색을 가능하게 한다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 온톨로지 모델링과 어노테이션에 대한 관련연구를 살펴보고, 3장에서는 콘텐츠 온톨로지의 반자동 모델링 방법과 VOD 어노테이션 원리, 그리고 이를 기반으로 구축한 시스템을 설명한다. 4장에서는 어노테이션 결과를 평가하여 본 연구의 타당성과 정확성을 검증한다. 마지막 5장에서는 결론을 맺고 향후 연구를 제시한다.

2. 관련 연구

온톨로지 모델링 선행 연구들[2,5,6]을 살펴보면, 어휘 온톨로지를 응용 프로그램의 서비스 도메인에 한정 없이 워드넷과 같은 어휘 사전에 정의된 대용량 어휘를 대상으로 분류체계를 만드는 방법에만 중점을 두었다. [2]는 워드넷에 정의된 동의어 집합에서 대표 단어를 추출하여 클래스로 정의하고, 나머지 단어들은 그 클래스의 인스턴스로 자동 모델링한다. [5]는 Wikipedia XML 코퍼스[7]를 어휘사전으로 사용하여 어휘 분류 체계를 반자동 모델링하며, 특정 도메인에 관련된 방대한 어휘를 모두 온톨로지에 정의하는 방법으로 응용 프로그램에 적용하기에는 무겁다. [6]은 한 데이터베이스에 저장된 어휘들(예: Java, Pascal, Basic)이 워드넷에 동일한 의미(예: 프로그래밍 언어)로 정의되어 있으면, 그 의미를 상위어를 갖도록 자동 모델링한다. 이 연구는 관련 있는 어휘들이 저장된 데이터베이스일 경우와 관련 어휘가 동일한 문장으로 의미가 정의되어 있을 때만 정확도가 높다.

어노테이션 연구들[2,8]을 살펴보면, [2]와 [8]은 텍스트에 포함된 어휘가 온톨로지에 인스턴스로 정의되어 있으면, 그 인스턴스의 클래스로 어노테이션한다. 어노테이션한 결과는 상-하위 클래스 또는 클래스-인스턴스 사이의 포함 관계를 찾는 것에만 국한한다. [8]은 어노테이션 시에 어휘 C가 온톨로지에 정의되지 않은 경우는 어휘 C와 인접하여 나타난 어휘 D가 온톨로지에 클래스로 정의되어 있으면, 그 클래스의 인스턴스로 어휘 C를 추가하여 온톨로지를 확장한다. 이 연구는 입력 데이터인 온톨로지, 텍스트, 그리고 텍스트에서의 어휘 추출 패턴에 따라 온톨로지와 어노테이션의 정확도가 좌우된다.

본 연구는 특정 도메인의 응용 서비스에 사용되는 중요 어휘들을 추출하여 온톨로지에 클래스로 정의하고, 워드넷에 정의된 어휘의 의미를 클래스에 기술로써 정의하여 분류체계가 아닌 계층구조를 생성한다. 그리고 기술을 기반으로 관련된 어휘들의 추론된 분류체계를 생성한다. 또한 워드넷에 정의된 상·하위어와 동의어의 집합으로 어휘를 계층구조에 추가하므로 어휘의 중의성이 해결되고, 워드넷에 정의되지 않은 어휘도 온톨로지에 추가하여 확장한다. 반자동 모델링은 온톨로지 설계자의 주관이 개입되지 않은 일관된 어휘 체계의 자동 작성을 돕고, 자동화로 인한 오류를 사용자의 검증을 통해 해결한다. 이 방법은 새로운 어휘가 추가되어도 계층구조에 삽입이 쉬우며, 어노테이션 시에 새로운 알고리즘을 적용할 필요가 없이 온톨로지 분류체계를 기반으로 자동화된 어노테이션을 지원한다.

3. VOD 어노테이션 위한 반자동 콘텐츠 온톨로지 모델링

본 연구는 IPTV의 지능형 검색을 위해서 VOD 도메인(또는 장르)별 콘텐츠에 대한 온톨로지를 반자동 구축하는 방법을 제안한다. 그리고 생성한 온톨로지를 기반으로 VOD 어노테이션을 자동화하는 방법을 보인다. 제안하는 방법은 콘텐츠 어휘 온톨로지 반자동 모델링과 VOD 어노테이션으로 구성된다. 그림 1은 본 연구에서 제안하는 시스템 구조이다. 시스템 구조를 간단히 살펴보면, 콘텐츠 온톨로지는 콘텐츠별 관련된 어휘에 매치되는 부분 트리를 워드넷에서 추출하여 클래스로 기술되고, 정의되지 않은 어휘는 의미를 확장하여 콘텐츠 클래스를 확장한다. 어노테이션은 새로운 VOD 데이터가 입력되었을 때 콘텐츠 온톨로지를 기반으로 콘텐츠를 추론한다. 3.1과 3.2절에서 제안하는 방법을 설명하고, 3.3절에서 이 방법을 이용하여 구축한 시스템을 보인다.

3.1 VOD 콘텐츠 온톨로지

콘텐츠 온톨로지 구축 과정은 네 단계로 이루어진다. 각 장르의 시놉시스(학습 데이터)로부터 콘텐츠 키워드 추출, 키워드와 워드넷의 연결, 워드넷의 부분 트리를 온톨로지에 표현, 그리고 워드넷에 정의되지 않은 키워드를 위한 앵커링이다.

3.1.1 VOD 장르, 콘텐츠, 콘텐츠 키워드의 정의

본 연구에서는 VOD 정보를 세 가지 특성으로 구분한다. VOD 장르, 콘텐츠, 콘텐츠 키워드이다. 장르 G는 VOD의 대분류로써, 예를 들면 영화 VOD는 SF, 코미디, 전쟁, 판타지, 미스터리 등으로 나눌 수 있다. 콘텐츠 C는 멀티미디어 데이터의 특징을 파악하는데 바탕이 되는 재료로써, 장소, 시대, 등장인물, 행동, 감정 따위가 모두 콘텐츠가 될 수 있다. 콘텐츠 키워드 K는

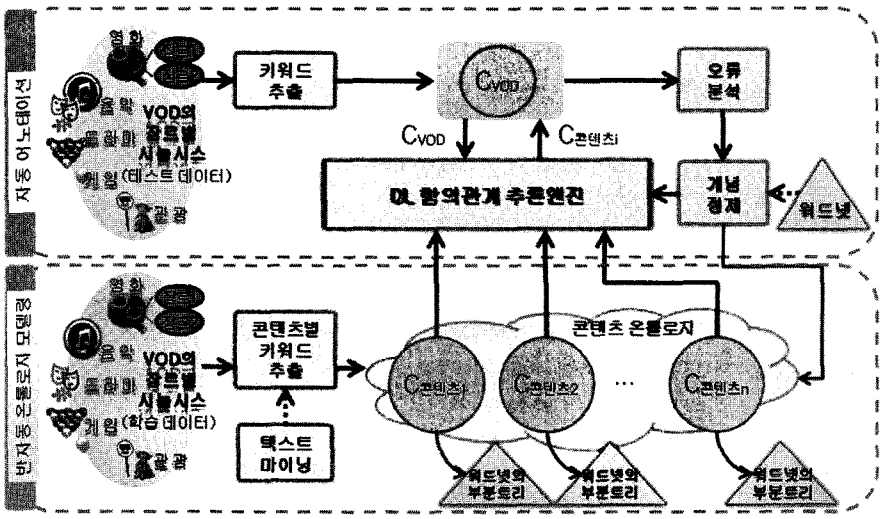

그림 1 시스템 구조

콘텐츠를 유추할 수 있는 어휘들의 집합이다. 영화 VOD의 SF 장르를 예로 들어 보면, 콘텐츠는 우주, 재난, 로봇, 괴수 등이 될 수 있으며, 이 중 콘텐츠 '우주'의 콘텐츠 키워드는 지구, 외계, 우주선, 행성 등이 될 수 있다. 장르는 현재 영화 또는 드라마 웹 서비스에서 통상적으로 제공되는 분류이다. 하지만 콘텐츠에 대한 검색을 제공하는 일반적인 서비스는 없다. 따라서 본 연구에서 제안하는 방법은 IPTV 뿐만 아니라 콘텐츠 기반 검색이 필요한 개인 미디어 관리 시스템에도 활용되어 사용자 자신이 사용하는 어휘로 개인 미디어 메타데이터를 생성하고 자신의 어휘를 사용하여 질의하도록 지원 할 수 있다.

3.1.2 온톨로지를 이용한 VOD 장르, 콘텐츠, 콘텐츠 키워드의 정의

본 논문에서는 온톨로지 Ot의 Cl은 클래스(class)의 집합, P는 속성 집합으로 표현한다. 온톨로지 Ot는 $Ot=\{cl_1, cl_2, \cdots, cl_n\}$, 각 클래스 $cl_1, cl_2, \cdots, cl_n$의 집합으로 표현한다.

$$G=\{g \mid g\in G, g\in Cl, Cl\in Ot\} \quad (1)$$

$$C=\{c \mid c\in C, c\in Cl, Cl\in Ot, c\subseteq G\} \quad (2)$$

$$K=\{k \mid k\in K, k\in Cl, Cl\in Ot, k\in S\} \quad (3)$$

장르, 콘텐츠, 콘텐츠 키워드는 의미기반 계층구조로 정의하기 위해 클래스로 정의한다. 장르 G는 (식 (1)), 콘텐츠 C는 장르의 하위 클래스로 정의한다(식 (2)). 콘텐츠 키워드 K는 VOD 시놉시스(synopsis) S에 포함된 단어의 집합이다(식 (3)).

3.1.3 각 장르의 학습 데이터로부터 키워드 추출

콘텐츠 온톨로지 구축을 위해서는 VOD별 장르를 구분하고 장르별 콘텐츠를 추출하는 과정이 필요하다. 이 자연언어 처리 과정은 언어 전문가가 개입하는 것이 바람직하다. 하지만 전문가가 작성하여 활용 가능한 콘텐츠별 분류체계는 없고, 있다고 해도 전문가에 따라 주관이 개입되어 분류 체계가 일관적이지 못하다. 본 논문에서는 전문가의 지식수준을 대신하기 위해 정확도와 신뢰도가 높은 텍스트 마이닝 기법을 사용한다.

학습 데이터로는 장르별 VOD 시놉시스를 수집한다. 그리고 시놉시스에 포함된 단어들 간의 군집화 실험을 하여 군집에서 대표 단어들을 추출하여 키워드 집합을 만든다. 다음으로 그 집합을 특징지을 수 있는 단어를 선정하여 콘텐츠로 정의한다. 텍스트 마이닝에 사용한 K-means 클러스터링 알고리즘[9]은 군집 영역에 속하는 모든 점으로부터 군집 중심까지의 거리의 제곱의 합으로 정의되는 성능 지표를 최소화하는데 바탕을 둔 방법이다. K-means 알고리즘에 다음과 같은 성능 개선 방법을 적용하여 중심벡터를 구한다.

① 초기 중심벡터를 선정 : 중심벡터 위치에 따라 결과가 크게 좌우되지 않도록 한다.
- 랜덤하게 하나의 중심벡터를 선정한다.
- 바로 전 단계에서 선정된 중심벡터에서 가장 멀리 떨어진 노드를 다음 중심벡터로 선정한다.
- k개의 중심벡터가 선정될 때 까지 반복한다.

② 단어 필터링 : 특징 벡터 생성 시 변별력 있는 단어들을 추출한다.
- DF(Document Frequency)가 높은 상위 5% 단어와 DF가 낮은 하위 5%의 단어를 제거한다.
- 한 글자로 이루어진 단어를 제거한다.
- 기타 불용어 목록에 포함된 단어를 제거한다.

③ 노드 평준화 : 하나의 노드에 의해 대표단어가 좌우되지 않도록 한다.
- 클러스터에 스케일이 큰 이상값이 포함될 경우 이상값에 의해 대표단어가 좌우될 수 있으므로 문서의 스케일을 다음 식으로 평준화하여 해결한다.

$$N_{new} = N/(\log(|N|)+1)$$

다음으로 각 클러스터의 중심벡터에서 가장 높은 값을 갖는 상위 10개의 단어를 클러스터의 대표단어, 즉 콘텐츠 키워드로 추출한다. 그리고 대표단어들을 포괄하는 단어를 콘텐츠로 정의한다. 이 전처리 작업이 끝나면 선정 키워드들의 의미를 분석하는 과정이 필요하다. 이 과정은 다음 절에서 살펴본다.

3.1.4 키워드와 워드넷 어휘의 연결

본 연구는 한국 VOD 데이터를 대상으로 연구하기 위해 워드넷과 함께 코렉스(KorLex)[10]를 사용한다. 3.1.3절에서 추출한 콘텐츠와 콘텐츠 키워드는 3.1.2절에서 언급한 방법으로 온톨로지에 클래스로 정의된다. 다음으로 콘텐츠 키워드는 워드넷에 정의된 어휘로 매치되어 그 어휘의 상·하위어와 동의어의 부분 트리를 가져와서 온톨로지에 클래스로 추가하는 과정을 반복한다. 하지만 워드넷은 영어 어휘사전이고, 코렉스는 이를 번역해놓았기 때문에 한국어 어휘를 모두 포함하지 않으며 언어 체계 역시 다르다. 따라서 본 논문에서는 이를 고려하여 콘텐츠 키워드가 워드넷에 정의된 경우와 워드넷에 정의되지 않은 경우로 나누어 처리한다. 워드넷의 부분 트리를 온톨로지로 표현하기 위해서는 워드넷의 의미 관계를 온톨로지의 공리에 대입해야한다. 키워드 클래스를 기준으로 상위어는 상위 클래스, 하위어는 하위 클래스, 동의어는 등가(equivalent) 클래스로 정의한다. 다음으로 3.1.4.1절과 3.1.4.2절에서 위에서 언급한 두 가지 경우에 대한 온톨로지 표현 방법을 각각 설명한다.

3.1.4.1 워드넷의 부분 트리를 온톨로지에 표현

온톨로지 지식 베이스는 TBox와 ABox로 구분된다. TBox는 어플리케이션 도메인의 어휘를 정의하며, ABox

는 TBox 어휘의 용어로 구성된다. 본 연구의 콘텐츠 어휘는 TBox에 정의하며, TBox 기술 구축과 TBox 추론의 두 단계로 구성된다(그림 2). 기술은 클래스에 대한 클래스명, 제약조건(⊑, ∃, ∀ 등), 컨스트럭터(constructor, ⊓, ⊔, ¬ 등), 인스턴스 집합을 포함한다[11]. 각 단계는 서로 다른 어휘 의미망을 형성한다. 보편적 계층구조와 특정 계층구조이다.

그림 2의 TBox 기술 구축은 TBox 기술 구축 알고리즘을 이용하여 보편적 계층구조를 생성한다. 생성 과정을 살펴보면, ① 콘텐츠 키워드 k_i를 온톨로지에 클래스로 정의하고, ② 3.1.3절의 알고리즘을 이용해서 추출한 키워드에 해당하는 콘텐츠 C_k를 도메인으로 하는 속성을 키워드 클래스에 존재양화사 기술로 추가한다. 그리고 ③ 워드넷에서 키워드 k_i에 대응하는 어휘의 부분 트리를 가져온다. 이 구조는 다른 응용 도메인에도 적용할 수 있는 어휘 계층구조이다. 다음으로 TBox 추론은 특정 계층구조를 생성한다. TBox 기술 구축을 통해 정의된 클래스 사이의 암시적인 함의관계 구조를 명시적으로 하여 어휘 계층구조를 확장한다. 즉, ④ 키워드 클래스를 키워드와 연관된 콘텐츠로 추론하고, 역으로 콘텐츠 클래스는 관련된 어휘들의 계층구조를 하위 클래스 계층으로 포함한다. 이 구조는 해당 VOD 도메인에만 종속된 체계이며, 추론 기능을 이용한 의미검색을 가능하게 한다. 즉, 사용자가 검색하고자 하는 VOD의 질의 q, 콘텐츠 C_1의 키워드가 $\{k_1, k_2, k_3, k_4, k_5\}$가 있을 때, $q \notin C_1$이더라도 q가 의미적으로 C_1의 멤버가 되면 콘텐츠 C_1을 반환한다. 따라서 질의가 키워드에 정확하게 매치 되지 않아도 시맨틱 검색을 통해 의미를 확장하여 근접한 콘텐츠의 VOD를 추천해 줄 수 있다.

TBox 기술 구축 알고리즘은 우선 3.1.2에서 언급한 대로 장르, 콘텐츠, 그리고 키워드를 생성한다. 다음으로 콘텐츠와 콘텐츠 키워드에 제약사항을 추가한다. 콘텐츠 C'는 C에 속성 도메인 제약[12] 기술을 추가한다(식 (4)). 즉, 각 콘텐츠 c_k에 자신을 도메인으로 갖는 속성 p_o를 정의한다. 키워드 K'는 K에 자신이 대표하는 콘텐츠의 속성에 대한 존재 양화사 기술을 추가한다(식 (5)). 즉, 3.1.3절에서 추출한 콘텐츠 키워드 k_i는 자신이 대표하는 콘텐츠 클래스의 속성 p_o를 하나 이상 갖도록 존재 양화사 기술을 추가한다. 그리고 워드넷으로부터 키워드의 상·하위어, 동의어를 3.1.4절에서 언급한 방법으로 추가하여 각 콘텐츠를 워드넷의 부분 트리로 표현한다. 그림 2는 이 과정을 도식화한 것이다.

$$C' = \{ c', p \mid c' \in C',\ c' \in Cl,\ Cl \in Ot,\ c' \subseteq g,\ Domain(p, c'), p \in P \} \quad (4)$$

$$K' = \{ k', p \mid k' \in K',\ k' \in Cl,\ Cl \in Ot,\ k' \in S,\ \exists p.T \subseteq K' \} \quad (5)$$

부분트리 추출 범위는 상위 어휘로 올라갈수록 너무 포괄적인 개념이어서 시놉시스에 포함될 확률이 낮고 콘텐츠 추론의 정확도를 떨어뜨리므로 상위어는 키워드의 바로 위의 어휘만 추출한다. 하위어도 바로 밑의 어휘만 추출하지만, 키워드를 구체화 하므로 다중 선택을 허용한다. 여기서 추가로 고려해야 할 점은 키워드, 상위어, 하위어 또는 동의어가 이미 온톨로지에 정의되어 있는 경우이다. 따라서 위에서 설명한 TBox 기술 구축 알고리즘에 표 1의 키워드 부분트리 구축 알고리즘을 추가한다.

표 1의 확장된 알고리즘에서 하나의 경우만 살펴보자. 키워드가 온톨로지에 정의된 경우, 키워드 k_i의 부분 트리를 확장한다. 먼저 k_i의 상위어 w_{sp}의 상위어 집합 W_{sp2}와 하위어 w_{sb}의 하위어 집합 W_{sb2}를 추출하여 k'_i를 만든다. 다음으로 k'_i의 계층구조에 함의되는 클래스 k^+가 온톨로지에 이미 정의되어 있으면, k_i 클래스를 새로 정의하지 않고 k^+에 k_i의 기술을 추가하여 수

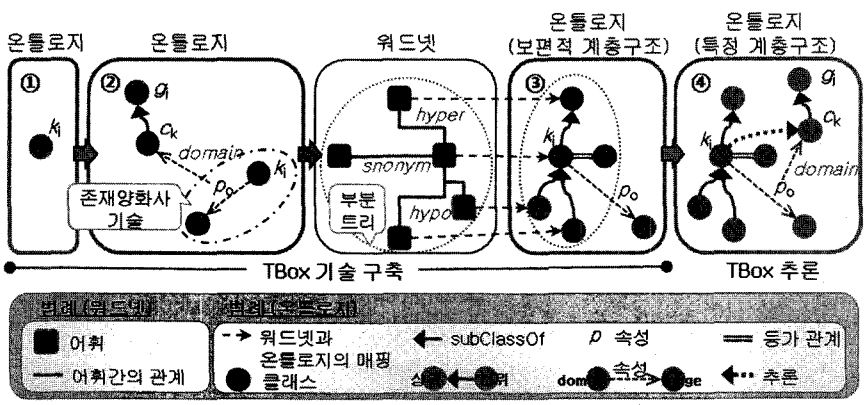

그림 2 콘텐츠 계층구조 구축 과정

표 1 TBox 기술 구축 알고리즘 확장

① 키워드가 온톨로지에 정의된 경우
redefine $k'_i = \{ W_{sb2} \subseteq w_{sb} \subseteq k_i \subseteq w_{sp} \subseteq W_{sp2} \}$
if $k^+ \subseteq k'_i$, $k^+ \in Ot$ then $k^+ = k^+ \cup k_i$
② 키워드의 상위어가 온톨로지에 정의된 경우
redefine $w'_{sp} = \{ W_{sb3} \subseteq w_{sp} \subseteq W_{sp3} \}$
if $w_{sp}^+ \subseteq w'_{sp}$, $w_{sp}^+ \in Ot$ then $w_{sp}^+ = w_{sp}^+ \cup w_{sp}$
③ 키워드의 하위어가 온톨로지에 정의된 경우
redefine $w'_{sb} = \{ W_{sb4} \subseteq w_{sb} \subseteq W_{sp4} \}$
if $w_{sb}^+ \subseteq w'_{sb}$, $w_{sb}^+ \in Ot$ then $w_{sb}^+ = w_{sb}^+ \cup w_{sb}$
④ 키워드의 동의어가 온톨로지에 정의된 경우
redefine $w'_s = \{ W_{sb5} \subseteq w_s \subseteq W_{sp5} \}$
if $w_s^+ \subseteq w'_s$, $w_s^+ \in Ot$ then $w_s^+ = w_s^+ \cup w_s$

정한다. 반면에 키워드 어휘는 같지만 표 1의 조건을 만족하지 않는다면, 다의어로 간주하고 k_i를 새로운 클래스로 정의한다. 키워드의 상위어, 하위어, 그리고 동의어에 대해서도 동일한 처리를 한다. 표 1은 동음이의어에 대해 다른 부분트리를 생성하므로 이 알고리즘은 어휘의 중의성을 고려한다.

TBox 추론은 TBox 기술에 온톨로지 함의관계 추론을 적용하여 추론된 VOD 분류 체계(classification)를 보여준다. 함의관계란 클래스 C(subsumee)가 클래스 D(subsumer)에 포함되는지를 검사하는데 사용된다. 즉, 모든 interpretation I에 대해 $C \sqsubseteq D \; iff \; C^I \subseteq D^I$이다. 분류 체계는 단일 클래스들의 부분적인 함의관계와 계층구조를 추론한다. 이 방법은 서술논리 추론 시스템에 의해 제공된다[13].

TBox 기술 구축을 통해 얻은 추론된 VOD 분류 체계는 VOD 콘텐츠와 콘텐츠 키워드에 기술을 추가한 계층구조이며, 명시적으로 정의한 장르에 속하는 콘텐츠의 함의관계를 보인다. TBox 추론을 통해 얻은 분류 체계는 암시적인 콘텐츠와 콘텐츠 키워드의 함의관계를 통해 콘텐츠에 포함되는 콘텐츠 키워드를 추론한다. 이 방법은 일반적인 어휘 체계를 기반으로 서비스 도메인에 필요한 어휘 체계를 구축할 수 있게 한다(그림 2).

3.1.4.2 워드넷에 정의되지 않은 키워드를 위한 앵커링

앵커링이란 한 자원으로부터 다른 자원에 연결하는 것을 의미한다. 워드넷은 영어 어휘를 기반으로 하므로 한국어 고유명사나 한자어, 그리고 외래어 등은 정의되어 있지 않다. 본 논문은 VOD 도메인에 적합한 어휘 체계의 부분 트리를 온톨로지로 구축하기 위해 워드넷에 정의되지 않은 중요 키워드에 대해서는 그림 3의 과정을 통해 VOD 온톨로지를 확장한다. 워드넷에 정의

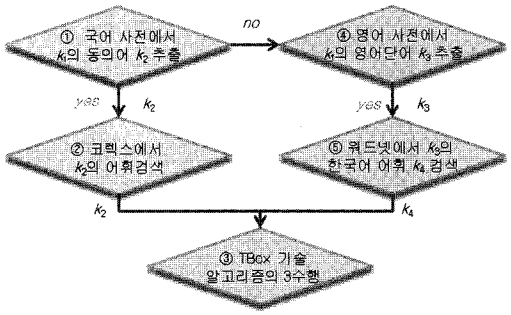

그림 3 워드넷에 정의되지 않은 키워드의 이음동의어 추출 과정

되지 않은 키워드 k_1은 워드넷에 정의된 키워드의 이음동의어와 연결하여 앵커링한다. 예를 들어, '초능력'은 코렉스에서 검색되지 않는다. 그림 3의 과정을 수행하면, ① '초능력'의 동의어 '텔레파시'를 추출, ② '텔레파시' 어휘 검색, ③ TBox 기술 구축 알고리즘을 수행한다. 또 다른 예로 '야생'은 ① '야생'의 동의어를 검색 못함, ④ 'wildness' 검색, ⑤ '맹렬함' 검색, ③ TBox 기술 알고리즘을 수행한다.

3.2 VOD의 콘텐츠 어노테이션

본 장에서는 구축된 콘텐츠 온톨로지를 이용하여 VOD 어노테이션이 자동으로 되는 방법을 설명한다. VOD 어노테이션은 콘텐츠 온톨로지를 기반으로 수행되며, 입력은 VOD 시놉시스이고 출력은 콘텐츠 클래스이다. 어노테이션 과정은 다음과 같다.

① VOD 시놉시스 S가 입력되면, 시놉시스로부터 온톨로지에 정의된 키워드의 집합 K를 추출한다.

② 콘텐츠 온톨로지에 익명의 클래스 A를 생성하고, 추출된 키워드들을 이 클래스의 기술로 생성한다. $A \subseteq k_1 \cap k_2 \cap \ldots \cap k_j$

③ 온톨로지 함의관계 추론을 이용해서 클래스 A의 상위 클래스로 추론되는 콘텐츠 클래스의 집합 M을 구한다.

추론된 콘텐츠 클래스는 이 VOD의 콘텐츠가 되고, 콘텐츠의 장르가 이 VOD의 장르가 된다. 이와 같은 장르와 콘텐츠는 이 VOD의 메타데이터가 된다. 그림 4는 어노테이션의 예로써 영화 VOD의 시놉시스가 입력되었을 때, 콘텐츠 온톨로지를 기반으로 콘텐츠가 추출되는 원리를 그래프 형태로 보여준다. 이 예는 학습데이터로 사용되지 않은 최신 개봉 영화 '국가대표'의 콘텐츠로 '스포츠'와 '우정'을 어노테이션 예이다. 온톨로지 추론을 이용한 시맨틱 매칭이 되는 부분만 설명한다. 우선, '대표', '경기', '올림픽', '코치' 등의 키워드는 TBox 추론에 의해 '드라마' 장르의 '스포츠' 콘텐츠로 추론된다. (1)

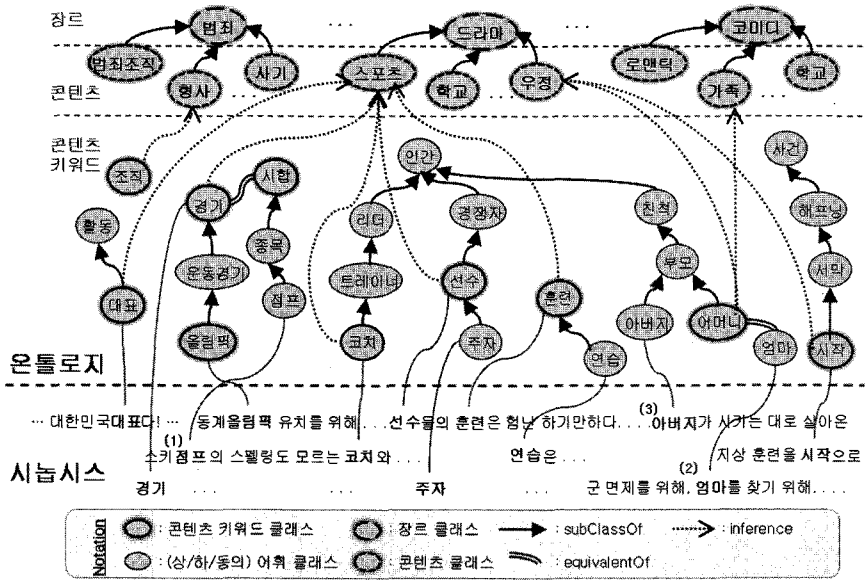

그림 4 VOD의 콘텐츠 어노테이션의 예

'점프'는 함의관계 추론에 의해 '스포츠'의 키워드의 '시합'의 하위어로 매칭된다. (2) '엄마'는 '어머니'와 동의어로 추론되며, (3) '아버지'는 '어머니'의 형제 클래스로 매칭되어 '우정' 콘텐츠로 분류된다. 추출된 콘텐츠 중 가중치가 가장 높은 콘텐츠가 메타데이터로 생성된다.

3.3 응용 프로그램을 활용한 VOD 온톨로지 모델링 및 어노테이션 방법

본 장에서는 제안한 방법을 검증하기 위해 구축한 응용 프로그램을 설명한다.

3.3.1 반자동 VOD 콘텐츠 온톨로지 모델링 결과

그림 5는 제안한 방법을 통해 구축한 온톨로지 모델링(①) 및 어노테이션(②) 도구이다. 이 도구는 두 가지 기능을 제공한다. 온톨로지 반자동 모델링과 어노테이션이다. 우선, 이 절에서는 온톨로지 반자동 모델링에 대

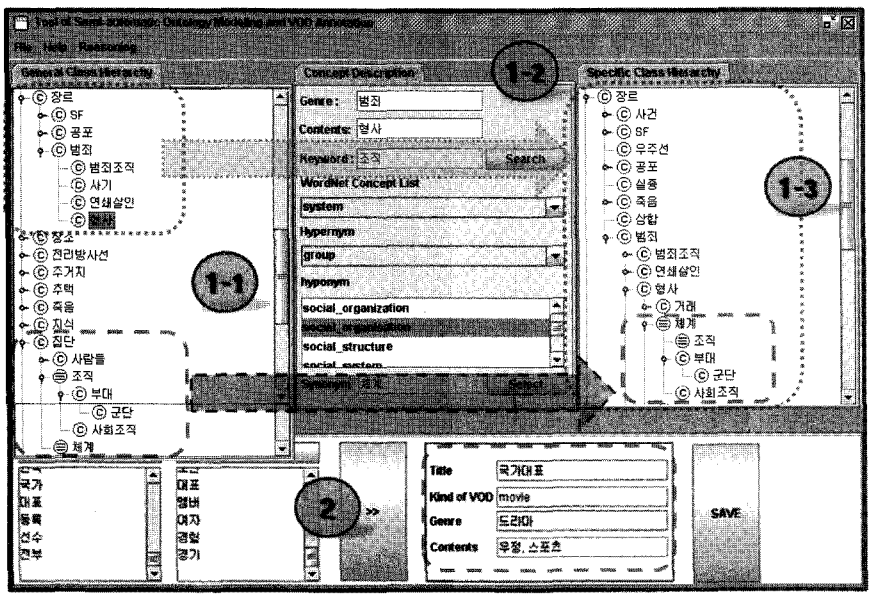

그림 5 VOD 온톨로지 반자동 구축 및 VOD 어노테이션의 예

해 살펴본다. 온톨로지 모델링은 두 가지 계층구조를 생성한다. 보편적 계층구조와 특정 계층구조이다. 보편적 계층구조는 서비스 도메인에 독립적인, 즉 어느 서비스에서나 활용될 수 있고, 누구나 동의할 수 있는 보편적인 어휘 분류 체계를 보여준다(그림 5의 ①-1). 특정한 계층구조는 온톨로지 추론 방법에 의해 형성된 어휘 계층구조로 서비스 도메인에 종속적이다(그림 5의 ①-3).

보편적 계층구조는 그림 5의 ①-2의 편집기를 통해서 키워드의 부분 트리를 추가한다. 이 때 3.1.4절에서 언급한 TBox 기술 구축 알고리즘이 적용된다. 예를 보면, 설계자가 자동으로 추출된 어휘 중 '조직'의 상위어로 '집단', 하위어로 '사회조직'과 '부대', 동의어로 '체계'를 선택함에 따라 '조직' 클래스의 기술은 다음과 같이 보편적 계층구조에 추가된다. 조직 $\subseteq$ 집단 $\cap \exists$ *hasKey_7_2* some *out* : *Thing*, 조직 $\equiv$ 체계

그림 5의 ①-3은 보편적 계층구조에 온톨로지 추론을 적용하여 특정 계층구조를 출력한 화면이다. 특정 계층구조는 함의관계 추론을 이용하여 콘텐츠 키워드를 키워드의 기술이 함의되는 콘텐츠 클래스의 하위 개념으로 분류한다. 위의 '조직' 클래스의 기술을 보면, '조직'은 '체계'와 같고, hasKey_7_2의 속성을 가진다. hasKey_7_2의 도메인은 '형사' 이므로 그림 5의 ①-3에서 '조직'은 '형사'의 하위어이고 '조직'과 등가 개념으로 추론된 것을 볼 수 있다. 결론적으로 특정 계층구조 ①-3은 콘텐츠에 속하는 콘텐츠 키워드를 모두 추론하여 보여준다. 이 구조는 영화 VOD 도메인에 종속적이다.

3.3.2 VOD 어노테이션의 구축 결과

이 절에서는 3.3.1절의 온톨로지 모델링을 기초로 어노테이션된 결과를 살펴본다. 그림 5의 ②는 그림 4에서 살펴본 예제 VOD가 어노테이션된 결과를 보여준다. '국가대표'의 콘텐츠로 '우정'과 '스포츠'가 추출된 것을 볼 수 있다.

4. 실험 및 평가

제안한 온톨로지 반자동 모델링 방법은 워드넷의 부분 트리를 활용하여 응용 도메인에 사용되는 콘텐츠 어휘의 의미체계를 온톨로지로 구축하고, 구축된 온톨로지에 추론 기법을 적용하여 의미적으로 가장 적합한 콘텐츠를 추출하여 VOD의 메타데이터를 생성한다. 제안한 방법의 정확도 검증을 위해 학습 데이터에 포함되지 않은 최신 개봉 영화를 대상으로 이 영화의 콘텐츠를 사용자들에게 선택하게 하여 구축한 시스템의 메타데이터 결과와 비교 분석하였다.

4.1 실험 데이터

실험 도메인으로는 대부분의 웹 사이트에서 동일한 장르체계를 가지고, 사람들이 선호하는 장르가 분명한 영화 VOD를 선택하였다. 콘텐츠의 키워드 추출을 위한 학습 데이터로는 영화의 각 장르 별 VOD의 시놉시스를 수집하였다. 3.1.3절에서 언급한 콘텐츠 키워드 추출 방법에 적용한 학습 데이터는 총 12개의 장르, SF(1,271건), 공포(1,861건), 범죄(1,290건), 어드벤처(1,296건), 코미디(5,861건), 전쟁(771건), 멜로(3,031건), 스릴러(1,152건), 판타지(1,021건), 미스터리(546건), 드라마(3,836건), 액션(1,321건)이다. 이 학습 데이터를 3.1절에서 언급한 방법을 통해 반자동 온톨로지 모델링한 결과, 클래스 778개, 속성 71개, subClassOf 관계는 1,262개, 등가 클래스 관계는 107개를 포함하는 서술논리 AL의 표현력을 가지는 온톨로지를 구축하였다. 메타데이터 생성에 사용된 테스트 데이터는 학습 데이터에 포함되지 않은 최신 영화로 각 장르별 2개의 샘플 데이터를 선택하여 총 24편의 영화를 추출하였다.

4.2 성능 평가 방법

본 시스템의 성능 평가 방법은 정답 평가 담당자간의 견해 차이에 따라 정답 기준이 다 다르므로 샘플 데이터로 추출한 24건의 영화에 대해 100명에게 영화별 콘텐츠를 선택하게 하여 정답 집합을 만들었다. 그리고 실험 결과로 생성된 메타데이터(콘텐츠)와 비교하여 시스템을 평가하였다. 평가방법은 실험 대상 영화의 메타데이터에 대하여 정확률, 재현율, 그리고 F-measure를 계산한다. 정확률 P는 시스템이 추출한 콘텐츠(CE) 중에서 몇 개가 사용자도 그 영화의 콘텐츠라고 선택(AE: 맞은 수)했는지 비율을 계산하며 식 6과 같다. 재현율 R은 사용자가 만든 정답(C) 중 본 시스템이 맞춘 개수(AE)를 평가하며 식 (7)과 같다. F-measure는 정확률과 재현율에 동등한 중요도를 부여하기 위해 식 (8)과 같이 구한다.

$$P = \frac{AE}{CE} \tag{6}$$

$$R = \frac{AE}{C} \tag{7}$$

$$F = \frac{2RP}{P+R} \tag{8}$$

4.3 실험 결과

그림 6은 본 논문에서 제안한 방법을 평가하기 위해 어노테이션 결과의 정확도를 나타낸 그래프이다. 그림 6의 결과를 보면 정확률 82%, 재현율 72%, F-measure 71%의 우수한 성능을 보임을 알 수 있다. 정확률은 높지만 비교적 재현율이 낮은 이유는 VOD 도메인의 특성상 사람들이 생각하는 VOD 콘텐츠를 모두 추출할 수 없기 때문이다. 즉, 본 시스템은 가장 적합하다고 생각하는 상위 몇 개의 콘텐츠를 추출하고, 사람들은 더 폭넓은 콘텐츠를 정답으로 원할 수 있게 때문이다.

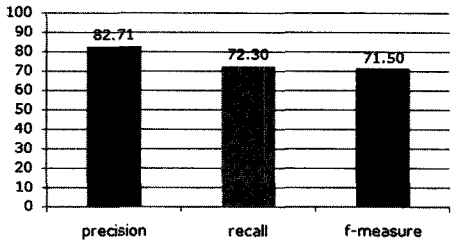

그림 6 실험 평가 결과

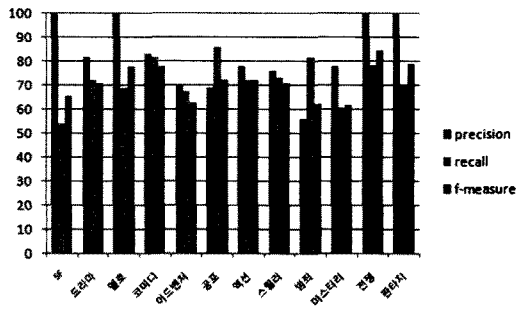

그림 7 장르별 어노테이션 실험 결과

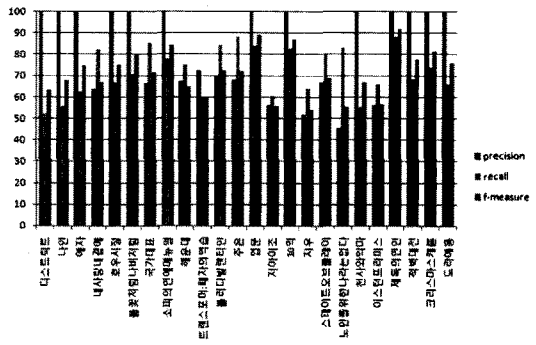

그림 8 영화별 어노테이션 실험 결과

그림 7은 장르별 어노테이션 실험 평가 결과를 보여준다. 이 결과는 시놉시스에 포함된 키워드들이 유일 할수록 정확도가 높게 나왔다. 어드밴처와 범죄는 콘텐츠가 다양하므로 정확도가 대체적으로 낮게 나왔으며, 멜로는 해당 콘텐츠가 '섹스', '댄스', '음악', '결혼' 등으로 폭이 넓어서 코미디의 '로맨틱' 콘텐츠를 유사 콘텐츠로 처리하였고, 따라서 높은 정확도를 보여준다.

그림 8은 영화별로 살펴본 어노테이션 결과이다. 그림 4에서 예제로 살펴본 '국가대표'의 결과를 살펴보면, 재현율이 정확도에 비해 많은 차이로 높은 것을 볼 수 있다. 이 이유는 제안한 시스템은 우선순위가 동등하게 높게 나온 '스포츠'와 '우정' 콘텐츠로 '국가대표'를 분류하지만, 대부분의 사용자는 '스포츠'만 정답으로 선택하였기 때문이다. 이와 같이 정답 집합이 시스템이 추출한 콘텐츠의 부분 집합이 되는 VOD는 비슷한 분포를 보여준다.

결론적으로 어노테이션 실험 평가 결과는 '축구', '슬래셔', '우주' 등과 같이 콘텐츠가 분명 할수록 정확도가 높았다. 온톨로지 모델링과 어노테이션은 사람들이 사용하는 어휘를 다루는 분야이기 때문에 사람들의 지식과 주관이 개입되어 변수가 많이 발생한다. 하지만 본 논문에서 제안한 방법은 워드넷의 어휘 분류체계를 기반으로 어휘의 동음이의어와 이음동의어를 고려하여 의미기반의 어휘 온톨로지를 모델링한다. 이 모델은 서비스 도메인에 따른 계층구조를 생성하며 이를 기반으로 어노테이션 하기 때문에 사람들에게 더 신뢰도가 높은 결과를 얻을 수가 있었다.

5. 결론 및 향후 연구

본 논문은 검증된 어휘 분류체계인 워드넷에서 부분 트리를 추출하여 서비스 도메인에 적합한 온톨로지를 모델링하는 방법을 제안하였다. 제안한 방법은 지능형 검색을 위한 데이터 어노테이션을 목표로 설계되었다. 본 논문에서는 IPTV의 지능형 검색을 위해 VOD에 어노테이션 하는 것을 목표로 서술되었지만, 의미검색 서비스의 모든 도메인에 적용될 수 있으며 구축한 온톨로지를 다른 도메인에 적용할 수도 있는 일반적인 접근법을 제안하였다.

본 연구에서 온톨로지 구축과 메타데이터 생성을 반자동으로 한 이유는 모든 어휘는 의미 중의성이 존재하기 때문이다. 그래서 본 연구에서는 워드넷을 이용하여 상·하위어의 매칭을 통해 이 문제를 고려하였다. 향후 연구로는 더 지능적인 의미 모호성 알고리즘을 이용한 온톨로지 구축과 어노테이션 도구를 연구하고자 한다. 또한 온톨로지에 정의된 개념들에 확률을 적용하여 어노테이션 결과의 정확성을 향상시킬 수 있는 방법의 개선이 필요하다.

참 고 문 헌

[1] B. A. Emilio, "Real-Time Metadata for IPTV Systems," *Network Division*, NEC Labs Europe, 2007.

[2] A. Sanfilippo, S. Tratz, M. Gregory, A. Chappell, P. Whitney, C. Posse, P. Paulson, B. Baddeley, R. Hohimer, A. White, "Automating Ontological Annotation with WordNet," *In Proc. of the Third International Global WordNet Conference (GWC-06)*, pp.85-93, Jeju Island, South Korea, January 2006.

[3] V. Snášel, P. Moravec, J. Pokorný, "WordNet Ontology Based Model for Web Retrieval," *In Proc. of WIRI'05 Workshop*, Tokyo, Japan, 2005.

IEEE Press.
[4] C. Fellbaum, "WordNet: An Electronic Lexical Database," *MIT Press*, 1998. http://wordnet.princeton.edu.
[5] L. De Silva and L. Jayaratne, "Semi-automatic extraction and modeling of ontologies using Wikipedia XML Corpus," *In Proc. of the Applications of Digital Information and Web Technologies (ICADIWT 2009)*, pp.446-451, 2009.
[6] S. N. Lee, S. Y. Huh, R. D. McNiel, "Automatic generation of concept hierarchies using WordNet," *Expert Systems with Applications: An International Journal*, vol.35, no.3, pp.1132-1144, 2008.
[7] L. Denoyer and P. Gallinari, "The Wikipedia XML Corpus," SIGIR Forum, 2006.
[8] M. Laclavik, M. Seleng, and M. Babik, "OnTeA: Semi-automatic Ontology based Text Annotation Method," *In Proc. of the Tools for Acquisition, Organisation and Presenting of Information and Knowledge*, pp.49-63, ISBN 80-227-2468-8, 2006.
[9] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Machine Intelligence*, vol.24. no.7, pp.881-892, 2002.
[10] A. S. Yoon, S. H. Hwang, E. R. Lee, H. C. Kwon, "Construction of Korean Wordnet 「KorLex 1.5」," *Journal of KIISE : Software and Applications*, vol.36, no.1, pp.92-108, Jan. 2009. (in Korean)
[11] M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, "OWL web ontology language reference," *W3C Working Draft*, 31 March 2003. Available at http://www.w3.org/TR/2003/WD-owl-ref-20030331.
[12] D. Tsarkov and I. Horrocks. "Efficient reasoning with range and domain constraints," *In Proc. of the 2004 Description Logic Workshop (DL 2004)*, pp.41-50, 2004.
[13] I. Horrocks, S. Tesssaris, "Querying the Semantic Web: a Formal Approach," *In Proc. of the 2002 International Semantic Web Conference, Lecture Notes in Computer Science*, no 2342, 2002.

최 정 화
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 4 호 참조

허 길
2009년 2월 삼육대학교 컴퓨터과학과 졸업(학사). 2009년 3월~현재 숭실대학교 대학원 컴퓨터학과 석사과정. 관심분야는 시맨틱웹, 상황인지, 온톨로지 추론, 유비쿼터스 컴퓨팅, 개인화 에이전트 등

박 영 택
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 4 호 참조