제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

eBook Annotation 시스템을 위한 온톨로지 설계

김종석*, 고승규*, 임순범**, 최윤철*
*연세대학교 컴퓨터과학과
**숙명여자대학교 멀티미디어학과
e-mail: himang01@rainbow.yonsei.ac.kr

Design of an Ontology for eBook Annotation System

Jong-Suk Kim*, Seung-Kyu Ko*, Soon-Bum Lim**, Yoon-Chul Choy*
*Dept. of Computer Science, Yonsei University
**Dept. of Multimedia Science, Sookmyung Women's University

요 약

본 연구에서는 온라인 다중 사용자 환경의 eBook annotation 시스템 개발에서 데이터를 의미 기반으로 관리하고, 데이터에 대하여 상호 공통적인 이해를 표현하며, 그리고 데이터에 대한 무결성 검사 등을 지원하기 위해서 eBook annotation 온톨로지를 설계하였다. eBook annotation 데이터에 대한 상호 공통적인 이해를 표현을 위해서 한국 전자책 문서 표준인 EBKS(Electronic Book of Korea Standard)를 기반으로 설계 하였으며 설계 된 온톨로지는 Conceptual Graph(CG)를 사용하여 표현하였다. 의미 기반의 처리를 위해서 본 온톨로지에서는 동의어(Synonym) 관계와 다국어(Interlingua) 관계를 고려하였으며 또한 annotation 데이터 생성시 오류 방지와 중요도를 표현 하기 위해서 integrity, important axiom을 고려 했다. 제안된 온톨로지는 annotation 데이터의 재사용성을 높일 수 있고 의미 정보를 활용함으로써 eLearning, cyberclass과 같은 다중 사용자 환경에서 효과적인 협업을 가능하게 한다.

1. 서론

제조업 중심의 산업사회에서 정보화 사회로 넘어가면서 우리의 환경은 디지털화하고 있다. 이러한 디지털화의 환경에 따라 전통적인 정보 전달 매체 중 하나인 책이 디지털화한 것을 전자책(eBook)이라고 한다.

기존의 종이 책에서 독자의 생각을 표현하기 위해서 사용 되었던 annotation은 웹스터 사전에서 "의견이나 설명을 위해 추가되는 노트"라고 정의 되어 있다. 본 논문에서는 전자책 이라는 환경을 고려하여서 annotation을 "원본 문서에 추가되는 모든 정보"로 정의한다. [8]

전자책에서는 기존의 종이책에서 사용하였던 annotation을 적용함으로써 다중 사용자 환경에서 사용자의 생각이나 의견을 추가할 수 있다. 전자책 annotation이 기존의 종이책 annotation과 구별되는 특징[1]은 annotation에 대한 키워드 검색이 가능하고 생성한 annotation에 대한 공유가 가능하다는 점이다.

전자책에서의 annotation 시스템의 구성은 인터페이스 측면과 원본 문서와 annotation 관리 측면으로 나눌 수 있다. Annotation 인터페이스 연구는 다중 사용자 환경에서 효과적으로 annotation을 생성하고 출력하는 것에 관한 것이다. 원본 문서와 annotation 데이터 관리에 관한 연구는 입력된 annotation 데이터를 어떻게 원본 문서와의 관계를 고려하여 저장하고 관리하는지에 대한 연구이다.

본 연구는 annotation 데이터의 관리에 관한 연구로써 annotation 데이터를 관리하는데 있어서 상호 공통적인 이해를 표현 하며, 의미 기반의 데이터 처리와 저장된 데이터 간의 관계에서 데이터 무결성 검사, important등의 추론이 가능 하도록 온톨로지[5]를 설계한다. 이를 위해 전자책에 대한 상호 공통적인 이해를 표현하는 국내 전자책 표준인 EBKS[2]을 기반으로 온톨로지를 설계하였다.

2. 관련 연구

본 절에서는 온톨로지의 의미에 대한 정의와 설계 방법을 살펴 본다. 또한 대표적인 온톨로지 표현 언어인 CG에 대하여 소개한다. 그리고 eBook annotation 온톨로지를 설계하기에 앞서 기존의 개발된 온톨로지와 온톨로지 기반의 annotation 시스템을 비교, 분석하였다.

2.1 온톨로지

정보 시스템에서의 온톨로지는 인공지능의 지식 표현, 지식 베이스, 자연어 처리 등의 분야에서 1990년대 초부터 연구되어 왔다.[6] 인공지능에서 온톨로

지는 특정 도메인의 지식에 대한 개념화의 명시적인 명세라고 말한다.

온톨로지가 로직, 의미 망, 프레임, 생성 규칙 등의 지식 표현 방법이 있음에도 불구하고 최근에 활발히 연구되고 있는 이유는 지식의 재사용을 가능하게 하고, 사람들이나 시스템간에 서로 다르게 이해할 수 있는 정보의 구조에 대하여 공통적인 이해를 갖게 하며, 특정 도메인의 가정이나 사실들을 가시화하여 보여주고, 그리고 암묵적으로 표현되지 않은 도메인 지식을 분석하여 표현하기 때문이다.[4]

온톨로지 설계 방법은 여러 가지가 있을 수 있는데 본 연구에서 사용한 설계 방법[3]은 다음과 같다.

• 온톨로지의 목적과 범위를 설정

온톨로지를 왜 만들며 무엇에 사용하고자 만드는지를 명확히 정한다. 이 온톨로지를 개발 이후에 누가 사용할 것인가 정하는 것도 중요하다.

• 온톨로지 생성

-온톨로지 캡쳐: 범위 안에서 객체와 관계를 추출한다.
-온톨로지 코딩: 추출된 객체와 관계를 자연어나 formal한 언어로 표현한다.
-만약 기존에 개발된 온톨로지가 있다면 통합한다.

• 온톨로지 평가

개발 목적과 범위, 그리고 개발된 온톨로지가 대답할 수 있는 질문들에 대하여 어느 정도의 수준으로 충실하게 만족하고 있는가에 따라서 온톨로지를 평가할 수 있다.

• 온톨로지 문서화

소프트웨어 개발과 유사하게 온톨로지의 사용을 위해서 문서화 작업은 필요하다. 개발된 온톨로지에 대한 문서를 통해서 지식의 재사용과 상호 운용성은 높아질 수 있다.

2.2 기존의 개발된 온톨로지 비교 분석

eBook annotation 온톨로지를 개발하기에 앞서 이미 개발된 온톨로지들을 비교 분석하였다. 비교 분석 대상 온톨로지로는 상식적 지식에 대한 온톨로지인 CYC, 기업 및 상거래를 위한 온톨로지인 TOVE, 그리고 의학 분야의 온톨로지인 UMLS이다. 이들 온톨로지를 비교 분석한 차원은 온톨로지 구성의 일반적인 요소들로서 다음과 같다.[9]

• 영역/일반 온톨로지인가?
• 개념의 분류법(Taxonomy)이 존재하는가?
• 개념 구조 및 개념간의 관계를 제공하는가?
• Axiom을 사용하였는가?

[표 1]은 위의 4가지 기준에 따라 CYC, TOVE, UMLS를 분석한 결과를 보여주고 있다.

온톨로지	영역/일반	개념 분류법	개념 구조	Axiom	표현 언어
CYC	일반	O	O	명시적 사용	F.O.L
TOVE	기업	O	O	명시적 사용	F.O.L
UMLS	의학	O	X	사용 않함	ASN.1

[표1] CYC, TOVE, UMLS 비교 분석

위의 차원에 따라 본 연구에서 개발한 eBook annotation 온톨로지를 평가하면 eBook annotation에 대한 영역 온톨로지이며 EBKS를 중심으로 개념 분류법이 존재하고 개념구조 및 개념간의 관계가 존재하며 axiom이 제공된다.

2.3 Conceptual Graph(CG)

본 연구에서 온톨로지를 표현하기 위해서 CG를 사용 하였다. CG(Conceptual Graph)[7]는 Sowa에 의해 1984년 개발 되었으며 Chen의 ER 다이어그램(Chen, 1981)과 비슷하지만 철학적, 심리학적, 언어학적, 객체 지향의 원리에 기반을 두고 비슈얼한 더욱 진보된 지식 기반 표현 방법이다. CG는 F.O.L.(First Order Logic)에 기반을 두고 있으나 다이어그램을 사용하여서 로직을 보여주고 있다.

• CG의 구성

-AS (Abstract Syntax): AS는 concept, conceptual relation, lambda expression, type, referent, context 등의 CG의 핵심이 되는 구성 요소들을 말한다.
-CGIF (Conceptual Graph Interchange Format): CGIF는 CG를 시스템간의 교환을 위해서 고안된 엄격한 syntax를 갖는 포맷이다. AS에서 설명된 것들과 추가적인 개념들에 대하여 EBNF로 기술하고 있다.
-DF (Display form): DF는 사람간의 CG 교환 시 가독성을 높이기 위해서 개발되었다. 기존의 로직은 온톨로지를 표현하는 데 있어서 의미 정보는 잘 표현할 수 있으나 사람이 읽기 어려워 가독성이 좋지 못하였다. 그러나 CG는 F.O.L에 기반한 의미 정보의 표현력을 갖고 있음과 동시에 DF를 사용하여서 logic을 그래픽하게 보여줌으로써 가독성을 높였다.

-LF (Linear form): LF는 CGIF의 간략화된 표현으로써 사람이 CG를 사용할 때 가독성을 높이기 위해서 개발되었다.

2.4 온톨로지 기반 annotation 시스템

웹 문서에 대한 annotation을 지원 하는 기존의 시스템들도 온톨로지의 사용 목적은 의미 기반으로 데이터를 처리하도록 하기 위함이었다.

• CREAM

HTML 문서에 대하여 annotation 을 하는 데 있어서 온톨로지 기반으로 할 수 있도록 하는 시스템이다. 사용자 인터페이스에서 온톨로지를 그래픽하게 보여 줌으로써 사용자가 쉽게 온톨로지 기반으로 annotation 을 하도록 하고 있다. 웹 기반의 annotation 시스템이기에 온톨로지를 표현하기 위해서 DAML+OIL 을 사용하였다.

• Ontology-based Text Annotation Tools (in KA^2-initiative)

Web 문서에 대하여 온톨로지 기반으로 annotation 을 할 수 있도록 하는 KA^2-initiative 에서 annotation 생성 시 구문 오류 방지와 annotation 이 되는 목적 대상에 대한 정확한 concept 을 선택할 수 있도록 하기 위해서 annotation 다이얼로그를 개발하였다.

3. eBook annotation 온톨로지의 설계

지금까지 개발 된 annotation 시스템에서는 의미 기반의 정보 저장 및 관리, annotation 데이터에 대한 이해의 공통적 표현 등을 지원하지 않았다. 이러한 문제를 해결하고자 eBook annotation 시스템에서 온톨로지를 사용하여 데이터를 표현하고 저장하도록 하였다.

3.1 eBook annotation 온톨로지 설계 목적과 범위 설정

• 목적

-eBook annotation 시스템에서 사용자와 개발자, 개발자와 개발자, 시스템과 시스템간에 상호 이해를 공통적으로 표현하도록 한다.

-eBook annotation 시스템에서 데이터가 다른 언어로 표현 될 때 의미 해석의 차이가 발생하는 것에 대하여 일관성 있는 해석이 가능하도록 한다.

-각각의 언어들은 각 단어와 같은 의미를 갖는 동의어가 있다. 정보 시스템에서도 동의어에 대한 의미적 해석을 지원해야 한다.

-생성되는 annotation 데이터에 대하여 데이터간의 관계 속에서 중요도를 체크할 수 있게 한다.

-생성되는 annotation 데이터에 대하여 미리 설정해 놓은 검사 규칙으로써 데이터 무결성을 검증한다. 이것으로써 불필요한 데이터의 오류가 발생하는 것을 막을 수 있다.

• 범위

-온라인 다중 사용자 환경의 eBook annotation 시스템을 대상으로 한다.

3.2 eBook annotation 온톨로지 생성

• 온톨로지 캡쳐

EBKS에서 사용하는 용어들은 eBook 분야에서 공통적으로 사용하는 용어들이며 서로간의 이해를 공통적으로 표현해 주기에 EBKS를 중심으로 concept과 relation을 추출한다. Annotation에 관한 concept과 relation은 온라인 다중 사용자 환경을 고려하여서 추출하였다.

• 온톨로지 코딩

Formal한 언어인 CG를 사용한다. Formal한 언어를 사용함으로써 생기는 장점[3]은 서로 다른 온톨로지간의 비교가 가능하며 잠재적인 가정들을 명시적으로 드러내고 설계된 온톨로지가 목적과 범위에 적합한지를 평가하는 기준을 제공하는 것이다.

• 기존의 온톨로지 통합

아직 eBook 관련 온톨로지가 없다. 상위 레벨의 온톨로지와의 통합은 본 연구에서는 고려하지 않는다.

3.3 온톨로지 평가

설계한 eBook annoation온톨로지가 목적에 맞는지를 체크하기 위해서 다음과 같은 질문들을 사용하였다.

Q1: eBook annotation 시스템에서 사용되며 생성되는 데이터에 대한 상호 이해의 공통적인 표현을 제공하는가?

Q2: eBook annotation 데이터에 대하여 언어간의 차이를 해결해 주는가?

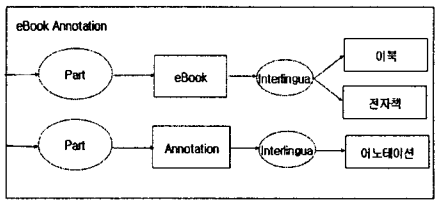
Q3: 정보 시스템이 동의어를 의미적으로 해석하도록 하는가?

Q4: 생성되는 annotation 데이터에 대하여 중요도를 줄 수 있는가?

Q5: 데이터가 잘못 입력되는 경우를 경고하고 사전에 막을 수 있는가?

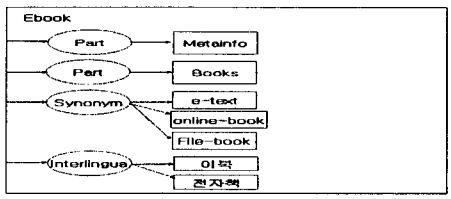
3.4 CG의 DF로 표현된 eBook annotation 온톨로지

[그림 1]은 eBook annotation 온톨로지의 상위 구조이다. 'eBook'이라는 것에 대한 한국어 표현으로써 '이북', '전자책'이 쓰이고 있다. 영어에 대한 한국어 표현의 관계를 사용함으로써 annotation 데이터 검색 등의 사용시 같은 의미의 용어에 대한 언어적 표현의 차이로 인해서 의미 해석의 결과가 다르게 나오는 것을 막을 수 있다.



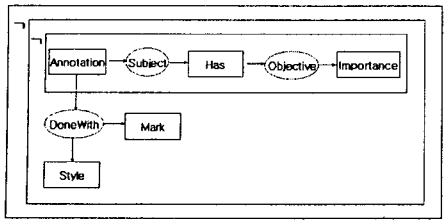
[그림 1] eBook annotation 온톨로지의 상위 구조

[그림 2]는 eBook의 구성에 관한 온톨로지의 한 부분이다. 여기서 우리는 eBook에 대한 동의어 관계를 볼 수 있다. 동의어 관계를 통해서 언어간의 (Interlingua) 관계와 같이 동의어에 대한 의미 해석의 차이가 발생하는 것을 막을 수 있다. eBook의 하위 구성 및 annotation의 구성도 part, 다국어, 동의어 관계를 통해서 이루어 진다.



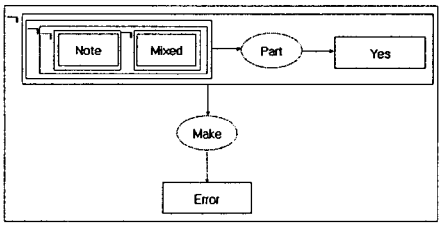
[그림 2] eBook의 구성 요소

[그림 3]은 온톨로지의 axiom 중 important를 보여주고 있다. Annotation이 될 때 mark와 style이 같이 되었을 경우에 그렇지 않은 경우에 비해서 중요성을 갖기에 important라는 axiom이 필요하다.



[그림 3] important라는 axiom

[그림 4]는 데이터 무결성 검사에 대한 axiom을 표현한 것이다. Annotation type이 note나 mixed로 선택되었는데 입력된 텍스트가 없을 때 에러를 발생하게 된다. 이러한 데이터 무결성 검사 axiom은 annotation 인터페이스와의 관계에서 의미 있게 사용된다.



[그림 4] 데이터 무결성 검사 axiom

[표 2]은 위에서 설명한 온톨로지의 part, synonym, interlingua, importance, integrity 관계에 대하여 CGIF로 표현한 것이다. CGIF로 표현함으로써 향후 eBook annotation시스템 개발 시 온톨로지에 대한 정확한 이해를 상호 간에 할 수 있다.

Relation	CGIF
Part	[eBook Annotation *x] [eBook *y] [Annotation *z] (part ?x ?y) (part ?x ?z)
Synonym	[eBook *y] [e-text *y1] [synonym ?y ?y1]
Interlingua	[eBook *y] (interlingua ?y [전자책])
Importance	[if [[Annotation *z] [Style *a] [Mark *b] (donewith ?z ?a ?b)] [then [[Annotation *z] (subject ?z [has]) (object [importance] [has])]]]
Integrity	(make [~[either [or (part [yes] [note])] [or (part [yes] [mixed])]]] [error])

[표 2] part, synonym, interlingua, importance, integrity 관계에 대한 CGIF 표현

4. 결론 및 향후 연구방향

본 연구에서는 eBook annotation 시스템 개발 시 eBook annotation 데이터에 대한 상호 이해의 공통적인 표현, 데이터의 상호 운용, 의미 기반의 데이터 저장 및 관리, 추론을 지원하기 위해서 eBook annotation 온톨로지를 설계하였다. 설계된 eBook annotation 온톨로지는 EBKS를 기반으로 온톨로지를 설계하였기에 상호 이해의 공통적 표현을 지원한다. 데이터 처리 시 동의어와 표현 언어의 차이로 인해 의미가 다르게 해석 되는 것을 방지하기 위해 동의어와 다국어(Interlingua) 관계를 추가하였다. 그리고 데이터 무결성 검사와 annotation입력 시 특정 데이터에 대한 중요성을 반영하기 위해서 axiom을 추가하였다. 설계 된 온톨로지는 의미 기반으로 데이터를 처리하도록 하기 때문에 eLearning, cyberclass과 같은 다중 사용자 환경에서 협업을 가능하게 하며 annotation 데이터의 재사용성을 높일 수 있다.

향후 eBook 온톨로지 설계와 관련하여서 더 연구해야 할 것은 이미 개발된 다른 온톨로지와의 통합으로써 이는 eBook annotation 시스템이 온톨로지를 사용하는 다른 시스템이나 Semantic Web과의 상호 데이터 교환에 필요하다. 특히 Semantic Web과의 연동을 위해서는 RDF(S)나 웹 온톨로지 표현 언어인 WOL(Web Ontology Language)로 본 연구에서 설계된 온톨로지를 표현하는 것이 필요하다.

참고문헌

[1] Ilia A. Ovsiannikov, "Annotation Technology", International Journal of Human-Computer Studies v.50 n.4, 1999

[2] EBK(e-Book of Korea) Consortium, "A Study of Korean Standardization of eBook documents", Technical Report, 2001

[3] Mike Uschold, Ontologies: principle, methods and application, The Knowledge Engineering Review, 1996, 11(2):93-136.

[4] Noy, Natalya Fridman and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology

[5] Maedche, A.; Staab, S., Ontology learning for the Semantic Web, IEEE Intelligent Systems [see also IEEE Expert] , Volume: 16 Issue: 2 , March-April 2001, Page(s): 72 -79

[6] Nicola Guarino, Formal Ontology and Information Systems, Proceedings of the First International Conference (FOIS'98), June 6-8, 1998, Page(s): 3-15

[7] Sowa, John F., ed. (1998) Conceptual Graphs, draft proposed American National Standard, NCITS.T2/98-003.

[8] 고승규, 이현찬, 최윤철, 임순범, "RDF 에 기반한 전자책 Annotation 모델링", HCI 2002 학술발표논문집, 2002

[9] Noy, Natalya Fridman and Carole D. Hafner. 1997. "The State of the Art in Ontology Design: A Survey and Comparative Review". AI Magazine. Fall 1997.