RDBMS 테이블에서의 온톨로지 자동 추출에 관한 연구

A Study on Extracting Ontology from RDBMS Tables

이기정, 황보택근
경원대학교

Lee Ki-Jung, Whangbo Taeg-Keun
Kyungwon University

요약

유형문화재 데이터는 전국 각지의 박물관 혹은 유관기관에서 소유하고 있으며, 각 기관마다 서로 다른 형태로 수집하고 관리한다. 현재 각 기관간의 연동 서비스가 제공되고 있지 않아서 사용자는 자신이 원하는 데이터를 검색하기 위해서 각 기관별 홈페이지에 접속하여 정보를 검색하여야 한다. 각 기관별로 소유하고 있는 정보를 통합하여 검색할 수 있으면 사용자의 편리성을 보장할 수 있고, 각 기관별 연동 서비스도 제공할 수 있다.
일반적인 유형문화재 데이터는 관계형 데이터베이스 형태로 존재하고 있으며, 각 기관별로 서로 다른 형태의 데이터베이스와 구조를 가지고 있다. 통합 검색 서비스를 제공하기 위해서는 통합된 형태의 온톨로지 제공이 필요하며, 통합된 온톨로지와 각 기관별 데이터베이스 스키마를 매칭하는 과정이 필요하다. 하지만, 온톨로지에 대한 전문적인 지식을 가지고 있지 않은 관리자가 온톨로지를 생성하는 것은 많은 시간과 비용이 추가적으로 요구된다.
본 논문에서는 각 기관별 관계형 데이터베이스에서 온톨로지를 자동으로 추출하는 방법을 제시하여 통합 검색 시스템을 위한 기초를 제공하고, 이를 통한 통합 검색 시스템을 제안한다.

Abstract

A number of museums and related parties have tangible heritage data and those data are collected by different types and formats. In these days, there are no service which connect each others so if a user wants to search a data, user have to log on certain homepage and search data.
In general relational database systems have used for tangible heritage but sometime there is no ontology information for such relational database system. Therefore, for efficient searching of tangible heritage, we need a method which is extracting ontology information from relational database system. And we need a method which makes alignment between local ontologies extracted from relational database system and global ontology which has global information of tangible heritage.
In this paper, we propose a system which can search tangible heritage efficiently using a method of extracting ontology from RDBMS and a method of aligning between local ontology and global ontology.

I. 서 론

유형문화재는 전국 각지의 박물관 및 유관기관에서 소장하고 있으며, 최근의 많은 프로젝트에서 디지털화 작업이 진행되고 있다[1]. 유형문화재 검색 시스템은 각 기관별로 구축되어 있으며, 일반적으로 관계형 데이터베이스에 저장되어 있다.

각 기관별로 구축된 유형문화재 데이터를 통합하여 검색할 수 있는 시스템은 아직 구축되지 못하였고, 일부 박물관과 문화재청의 소장 자료가 통합 검색 서비스를 제공하고 있다. 이 서비스의 경우 키워드 검색 혹은 디렉토리 검색 등과 같은 제한적인 형태의 서비스만을 제공하고 있다[1]. 예를 들어, “조선시대 석탑을 소유한 신라시대 사찰을 찾아라“ 라는 질의어에 대해서는 각 단어별로 분리하여 단어를 포함한 데이터만을 제공하고 있다.

유형문화재에 대한 시맨틱 검색을 수행하기 위해서는 데이터간의 관계를 표현한 온톨로지 구축이 필요하다. 그러나, 온톨로지에 전문적이지 못한 시스템 운영자가 온톨로지를 구축하는 것은 많은 시간과 노력이 필요하다. 따라서, 본 논문에서는 각 기관별로 구축된 관계형 데이터베이스의 테이블을 온톨로지로 변환하고, 변환된 온톨로지와 유형문화재 온톨로지를 매칭하여 이를 통하여 유형문화재 검색을 수행하도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 관계형 데이터베이스 테이블을 온톨로지로 변환하는 과정에 대하여 살펴본다. 4장에서는 통합된 온톨로지를 기반으로 한 통합 검색 시스템을 제안한다. 마지막으로 5장에서는 결론 및 향후 연구 과제에 대하여 언급한다.

Ⅱ. 관련 연구

관계형 데이터베이스 테이블의 온톨로지 변환에 관련된 연구는 ER 데이터모델을 이용하는 방법과 테이블 명칭을 이용하는 방법으로 구분된다. An[2]과 Gramajo[3]는 ER 데이터모델을 이용하여 엔티티의 종류를 strong과 week로 분류했으며, 이들의 관계를 함수로 표현하여 온톨로지를 구축하는 방법을 사용하였다. 그러나, 관계형 데이터베이스를 구축하였다고 하더라도 ER 데이터모델을 사용하지 않았거나 현재 가지고 있지 않은 경우에는 이를 적용하기 어려운 단점이 있다.

테이블 명칭을 이용하는 방법[4]은 테이블의 명칭과 온톨로지 클래스 레이블들을 비교하여 유사한 것들끼리 매칭하는 알고리즘을 사용한다. 하지만 이 경우 명칭이 유사하지 않거나, 같은 명칭이라도 다른 의미의 데이터일 경우 매칭의 정확도가 저하된다.

본 논문에서는 관계형 데이터베이스의 테이블에서 ER 데이터모델을 추출하고, 테이블의 인스턴스와 유형문화재 온톨로지의 인스턴스를 비교하여 클래스를 생성하는 방법을 사용하였다.

Ⅲ. RDBMS 테이블에서 온톨로지 추출

RDBMS 테이블에서 온톨로지를 추출하는 과정은 통합 검색을 제공하기 위한 사전 단계이다. 이 과정을 통해서 추출된 로컬 온톨로지들은 향후 전문가의 도움을 얻어 구성할 유형문화재 온톨로지와 매칭된다. 이 매칭 관계가 통합 검색의 연결고리가 된다.

1. 온톨로지 추출 과정

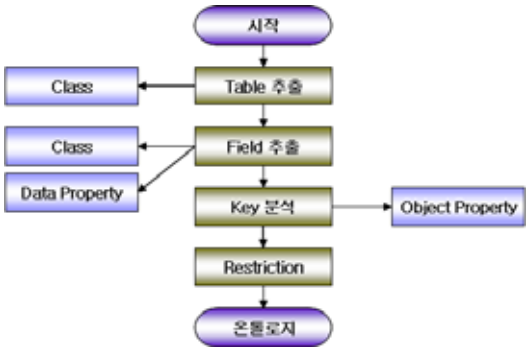
각 기관별 RDBMS 테이블에서 로컬 온톨로지를 추출하는 과정을 그림 1에 도시하였다. RDBMS 테이블은 ER 데이터모델추출기와 개체 매칭기를 거쳐서 로컬 온톨로지로 변환된다.

▶▶ 그림 1. 로컬 온톨로지 추출 과정

모델 추출기(ER Model Extractor)는 관계형 데이터베이스의 테이블들을 분석하여 클래스를 생성하고, 테이블들간의 관계를 설정하는 역할을 수행한다. 개체 매칭기(Instance Matcher)는 테이블의 필드 정보와 필드 인스턴스들을 이용하여 클래스와 데이터 프로퍼티로 분류하는 역할을 수행한다. 이 과정을 그림 2에 도시하였다.

모델 추출기는 먼저 테이블의 메타정보를 분석하여 테이블 명칭을 추출하여, 클래스로 분류한다. 데이터베이스내의 모든 테이블은 각각의 필요성에 의하여 작성되었기 때문에 이들 모두를 클래스로 분류한다.

다음으로 각 테이블의 필드를 분석하여, 클래스와 데이터 프로퍼티로 분류한다. 필드의 정보들은 대부분 데이터 프로퍼티로 분류할 수 있지만, 본 논문에서는 데이터의 의미 관계를 포함하기 위해서 테이블의 데이터들을 유형문화재 온톨로지의 인스턴스와 비교하였다.



▶▶ 그림 2. ER Model 추출기

이 과정은 개체 매칭기가 수행한다. 인스턴스 매칭에서의 유사도 측정을 위하여 Levenshtein Distance(LD)와 Longest Common Prefix-Suffix (LCPS)를 이용하였다. LD[5]는 문자 s에서 t로의 변환을 위해서 삽입, 삭제, 대체 등에 필요한 회수를 계산하는 방법으로 Edit Distance로 표현되기도 한다.

$$LD(s,t) = 1 - \frac{changeLength}{\max Length} \qquad (1)$$

식(1)에서 maxLength는 max(length(s), length(t))로 표현되며, changeLength는 변환을 위해 필요한 회수를 의미한다.

LCPS는 접두사 혹은 접미사의 일치하는 숫자를 이용하여 비교하는 방법이다.

$$\begin{aligned} LCP(s,t) &= \frac{prefixLength}{\min Length} \\ LCS(s,t) &= \frac{suffixLength}{\min Length} \\ LCPS(s,t) &= \max(LCP(s,t), LCP(s,t)) \end{aligned} \qquad (2)$$

식(2)에서 prefixLength는 일치하는 접두어의 길이이고, suffixLength는 일치하는 접미어의 길이, min- Length는 min(length(s), length(t))를 의미하며, 두 길이 중 최대값을 선택한다.

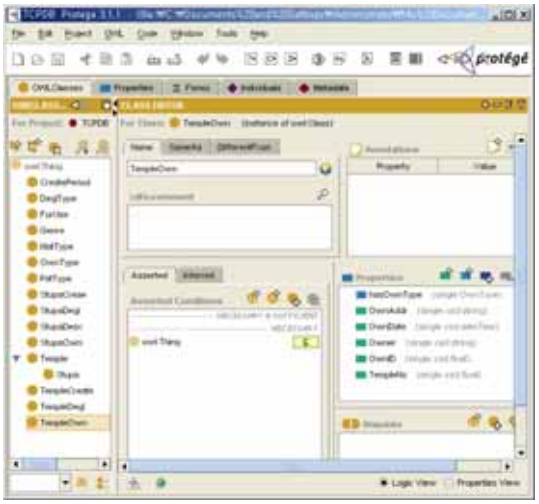
$$GSD_k = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} SD(s_i, t_j)$$
$$SD(s,t) = \alpha \times LD(s,t) + (1-\alpha) \times LCPS(s,t) \quad (3)$$
$$\text{단}, 0 \le \alpha \le 1$$

최종적으로 식(3)을 이용하여 유사도를 계산한다. $SD(s_i, t_j)$는 각 매칭쌍에 대한 유사도이며, 이는 LD와 LCPS에 별도의 가중치를 적용하여 계산한다. 유사도가 일정 정도를 넘는 요소들은 클래스로 분류되고, 그렇지 않은 요소들은 데이터 필드로 분류된다.

테이블의 primary key와 foreign key는 테이블간의 관계를 나타내는 중요한 요소이다. Key 분석 단계에서는 이 키들을 이용하여 테이블간의 관계 설정을 하고, 이 관계는 온톨로지의 오브젝트 프로퍼티가 된다. 테이블간의 관계 혹은 테이블과 필드의 관계는 향후 로컬 온톨로지와 유형문화재 온톨로지간의 온톨로지 구조 매칭을 위해서 사용될 수 있다.

마지막으로 테이블의 카디날리티 등의 정보를 이용하여 온톨로지의 제약조건으로 삼는다.

지금까지의 과정을 거쳐서 만들어진 로컬 온톨로지를 그림 3에 나타내었다.

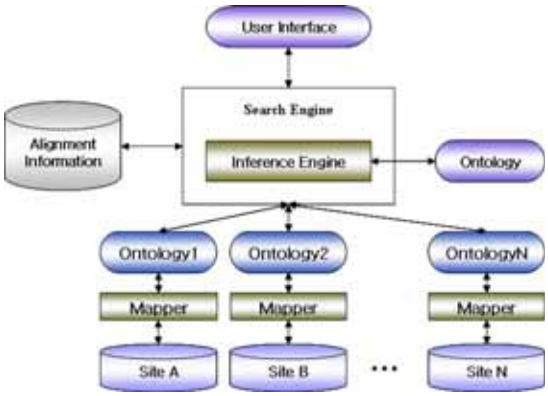


▶▶ 그림 3. 추출된 로컬 온톨로지

Ⅳ. 유형문화재 검색 시스템

본 장에서 제안하는 검색 시스템은 향후 로컬 온톨로지와 유형문화재 온톨로지간의 매칭 작업을 수행한 후의 검색 방법을 제안한 것이다.

유형문화재 검색 시스템의 구성은 그림 4와 같다. 사용자가 검색어를 입력하면 검색 엔진은 이를 추론 엔진으로 보낸다. 추론 엔진은 유형문화재 온톨로지와 동의어 사전을 이용하여 검색 질의어에 대한 추론을 수행한다. 추론된 결과는 다시 검색 엔진으로 보내지고, 검색 엔진은 유형문화재 온톨로지와 로컬 온톨로지간의 매칭 정보를 이용하여 해당 기관의 데이터베이스에서 데이터를 검색하여 사용자에게 출력한다.



▶▶ 그림 4. 제안한 검색 시스템 구성도

Ⅴ. 결 론

본 논문에서는 각 기관에서 가지고 있는 유형문화재 정보를 활용하여 유형문화재 통합 검색을 위한 방법론을 제시하였다.

온톨로지에 대한 전문적인 지식없이 관계형 데이터베이스의 테이블을 온톨로지로 변환하는 과정에 대한 제안을 하였으며, 이를 통한 통합 검색 시스템을 제안하였다. 이 방법론은 향후 로컬 온톨로지와 유형문화재의 온톨로지 매칭 과정을 거쳐서 통합 검색 시스템의 기틀을 제공한다.

사용자가 유형문화재에 대한 검색 질의를 검색 엔진에게 보내면, 검색 엔진은 유형문화재 온톨로지를 검색하여 결과를 추론하고, 매칭 정보를 이용하여 해당 데이터베이스에서 데이터를 추출하여 사용자에게 제공한다.

본 논문에서 제시한 방법론을 구현하고, 이에 대한 평가방법에 대한 연구가 향후 연구해야 할 과제이다.

▌참 고 문 헌▐

[1] http://www.heritage.go.kr(국가문화유산종합정보서비스)
[2] Y. An, A. Borgida, J. Mylopoulos, Inferring Complex Semantic Mappings between Relational Tables and Ontologies from Simple Correspondences, In Proceedings of ODBASE, 2005.
[3] Gramajo Javier and David Riano, Meta-data and ER Model Automatic Generation from Unstructured Information Resources, 5th Joint Conference on Knowledge-Based Software Engineering, September 2002.
[4] J. Kang, J. F. Naughton, On Schema Matching with Opaque Column Names and Data Values, In Proceedings of SIGMOD, 2003.
[5] J. De Bo, P. Spyns, and R. Meersman, Assisting ontology integration with existing thesauri, In Proc. of ODBASE04, pp.801-818, 2004.