pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.3.173

Hybrid Schema Matching (HSM): Schema Matching Algorithm for Integrating Geographic Information

Jiyoon Lee[†] · Sukhoon Lee[††] · Jangwon Kim[†††] · Dongwon Jeong[††††] · Doo-Kwon Baik[†††††]

ABSTRACT

Web-based map services provide various geographic information that users want to get by continuous updating of data. Those map services provide different information for a geographic object respectively. It causes several problems, and most of all various information cannot be integrated and provided. To resolve the problem, this paper proposes a system which can integrate diverse geographic information and provide users rich geographic information. In this paper, a hybrid schema matching (HSM) algorithm is proposed and the algorithm is a mixture of the adapter-based semantic processing method, static semantic management-based approach, and dynamic semantic management-based approach. A comparative evaluation is described to show effectiveness of the proposed algorithm. The proposed algorithm in this paper improves the accuracy of schema matching because of registration and management of schemas of new semantic information. The proposal enables vocabulary-based schema matching using various schemas, and it thus also supports high usability. Finally, the proposed algorithm is cost-effective by providing the progressive extension of relationships between schema meanings.

Keywords : Schema Matching, Integrating Geographic Information, Integrating Non-spatial

Hybrid Schema Matching (HSM): 지리정보 통합을 위한 하이브리드 스키마 매칭 알고리즘

이 지 윤[†] · 이 석 훈[††] · 김 장 원[†††] · 정 동 원[††††] · 백 두 권[†††††]

요 약

웹 기반 지도서비스들은 지속적인 업데이트를 통해 사용자가 원하는 지리정보를 다양하게 제공해준다. 그러나 이러한 지도서비스들은 하나의 지리객체에 대해 각각 다른 정보를 제공한다. 이는 여러 가지 문제를 야기하며, 특히 사용자에게 다양한 정보를 통합적으로 제공하지 못하는 문제점을 지닌다. 이 논문에서는 이러한 문제점을 해결하기 위해 웹에 존재하는 다양한 지리정보들을 통합하여 사용자에게 풍부한 지리정보를 제공할 수 있는 시스템을 제안한다. 이 논문에서는 다양한 비공간정보 스키마를 통합하기 위해 어댑터 기반 의미 처리방법과 정적·동적 의미 관리 기반 접근방법을 혼합한 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 알고리즘을 제안한다. 또한 제안한 알고리즘의 평가를 위해 기존 스키마 매칭 방법들과의 비교평가를 수행한다. 이 논문에서 제안한 알고리즘은 새로운 의미정보 스키마들을 등록하여 관리하기 때문에 스키마 매칭의 정확성을 향상시킨다. 또한 다양한 스키마를 활용한 어휘 기반 스키마 매칭이 가능하므로 높은 범용성을 제공한다. 마지막으로, 제안한 알고리즘은 스키마 의미 간 관계성을 점진적으로 확장함으로써 비용의 효율성을 제공한다.

키워드 : 스키마 매칭, 지리정보 통합, 비공간정보 통합

※ 이 연구에 참여한 연구자는 '2 단계 BK21 사업'의 지원을 받았으며 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NO.2011-0004911)의 결과물임을 밝히며, 또한 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. 2012M3C4A7033346).

† 준 회 원 : 고려대학교 컴퓨터·전파통신공학과 석사과정
†† 준 회 원 : 고려대학교 컴퓨터·전파통신공학과 박사과정
††† 준 회 원 : 고려대학교 컴퓨터·전파통신공학과 박사
†††† 종신회원 : 군산대학교 통계컴퓨터과학과 교수
††††† 종신회원 : 고려대학교 컴퓨터·전파통신공학과 교수

논문접수 : 2012년 9월 13일
수 정 일 : 1차 2012년 11월 8일, 2차 2012년 12월 4일
심사완료 : 2012년 12월 5일

* Co-corresponding Author : Doo-Kwon Baik(baikdk@korea.ac.kr)
Dongwon Jeong(djeong@kunsan.ac.kr)

1. 서 론

웹에서 제공되는 지도서비스들은 다양한 매쉬업(Mash-up) 서비스를 제공하기 위해서 여러 가지 오픈 API를 제공한다[1]. 이러한 매쉬업 서비스를 이용함으로써 사용자들에게 보다 풍부한 지리정보를 제공할 수 있다. 또한 웹 기반 지도서비스들은 환경의 변화에 따른 새로운 서비스를 지속적으로 업데이트하여 사용자가 원하는 데이터를 제공한다[2]. 구글에서 제정한 KML[3]은 다양한 지리정보를 표현하고 지리정보간 교환을 위한 기술로 구글 지도 및 구글 어스 등에서 사용된다. GML[4]은 OGC에서 제정한 국제 표준으로서, 다양한 지리정보를 표현하기 위하여 정의된 마크업 언어이다[5]. GeoRSS[6]는 XML 기반의 지리정보 데이터를 GML과 같은 포맷을 활용하여 위치정보를 RSS피드로 기술하는 것이다. 이 기술을 사용하면 사용자가 등록한 위치정보를 자동으로 갱신할 수 있고 다른 지도시스템에 위치정보를 추가할 수 있다[7]. 이러한 지리정보를 표현하는 기술들은 서로 만들어진 배경과 목적이 다르기 때문에 각각의 기술들에 기반한 다양한 지리정보들이 생성되어 왔다. 그러나 이러한 환경으로 인해 각각의 기술에 기반한 지리정보 및 지도서비스 간 상호운용이 어려워지게 되었다. 예를 들어 구글 지도는 해당 장소의 주소, 전화번호를 제공해주는 반면에, 네이버 지도는 주소, 전화번호, 대중교통정보를 제공해준다. '코엑스'에 대해서 구글 지도는 '서울특별시 강남구 삼성동', '02-6000-0114'라는 정보를 제공해주고, 네이버 지도는 '서울특별시 강남구 삼성1동 159', '02-6000-0114', '2호선 삼성역'이라는 정보를 제공해준다. 만약 구글 지도서비스를 이용하고 있는 사용자가 '코엑스'의 대중교통정보를 제공받으려면 현재 이용중인 구글 지도서비스가 아닌 해당 장소의 주소, 전화번호, 대중교통정보를 제공해주는 네이버 지도 서비스에서 정보를 제공받아야 하는 번거로움이 발생한다[8]. 따라서 이 논문에서는 해당 장소에 대해 서로 다르게 제공해주는 지리정보를 통합함으로써, 사용자들에게 하나의 지도서비스에서 통합된 지리정보를 제공해 주고자 한다.

기존의 지리정보 통합 연구들은 점, 선, 면 또는 입체 등의 지형적 특성을 갖는 공간정보들을 대상으로 지리정보 통합을 수행한다[9]. Stefanakis 외 1명은 GML 기반인 공간정보를 KML 좌표에 맞게 변환하여 웹 지도서비스에 게시할 수 있도록 해당 장소의 공간정보를 통합한다[10]. Francis 외 2명은 KML, GML 간 지리정보간 상호운용의 향상을 위해 공간정보 간 관계분석이 가능한 데이터 모델을 제안한다[11]. 그러나 이러한 연구들은 공간정보만을 대상으로 하고 비공간정보, 즉 특정 건물이나 해당 장소에 대한 이름 및 사진과 같이 웹에서 생성되는 방대한 정보에 대한 확장성은 제공하지 않는다. 따라서 다양한 지리정보를 표현하고 해석 및 확장, 공유하려고 할 때 의미정보 간 불일치가 발생하는 문제점을 지닌다.

이러한 문제점을 해결하기 위해 지리정보 스키마의 의미정보를 해석하고, 이를 체계적으로 관리할 수 있는 다음과 같은 네 가지 연구가 진행되어 왔다[12]. 먼저 유사도 기반 의미 처리방법을 활용한 스키마 통합은 스키마 간 유사성을 측정하여 유사도 값이 큼에 따라 스키마 매칭이 이루어진다. 이 방법은 어휘에 기반하여 대부분의 스키마에 적용이 가능하지만 결과가 정확하지 않아 정확한 의미 해석이 어렵다는 문제점이 발생한다[13-14]. 두 번째 접근 방법인 어댑터 기반 의미 처리방법을 활용한 스키마 통합은 규칙 등과 같이 서로 상이한 스키마 간의 관계성을 사전에 정의하고, 의미정보 간 관계성을 정의할 수 있는 기능을 제공한다. 이 방법은 스키마 간 의미 관계성을 사전에 정의하므로 추가적인 오버헤드가 발생하지 않으나 새로운 서비스를 처리할 때 시간이 오래 걸리고 이에 따른 비용이 높아진다는 문제점이 발생한다[11,15]. 세 번째 방법으로서, 정적 의미 관리 기반 접근방법을 활용한 스키마 통합은 표준화된 의미정보를 정의하고 정의된 정보를 이용하여 지리정보 스키마의 의미의 불일치를 해결할 수 있다. 이 방법은 정확한 의미 처리가 가능하지만 의미정보가 한정되어 있어서 새로운 의미를 확장하는데 어려움이 발생한다[16-17]. 마지막으로, 동적 의미 관리 기반 접근방법을 활용한 스키마 통합은 의미정보를 등록하고 등록된 의미정보를 이용할 수 있는 기능을 제공한다. 이 방법은 여러 가지 의미정보를 해석할 때 사전에 의미적 사상관계를 구축할 필요가 없다. 그러나 서비스 기반 구조상 현실적 적용함에 제약이 따르는 문제점이 발생한다[18].

이 논문에서는 비공간정보 스키마를 통합하기 위해 어댑터 기반 의미 처리방법과 정적 의미 관리 기반 접근방법, 동적 의미 관리 기반 접근방법의 장점들을 혼합한 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 알고리즘을 제안한다. 제안방법은 각각의 비공간정보 스키마 통합 방법들이 가지는 단점을 보완하여 정확하고 효율적인 비공간정보 스키마 통합을 가능하게 한다.

이 논문의 구성은 다음과 같다. 제2장에서는 관련연구에 대해 언급하고 제3장에서는 이 논문에서 제안하는 알고리즘을 위한 전체적인 프레임워크를 기술한다. 제4장에서는 이 논문에서 제안하는 알고리즘을 정의하고 전반적인 절차를 기술한다. 제5장에서는 제안한 알고리즘을 구현하고 기술한다. 제6장에서는 제안한 알고리즘을 평가하기 위해 기존의 스키마 매칭 방법들과의 정량적 비교 평가를 기술한다. 마지막으로 제7장에서는 결론 및 향후 연구에 대하여 기술한다.

2. 관련 연구

이 장에서는 스키마의 의미정보를 체계적으로 관리하기 위하여 스키마 매칭 기법을 유사도 기반 의미 처리, 어댑터 기반 의미 처리, 정적 의미 관리 기반, 동적 의미 관리 기반 접근방법으로 분류하고 각 분류에 기반한 지리정보 통합연구들을 기술한다.

Laura 외 1명은 유사도 기반 의미 처리방법을 활용한 연

구로서, 인스턴스 간 유사성 측정을 통해 의미가 비슷한 스키마들간의 매칭을 수행하는 방법을 제안한다[13]. 이 방법은 스키마가 지니는 인스턴스 간 유사도를 측정하고, 이 값이 클수록 해당 스키마와 일치한다고 판단하여 스키마 매칭을 수행한다. 이 후 스키마 간 유사도 측정을 통해 최종적으로 스키마 매칭을 완료한다. 또한 Jeffrey 외 3명은 XML 스키마 간의 가중치를 측정하여 가중치가 높은 스키마 간 매칭 연산을 수행하는 방법을 제안한다[14]. 이 방법은 스키마가 지니는 의미정보를 기반으로 스키마 간의 거리를 측정하고 이 값을 통해 거리에 대한 가중치 값이 높은 스키마들을 식별하여 스키마 매칭을 완료한다. 그러나 이러한 방법은 측정된 유사도 값의 상대적인 수치로 스키마 매칭 연산을 수행하기 때문에 낮은 유사도 값 간의 비교가 이루어졌을 때 정확도가 떨어지는 문제점을 지닌다.

Francis 외 2명은 어댑터 기반 의미 처리방법을 활용한 연구로서, 구조화된 데이터를 사용하여 서로 다른 다양한 데이터를 하나로 통합하기 위한 스키마 매칭 방법을 제안한다[11]. 이 방법은 데이터 구조를 사전에 정의하여 매핑 테이블을 생성하고 이를 기반으로 서로 다른 스키마들을 매칭한다. 또한 Sanjay 외 2명은 여러 개의 XML 데이터를 하나로 통합하기 위하여 매핑 테이블과 매핑 규칙을 사용한다[15]. 매핑 테이블은 서로 다른 스키마 간의 의미를 사전에 정의하여 나열한 테이블로, 이를 기반으로 매핑 규칙을 생성하여 스키마 매칭을 수행한다. 매핑 테이블과 매칭 규칙을 이용하여 스키마 매칭을 수행할 때 의미 불일치가 발생하게 되면 쿼리 질의를 통해 불일치되는 의미를 제거한다. 이를 통해 스키마 간 상호운용이 가능하게 된다. 그러나 이러한 방법들은 스키마의 수가 많아질수록 재정의 해야 하는 의미 관계들이 많아지므로 스키마 간 관계 정의를 위한 시간이 증가하며 스키마를 매칭하는데 많은 시간을 소모한다는 문제점을 지닌다.

Nengcheng 외 3명은 정적 의미 관리 기반 접근방법을 활용한 연구로서, 이미 정의된 기준 스키마에 기반하여 스키마 간 매칭을 수행하는 방법을 제안한다[16]. 스키마 매칭이 완료된 후에 매칭된 스키마 간의 관계가 올바르게 매칭되었나 검증을 수행한다. Jaewook 외 3명은 입력 받은 스키마 파일을 트리 형태의 구조로 분리하여 사전에 정의된 트리 구조를 통해 스키마 매칭을 수행하는 방법을 제안한다[17]. 스키마 파일은 네임스페이스를 기준으로 트리 구조로 분리 한 후 이를 이용하여 스키마 간 매칭을 완료한다. 그러나 이러한 방법들은 사전에 정의되지 않은 스키마에 대한 매칭이 불가능하다는 문제점을 지닌다. 또한, 새로운 스키마가 추가 될 경우 추가된 스키마에 대해 유연한 매칭이 어렵다는 문제점을 지닌다.

Lee 외 4명은 동적 의미 관리 기반 접근방법을 활용한 연구로서, 서로 다른 스키마를 메타데이터 레지스트리 기반으로 스키마 매칭을 수행하는 방법을 제안한다[18]. 이 방법은 메타데이터 레지스트리 내에 존재하는 다양한 스키마의 의미정보를 관리 및 생성하여 스키마를 확장하고 이를 기반으로 스키마 간 매칭을 수행한다. 그러나 이 방법은 현재 제공되는 서비스 기반 구조상 제약이 따른다는 문제점을 지닌다.

3. 지리정보 통합 시스템

이 장에서는 지리정보 통합을 위한 전체적인 시스템 구조를 기술하고, 공간정보 통합과 비공간정보 통합에 대해 전반적인 절차를 기술한다.

Fig. 1은 지리정보 통합을 위한 전체적인 시스템 구조를 보인다.

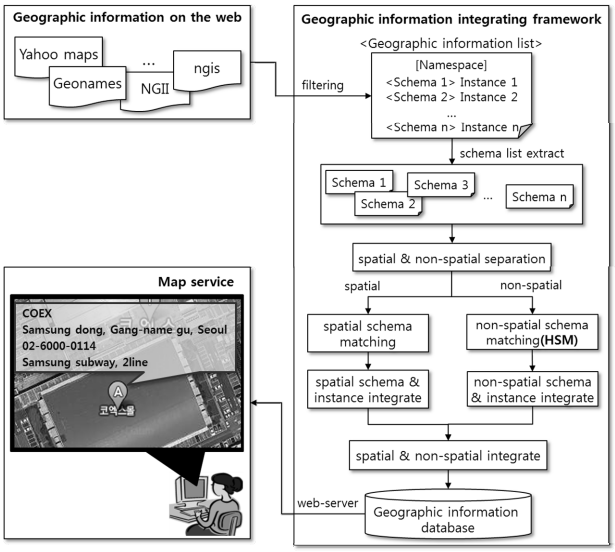


Fig. 1. System architecture for Integrating Geographic Information

우선 웹에 존재하는 여러 가지 지리정보들을 수집하고, 수집한 지리정보들을 공간정보와 비공간정보로 각각 통합한다. 비공간정보 통합은 제안하는 알고리즘인 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 방법을 이용하며, 자세한 내용은 4장에서 기술한다. 통합된 공간정보와 비공간정보를 지리정보 데이터베이스에 저장하고 웹 서버를 통해 사용자에게 하나로 통합된 지도서비스를 제공한다.

3.1 지리정보 통합 절차

이 절에서는 지리정보 통합을 위한 전체적인 시스템 절차를 보인다. 이 시스템은 KML, GML, GeoRSS 등으로 표현된 다양한 지리정보들을 통합하여 웹 기반 지도서비스를 사용하는 사용자에게 풍부한 지리정보 제공을 목적으로 한다. Fig. 1에서 지리정보 통합 시스템은 크게 다음과 같은 절차를 통해 지리정보 통합 연산을 수행한다.

- 지리정보 수집: 다양한 지리정보의 리스트 생성을 위하여 웹에 존재하는 지리정보 수집 및 수집한 지리정보를 파싱하여 각각의 지리정보 리스트 생성

- 스키마 리스트 추출: 생성된 리스트를 이용하여 스키마를 하나의 단위로 스키마 분리 및 스키마 리스트 추출
- 공간정보 및 비공간정보 분리: 분리된 스키마간의 매칭 및 통합이 이루어 지도록 공간정보와 비공간정보로 분리
- 공간정보 통합: 기존 기술들을 이용하여 공간정보 스키마 매칭 및 매칭된 스키마에 따라 자동적으로 인스턴스 통합
- 비공간정보 통합: 제안하는 알고리즘인 HSM을 이용하여 비공간정보 스키마 매칭 및 매칭된 스키마에 따라 자동적으로 인스턴스 통합
- 공간정보 및 비공간정보 통합: 해당 정보에 맞게 통합 작업이 완료된 후 각각 변환된 공간정보와 비공간정보 통합 및 지리정보 데이터베이스에 저장

3.2 공간정보 및 비공간정보 통합

이 절에서는 공간정보와 비공간정보의 스키마 통합 방법을 기술한다.

GeoRSS, GML, YAHOO, VIRTUAL, KML 이 5가지는 대표적인 지리정보를 표현하는 기술들로, Table 1은 5개의 공간정보에 대한 특징들을 보여준다.

Table 1. Spatial feature table

	Coordinate System	Scope	Development Organization
GeoRSS	WSG84	Point, Polyline, Polygon, Feature	Atom, IETF
GML	WSG84	Point, Polygon	OGC
YAHOO	WSG84	Point, Polyline, Feature	Yahoo
VIRTUAL	WSG84	Point, Polyline, Polygon, Feature	Microsoft
KML	WSG84	Polyline, Polygon, Feature	Google

공통적으로 5개의 공간정보들은 WGS84 타원체상의 위도, 경도좌표, 즉 표준좌표체계인 WSG84를 사용한다. GeoRSS는 Atom, IETF에서 제정된 최신 표준으로, '점(Point)', '선(Polyline)', '면(Polygon)', '특징(Feature)'의 모든 범위를 표현할 수 있다. GML은 국제 표준기구인 OGC에서 제정되었으며 '점', '면' 범위까지 표현할 수 있고 지리정보 시스템에 제공하기 위한 특징을 지닌다. Yahoo에서 개발한 YAHOO Map[19]은 '점', '선', '특징'의 범위를 표현할 수 있다. VIRTUAL Earth[20]는 마이크로소프트에서 개발한 좌표체계로 이는 좌표체계의 모든 범위인 '점', '선', '면', '특징'을 표현할 수 있다. 구글에서 제정한 KML은 '선', '면', '특징'의 범위까지 표현할 수 있는 특징을 지닌다. 이러한 특징들을 고려하고 기존에 연구된 공간정보 스키마 매칭 및 통합에 대한 기술들을 사용하여 공간정보를 통합한다. GML2KML Conversion[10]과 Geo Parsing MEDIATOR[21]는 해당 지도시스템의 좌표에 맞게 공간정보를 변환시켜주는 시스템 제공하는 기술로, 이와 같은 기술들을 이용하여 각각의 좌표체계를 변환한다. 변환된 스키마를 기준으로 스키마가 통합되면, 통합된 스키마에 따라 자동적으로 인스턴스가 통합되어 공간정보 통합을 완료한다.

비공간정보 통합의 기존 연구들은 공통적으로 비공간정보의 다양한 지리정보를 표현하고 해석 및 확장, 공유하려고 할 때 의미정보 간 불일치가 발생한다는 문제점을 지닌다. 따라서 이 논문에서는 비공간정보의 다양한 스키마를 해석하여 공유 및 확장할 수 있는 비공간정보 스키마 통합 방법을 제안한다. 비공간정보 스키마 통합 시 의미정보에 대한 관계성을 사전에 정의하여 스키마에 대한 정확성을 보장할 수 있는 어댑터 기반 의미 처리방법과 표준화된 의미정보를 활용하여 각 스키마들을 동일하게 정의하는 정적 의미 관리 기반 처리방법, 동일한 의미정보를 활용하지 않더라도 어떠한 형태로든지 의미정보를 활용할 수 있는 동적 의미 관리 기반 처리방법을 혼합한 HSM 방법을 제안한다. 제안하는 HSM 방법을 통해 입력 받은 비공간정보 스키마 매칭을 수행하고 매칭된 스키마를 기준으로 자동적으로 인스턴스가 통합되어 비공간정보 통합을 완료한다. HSM에 대한 자세한 설명은 4장에서 기술한다.

4. 하이브리드 스키마 매칭 알고리즘

HSM 알고리즘은 비공간정보 스키마 매칭 알고리즘으로서 정적 의미 관리 기반 처리방법, 어댑터 의미 기반 처리방법, 동적 의미 관리 기반 처리방법을 혼합한 알고리즘이다. HSM 알고리즘은 정적 스키마 매칭 알고리즘, 어댑터 기반 스키마 매칭 알고리즘, 동적 스키마 매칭 알고리즘으로 구성된다.

Fig. 2는 입력 받은 지리정보 스키마를 변환하여 출력하기 위해 수행되는 알고리즘이다.

Hybrid_Schema_Matching_Algorithm *(T)*

1: Schema ← {*Geo_S*$_1$, ⋯, *Geo_S*$_{i-1}$, *Geo_S*$_i$}
2: IF Schema$_k$ is Standard_Schema_Table(k) THEN
3: *Static_Schema_Matching_Algorithm()*
4: ELSE IF Schema$_k$ is Meaning_Relation_Table(k) THEN
5: *Adapter_Schema_Matching_Algorithm()*
6: ELSE
7: *Dynamic_Schema_Matching_Algorithm()*
8: END_IF

Fig. 2. HSM algorithm

입력 받은 스키마가 기준 스키마 정의 테이블에 존재할 경우 정적 스키마 매칭 알고리즘에서 수행한다. 입력 받은 스키마를 정의되어 있는 기준 스키마로 변환하고 스키마 리스트에 저장한다. 만약 입력 받은 스키마가 정적 스키마 매칭 알고리즘에서 수행되지 못했을 경우, 어댑터 기반 스키

마 매칭 알고리즘에서 수행된다. 어댑터 기반 스키마 매칭 알고리즘에서 수행되는 스키마는 사전에 스키마 간 관계 정의 되어있는 해당 스키마로 변환하고 스키마 리스트에 저장한다. 만약 입력 받은 스키마가 정적 스키마 매칭 알고리즘과 어댑터 기반 스키마 매칭 알고리즘에서 수행되지 못했을 경우, 동적 스키마 매칭 알고리즘에서 수행한다. 동적 스키마 매칭 알고리즘에서는 의미가 비슷한 지리정보 스키마가 의미정보 관계성 정의 테이블에 존재할 경우 어댑터 기반 스키마 매칭 알고리즘에서 스키마 매칭이 수행된다. 반면에, 의미가 비슷한 지리정보 스키마가 의미정보 관계성 정의 테이블에 존재하지 않을 경우 정적 스키마 매칭 알고리즘에서 수행한다. 이 때 스키마의 형태가 완전할 경우 정확한 의미 해석이 가능하다고 판단하여 기준 스키마 정의 테이블에 입력 받은 스키마를 등록한다. 각 알고리즘에 대한 설명은 다음절에서 상세하게 기술한다.

4.1 정적 스키마 매칭 알고리즘

정적 스키마 매칭 알고리즘은 정적 의미 관리 기반 처리 방법으로서, 제일 처음 수행되는 알고리즘이다.

Table 2는 웹에 존재하는 여러 가지 지리정보들 중 기준으로 사용할 수 있는 스키마를 정의한 지리정보 관계성 정의 테이블이다. 이 테이블은 구글에서 제정한 KML, OGC에서 제정한 국제 표준인 GML, GML과 같은 포맷을 활용하여 위치정보를 RSS 피드로 기술하는 GeoRSS, 비공간정보 데이터를 다루는 FOAF[22], JSON을 기반으로 공간정보 데이터를 교환하는 형식인 GeoJSON[23], 국가별 위치정보를 제공해주는 Geonames[24], Yahoo에서 개발한 YAHOO Map[19]에서 지리정보를 수집하였다.

지리정보 통합 시스템에서 지리정보 리스트를 구성하고 그 중 추출된 스키마 중 스키마를 공간정보와 비공간정보로 분리하기 위한 기준으로 사용한다. 또한 여러 가지 지리정보들 중에서 기준으로 사용할 수 있는 스키마를 정의하는 표로 사용한다. 웹에 존재하는 다양한 스키마 중 기준 스키마로 정의할 수 있는 조건은 공통적으로 스키마의 의미 해석이 분명할 수 있도록 어휘의 풀 네임을 사용해야 한다.

예를 들어 '전화번호'를 의미하는 스키마 'number'는 또 다른 의미 '숫자'를 포함하고 있기 때문에 의미의 해석이 불분명하다. 따라서 'phoneNumber'스키마를 사용하면 '전화번호'라는 정확한 의미 해석이 가능하다. 지리정보 관계성 정의 테이블의 'longtitude', 'latitude'와 같은 스키마들은 공간정보 스키마를 분리하는데 사용된다.

Table 3은 Table 2의 지리정보간 관계성 정의 테이블에서 기준 지리정보 스키마로 정의된 테이블이다.

Table 3. Static schema table

Static schema
kml:name
kml:phoneNumber
kml:address
kml:description
geonames:countryCode
geonames:countryName
geonames:temperature
geonames:humidity
…
foaf:image
foaf:homepage

Table 2. Geographic information relation table

Static geographic information schema	KML	GML	GeoRSS	FOAF	GeoJSON	Geonames	…	YAHOO Maps
kml:name	name	title name	title	name	name	placename		–
kml:phoneNumber	phoneNumber	–	–	phone	–			–
kml:address	address	where	–	–	–	postalcode		–
kml:description	description	description	summary content	–	–	–		–
geonames:countryCode	–	–	–	–	–	countryCode	…	–
geonames:countryName	–	–	–	–	–	countryName		–
geonames:temperature	–	–	–	–	–	Temperature		–
geonames:humidity	–	–	–	–	–	humidity		–
…	…	…	…	…	…	…		
foaf:image	–	–	–	image	–	–		YImage
foaf:homepage	–	–	–	homepage	–	–		–
…	…	…	…	…	…	…	…	…
–	longtitude	coordinates	coords	–	lat	–	…	Lat
	latitude	…	…		long			Lon
	…	point	pos		…			…

'kml:name'은 해당 장소의 이름이나 건물명이고, 'kml:phoneNumber'은 음식점이나 공공시설과 같은 특정 공간의 전화번호를 의미한다. 'kml:address'는 해당 공간의 주소와 우편번호를 의미한다. 'kml:description'은 장소에 관한 부가적인 설명으로 예를 들어, 회사 연혁과 주차공간, 부대

시설 등이 있다. 'geonames:countryCode'는 ISO-3116을 사용하는 국가코드이다. 'geonames:countryName'은 국가 이름과 수도를 영문으로 나타낸 것이고 'geonames:temperature'와 'geonames:humidity'는 해당 지역의 온도와 습도를 의미한다. 'foaf:image'는 해당 지리정보에 대한 사진을 의미하고 'foaf:homepage'는 특정 시설에 대한 홈페이지 주소를 의미한다. 기준 스키마 정의 테이블은 스키마 리스트에서 추출한 스키마를 입력 받아 기준 스키마 정의 테이블 내에 스키마가 존재하는지에 대한 질의쿼리를 보낸다. 이 때, 입력 받은 스키마가 테이블 내에 존재할 경우 이 스키마는 기준 스키마라고 판단되어 입력 받은 스키마를 그대로 변환된 스키마 리스트에 저장한다. 만약 입력 받은 스키마가 기준 스키마 정의 테이블 내에 존재하지 않을 경우 어댑터 기반 스키마 매칭 알고리즘을 수행한다. 입력 받은 스키마가 기준 스키마 테이블 내에 존재하는지 스키마 비교를 수행하는 알고리즘은 Fig. 3과 같다.

Static_Schema_Matching_Algorithm(*T*)

1: Standard ← $\{S_1, \cdots, S_{i-1}, S_i\}$
2: T_s ← *getTagName(T)*
3: Target ← $\{T_{s1}, \cdots, T_{s(i-1)}, T_{si}\}$
4: FOR $i \in$ *getLength(target)* DO
5: FOR $k \in$ *getLength(Standard)* DO
6: IF Standard_k is Target_k THEN
7: TransSchema ← Standard_k
8: END_IF
9: END_FOR
10: *Adapter_ Schema_Matching_Algorithm(k)*
11: END_FOR

Fig. 3. Static schema matching algorithm

정적 스키마 매칭 알고리즘은 제일 처음 수행되는 알고리즘으로 입력 받은 스키마에 대한 매칭이 수행되지 못할 경우 어댑터 스키마 매칭 알고리즘에서 수행된다. 어댑터 스키마 매칭 알고리즘에 대한 설명은 다음절에서 상세하게 기술한다.

4.2 어댑터 스키마 매칭 알고리즘

어댑터 스키마 매칭 알고리즘은 어댑터 의미 기반 처리 방법으로서, 기준이 되는 스키마를 기반으로 지리정보 스키마들을 일대일 관계로 매핑한 테이블을 포함하는 알고리즘이다. 입력 받은 스키마가 기준 스키마 테이블 내에 존재하지 않을 때, 이 스키마가 의미정보 관계성 정의 테이블 내에 존재하는지 쿼리 질의를 통해 존재 여부를 수행한다.

Table 4. Semantic relation table

Static schema	Target schema
kml:name	gml:title
	foaf:name
	geonames:placename
kml:phoneNumber	foaf:phone
kml:address	geonames:postalcode
	gml:where
kml:description	georss:summary
	gml:description
…	…
foaf:image	yahoomaps:yimage

Table 4는 의미정보 관계성 정의 테이블로서, 기준이 되는 스키마와 서로 다른 지리정보 스키마 간의 관계성을 일대일로 정의한다.

입력 받은 스키마가 의미정보 관계성 정의 테이블 내에 존재하면 사전에 정의된 해당 기준 스키마로 변환하고 변환된 스키마를 스키마 리스트에 저장한다. 입력 받은 스키마가 기준 스키마 정의 테이블과 의미정보 관계성 정의 테이블 내에 존재하지 않을 경우, 정의 되지 않은 스키마로 간주하여 동적 스키마 매칭 알고리즘에서 수행한다. 입력 받은 스키마가 의미정보 관계성 테이블 내에 존재하는지 스키마 비교를 수행하는 알고리즘은 Fig. 4와 같다.

Adapter_Schema_Matching_Algorithm(*T*)

1: Standard ← *Static_Schema_Matching_Algorithm.Standard*
2: Relation ← $\{R_1, \cdots, R_{i-1}, R_i\}$
3: Target ← *Static_Schema_Matching_Algorithm.target*
4: FOR $i \in$ *getLength(target)* DO
5: FOR $k \in$ *getLength(relation)* DO
6: IF Relation_k is Target_k THEN
7: TransSchema ← Relation_k
8: END_IF
9: END_FOR
10: *Dynamic_Schema_Matching_Algorithm(k)*
11: END_FOR

Fig. 4. Adapter schema matching algorithm

어댑터 기반 스키마 매칭 알고리즘은 4.1절의 정적 스키마 매칭 알고리즘에서 수행되지 못한 경우 수행되는 알고리즘으로, 입력 받은 스키마에 대한 매칭이 수행되지 못했을 경우 동적 스키마 매칭 알고리즘 에서 수행된다. 동적 스키마 매칭 알고리즘에 대한 설명은 다음절에서 상세히 기술한다.

4.3 동적 스키마 매칭 알고리즘

동적 스키마 매칭 알고리즘은 동적 의미 관리 기반 처리 방법으로서, 입력 받은 스키마가 기준 스키마 테이블과 의

미정보 관계성 테이블에 존재하지 않을 경우 수행하는 알고리즘이다. 정적 스키마 매칭 알고리즘과 어댑터 기반 스키마 매칭 알고리즘에서 수행되지 못한 지리정보는 미등록 스키마로 간주된다. 미등록 스키마는 테이블에 등록되지 않거나 존재하지 않은 스키마 또는 적절하지 않은 어휘, 축약되거나 생략된 언어 즉, 새롭게 추가되는 스키마이다. 이 때, 스키마 관리자는 미등록 된 스키마의 의미를 등록하고 관리하여 지리정보들과의 관계성을 정의하거나 기준이 되는 스키마로 정의할 수 있다.

Fig. 5와 Fig. 6은 미등록 된 스키마를 처리하는 과정을 보여준다. 입력 받은 미등록 스키마와 의미가 비슷한 지리정보 스키마가 기준 스키마 테이블에 존재할 경우, 미등록 된 스키마를 해당하는 지리정보 스키마와 일대일 관계성 정의를 수행하고 의미정보 관계성 테이블에 미등록 된 스키마를 등록한다. 그 후 의미정보 관계성 테이블을 통해 미등록 된 스키마를 해당 기준 스키마로 변환하여 변환된 스키마 리스트에 저장한다.

예를 들어 Fig. 5에서 입력 받은 스키마가 GeoRSS 기술을 이용하고 해당 장소의 이름이나 건물명을 의미 하는 'name'일 때, 먼저 의미가 비슷한 스키마가 기준 스키마 정의 테이블에 존재하는지 확인한다.

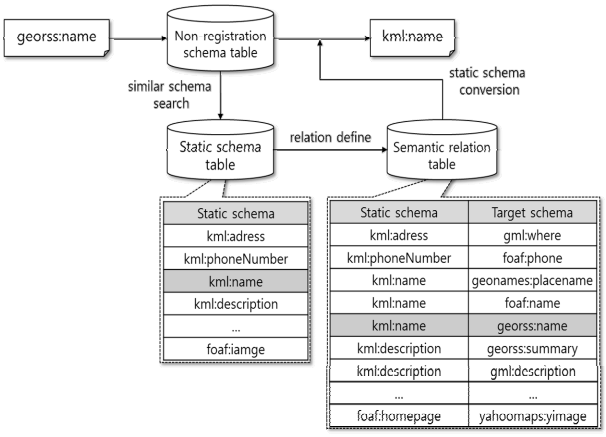


Fig. 5. Existence of a similar schema

그러나 미등록 스키마의 형태가 부적절하거나 어휘가 축약되어 스키마의 의미 해석이 어려우면 기준 스키마가 없는 상태로 의미정보 관계성 정의 테이블에 저장된다. 저장된 미등록 스키마와 의미가 비슷하고 의미 해석이 정확한 스키마의 형태를 가지고 있는 다른 지리정보 스키마가 들어오면 스키마 간의 관계성 정의를 통해 의미정보 관계성 정의 테이블에 등록한다. 또한 의미 해석이 분명한 형태의 스키마를 기준 스키마로 등록한다.

Fig. 6은 입력 받은 스키마 Geonames의 'cloud'가 기준 스키마에 존재하지 않을 경우, 기준 스키마로 정의하는 것을 보여준다. 'cloud'는 '해당 지역의 구름의 양'을 알려주기 위한 Geonames의 스키마로, 의미해석이 정확하고 분명한 형태로 되어 있어서 기준 스키마로 정의할 수 있다.

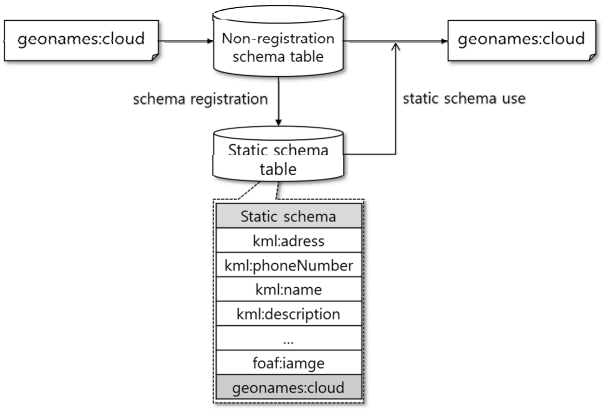


Fig. 6. Non-existence of a similar schema

Fig. 7은 입력 받은 스키마가 동적 스키마 매칭 테이블에서 수행되는 알고리즘이다. 동적 스키마 매칭 알고리즘은 4.1절의 정적 스키마 매칭 알고리즘과 4.2절의 어댑터 기반 스키마 매칭 알고리즘에서 수행되지 못한 경우 마지막으로 수행되는 알고리즘이다.

입력 받은 스키마와 의미가 비슷한 스키마가 의미정보 관계성 정의 테이블에 존재할 경우 4.2절의 어댑터 기반 스키마 매칭 알고리즘에서 스키마 매칭이 수행된다. 또한 입력 받은 스키마와 의미가 비슷한 스키마가 의미정보 관계성 정의 테이블에 존재하지 않을 경우 4.1절의 정적 스키마 매칭 알고리즘에서 수행된다.

Dynamic_Schema_Matching_Algorithm(*T*)

1: Standard ← *Static_Schema_Matching_Algorithm.Standard*
2: Target ← *Static_Schema_Matching_Algorithm.target*
3: SimilarSchema ← Search(Target$_k$, Adaptive_ Meaning _Table(k))
4: IF SimilarSchema$_k$ is *Static_Schema_Matching_Algorithm()* THEN
5: TransSchema ← SimilarSchema$_k$ /* relation definition */
6: ELSE IF SimilarSchema$_k$ is *Adaptive_Schema_Matching_Algorithm()*THEN
7: Add(SimilarSchema, Target$_k$, Adaptive_Meaning_Table(k))
8: TransSchema ← SimilarSchema$_k$ /* schema change */
9: ELSE
10: TransSchema ← SimilarSchema$_k$ /* target equal schema */
11: END_IF

Fig. 7. Dynamic schema matching algorithm

5. 구 현

이 장에서는 제안한 HSM 알고리즘 구현 결과에 대하여 서술한다. 구현을 위해 다음의 Daum Maps Open API[25]를 사용하고 해당 장소의 지리정보는 GML, KML, GeoRSS, Geonames 기술을 사용하여 수집한다. 구현을 위한 실험은 Windows 7 Enterprise K, Intel® CoreTM i3-2100 CPU @ 3.10GHz, 2.00GB Memory, JavaTM SE Development Kit 7, MySQL JDBC, MySQL Workbench 5.2CE 로 구성된 환경에서 수행한다.

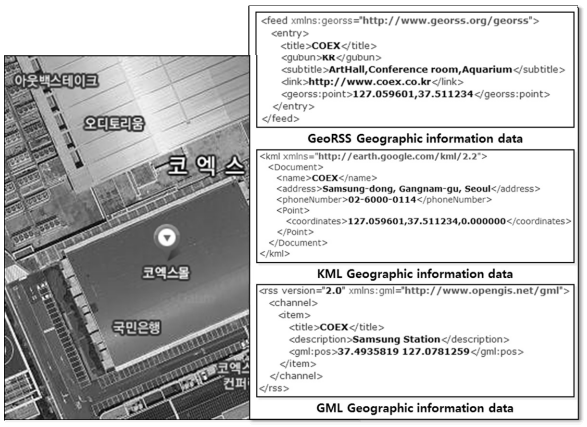


Fig. 8. Geographic information that exists on the web

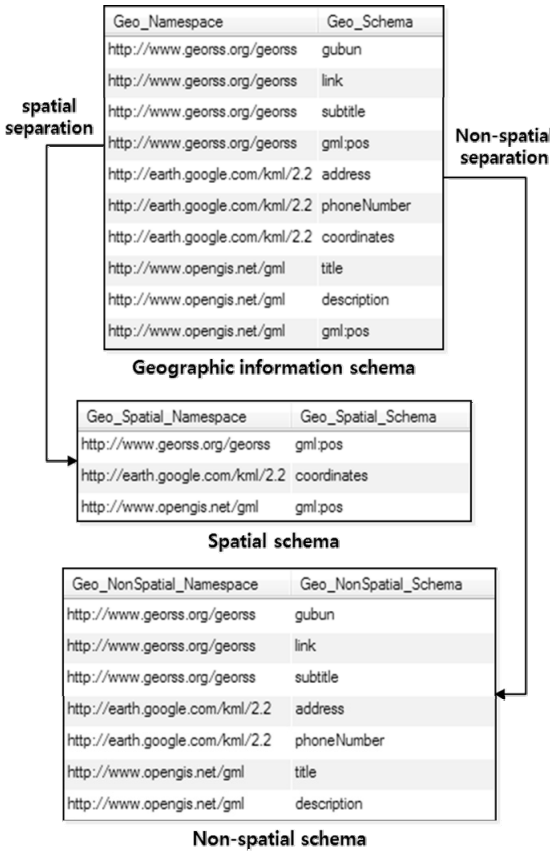


Fig. 9. Geographic information schema separation

Fig. 8은 웹에 존재하는 지리정보 예를 보여준다. 하나의 해당 장소 지리정보에 대해 각각의 웹 기반 지도서비스들이 제공해주는 데이터와 기술 그리고 구조가 다르다는 것을 알 수 있다. GeoRSS 기술을 사용하는 지도서비스는 지리정보의 국가네임과 해당 장소에 대한 관련시설, 해당 장소의 웹 사이트를 제공해준다. KML 기술을 사용하는 지도서비스는 해당 장소의 주소와 전화번호를 제공해준다. GML 기술을 사용하는 지도서비스들은 해당 장소의 부가적인 설명을 제공해준다.

Fig. 9는 웹에 존재하는 지리정보 중 인스턴스를 제외한 스키마를 추출하여 지리정보 스키마 리스트를 생성하고 스키마를 분리하는 과정을 보여준다. 지리정보 스키마 리스트를 기반으로 공간정보와 비공간정보를 분리한다. 공간정보와 비공간정보를 분리하는 기준은 4.1절에 있는 Table 2의 지리정보 관계성 정의 테이블을 이용하여 3개의 공간정보와 7개의 비공간정보를 분리한다. 공간정보 스키마는 기존의 공간정보 통합 방법을 사용하여 매칭을 수행한다.

이 때 3.2절의 Geo Parsing MEDIATOR 기술을 사용하여 3개의 공간정보 스키마는 'gml:pos'로 매칭한다. 비공간정보 스키마는 4장에 서술한 HSM 알고리즘을 사용하여 각각의 스키마가 수행되어야 할 알고리즘으로 분류한다. 'gubugn', 'link', 'subtitle'은 기준 스키마 정의 테이블과 의미정보 관계성 정의 테이블에 존재하지 않으므로 동적 스키마 매칭 알고리즘에서 수행된다. 'address'와 'phoneNumber'는 기준 스키마 정의 테이블에 존재하므로 정적 스키마 매칭 알고리즘에서 수행된다. 'title'과 'description'은 기준 스키마 정의 테이블에 존재하지 않고 의미정보 관계성 정의 테이블에 존재하므로 어댑터 기반 스키마 매칭 알고리즘에서 수행된다.

Fig. 10은 Fig. 9에서 분리된 비공간정보 스키마 중 동적 스키마 매칭 알고리즘에서 수행되는 과정을 보여준다. 입력받은 스키마 중 'gubun'은 해당 장소의 국가명을 의미하고 'link'스키마는 해당 장소의 홈페이지를 의미한다. 이 두 개의 스키마는 기준 스키마 정의 테이블에 의미가 비슷한 스키마가 존재한다. 따라서 의미가 비슷한 스키마들과의 관계성 정의를 수행하여 의미정보 관계성 정의 테이블에 저장한다. 해당 장소의 국가명을 의미하는 GeoRSS의 'gubun'은 Geonames의 'countryName'과 관계성 정의를 수행하고, 'title'은 FOAF의 'homepage'와 관계성 정의를 수행한다. 해당 장소와 관련된 시설을 의미하는 'subtitle'은 기준 스키마 정의 테이블에 의미가 비슷한 스키마가 존재하지 않으므로 기준 스키마로 정의한다. 미등록 처리 알고리즘을 통해 GeoRSS의 'gubun', 'link'로 매칭하고 'subtitle'은 Geonames의 'countryName'으로 매칭한다. FOAF의 'homepage'는 GeoRSS의 'subtitle'로 스키마 매칭을 완료한다.

Fig. 11은 Fig. 9에서 분리된 비공간정보 스키마 중 정적 스키마 매칭 알고리즘에서 수행되는 과정을 보여준다.

해당 장소의 주소를 의미하는 'address'와 해당 장소의 전화번호를 의미하는 'phoneNumber'는 기준 정의 스키마 테이블에 존재하므로 입력 받은 스키마를 그대로 사용한다.

Fig. 12는 Fig. 9에서 분리된 비공간정보 스키마 중 어댑터 기반 스키마 매칭 알고리즘에서 수행되는 과정을 보여준다.

'title'은 해당 장소의 이름을 의미하고 'description'은 해당 장소의 부가적인 설명을 의미한다. 이 두 개의 스키마는 의미정보 관계성 정의 테이블에 존재하므로 관계성 정의 되어있는 기준 스키마로 스키마 매칭을 수행한다. 이를 통해 GML의 'title'은 KML의 'name'으로 매칭되었고, 'description'은 KML의 'description'으로 매칭된다.

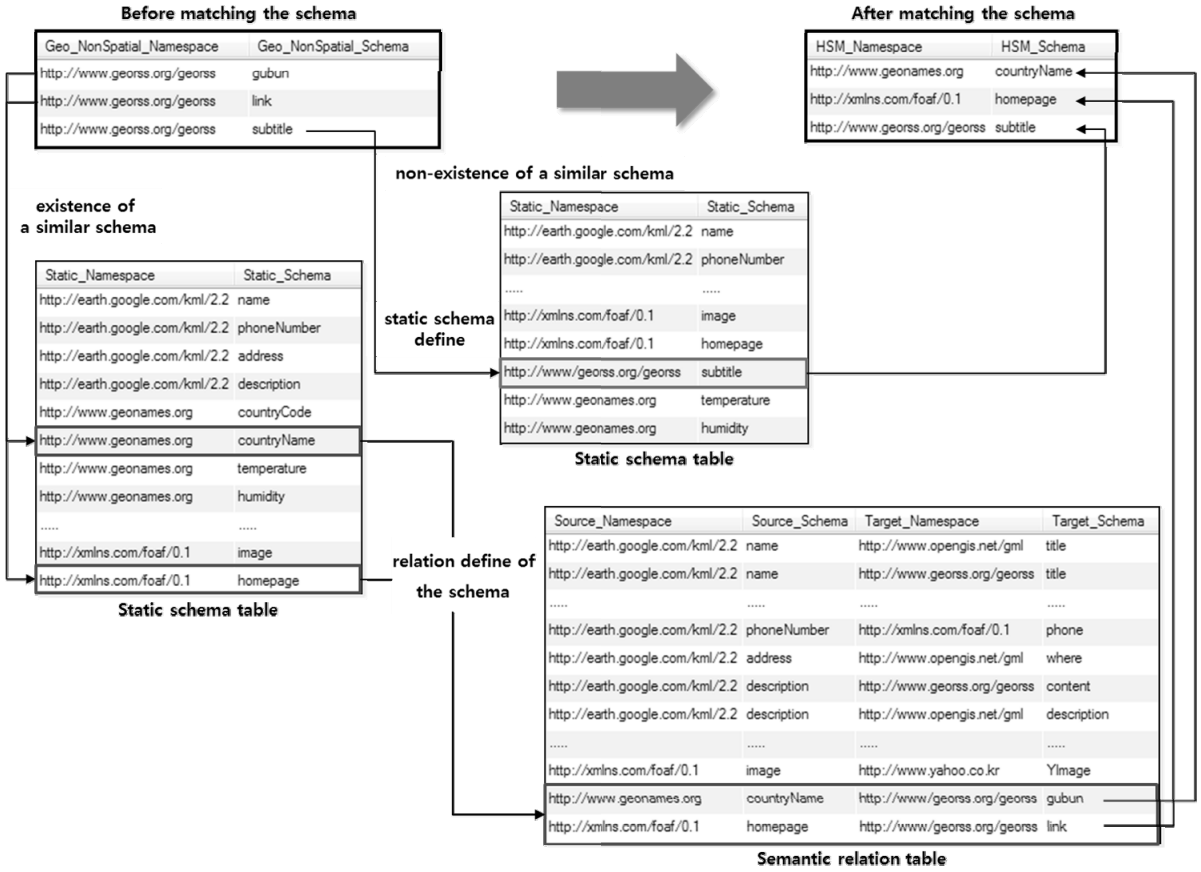


Fig. 10. Described on the Dynamic schema matching algorithm

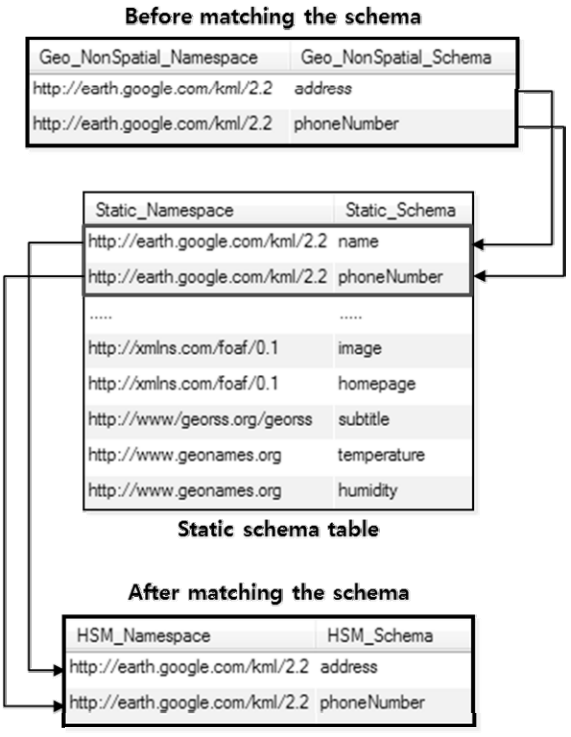


Fig. 11. Described on the static schema matching algorithm

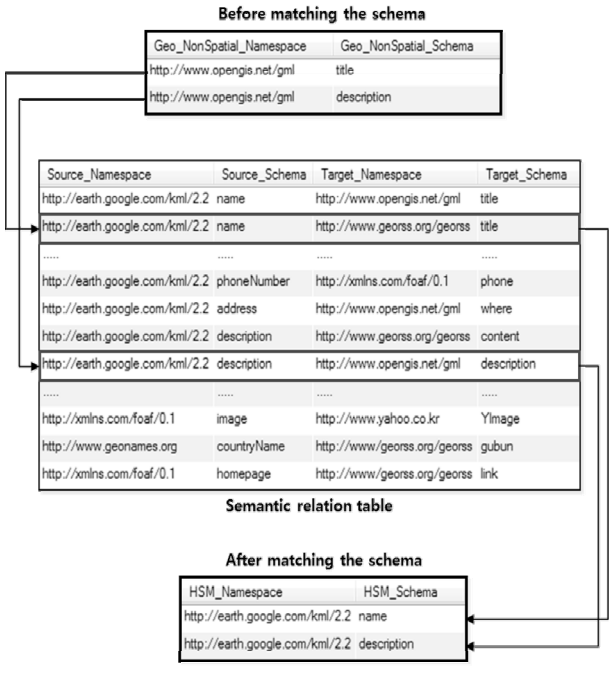


Fig. 12. Described on adapter schema matching algorithm

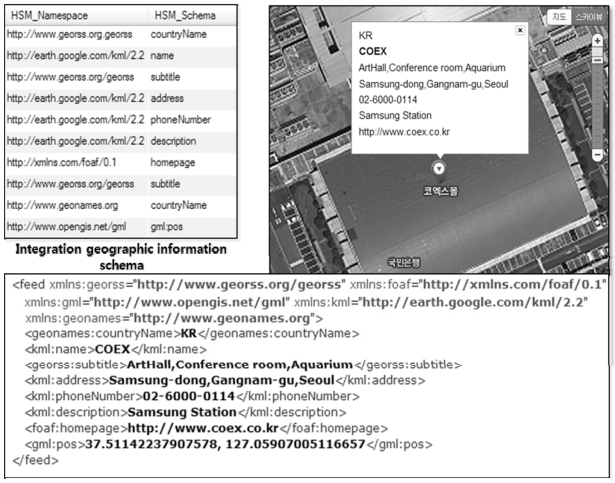


Fig. 13. Integrating Geographic Information

Fig. 13은 매칭이 완료된 스키마와 해당 인스턴스와 통합을 통해 하나의 지도서비스에 다양해진 지리정보를 보여준다.

매칭이 완료된 스키마를 기반으로 해당 인스턴스를 통합하여 하나의 지리정보를 만든다. 만들어진 지리정보를 지리정보베이스에 저장 한 후 웹 서버를 통해 사용자에게 하나로 통합된 지도서비스를 제공해 줄 수 있다. 이를 통해 하나의 장소에 대한 지리정보가 풍부해 짐을 알 수 있다.

6. 실험 및 평가

이 장에서는 제안한 알고리즘과 기존의 스키마 매칭 방법들과의 정량 평가 결과를 기술한다. 제안한 알고리즘과 비교평가를 수행할 대상은 다음과 같다.

- GSim[13] 스키마 매칭 방법: 유사도 기반 의미 처리방법을 활용한 방법으로써, 인스턴스 사이의 유사성을 측정하여 스키마 매칭을 수행하여 지리정보 스키마 매칭의 향상을 위한 방법
- XSDM[15] 스키마 매칭 방법: 어댑터 의미 기반 처리방법을 활용한 방법으로써, 여러 개의 XML 데이터간의 스키마 관계성을 사전에 정의하고 매핑 테이블과 매핑 규칙을 사용하여 스키마 매칭을 수행하는 방법
- FRAG-BASE[16] 스키마 매칭 방법: 정적 의미 관리 기반 접근방법을 활용한 방법으로써, 사전에 정의된 XML 구조를 통해 스키마 간 매칭을 수행하는 방법

Table 5. Geographic information data

	GML	KML	Geo-RSS	Geo-names	Geo-JSON	Yahoo Maps
Schema Set	16	20	15	6	3	3
Schema data	106	143	102	43	13	12

3가지 기존 방법과 제안한 알고리즘과의 비교 평가를 위해 웹 상에 존재하는 지리정보를 이용하였다. Table 5는 실험에 사용될 지리정보 데이터를 보여주는 것으로 GML과 KML을 포함한 총 6개의 웹에서 수집하는 63개의 지리정보들과 419개의 스키마로 이루어진 데이터를 이용하여 비교평가를 수행한다.

6.1 스키마 매칭의 정확성

이 절에서는 스키마 매칭의 정확성에 대한 비교평가 결과를 기술한다. 스키마 매칭의 정확성은 전체 스키마 중에서 정확하게 매칭된 스키마의 개수를 의미한다. 스키마 매칭의 정확성에 대한 수식은 식(1)로, Table 6은 제안한 스키마 매칭 방법과 기존 스키마 매칭 방법과의 비교평가를 수행한 결과이다.

스키마 매칭의 정확성(Accuracy)
= *정확하게 매칭된 스키마 수/전체 스키마 수(%)* - 식(1)

Table 6. Schema matching accuracy

Accumlate schema data	Schema matching method			
	GSim	XSDM	FRAG-BASE	HSM
100	62.96%	66.62%	65.09%	67.31%
200	63.38%	68.05%	65.99%	68.86%
300	64.72%	68.67%	66.61%	69.04%
419	66.19%	69.95%	67.97%	70.55%
Average	64.68%	68.63%	66.71%	69.24%

GSim 스키마 매칭 방법은 정확한 스키마 간 매칭을 요구할 경우 매칭 결과가 정확하지 않은 문제점이 발생한다. 따라서 스키마 매칭 방법 중 가장 낮은 정확도를 보인다. XSDM 스키마 매칭 방법은 스키마 간 관계성을 사전에 정의하므로 스키마 매칭의 정확성이 보장됨을 보인다. FRAG-BASE 스키마 매칭 방법은 표준화된 의미정보를 활용하지 않은 스키마에 대한 정확성을 보장하지 못하므로 부분적 정확성을 보인다. 제안 방법인 HSM 스키마 매칭 방법은 기본적으로 표준화된 의미정보를 활용하여 사전에 스키마를 정의하고 의미정보 간 관계성을 정의하여 스키마를 매칭하는 방법이다. 이 방법은 동일한 의미정보를 활용하지 않더라도 정확한 스키마 매칭이 가능하므로 가장 높은 스키마 매칭의 정확성을 보인다. 정확성 100%를 보이지 않는 이유는 지리정보에서 같은 스키마들이 포함하고 있는 의미가 다르기 때문이다. 예를 들어 지리정보 A의 'description'과 지리정보 B의 'description'의 같은 스키마가 존재할 경우, 두 개가 같은 스키마이지만 지리정보 A에서는 해당 장소에 대한 설명을 의미하지만 지리정보 B에서는 해당 위치의 주소를 의미하기 때문에 의미 불일치가 발생한다. 따라서 정확한 스키마 매칭이 이루어지지 않는다.

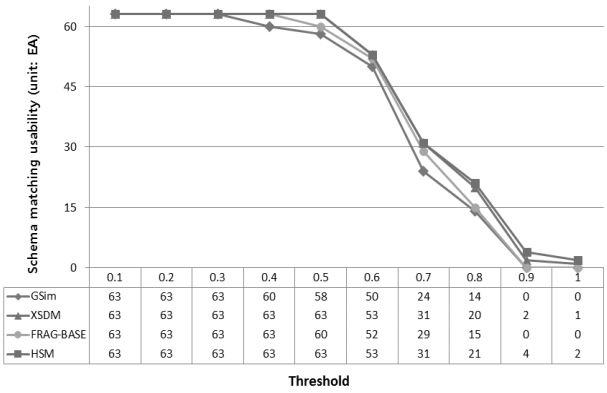
6.2 스키마 매칭의 범용성

이 절에서는 스키마 매칭의 범용성에 대한 비교평가 결과를 기술한다. 범용성이란 웹에서 수집한 여러 지리정보를 스키마 매칭 방법에 적용할 때 얼마나 다양한 스키마 집합에 적용 가능한지에 대한 적용 범위를 뜻한다. 이는 일정 수치 이상의 정확성을 가지는 지리정보 스키마 집합의 총 개수로, 범용성이 높을수록 지리정보에 대한 적용 범위가 넓다는 것을 의미한다. 식(2)는 스키마 매칭의 범용성에 대한 수식으로 스키마 매칭의 범용성은 지리정보 스키마 집합의 개수로 수치화하여 표현할 수 있다.

$$\text{스키마 매칭의 범용성} = \sum_{i=1}^{n} \begin{cases} 1 & (Accuracy(g_i) > h) \\ 0 & (otherwise) \end{cases} \text{-식(2)}$$

지리정보 $G = \{g_1, g_2, g_3 \ldots g_n\}$ 일 때,
g_i : 지리정보 스키마 집합 중 i번째 지리정보 스키마 집합
n : 지리정보 스키마 집합의 개수
h : 임계 값 (지리정보 스키마 집합의 스키마 매칭 정확도)

Fig. 14는 스키마 매칭의 범용성에 대한 결과로 제안한 스키마 매칭 방법과 기존 스키마 매칭 방법과의 매칭 방법들과의 큰 격차를 보이기 시작한다. X축의 표는 그래프의 가독성을 높이기 위해 수치를 작성한 것으로 임계 값을 높게 설정할수록 정확하게 매칭되는 지리정보의 스키마 수가 줄어들지만 이는 여러 개의 지리정보 스키마 집합에서 매칭되는 스키마의 수가 많다는 것을 보여준다.



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
GSim	63	63	63	60	58	50	24	14	0	0
XSDM	63	63	63	63	63	53	31	20	2	1
FRAG-BASE	63	63	63	63	60	52	29	15	0	0
HSM	63	63	63	63	63	53	31	21	4	2

Fig. 14. Schema matching usability

GSim 스키마 매칭 방법은 유사성 측정을 통해 스키마 매칭을 수행하므로 임계 값이 높아질수록 정확하게 매칭되는 스키마가 줄어들고, 임계 값이 0.9를 넘어가면 정확하게 매칭되는 스키마 수는 0개로 스키마 매칭의 범용성은 매우 낮다. XSDM 스키마 매칭 방법은 사전에 스키마 간 의미관계성 정의를 수행하고 스키마 간 상호 의미관계를 정의해 놓기 때문에 임계 값이 높아지더라도 정확하게 스키마가 매칭될 수 있다. 따라서 스키마 매칭의 범용성은 높다. FRAG-BASE 스키마 매칭 방법은 사전에 정의된 부분만 정확하게 의미 처리가 가능하기 때문에 임계 값이 0.9일 때까지는 정확하게 매칭되는 스키마가 존재하여 스키마 매칭의 범용성은 좋다. 그러나 임계 값이 1.0이 되면 정확하게 매칭되는 스키마 수가 0개로 스키마 매칭의 범용성은 보통이다. 제안한 알고리즘인 HSM 스키마 매칭 방법은 XSDM 스키마 매칭 방법과 유사하나 HSM은 다양하게 표준화된 스키마를 활용한다는 장점을 지닌다. 따라서 어떠한 형태로든 입력 받은 스키마 매칭을 수행할 수 있으므로 임계 값이 높아질수록 정확하게 매칭되는 스키마 수는 보장되고, 이는 XSDM 스키마 매칭 방법보다 조금 더 높은 스키마 매칭의 범용성을 보인다.

6.3 스키마 데이터 통합 성능

이 절에서는 스키마 데이터 통합 성능을 비교하기 위하여 Melnik[26]의 연구에서 소개된 Overall과 매칭되지 않은 스키마를 처리하는데 걸리는 시간을 수치화하여 표현한다. 데이터 통합 성능은 스키마 매칭 연산을 수행할 때 필요한 시스템 자원 및 시간을 의미하는 것으로서 평가 계산식은 식(3)과 같다. 매칭되지 않은 스키마 처리속도는 스키마가 동적 스키마 매칭 알고리즘에서 수행 될 때 스키마당 걸리는 시간을 수치화한 값이다.

스키마 데이터 통합 성능(%)
= 매칭되지 않은 스키마 처리 속도 + Overall - 식(3)

*Overall = 재현율(Recall) * (2 - 1/정확도(Precision))* -식(4)

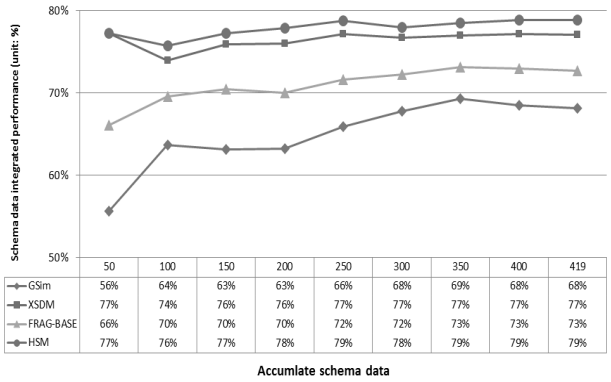
정확도(Precision) = 매칭 완료된 스키마 수 /
매칭이 기대되는 스키마 수 -식(5)

재현율(Recall) = 매칭 완료된 스키마 수 /
매칭 알고리즘이 찾아낸 스키마 수 - 식(6)

Overall은 알고리즘 매칭에서 쓰이는 평가 방법으로 매칭 후, 잘못된 매칭 결과를 제거하거나 매칭하지 못한 스키마의 관계성을 정의할 때 사용자에게 요구되는 후처리 비용을 수치화한 값이다. 이 방식의 식은 식(4)와 같으며 Overall값이 작을수록 후처리 비용을 더 많이 요구함을 의미한다. 식(5)와 식(6)은 Overall값을 구하기 위한 정확도와 재현율을 의미한다. 식(5)의 정확도는 매칭이 기대된 스키마 수 중 정확하게 매칭된 수를 의미한다. 또한, 식(6)은 재현율은 매칭이 수행될 스키마 중 매칭이 완료된 스키마 수를 의미한다.

평가를 위해 GSim, XSDM, FRAG-BASE 방법을 제안한 알고리즘으로 구현하여 스키마 데이터 통합 성능 평가를 수행하였다.

누적된 스키마 수를 기준으로 4개의 스키마 매칭 방법을 이용하여 데이터 통합 성능을 측정한 결과는 Fig. 15와 같다. X축의 표는 그래프의 가독성을 높이기 위해 수치를 작성한 것으로 이를 통해 제안 방법인 HSM의 스키마 데이터 통합에서 우수한 성능을 보임을 알 수 있다.



	50	100	150	200	250	300	350	400	419
GSim	56%	64%	63%	63%	66%	68%	69%	68%	68%
XSDM	77%	74%	76%	76%	77%	77%	77%	77%	77%
FRAG-BASE	66%	70%	70%	70%	72%	72%	73%	73%	73%
HSM	77%	76%	77%	78%	79%	78%	79%	79%	79%

Fig. 15. Schema data integrated performance

GSim 스키마 매칭 방법은 유사성 측정을 통해 스키마 매칭을 수행하므로 스키마 수가 증가함에 따라 정확성이 떨어지므로 매칭되지 않은 스키마 수는 증가한다. 따라서 스키마 데이터 통합 성능은 68%로 가장 낮다. XSDM 스키마 매칭 방법은 사전에 스키마 간 의미 관계 정의를 수행하기 때문에 스키마 매칭 연산을 수행할 때 추가적인 오버헤드가 발생하지 않는다. 그러나 새로운 의미 관계 정의가 요구되는 경우 추가적인 의미 관계 정의를 수행해야 하므로 스키마 데이터 통합 성능은 77%로 높다. FRAG-BASE 스키마 매칭 방법은 사전에 표준화된 의미정보를 활용하여 스키마를 매칭하므로 표준화된 의미정보가 아닌 다른 의미정보간의 매칭이 수행되지 않는다. 따라서 스키마 매칭을 수행할 때 스키마 데이터 통합 성능은 73%로 보통이다. 제안한 HSM 방법은 스키마 매칭 방법은 사전에 스키마 간 의미 관계성 정의를 수행하는 XSDM 스키마 매칭 방법과 유사하나 스키마 등록 기능을 지원하고, 동적 스키마 관리가 가능하기 때문에 XSDM 스키마 매칭 방법의 데이터 통합 성능보다 우수하다.

7. 결 론

이 논문에서는 다양한 스키마에 관계없이 표준화된 형태의 스키마로 비공간정보를 표현하기 위해 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 알고리즘을 제안하고 구현하였다. 어댑터 기반 의미 처리방법과 정적 의미 기반 처리방법, 동적 의미 기반 처리방법을 혼합한 HSM 방법은 스키마 매칭 후 스키마 관리자가 스키마 매칭이 올바르게 정의되었는지 확인 가능하기 때문에 스키마 매칭의 정확도를 향상시킨다. 제안한 알고리즘을 평가하기 위해 기존 스키마 매칭 방법들을 제안한 알고리즘 기반으로 구현하였으며 GSim 스키마 매칭 방법과 XSDM 스키마 매칭 방법, FRAG-BASE 스키마 매칭 방법과의 정량적 비교를 수행하였다. 제안한 HSM 방법은 새로운 스키마 등록 기능을 지원하기 때문에 동적 스키마 관리가 가능하며, 이를 통한 스키마 매칭의 정확성을 점진적으로 향상시킨다. 또한 다양한 표준화된 의미정보를 활용하므로 어떠한 형태로든 스키마 매칭이 가능하므로 스키마 매칭 범용성은 높고, 스키마의 의미 간 관계성을 점진적으로 확장함으로써 기존방법들보다 효율적인 스키마 데이터 통합 비용을 제공한다. 이를 통해 제안한 알고리즘이 기존 스키마 매칭 방법들보다 더 나은 성능을 보였다.

제안한 HSM 방법은 비공간정보의 다양한 지리정보를 표현하고 해석하여 공유 및 확장할 수 있다. 이를 통해 웹에 존재하는 다양한 지리정보들을 하나로 통합하여 사용자가 원하는 지리정보를 풍부하게 제공해 줄 수 있다. 그러나 의미정보들의 관계성을 사전에 정의해야 하므로 초기 알고리즘 구축 비용이 많이 드는 단점이 있다. 향후 연구로는 변환된 스키마들에 대한 인스턴스 통합을 통해 중복되는 데이터를 제외하여 사용자들에게 지리정보 서비스로 제공하기 위한 연구가 요구된다.

참 고 문 헌

[1] J. Elson, J. Howell, and J. R. Douceur, “Map-Cruncher: Integrating the World’s Geographic Information,” ACM SIGOPS Operating Systems Review, Vol.41, No.2, pp.50-59, 2007.

[2] S. Bastian, B. Johannes and J. Simon, “Integrating OGC Web Processing Services into Geospatial Mass-Market Applications,” in Proceedings of the International Conference on Advanced Geographic Information Systems & Web Services, pp.98-103, 2009.

[3] T. Wilson, “KML,” Open Geospatial Consortium Inc., 2008.

[4] M. Kyle, D. Burggraf, S. Forde and R. Lake, “GML in JPEG 2000 for Geographic Imagery (GMLJP2) Encoding Specification,” Open Geospatial Consortium Inc., 2006.

[5] Open Geospatial Consortium Inc., http://www.opengeospatial.org/standards/gml

[6] C. Reed, “An Introduction to GeoRSS: A Standards Based Approach for Geo-enabling RSS feeds,” Open Geospatial Consortium Inc., 2006.

[7] Wikipedia, GeoRSS (information theory), http://www.georss.org/

[8] J. Lee, S. Lee, J. Kim, D. Jeong, D. Baik, “A Hybrid Schema Matching Method for Integrating Geographic Information,” The KIPS Proceedings of The 36th conference of the KIPS Fall conference 2011, Vol.18, No.2, pp.1272-1275, 2011.

[9] S. Madria, K. Passi and S. Bhowmick, “An XML Schema integration and query mechanism system,” in Proceedings of the Data & Knowledge Engineering, Vol.65, pp.266-303, 2008.

[10] E. Stefanakis and K. Patroumpas, “Google Earth and XML: Advanced Visualization and Publishing of Geographic Information,” The Lecture Notes in Geo information and Cartography Part B, pp.143-152, 2008.

[11] M. Francis, D. Ludovic and G. Patrick, “Corpus-Based

Structure Mapping of XML Document Corpora A Reinforcement Learning Based Model," in Proceedings of the COMPARATIVE EVALUATION OF XML INFORMATION RETRIEVAL SYSTEMS, Vol.370, pp.249-266, 2011.

[12] D. Jeong, S. Lee, J. Kim, D. Baik, "A Study on Semantic Processing Methods for Smart Mobile Services," The KIISE Proceedings of the Korea Computer Congress 2011, Vol.38, No.1(c), pp.89-92, 2011.

[13] P. Laura and S. Serena, "Automatic generation of probabilistic relationships for improving schema matching," in Proceedings of the Information Systems, Vol.36, pp.192-208, 2011.

[14] P. Jeffrey, P. Pallabi, K. Latifur, B. Thuraisinghama and S. Shashi, "Enhanced geographically typed semantic schema matching," in Proceedings of the Science, Services and Agents on the World Wide Web, Vol.9, pp.52-70, 2011.

[15] M. Sanjay, P. Kalpdrum and B. Sourav, "An XML Schema Integration and query mechanism system," in Proceedings of the Data & Knowledge Engineering, Vol.65, pp.266-303, 2008.

[16] C. Nengcheng, H. Jie, W. Wei and C. Zeqiang, "Extended FRAG-BASE schema-matching method for multi-version open GIS Web services retrieval," International Journal of Geographical Information Science, Vol.25, No.7, pp.1045-1068, 2011.

[17] K. Jaewook, P. Yun, I. Nenad and F. Junho, "An Optimization Approach for Semantic-based XML Schema Matching", International Journal of Trade, Economics and Finance, Vol.2, No.1, pp.78-86, 2011.

[18] J. Lee, S. Lee, J. Kim, D. Jeong, D. Baik, "An Ontology Matching Method based on ISO/IEC 11179," The KIISE Proceedings of the Korea Computer Congress 2012, Vol.39, No.1(c), pp.95-97, 2012.

[19] YAHOO Map, http://developer.yahoo.com/maps/georss/

[20] VIRTUAL, http://dev.live.com/Virtualearth/sdk/

[21] K. Almaliotis and I. Diakakis, "A Preliminary Attempt to Create a Unified Model for Obtaining and Processing Geodata: Geodata Information Sharing," in Proceedings of the Systems, Signals and Image, 2009.

[22] FOAF Vocabulary Specification, FOAF, http://xmlns.com/foaf/spec/20100809.html

[23] The GeoJSON Format Specification, GeoJSON, http://geojson. org/geojson-spec.html

[24] Geonames, http://www.geonames.org/

[25] Open API of Daum Maps, http://dna.daum.net/apis/view_all

[26] S.Melnik, H. Garcia-Molina and E, Rahm, "Similarity Flooding - A Versatile Graph Matching Algorithm," in Proceeding of the Data Engineering, pp.117-128, 2002.

이 지 윤

e-mail : sheisljy@korea.ac.kr

2011년 중부대학교 정보통신학과(학사)

2011년~현 재 고려대학교 컴퓨터·전파통신공학과 석사과정

관심분야 : GIS, 데이터베이스, 클라우드 컴퓨팅, 메타데이터 레지스트리

이 석 훈

e-mail : leha82@korea.ac.kr

2009년 고려대학교 전자 및 정보공학부(학사)

2009년~2011년 고려대학교 컴퓨터·전파통신공학과(공학석사)

2011년~현 재 고려대학교 컴퓨터·전파통신공학과 박사과정

관심분야 : 온톨로지, 데이터마이닝, 메타데이타 레지스트리, 자율 컴퓨팅, 자가 적응형 소프트웨어

김 장 원

e-mail : ikaros1223@korea.ac.kr

2005년 상명대학교 소프트웨어공학과(학사)

2005년 한국과학기술연구원(KIST) 위촉연구원

2008년 고려대학교 컴퓨터학과(석사)

2008년~2012년 고려대학교 컴퓨터·전파통신공학과 박사과정

2012년 고려대학교 컴퓨터학과(박사)

관심분야 : 온톨로지, 시맨틱 웹, GIS, 데이터베이스, 메타데이터 등

정 동 원

e-mail : djeong@kunsan.ac.kr

1997년 군산대학교 컴퓨터과학과(이학사)

1999년 충북대학교 전산학과(이학석사)

2004년 고려대학교 컴퓨터학과(이학박사)

2004년~2005년 고려대학교 정보통신기술연구소 연구조교수

2005년 Pennsylvania State University PostDoc.

2002년~2004년 TTA 표준화위원회-데이터연구회(SG08.02) 특별위원

2004년~현 재 TTA 표준화위원회-메타데이타 표준화 프로젝트 그룹(PG406) 위원

2005년~현 재 군산대학교 통계컴퓨터과학과 교수

2006년~현 재 ISO/IEC JTC 1/SC 32 국내전문위원회 위원

2008년~현 재 ISO/TC 211 국내전문위원회 위원

2009년~현 재 TTA 표준화위원회-NGIS 프로젝트그룹 위원

2010년~현 재 인터넷윤리실천협회 이사

2010년~현 재 ICDL Korea 교수위원

2010년～현 재 전북지역 과학기술정보협의회 위원
2010년～현 재 한국과학기술정보연구원 자문위원
2010년～현 재 한국컴퓨터교육학회 이사
관심분야 : 데이터베이스, 시맨틱 웹, 시맨틱 GIS, 유비쿼터스 컴퓨팅, 시맨틱 모바일 서비스, 클라우드 컴퓨팅 등

백 두 권

e-mail : baikdk@korea.ac.kr

1974년 고려대학교 수학과(학사)
1977년 고려대학교 산업공학과(석사)
1983년 Wayne State Univ. 전산학과 (석사)
1985년 Wayne State Univ. 전산학과 (박사)
1986년～현 재 고려대학교 컴퓨터・전파통신공학과 교수
1991년～현 재 (사)한국시뮬레이션학회 (이사/부회장/감사/회장/고문)
1991년～현 재 ISO/IEC JTC1/SC32 전문위원회(위원장)
1999년～2000년 고려대학교 컴퓨터과학기술연구소(소장)
1999년～1999년 한국 DB 진흥센터 (표준연구위원)
2000년～2003년 소프트웨어 컴포넌트 표준화 포럼(부의장)
2001년～현 재 (사)도산아카데미(원장)
2002년～2004년 고려대학교 정보통신대학(초대학장)
2004년～2005년 (사)정보처리학회(부회장)
2004년～현 재 한국 프로젝트 관리 연구회(회장)
2009년～2010년 고려대학교 정보통신대학 학장
관심분야 : 메타데이터, 소프트웨어공학, 데이터공학, 컴포넌트 기반 시스템, 메타데이터 레지스트리, 프로젝트 매니지먼트 등