

지리정보 통합을 위한 하이브리드 스키마 매칭 방법

이지윤*, 이석훈*, 김장원*, 정동원**, 백두권*

*고려대학교 컴퓨터·전파통신공학과

**군산대학교 정보통계학과

e-mail : sheisljy@korea.ac.kr, leha82@korea.ac.kr, ikaros1223@korea.ac.kr,

djeong@kunsan.ac.kr, baikdk@korea.ac.kr

A Hybrid Schema Matching Method for Integrating Geographic Information

Jiyeon Lee*, Sukhoon Lee*, Jangwon Kim*, Dongwon Jeong**, Doo-Kwon Baik*

*Dept. of Computer and Radio Communications Engineering, Korea University

**Dept. of Informatics & Statistics, Kunsan National University

요 약

네이버, 다음, 구글 등의 웹 상에서 제공하는 지도 서비스와 KML, GML, GeoRSS와 같은 기술들을 이용하여 하나의 위치 정보로 통합한 지리정보를 사용자에게 제공해 줄 수 있는 연구들은 현재까지 활발히 진행되어 왔다. 그러나 이러한 연구들은 위치 정보만 통하여 줄 뿐 의미 정보까지 통합하여 사용자들에게 다양한 정보들을 제공해주지 못한다. 이 논문에서는 KML, GML, GeoRSS 등으로 표현된 풍부한 지리정보들을 통합하여 웹 기반 지도 서비스에 제공해주는 시스템을 제안한다. 또한 지리정보들의 스키마 통합을 위해 어댑터 기반 의미 처리 방법과 정적/동적 의미 관리 기반 접근 방법을 혼합한 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 방법을 제안하고, 제안 시스템의 평가를 위해 스키마 매칭을 위한 4가지 접근 방법과 비교 평가를 수행한다. 평가의 결과로 제안 시스템은 의미 해석에 대한 신뢰성이 보장되고 시스템 구축 비용과 데이터 통합 비용이 상대적으로 낮다는 특징을 지닌다.

1. 서론

현재 웹 기반 지도 서비스들은 사용자들에게 다양한 지리정보들을 제공해준다. 그 중 네이버, 다음, 구글 등의 웹 기반 지도 서비스를 제공해 주는 회사에서는 지도 서비스를 포함한 여러 가지 오픈 API를 제공함으로써 이를 이용하여 다양한 매쉬업(Mash-up) 서비스들이 등장하였다[1]. 또한 매쉬업 서비스에 공간정보 이외에 비공간정보를 포함하고, 보다 풍부한 데이터를 보여주기 위해 KML[2], GeoRSS[3], GML[4] 등과 같은 기술들이 사용되고 있다. 특히 GeoRSS는 RSS에 위치 정보가 포함된 것으로서 다양한 데이터들을 자동으로 수집하여 갱신할 수 있어 매쉬업 서비스에서 다양하게 적용되고 있다[5, 6].

이 논문은 KML, GeoRSS, GML 등으로 표현된 풍부한 지리정보들을 통합하여 웹 기반 지도 서비스에서 제공하는 시스템을 제안한다. 웹 기반으로 다양한 데이터를 학습하고 확장하기 위한 방법으로 크게 두 가지 이슈 사항

이 있다. 첫 번째는 지리정보들을 표현하기 위해 사용되는 XML 스키마의 의미를 파악하여 서로 같은 의미의 스키마를 통합하는 것이고[7], 두 번째는 여러 지리정보를 통합할 때 발생하는 중복되는 인스턴스들을 통합하는 것이다[8]. 그 중 이 논문은 첫 번째 이슈 사항인 스키마의 통합에 초점을 맞추어 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM) 방법을 제안하고, 이 방법을 이용하여 웹 상의 다양한 지리정보들을 통합하고 웹 기반 지도 서비스로 해당 정보들을 제공해 주는 시스템을 제안한다.

이 논문의 구성은 다음과 같다. 제2장에서는 스키마 통합을 위한 위치 정보 통합과 스키마 매칭에 관한 연구를 소개하고 제3장에서는 하이브리드 스키마 매칭 방법과 이를 이용한 지리정보 통합 시스템을 보여준다. 제4장에서는 제안한 방법에 대한 평가를 기술하고 제5장에서는 논문에 대한 결론을 기술한다.

2. 관련 연구

2.1 공간정보 통합 관련 연구

지리정보를 표현하기 위하여 XML 문법을 기반으로 하는 KML, GML 등이 사용된다. 각각의 기술들은 지리정

이 연구에 참여한 연구자는 '2단계 BK21 사업'의 지원을 받았으며, 지식경제부의 기술혁신사업-표준기술력향상사업(산업기술표준화 및 인증지원사업)의 일환으로 수행되었음. 또한 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2011-0004911).

† 공동책임저자 : 백두권, 정동원

보를 표현할 때 서로 상이한 좌표 체계를 기반으로 한다. 따라서 지리정보의 통합을 위하여 KML, GML 등이 가지고 있는 좌표 정보를 특정 좌표체계로 변환하기 위한 작업이 필요하다. GML2KML Conversion은 GML을 KML로 변환시켜 GML에 있는 지리정보들을 구글 지도에 보여준다[7]. 또한, Geo Parsing MEDIATOR는 KML, GML 간의 지리정보 데이터의 상호 운용을 가능하게 하기 위해 지리정보 데이터가 가질 수 있는 메타데이터 간의 관계를 분석하는 데이터 모델을 만들어 제공한다[9]. 그러나 이러한 연구들은 각 지리정보 표현을 위한 언어들이 지니는 위치 정보만을 상호 적용 시킬 뿐, 각각의 언어들이 가지고 있는 스키마들의 의미를 파악하여 변환하는 방법까지는 제시하지 않는다. 이는 지리정보의 통합 시 위치 정보는 통합 할 수 있지만 위치 정보 이외의 다른 공간정보 혹은 비공간정보들은 통합하지 못하는 문제점을 지닌다.

2.2 스키마 매칭 관련 연구

지리정보는 공간정보 뿐 만 아니라 해당 위치의 이름과 설명과 같은 비공간정보들도 포함하고 있다. 이러한 비공간정보들의 통합을 위하여 스키마 매칭 기법이 필요하다. [10]은 메타데이터 간의 의미의 상호운용성 증진을 위하여 스키마 매칭 기법들을 4가지 접근 방법으로 분류하고 각 분류에 따른 의미 처리 방법들을 기술한다.

첫 번째는 유사도 기반 의미 처리 방법으로, 이 방법은 스키마들 사이의 유사도 측정 알고리즘을 이용하여 스키마들을 매칭한다. 또한 [11]은 워드넷에 어휘 간 유사성 기법을 적용하여 스키마 매핑을 반자동적으로 수행한다.

두 번째는 어댑터 기반 의미 처리 방법으로, 이 방법은 매핑 테이블 혹은 매핑 규칙 등과 같이 서로 상이한 스키마 간의 관계성을 사전에 정의하여 의미적 상호 운용성을 해결하는 방법이다. [12]는 여러 개의 XML 데이터를 하나로 통합하기 위하여 매핑 테이블과 매핑 규칙을 이용한 XML 스키마 통합 및 질의 방법을 제안한다. [13]은 분산 환경에서 온톨로지 정렬을 위하여 스키마의 검색을 매핑 테이블을 생성에 이용하는 방법을 제안한다.

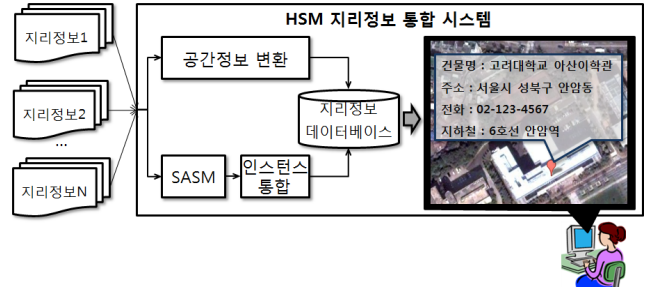
세 번째와 네 번째는 정적/동적 의미 관리 기반 접근 방법으로, 이 방법은 표준화 된 의미 정보를 정의하고 이를 이용하여 각각의 지리정보의 스키마를 정의함으로써 근본적인 의미의 불일치 문제를 해결하는 방법이다. 정적 의미 관리와 동적 의미 관리의 차이는 스키마 정의에 이용되는 표준화 된 의미정보가 새로운 의미 생성 및 이용 시 즉각적으로 반영할 수 있는가의 차이를 나타낸다. 정적 의미 관리 기반 접근 방법은 서지 정보 분야의 MARC[14], Dublin Core[15] 등을 예로 들 수 있으며, 동적 의미 관리 기반 접근 방법은 MDR 등에 기반한 스키마 정의에 이용될 수 있다[16, 17].

3. 제안 시스템

이 장에서는 제안 시스템을 기술하고 지리정보 통합을

위한 스키마 매칭 방법인 하이브리드 스키마 매칭(Hybrid Schema Matching, HSM)을 제안한다.

그림 1은 제안 시스템의 구조로 웹 상에 존재하는 다양한 지리정보를 가져와 통합된 지리정보 시스템을 구축하여 사용자에게 제공해 주기 위한 구조를 보인다.



(그림 1) 제안 시스템 구조

제안 시스템은 웹 상에서 KML, GML 등으로 표현된 지리정보를 GeoRSS와 같은 기술을 이용하여 가져온다. 각각의 지리정보는 공간정보와 비공간정보를 포함하고, 각각 공간정보 변환 모듈과 지리정보 모듈과 인스턴스 통합 모듈을 이용하여 통합된 지리정보를 얻는다. 통합된 지리정보는 지리정보 데이터베이스, 즉 통합된 좌표체계와 스키마를 가지는 지리정보 저장소에 저장된다. 사용자는 웹 기반 지도 서비스를 이용함으로써 지리정보 저장소에 저장된 지리정보를 제공받는다.

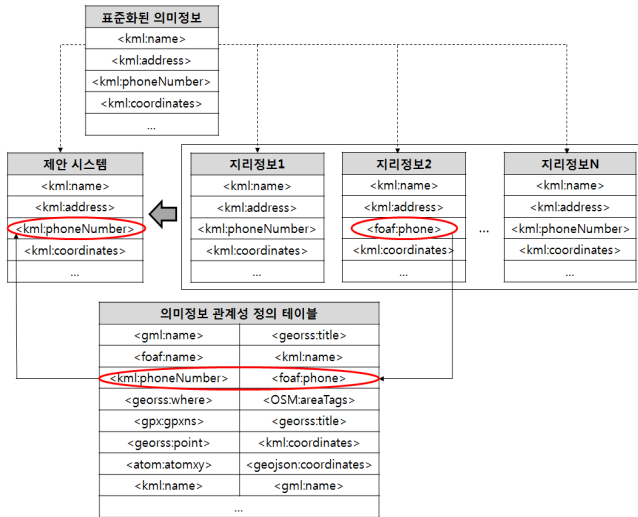
3.1 공간정보 변환 모듈

공간정보 변환 모듈에서는 지리정보에 포함된 공간정보를 통합한다. 제안 시스템은 표현의 풍부성, 구글 지도 적용의 편의성 등으로 인해 공간정보를 KML로 통합한다. 따라서 GML에서 표현된 좌표체계를 KML의 좌표체계로 변환하기 위하여 GML2KML Conversion, Geo Parsing MEDIATOR 등과 같은 기술들을 사용한다. 이러한 기술을 사용하여 통합된 좌표체계는 공간정보의 스키마 뿐 아니라 인스턴스의 실제 좌표까지 변환한다. 제안 시스템의 통합된 공간정보는 KML이 가지는 좌표체계인 WGS84로 변환된다.

3.2 HSM 모듈

HSM은 어댑터 기반 의미 처리 방법과 정적/동적 의미 관리 기반 접근 방법을 혼합한 하이브리드 스키마 매칭 방법이다. 이 방법은 정적/동적 의미 관리 기반 접근 방법과 같이 표준화된 의미 정보를 활용하여 제안 시스템과 지리정보들의 스키마를 동일하게 정의한다. 그러나 이러한 방법은 어떤 지리정보가 동일한 표준화된 의미 정보를 활용하지 않고, 다른 표준화된 의미 정보를 활용 할 경우 해당 지리정보를 통합하지 못한다. 이를 해결하기 위하여 어댑터 기반 의미 처리방법과 같이 표준화된 의미 정보 간 관계성을 정의함으로써 제안 시스템과 다른 표준화된 의미

정보를 활용한 지리정보의 통합이 가능해 진다.



(그림 2) 하이브리드 스키마 매칭 방법

예를 들어, 그림 2와 같이 제안 시스템과 N개의 지리 정보에 대한 통합을 위해 HSM 방법을 이용하는 상황을 가정한다. 먼저 KML이 가지는 의미정보들을 기반으로 표준화된 의미정보를 정의하고, 제안 시스템과 지리정보들은 모두 표준화된 의미정보를 활용하여 KML의 의미정보로 동일하게 정의한다. 그러나 실제로 모든 지리정보들이 동일하게 정의할 수 없으므로 이럴 경우 표준화된 의미정보 간 관계성을 정의한 의미정보 관계성 정의 테이블을 이용한다. '지리정보2'가 가지고 있는 세 번째 스키마인 '<foaf:phone>'은 제안 시스템이 가지고 있는 세 번째 스키마와 일치하지 않으므로 의미정보 관계성 정의 테이블을 참조하여 해당 스키마가 제안 시스템의 '<kml:phoneNumber>'와 동일하게 매치 되었다.

지리정보는 어댑터 기반 의미 처리 방법에서 사용하는 관계성을 정의할 때 일반적인 스키마 간의 관계성을 정의하는 것이 아닌 표준화된 의미정보 간 관계성을 정의하는 차이가 있다. 이는 새로운 지리정보가 추가될 때마다 스키마 간 관계성을 정의하는 것이 아닌 새로운 표준화된 의미정보를 활용한 지리정보가 추가될 때마다 표준화된 의미정보의 관계성을 정의하는 것을 의미한다.

3.3 인스턴스 통합 모듈

인스턴스 통합 모듈은 지리정보에 포함된 인스턴스들을 지리정보 모듈에서 통합된 스키마와 일치하도록 인스턴스의 스키마를 변환하는 작업을 한다. 통합된 스키마를 지니는 인스턴스들은 이후 통합된 지리정보 데이터베이스로 저장된다.

4. 평가

표 1은 스키마 매칭을 위한 4가지 접근 방법과 표준 어댑터 기반 스키마 매칭 방법의 정성적인 비교 평가 결과를 보인다. 평가 항목은 다음과 같이 4가지 항목으로 정의한다.

- 의미 해석의 정확성 : 스키마 매칭 후 매칭 결과의 정확성을 의미
- 시스템의 범용적 활용성 : 다른 시스템에서 스키마 매칭이 이루어지기 위한 지리정보의 적용 가능성 및 범위를 의미
- 시스템 구축 비용 : 스키마 매칭 방법을 이용하기 위한 시스템 구축 시 요구되는 초기 비용으로 스키마 매칭을 위한 사전 처리 비용을 의미
- 데이터 통합 비용 : 스키마 매칭이 이루어 질 때 필요한 시스템 자원 및 시간 등을 의미

유사도 기반 의미 처리 방법은 유사성 측정 기법을 적용하여 자동적인 스키마 매칭이 이루어지므로 의미 해석에 대한 정확성을 보장하지 못한다. 그러나, 어떠한 형태의 스키마에도 적용 가능하므로 시스템의 범용적 활용성은 매우 높다. 시스템 구축 비용과 데이터 통합 비용은 유사성 측정 기법에 따라 차이가 있지만, 시스템 구축 비용은 한 번 구축하면 추가적인 유지비용이 발생하지 않으므로 보통의 비용이 소요되고, 데이터 통합 비용은 스키마 매칭시 처리 시간이 비교적 오래 걸리므로 높다.

어댑터 기반 의미 처리 방법은 스키마 간 의미에 대한 관계성을 사전에 정의하므로 의미 해석의 정확성이 보장된다. 또한, 다양한 스키마에서 스키마 간 관계성을 사전 정의에 의해 적용 가능하므로 시스템의 범용적 활용성은 높다. 그러나 스키마 간 의미 관계 정의를 위한 사전 처리 비용이 발생하고, 새로운 의미 관계 정의가 요구될 때마다

<표 1> 스키마 매칭 방법 간 정성적 비교 평가

평가대상 / 평가항목	유사도 기반 의미 처리 방법	어댑터 기반 의미 처리 방법	정적/동적 의미 관리 기반 접근방법	하이브리드 스키마 매칭 방법
의미 해석의 정확성	정확성 보장 못함	정확성 보장	부분적 정확성 보장	정확성 보장
시스템의 범용적 활용성	매우 높음	높음	낮음	매우 높음
시스템 구축 비용	보통	높음	초기 비용 높음	보통
데이터 통합 비용	높음	보통	매우 낮음	낮음

상호 의미 관계를 추가 정의해야하므로 시스템 구축 비용은 높다. 그러나 의미 관계를 한 번 정의해 놓으면 추가적인 오버헤드가 발생하지 않으므로 데이터 통합 비용은 보통이다.

정적 의미 관리 기반 접근 방법과 동적 의미 관리 기반 접근 방법은 스키마 정의시 표준화된 의미 정보에 활용하여 각 스키마들을 동일하게 정의하므로 동일한 방법으로 본다. 정적/동적 관리 기반 접근 방법은 부분적으로 표준화된 의미 정보를 활용하지 않은 부분에 대한 정확성을 보장하지 못하므로 부분적 정확성을 보장한다. 또한, 사전에 정의된 부분만 정확한 의미 처리가 가능하다는 한계를 지니기 때문에 시스템의 범용적 활용성은 낮다. 그러나 동일한 표준화된 의미 정보를 활용하여 모든 스키마를 정의하므로, 각 시스템들의 높은 초기 비용이 요구된다. 또한, 시스템 자원과 스키마 매칭에 사용되는 시간에 따른 오버헤드가 발생하지 않으므로 데이터 통합 비용은 매우 낮다.

하이브리드 스키마 매칭 방법은 기본적으로 표준화된 의미 정보를 활용하여 사전에 스키마를 정의하며 표준화된 의미 정보간의 관계성 정의까지 포함하고 있어 동일한 의미 정보를 활용하지 않더라도 정확한 스키마 매칭이 가능하다. 다양한 표준화된 의미 정보를 활용하여 정의된 스키마에 대한 매칭이 가능할 뿐만 아니라 표준화된 의미 정보를 활용하여 어떠한 형태로든 스키마 매칭이 가능하므로 시스템의 범용적 활용성은 매우 높다. 또한, 각 스키마들이 어떤 표준화된 의미정보라도 활용하여 정의되며 각 표준화된 의미 정보끼리의 관계성 정의가 이루어지므로 상대적으로 보통의 시스템 구축 비용이 요구되며 추가적인 표준화된 의미 정보끼리의 관계성 정의도 어댑터 기반 의미 처리 방법 보다 적으므로 데이터 통합 비용이 낮다.

5. 결론 및 향후연구

이 논문은 웹 상에서 존재하는 풍부한 지리정보들을 통합하여 웹 기반 지도 서비스에서 제공하기 위한 시스템을 제안하였다. 제안 시스템을 위하여 하이브리드 스키마 매칭 (Hybrid Schema Matching, HSM) 방법을 제안하고, 4가지 스키마 매칭 방법들과 정성적 비교 평가를 통하여 제안 방법의 특징을 기술하였다.

HSM은 어댑터 기반 의미 처리 방법과 정적/동적 의미 관리 기반 접근 방법을 혼합한 하이브리드 스키마 매칭 방법으로 각 접근 방법의 장점들을 수용하여 정확성이 보장되고 시스템 구축 비용 및 데이터 통합 비용이 낮은 비교적 우수한 성능을 보였다.

향후 제안된 모델을 구현하고 정량 평가를 수행하기 위한 연구가 요구된다. 그리고 서론에서 언급한 중복된 인스턴스의 통합을 고려하여 지리정보 통합 서비스를 제공하기 위한 연구가 요구된다.

참고문헌

- [1] eWIDEPLUS, "Open API & Mashup", 2009.
- [2] T. Wilson, "KML", OGC 07-147r2, 2008.
- [3] C. Reed, "An Introduction to GeoRSS: A Standards Based Approach for Geo-enabling RSS feeds", OGC 06-050r3, 2006.
- [4] M. Kyle, D. Burggraf, S. Forde and R. Lake, "GML in JPEG 2000 for Geographic Imagery (GMLJP2) Encoding Specification", OGC 05-047r3, 2006.
- [5] 이법석, 황병연. "GeoRSS 분할기법 기반의 지도 서비스 구현", 정보과학회논문지, 컴퓨팅의 실제 및 레터 제14권 제7호, 2008.
- [6] 김문기, 최문용. "시민서비스 고도화를 위한 GeoRSS 구현", 한국지형공간정보학회 추계학술대회, pp.93-96, 2010.
- [7] E. Stefanakis and K. Patroumpas, "Google Earth and XML: Advanced Visualization and Publishing of Geographic Information", Lecture Notes in Geoinformation and Cartography Part B, pp.143-152, 2008
- [8] L. M. Haas, M. Hentschel, D. Kossmann and R. J. Miller, "Schema AND Data: A Holistic Approach to Mapping, Resolution and Fusion in Information Integration", Springer-Verlag Berlin Heidelberg, LNCS 5829, pp.27-40, 2009
- [9] K. Almaliotis and I. Diakakis, "A preliminary attempt to create a unified model for obtaining and processing Geodata", Systems, Signals and Image Processing, 2009.
- [10] 정동원, 이석훈, 김장원, 백두권. "스마트 모바일 서비스를 위한 의미 처리 방법에 관한 연구", 한국컴퓨터종합학술대회 논문집, 제38권, 제1(C)호, pp.84-87, 2011.
- [11] J. Li, "LOM: a lexicon-based ontology mapping tool", Proceedings of the Performance Metrics for Intelligent Systems, 2004.
- [12] S. Madria, K. Passi and S. Bhowmick, "An XML Schema integration and query mechanism system", Data & Knowledge Engineering, Vol.65, pp.266-303, 2008
- [13] I. F. CRUZ, W. G. SUNNA and K. AYLOO, "Concept Level Matching of Geospatial Ontologies", GIS Planet Second Conference and Exhibition on Geographic Information, 2005.
- [14] MARC, <http://www.loc.gov/marc/>, 2011.
- [15] Dublin Core Metadata Initiative (DCMI) and DublinCore (DC), <http://dublincore.org/>, 2011.
- [16] ISO/IEC JTC1/SC32, ISO/IEC 11179 Metadata Registries(MDR) Part 3, 3rd Edition. 2009.
- [17] 이석훈, 정동원, 김장원, 백두권, "XOB: XMDR 기반의 온톨로지 생성 시스템", 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제16권, 제9호, pp.904-917, 2010.