http://dx.doi.org/10.3745/KIPSTD.2011.18D.5.385

효과적인 공간 데이터 마이닝을 위한 SOA 기반 데이터 통합 프레임워크 설계

문 일 환[†] · 허 환[††] · 김 삼 근[†††]

요 약

최근 농업 분야에 IT를 접목시킨 농업-IT 융합 기술에 대한 연구가 주목 받고 있다. 특히, 공간 데이터 마이닝(spatial data mining, SDM)을 이용한 농작물 관련 예측 서비스들을 통해 자연재해에 대한 피해를 줄이고 농작물의 생산성을 높이고자 하는 연구들이 있어 왔다. 그러나 예측 서비스를 위한 SDM에 필요한 학습 데이터는 분산되어 있는 데이터간의 이질성으로 인해 데이터 변환과 통합과정에 많은 비용과 시간이 발생한다. 또한 공간 데이터와 비공간 데이터 간의 공간적 이웃 관계를 연산하기 위해 대용량의 데이터에 대한 복잡한 연산과정이 필요하다. 본 논문에서는 각각의 데이터 소스를 하나의 서비스 단위로 취급함으로써 분산된 이질적인 데이터를 효과적으로 통합 관리할 수 있고 SDM을 위한 학습 데이터의 생산성을 향상시켜 최적의 예측 서비스의 발견을 지원해 주는 SOA 기반의 데이터 통합 프레임워크를 제안한다. 실험을 통해 경기도 이천시의 복숭아나무의 동해 피해지역에 대한 최적의 예측 서비스의 발견을 위해 제안 프레임워크를 효과적으로 적용할 수 있음을 확인하였다.

키워드 : SOA, 데이터 통합, 데이터 통합 관리 서비스, 예측 서비스, 공간 데이터 마이닝

A Design of SOA-based Data Integration Framework for Effective Spatial Data Mining

Il-hwan Moon[†] · Hwan Hur[††] · Samkeun Kim[†††]

ABSTRACT

Recently, the concern of *IT-in-Agriculture* convergence technology that combines information technology and agriculture is increasing rapidly. Especially, the crop cultivation related prediction services by spatial data mining (SDM) can play an important role in reducing the damage of natural disaster and enhancing crop productivity. However, the data conversion and integration procedure to acquire the learning dataset of SDM for the prediction service need a lot of effort and time, because of their heterogeneity between distributed data. In addition, calculating spatial neighborhood relationships between spatial and non-spatial data necessitates requires the complicated calculation procedure for large dataset. In this paper, we suggest a SOA-based data integration framework that can effectively integrate distributed heterogeneous data by treating each data source as a service unit and support to find the optimal prediction service by improving productivity of learning dataset for SDM. In our experiment, we confirmed that our framework can be effectively applied to find the optimal prediction service for the frost damage area, by considering the case of peach crop cultivation in Icheon in Korea.

Keywords : SOA(Service Oriented Architecture), Data Integration, Data Integration Management Service, Prediction Service, Spatial Data Mining

1. 서 론

최근 세계적인 이상기후로 인한 농작물 피해 및 식량부족 현상에 대응하기 위해 농업 분야에 IT를 접목시킨 농업 IT 융합기술에 대한 연구가 주목 받고 있다[1]. 특히 농작물의 경쟁력을 갖추기 위하여 이상 기후에 대한 농작물 피해지역, 생태조건에 따른 농작물의 성장변화와 같이 농작물 재배에 큰 영향을 미칠 수 있는 공간 정보가 반영된 환경적인 요소를 분석하고 예측할 수 있는 서비스 개발이 필요하다[2]. 이

† 준 회 원 : 한경대학교 컴퓨터공학과 박사과정
†† 정 회 원 : 경기동부과수농협 상무
††† 종신회원 : 한경대학교 컴퓨터공학과 교수(교신저자)
논문접수 : 2011년 5월 2일
수 정 일 : 1차 2011년 7월 4일, 2차 2011년 8월 2일
심사완료 : 2011년 8월 3일

러한 공간 정보가 반영된 예측 서비스는 공간 데이터 마이닝 (spatial data mining, SDM)을 이용하여 예측이 가능하다. SDM이란 공간 데이터의 특수성을 고려하면서 대용량의 데이터로부터 의미 있는 상관관계, 패턴, 경향 등을 찾아내는 일련의 과정이다[3].

그러나 위치 및 면적, 토양의 물리적 성질 및 화학적 성질, 고도, 강수량, 일조량, 풍속, 유통정보, 동해 피해정보, 하천과의 인접 거리 등과 같은 농작물 재배 환경과 관련된 데이터들은 공간 데이터와 비공간 데이터들로 구성되어 있다. 이와 같은 이질적인 데이터를 SDM의 학습 데이터로 사용하기 위해서는 공간데이터를 포함한 비공간 데이터 간의 공간적 이웃 관계를 연산하여 데이터를 통합한 후 별도의 저장 공간에 저장되어야 한다[4]. 특히 작물 재배 환경과 같은 공간 데이터를 포함한 이질적인 데이터들을 이용하여 SDM에 적용하기 위해서는 속성 데이터간의 공간 정보가 반영되도록 데이터 변환을 통한 통합과정이 필요하다. 이러한 통합과정을 위해 분산되어 있는 데이터를 수집하고 EAI(Enterprise Application Integration)[5] 또는 ETL(Extraction, Transformation, Loading)[6]과 같은 데이터 변환 솔루션을 이용하여 통합하는 과정은 많은 시간과 비용을 발생시키는 비효율적인 작업이다.

Service Oriented Architecture (SOA) 는 분산 객체의 형태로 존재하는 컴포넌트들 간의 메시지 통신을 통해 정보를 교환하는 느슨하게 연결된 형태의 S/W Architecture이다[7]. SOA를 이용하여 이질적인 데이터를 서비스 단위로 관리하고 Enterprise Service Bus (ESB)를 통해 서비스들 간의 연결을 지원하고 일관된 인터페이스를 제공할 수 있다[8]. 위와 같은 SOA의 특징을 이용하여 SDM에 필요한 데이터 전처리(preprocessing) 작업을 분석 데이터들에 대한 서비스들의 조합을 통해 학습 데이터에 필요한 형태로 변환이 가능하다. 또한 일관된 인터페이스의 서비스 단위로 데이터를 관리할 수 있어 데이터의 통합이 용이하고, 분석 데이터의 확장이 가능하다.

따라서 본 논문에서는 각각의 원시 데이터를 하나의 서비스 단위로 취급함으로써 분산된 이질적인 데이터를 효과적으로 통합 관리할 수 있고 SDM을 위한 학습 데이터의 생산성을 향상시켜 최적의 예측 서비스의 발견을 지원해 주는 SOA 기반의 데이터 통합 프레임워크를 제안한다. 제안 프레임워크는 각 원시 데이터를 일종의 서비스 단위로 간주하기 때문에 새로운 서비스 도출을 위한 과정에 일관된 서비스 기반 인터페이스를 제공할 수 있다. 따라서 이러한 인터페이스를 통하여 SDM을 위한 다양한 학습 데이터를 효과적으로 생성할 수 있고 이러한 데이터 셋을 이용하여 최적의 예측 서비스를 발견할 수 있도록 지원할 수 있다.

본 논문의 전체적인 구성은 다음과 같다. 2장에서는 관련 연구에 대해 알아보고, 3장에서는 제안한 데이터 통합 프레임워크를 설계하고 데이터 통합관리 프로세스를 기술한다. 4장에서는 제안한 프레임워크를 이용하여 SDM에 적용한 결과를 기술한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대해서 기술한다.

2. 관련 연구

2.1 SOA 기반 데이터 통합

SOA 기반의 웹 서비스는 서비스 단위로 재사용이 가능하므로 서비스들의 조합을 통해 새로운 서비스를 창출할 수 있다. 또한 서비스들 간의 느슨한 결합을 통해 내·외부 서비스들 간의 통합이 가능하다. 이러한 SOA의 장점을 이용하여 서비스들 간의 조합을 통한 데이터 통합이 가능하다[9].

Sha와 Xie[10]은 분산되어 있는 다양한 종류의 공간 데이터를 통합하고 WebGIS 애플리케이션 서비스를 지원하기 위한 SOA를 설계한다. Sha와 Xie는 공간 데이터 소스에 대한 Data provider 서비스와 Data Integration 모듈을 이용하여 데이터를 통합한다. 또한 통합된 데이터는 Data provider 서비스를 통하여 WebGIS 애플리케이션 서비스를 지원하는 Backend service에 이용된다.

[11]에서는 서비스 기반의 데이터 마이닝 서비스의 장점을 활용하기 위해 SOA 기반의 데이터 마이닝 시스템 설계를 위한 프레임워크를 제안한다. 제안한 프레임워크는 데이터 수준(data level)의 Entity class에 수집된 데이터를 Component level의 Software component를 이용하여 데이터 전 처리 작업을 하게 된다. Service level의 Software service는 software component를 통해 전 처리된 데이터를 이용하여 데이터 마이닝 서비스를 제공하고 Interface level의 User interface를 통해 클라이언트에게 분석 정보를 제공한다.

[12]는 특정 프레임워크에 한정되어 있어 분산된 컴포넌트간의 상호운영이 어려운 전통적인 ETL방식에서 SOA를 이용하여 분산된 컴포넌트간의 느슨한 연결을 통한 상호운영이 가능한 ETL 프레임워크를 제안하였다. [12]에서 제안한 프레임워크는 서비스 오케스트레이션 포인트, 서비스 공급자, 서비스 소비자, 서비스 인터페이스로 구성된 SOA 기반의 비즈니스 계층과 데이터 계층으로 구성되어 있다. 클라이언트가 ETL 서비스를 요청하게 되면 오케스트레이션 포인트는 활성화가 되고 ETL 관련 서비스들에 대해 변환 요청을 하게 되고 변환 결과를 클라이언트에게 제공한다.

2.2 SDM

데이터베이스에서 패턴을 찾아내는 과정으로서의 데이터 마이닝 (data mining)은 연관규칙 (association rules), 분류 (classification) 및 예측 (prediction), 클러스터링 (clustering), 경향분석 (trend analysis) 등의 다양한 기능들을 제공한다[3]. 최근 데이터 마이닝 기법들을 다양한 응용분야에 쉽게 적용할 수 있는 연구들이 점진적으로 이루어져 왔다[13, 14, 15]. SDM은 기존 데이터 마이닝에 대규모의 공간관련 데이터의 통합을 요구한다. 따라서 이것은 공간 데이터를 이해하고 공간 및 비공간 데이터 사이의 공간관계 또는 관계들을 발견하는데 이용될 수 있다[16]. 본 논문에서는 이러한 분류 학습(classification learning)을 위한 알고리즘으로 의사 결정 트리(decision tree) 알고리즘과 신경망(neural network)의 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 학습시키기 위한 오류 역전파(backpropagation) 알

고리즘을 채택한다. 의사결정 트리 알고리즘으로는 C4.5를 선택하였다. 이 알고리즘은 목표 값(target value)을 이분화하기 위해 유력한 속성들(influential attributes)과 임계값들(thresholds)로 구성된 분류 트리를 생성해준다. 이러한 분류 트리는 다양한 분야에서 유력한 속성들을 선택하는데 이용되어 왔다[17, 18].

3. SOA 기반 데이터 통합 프레임워크

본 장에서는 SDM을 위한 SOA기반의 데이터 통합 프레임워크를 설계하고 데이터 통합을 위한 서비스 모델을 제안한다. 제안한 프레임워크는 분산되어 있는 데이터를 서비스 단위로 관리하고 SDM을 위한 학습 데이터에 대한 전 처리 작업 및 통합과정을 서비스 단위로 제공한다.

3.1 데이터 통합 관리 프레임워크

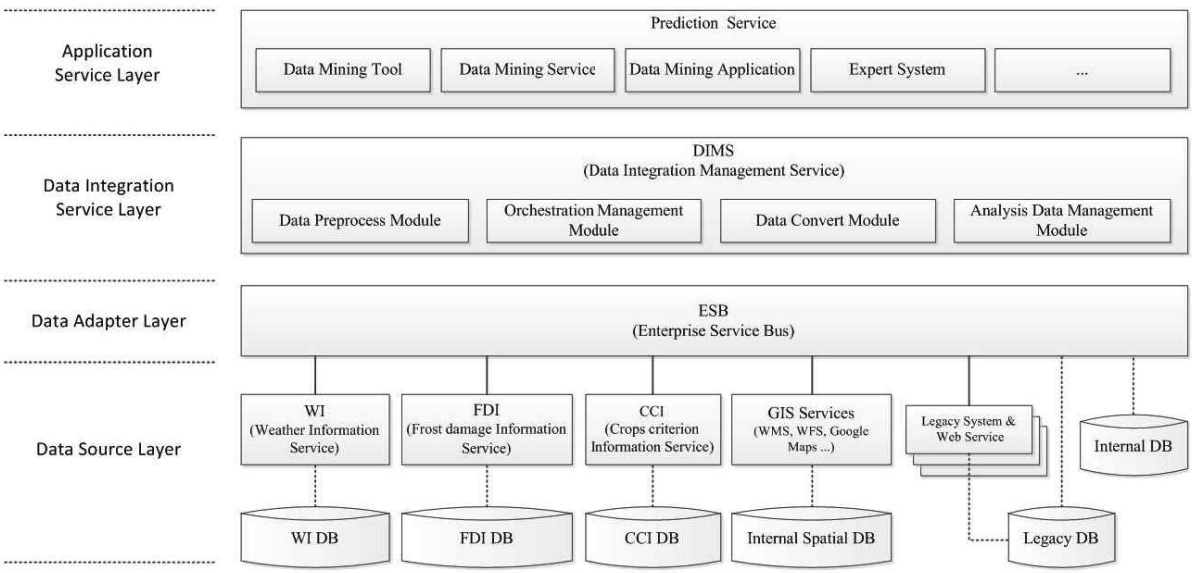
농작물 재배 환경정보와 같이 분산되어 있는 이질적인 데이터의 통합을 통해 데이터간의 공간적 연관 관계 추출과 같은 전 처리 작업을 효과적으로 지원하기 위하여 SOA기반 데이터 통합 프레임워크를 (그림 1)과 같이 제안한다.

제안한 프레임워크는 총 4개의 계층(layers)으로 구성되며 각각은 다음과 같은 기능을 가진다.

데이터 소스 계층(Data Source Layer) - SDM에 필요한 기초 데이터 및 데이터 통합을 위한 내부 데이터베이스를 제공해주는 계층이다. 데이터 분석에 필요한 비공간 데이터는 특정 지역에 대한 과거의 기상 데이터를 통해 기상변화 정보를 제공하는 WI(Weather Information) 서비스, 연도별 특정 농작물의 동해 피해지역에 대한 위치 정보를 제공하는 FDI(Frost Damage Information) 서비스, 농작물 재배적지에 대한 다양한 환경 및 재배 조건에 필요한 항목들을 수치화하여 기준정보를 제공하는 CCI(Crop Criterion Information) 서비스 등이 해당된다. 또한 분석에 필요한 다양한 데이터 획득을 위하여 레거시 시스템 및 서비스, 레거시 데이터베이스도 포함될 수 있다. 공간데이터는 OGC(Open Geospatial Consortium)의 WMS(Web Map Service), WFS(Web Feature Service) 등과 같은 Map Services와 Google Map API Service와 같은 공개된 공간 데이터 서비스 및 내부 공간 데이터베이스를 이용하여 공간 데이터에 대한 전 처리 작업을 지원하는 GIS(Geographic Information Systems) 서비스 등이 포함된다. 즉 Data Source Layer는 농작물 재배적지 또는 동해 피해지역 예측과 같은 분석에 필요한 기초 데이터를 수치화하고 분산되어 있는 데이터를 서비스 단위로 제공한다.

데이터 어댑터 계층(Data Adapter Layer) - 데이터 소스 계층에서 제공되는 다양한 데이터에 대한 일관된 인터페이스를 제공하고 계층 간의 통합이 가능하도록 연결해주는 역할을 하는 계층이다. ESB(Enterprise Service Bus)를 이용하여 느슨한 연결(loosely coupling)을 지원하며 유연성과 민첩성을 보장해주는 역할을 수행한다.

데이터 통합 서비스 계층(Data Integration Service Layer) - 데이터 어댑터 계층을 통해 제공되는 서비스들의 조합을 통해 SDM에 필요한 데이터 전 처리 작업을 지원할 수 있는 DIMS(Data Integration Management Service)를 제공하는 계층이다. DIMS는 ADM(Analysis Data Management) 모듈을 이용하여 SDM에 필요한 분석 데이터를 생성하고 관리할 수 있다. ADM은 서비스 요청에 대한 데이터를 생성하고 새로운 서비스 조합 생성과 데이터 변환 규칙을 설정한다. DP(Data Preprocess) 모듈은 ADM 모듈을 통해 설정된 변환 기준 값을 이용하여 분석 조건에 맞는 데이터 변환 기능을 제공한다. OM(Orchestration Management) 모듈은 데이터 전 처리 과정에 필요한 서비스의 조합을 통해 SDM에 필요한 통합 데이터를 제공한다. DC(Data Convert) 모듈



(그림 1) SOA 기반 데이터 통합 프레임워크

은 OM 모듈에 의해 생성된 통합 데이터를 분석에 필요한 특정 포맷의 데이터로 변환하는 기능을 제공한다.

응용 서비스 계층(Application Service Layer) - DIMS에서 제공되는 전 처리된 통합 데이터를 데이터 마이닝 툴(data mining tools), 데이터 마이닝 서비스(data mining services), 데이터 마이닝 애플리케이션(data mining applications, 전문가 시스템(expert systems) 등과 같은 분석 시스템에 적용하여 예측서비스와 같은 응용 서비스를 제공하는 계층이다.

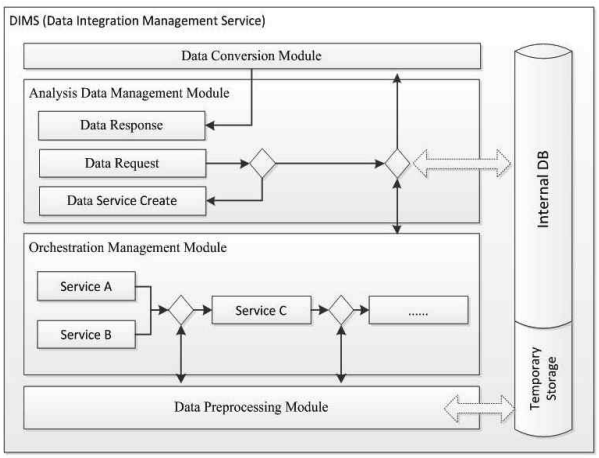
3.2 데이터 통합 관리 프로세스

데이터 통합과 관련하여 EAI 및 ETL 솔루션과 같이 특정 애플리케이션에서 산출되는 데이터를 변환하고 통합하는 기술은 각 솔루션 벤더마다 독자적인 기술을 사용하므로 사용자 요구사항에 맞는 솔루션을 도입해야한다. 또한 레거시 시스템과의 통합을 통해 데이터 추출, 변환 작업을 거쳐 타겟 시스템으로 전송 및 적재(loading)하는 과정이 이루어지므로 이를 위한 데이터 웨어하우스가 구축되어야 한다. 데이터 웨어하우스를 구축하는 과정에는 많은 가정과 추론이 필요하며, 전체적인 시스템에 영향을 주지 않으면서 분석을 위한 데이터가 수집되어야 한다. 위와 같은 분석 데이터 수집과정은 서로 다른 운영체제, 데이터베이스, 하드웨어 플랫폼 및 네트워크 환경을 고려해야하므로 많은 인력과 비용이 발생할 수 있다. 본 논문에서 제안한 프레임워크의 데이터 통합 관리 프로세스는 EAI 및 ETL과 같은 솔루션을 이용하여 데이터 웨어하우스를 구축하지 않고 데이터 전 처리 및 통합 과정을 서비스 단위로 관리하여 사용자의 요구사항에 맞는 분석 데이터를 효과적으로 제공한다. 또한 서비스 조합을 통하여 새로운 분석 데이터 서비스를 제공할 수 있다.

서비스 단위로 관리되는 분석 데이터를 통합하기 위해 데이터 어댑터 계층을 통해 서비스들을 연결한다. 데이터 어댑터 계층에 연결된 서비스들은 DIMS를 통해 새로운 분석 데이터 서비스를 제공하게 된다. DIMS는 데이터 변환 기준 값을 서비스 단위로 제공받을 수 있으며 서비스 확장을 통해 다양한 데이터 변환이 가능하다. 특히 SDM을 위한 공간 데이터와 비공간 데이터의 연관 관계를 통합하기 위해 GIS 서비스를 이용하여 변환과정이 이루어진다. SDM의 분석을 위해 생성되는 일련의 반복된 데이터 변환과정은 서비스를 요청하는 데이터 변환 기준 값을 달리하면서 다양한 분석 데이터를 쉽게 생성할 수 있다.

제안한 프레임워크에서 서비스 단위로 관리되는 데이터들을 이용한 데이터 전 처리 작업과 변환된 데이터의 통합 처리 과정은 (그림 2)와 같다.

1) 응용 서비스 계층의 서비스 사용자는 DIMS의 ADM 모듈에게 필요한 데이터를 요청한다.
2) ADM 모듈은 요청한 데이터에 대한 내부 데이터베이스의 존재여부를 판단한다. 데이터가 존재할 경우 DC 모듈을 이용하여 사용자가 요청한 데이터 형식에 맞게 전송한다.
3) 데이터가 존재하지 않을 경우에는 데이터 제공 서비스 존재 여부를 판단하여 서비스가 존재 하지 않을 경우에는 새로운 서비스를 생성할 수 있다.
4) 데이터 제공 서비스가 존재할 경우에는 OM 모듈을 이용한 서비스 조합을 통해 요청된 데이터를 생성한다. OM 모듈은 DP 모듈을 이용하여 변환 조건에 맞는 새로운 데이터를 생성하며 중간 단계에서 변환된 데이터는 임시저장 공간에 저장된다. 대용량의 공간 데이터를 변환할 경우 데이터 서비스의 연결이 중단될 수 있다. 이와 같은 문제를 해결하기 위한 방법으로 데이터 서비스에 대한 변환 성공 여부와 순서를 기록하여 서비스가 재개되었을 경우 중단된 이전 시점부터 다시 시작할 수 있도록 한다.
5) 변환된 데이터는 내부 데이터베이스에 저장되며 DC 모듈을 통해 서비스 사용자가 요청한 데이터 형식에 맞게 변환되어 제공된다. 내부 데이터베이스는 데이터 변환 서비스를 코드 값으로 관리하기 위한 TB_CONVERT_SERVICE 테이블, 서비스 코드 값을 기준으로 변환 데이터의 속성 이름을 저장하는 TB_ATTRIBUTE_NAME 테이블, 검색조건에 따른 데이터 변환 서비스를 관리하기 위한 TB_CONVERT_CONDITION 테이블, 변환 서비스 코드 값과 검색 조건 값에 대한 인덱스 코드 값을 이용하여 변환 데이터의 속성 값을 저장하는 TB_ATTRIBUTE_VALUE 테이블로 구성되어있다.



(그림 2) 데이터 통합 관리 프로세스

위와 같이 데이터 통합을 위해 서비스 단위로 관리하는 제안한 프레임워크의 장점은 다음과 같다.

- 이질적인 데이터들을 일관된 인터페이스를 통한 서비스 단위로 관리할 수 있어 데이터 통합 및 확장이 용이하다.
- 제안한 프레임워크에서 제공하는 웹 서비스들을 이용하여 데이터 변환도구를 이용하지 않고 쉽게 데이터 변환 작업이 가능하다.
- 공간 데이터를 이용하여 데이터간의 공간적 관계요소를 변환하는 복잡한 작업을 데이터 서비스 조합을 통해 설정할 수 있으며, 변환과정을 쉽게 추가하고 변경할 수 있다.
- 데이터 변환 서비스를 재사용하여 다양한 분석 데이터를 쉽게 생성할 수 있다.
- 웹 서비스 기반으로 데이터 서비스를 제공하므로 플랫폼에 상관없이 데이터 서비스가 가능하다.

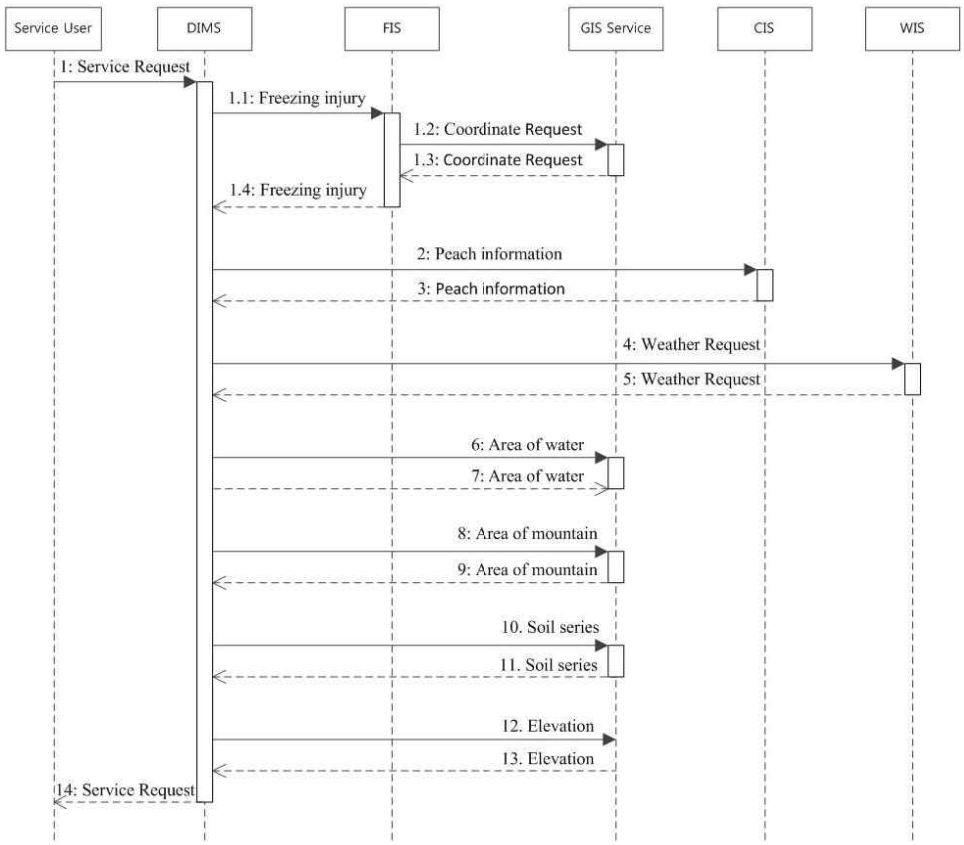
4. 실험 및 고찰

제안한 프레임워크는 DIMS를 통해 SDM에 필요한 다양한 데이터 셋을 효과적으로 생성할 수 있다. 즉 DIMS에 미리 정의된 서비스들에게 다양한 속성 값을 적용하여 대응되는 적절한 데이터 셋을 생성할 수 있으며, 이처럼 생성된 다양한 데이터 셋은 SDM의 학습 데이터로 활용될 수 있고 서로 비교 분석될 수 있다. 즉 속성 값의 변경에 따라 반복되는 일련의 데이터 통합 및 추출과정이 DIMS를 통해 재사용 가능하다. 이러한 특징은 속성 값의 변화에 따라 SDM의 예측 성능에 영향을 미칠 수 있는 도메인 지식(domain knowledge)을 효과적으로 찾을 수 있도록 지원할 수 있다. 제안한 프레임워크의 이러한 특징을 이용하여 획득된 데이터 셋을 경기도 이천시의 복숭아나무의 동해피해 지역 예측을 위해 SDM에 적용한다.

실험을 위해 Windows Communication Foundation(WCF), IIS 7.0, Microsoft SQL Server 2008, Microsoft BizTalk 2009를 이용하여 제안한 프레임워크를 구축하였다. 복숭아 동해피해 지역을 예측하기 위한 속성 데이터를 생성하기 위하여 FI, CCI, WI, GIS 서비스들을 제안 프레임워크의 데이터 소스 계층에 연결하였다. 복숭아나무의 동해피해 지역을 예측하기 위한 학습 데이터의 통합 및 추출과정에 대한 서비스를 DIMS에 추가하였으며, (그림 3)과 같이 내부 데이터 서비스 연결을 통하여 사용자 요청에 대해 응답하도록 하였다.

(그림 3)에서 서비스 이용자가 DIMS에게 경기도 이천시의 복숭아나무에 대한 동해피해지역의 분석 데이터를 요청하게 되면 DIMS 는 FDI 서비스에게 동해피해 지역정보를 요청하게 된다. FDI 서비스는 동해피해 지역의 주소에 대한 좌표정보를 GIS 서비스에게 요청하고 응답결과를 주소정보에 대한 좌표정보로 변환하여 DIMS에게 전달한다. DIMS는 복숭아의 동해피해 분석에 필요한 기준정보를 CCI 서비스에게 요청한다. CCI 서비스에서 응답한 결과 값에는 데이터 마이닝에 필요한 도메인 지식 정보가 포함되어 있다. DIMS는 CCI 서비스로부터 받은 기준정보와 동해피해지역에 대한 좌표 값을 이용하여 GIS 서비스로부터 주변지역의 하천영역, 주변지역의 임야면적, 동해피해지역에 대한 토양통 및 피해지역의 고도 정보를 요청하여 공간 데이터에 대한 정규화 작업을 처리한다. 각 속성 데이터에 대한 전 처리 작업이 끝나면 DIMS는 서비스 이용자가 요청한 데이터 타입에 맞게 변환하여 전달한다.

DIMS에 의해 획득된 데이터의 정보는 <표 1>과 같다. 생성된 데이터의 샘플 수는 808개이며 동해피해지역의 개수는 646, 피해를 받지 않은 지역은 162개이다. 클래스 속성 값인 동해 피해를 Yes, No 값으로 설정하였으며 나머지 데이터는 Real 데이터 값이다.



(그림 3) 데이터 통합 시퀀스 다이어그램

〈표 1〉 실험 데이터의 구성

구분	인자 수	인자 형태	클래스 종류	샘플 수
내용	10	실수형	2(Yes/No)	808(646/162)

〈표 2〉 후보 속성의 구성

속 성	내 용
Elevation	복숭아 재배지의 고도
Soil_Series	복숭아 재배지의 토양통 분포도에 따른 동해 피해율
Mountain_Count	복숭아 재배지의 지정된 반경 내에 포함된 산의 개수
Mountain_Area	복숭아 재배지의 지정된 반경 내에 포함된 산의 면적
Water_Count	복숭아 재배지의 지정된 반경 내에 포함된 하천의 개수
Water_Area	복숭아 재배지의 지정된 반경 내에 포함된 하천의 면적
Wind_Speed	복숭아 재배지의 평균 풍속
Max_Wind_Speed	복숭아 재배지의 최대 풍속
Rainfall	복숭아 재배지의 강수량
Frost_Damage	복숭아 재배지의 동해피해 유무

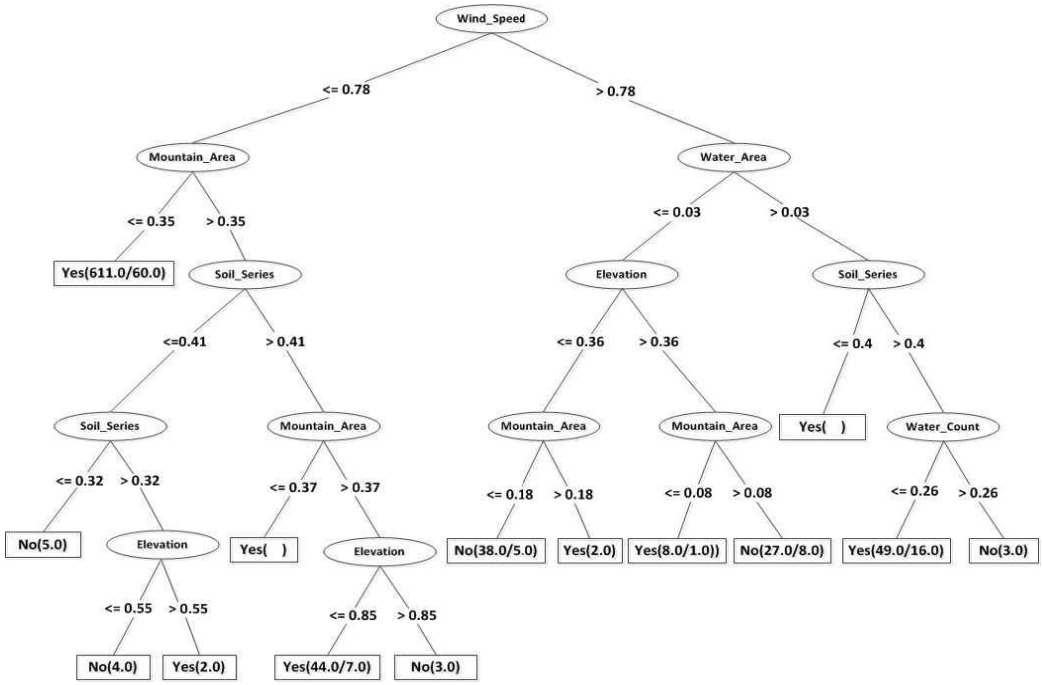
DIMS에 의해 생성된 후보 속성 데이터의 구성은 <표 2>와 같다. 각 속성 데이터들은 동해 피해에 큰 영향을 줄 수 있는 속성 값들이며 CCI 서비스의 작물 재배 기준 정보를 이용하여 전 처리 작업된 결과이다. 각 속성 값들은 해당 항목의 최대값을 기준으로 정규화 하였다. Soil_Series는 토양통이 동해피해에 미치는 영향을 수치 값으로 변환하기 위하여 DIMS에서 토양통 값을 기준으로 복숭아 재배지를 그룹화하고 그룹화한 지역의 동해피해율을 더한 값에 재배지의 개수를 나누어 토양통에 대한 동해 피해 값을 수치화한 것이다. <표 2>의 *Mountain_Count, Mountain_Area, Water_Count, Water_Area*의 속성들은 면적의 넓이와 관련된 속성들로 동해 피해에 대한 도메인 지식을 얻을 수 없다. 이와 같이 도메인 지식을 이용할 수 없는 경우에 제안한 프레임워크는 더 효과적으로 활용될 수 있다. 즉 단순히 DIMS는 ADM의 데이터 변환 설정 값만을 다르게 함으로써 별도의 전 처리 작업 없이 다양한 데이터 셋을 생성할 수 있다.

이와 같이 DIMS로부터 획득된 데이터 셋은 제안 프레임워크의 Application Service Layer 영역의 데이터 마이닝 툴인 Weka3.6.4[16]을 이용하여 동해발생지역에 대한 예측 서비스 구축을 위해 적용될 수 있다. 즉 DIMS로부터 획득한 다양한 데이터 셋을 이용한 비교 실험을 통하여 최적의 예측 성능을 제공하는 서비스를 구축할 수 있다.

서로 다른 데이터 셋에 대한 예측 성능 비교 실험을 위해 <표 2>의 4개의 속성들인 *Mountain_Count, Mountain_Area, Water_Count, Water_Area*의 속성 값을 물과 임야의 반경 500m와 1000m 값을 기준으로 하여 두 개의 데이터 셋, Dataset A(500m)와 Dataset B(1000m)를 추출하였다. 한편, SDM의 예측 성능은 학습 데이터의 속성들에 의해서도 영향을 받는다. 따라서 높은 예측 성능을 줄 수 있는 영향력 있는 속성을 선택하는 것이 중요하다. 그러나 <표 2>의 후보 속성들은 단순히 경험에 의한 도메인 지식에 의해 도출된 것이며 그 유용성은 증명되지 않은 것이다. 또한 영향력 있는 속성들을 추출하였다 해도 각 속성들의 임계값(threshold values)을 결정하는 것이 어렵다. 본 연구에서는 의사 결정 트리 (C4.5 알고리즘) 알고리즘을 이용하여 이론적으로 의미 있게 영향력 있는 속성들을 추출한다. (그림 4)는 DIMS에 의해 생성된 임의의 데이터 셋(Dataset A)에 대해 결정 트리 알고리즘을 적용한 결과이다. 그림에서 보면 <표 2>의 속성들 중에서 *Max_Wind_Speed*와 *Rainfall* 속성이 상쇄되어 사라진 것을 관찰할 수 있다. 이것은 *Max_Wind_Speed*와 *Rainfall* 속성들이 동해피해에 크게 영향을 주지 않는다는 사실을 의미할 수 있다. 본 연구에서는 예측 성능의 비교 실험을 위해 <표 2>의 전체 속성을 대상으로 하여 의사 결정 트리 알고리즘(C4.5)과 MLP 신경망 학습을 위한 오류 역전파 알고리즘을 실험하였고, (그림 4)의 의사 결정 트리에서 상쇄된 두 속성 *Max_Wind_Speed*와 *Rainfall*을 제외한 속성들만을 입력으로 하여 MLP 신경망에 적용하는 Combined 알고리즘을 추가적으로 비교 실험하였다.

<표 3>은 의사 결정 트리 (Weka[16]의 *J4.8* 알고리즘) 알고리즘과 MLP 신경망 (Weka[16]의 *Multilayer Perceptron* 알고리즘), 그리고 부가적으로 두 알고리즘을 결합한 Combined 알고리즘을 이용하여 Dataset A(500m)와 Dataset B(1000m)에 대해 동해피해지역을 예측하기 위해 실험한 오인식률의 결과 값을 보여준다. MLP 신경망의 학습 알고리즘(backpropagation)의 학습 환경은 학습률(learning rate) 0.3, 관성항(momentum term) 0.2, 그리고 종료조건으로 최대 epoch 수는 500으로 설정하였다. 은닉층(hidden layer)의 노드 개수는 MLP 신경망과 Combined 알고리즘 모두에서 6개이다. 크로스 밸리데이션(cross validation, CV)은 10 폴드로 실행했고, Test는 데이터 셋을 훈련 : 테스트 = 7 : 3으로 분할하여 10회 실행한 후 평균값을 계산한 것이다. 10-fold-CV는 전체 데이터 셋에서 표현된 클래스를 근사적으로 동일한 비율로 10개의 그룹으로 나누어서, 차례로 한 개의 그룹을 선택하여 테스트 데이터로 사용하고 나머지 9개의 그룹을 학습 데이터로 사용하는 과정을 10회 반복하여 얻은 오인식률을 평균한 값이다.

<표 3>에서 보이는 바와 같이 Dataset A(500m)에 Combined 알고리즘을 적용한 결과 값이 오인식률이 가장 낮은 것을 알 수 있다. 본 연구에서는 이와 같이 DIMS로부터 획득한 다양한 데이터 셋에 대해 최적의 예측 성능을 제공하는 데이터 마이닝 서비스를 제안한 프레임워크에 구축할 수 있다. 또한 이렇게 구축된 동해피해 예측 서비스는 다른 지역의 예측 서비스에 쉽게 확장될 수 있다.



(그림 4) 생성된 의사 결정 트리

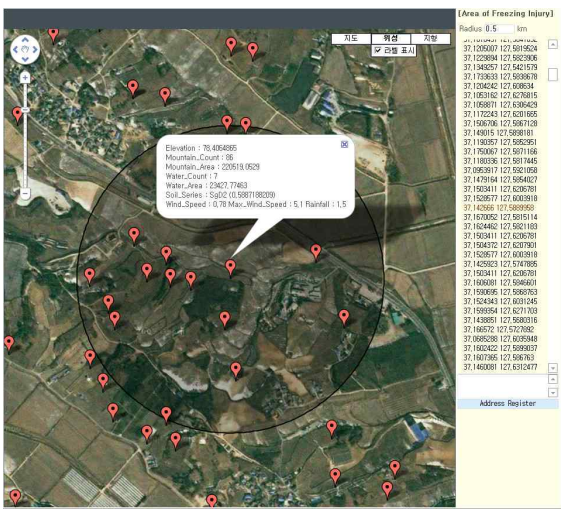
〈표 3〉 데이터 마이닝 알고리즘의 오인식률

		Decision Tree	Neural Network	Combined
Dataset A (500m)	10-fold-CV	0.1658415	0.1633663	0.1606436
	Test 평균	0.1698182	0.1516365	0.1516364
	Test 표준편차	2.17	1.98	1.46
Dataset B (1000m)	10-fold-CV	0.1803218	0.1780941	0.1788366
	Test 평균	0.1872727	0.1876364	0.1887272
	Test 표준편차	1.57	1.51	1.79

마지막으로, 제안한 프레임워크의 DIMS가 제공하는 데이터 셋은 서비스 형태로 제공되므로 분석 데이터에 대한 시각화 및 데이터 유효성 검사를 통해 실험결과를 다른 차원에서 분석할 수도 있다. (그림 5)는 복숭아나무 동해피해지역 예측을 위해 DIMS에서 제공되는 예측 서비스를 구글맵을 이용하여 시각화 한 것이다. 위치 검색을 통하여 동해발생 피해지역에 대한 속성정보를 시각적으로 확인할 수 있으며, 피해지역에 대한 분포 및 주변 환경에 대한 시각화를 통해 다차원 분석에도 활용될 수 있다.

5. 결 론

본 연구에서는 SDM의 학습 데이터 셋 생성을 위한 효과적인 전 처리 작업과 통합 환경을 제공하는 SOA 기반의 데이터 통합 프레임워크를 제안한다. 제안 프레임워크는 DIMS 모델의 내부 서비스들을 결합하여 적절한 데이터 셋을 효과적으로 생성할 수 있다. DIMS에 의해 생성된 다양한 데이터 셋들은 예측 성능 비교를 위해 SDM에 적용될 수 있으며, 최적의 예측 성능을 제공하는 데이터 마이닝 서비스를 제안한 프레임워크에 구축할 수 있음을 확인하였다. 이와 같이 구축된 동해피해지역 예측 서비스는 다른 지역의 예측서비스에 쉽게 확장될 수 있다. 또한 제안한 프레임워크



(그림 5) DIMS에 기반한 데이터 시각화

의 DIMS가 제공하는 데이터 셋은 서비스 형태로 제공되므로 분석 데이터에 대한 시각화 및 데이터 유효성 검사를 통해 실험결과를 다른 차원에서 분석하는데 이용될 수 있다. 향후 연구과제로는 제안한 프레임워크의 데이터 통합 관리 서비스와 SDM 기술의 유기적인 연결을 통한 서비스를 제공하여 자연 재해 예측 시스템, 농작물 재배적지 예측 시스템과 같은 예측 시스템에 활용할 수 있도록 하는 것이다.

참 고 문 헌

[1] Reddy, P. K. and Ankaiah, R., "A framework of information technology based agriculture information dissemination system to improve crop productivity," *Current science*, Vol.88, No.12, pp.1905 - 1913, 2005.

[2] Fraisse, C.W., Breuer, N.E., Zierden, D., Bellow, J.G., Paz, J., Cabrera, V.E., Garcia y Garcia, A., Ingram, K.T., Hatch, U., Hoogenboom, G., Jones, J.W., "AgClimate: A climate forecast information system for agricultural risk management in the southeastern USA," *Computers and electronics in agriculture*, Vol.53, No.1, pp.13 - 27, 2006.

[3] Han, J. and Micheline, K., "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2001.

[4] Ester, M., Kriegel, H. P., and J. Sander, "Algorithms and Applications for Spatial Data Mining," In H. J. Miller and J. Han, editors, Geographic *Data Mining and Knowledge Discovery*, 2001.

[5] Scheibler, T., Mietzner, R. and Leymann, F., "EAI as a Service - Combining the Power of Executable EAI Patterns and SaaS", Enterprise Distributed Object Computing Conference, pp.107-116, 2008.

[6] Huamin, W. and Zhiwei, Y., "An ETL Services Framework Based on Metadata", Intelligent Systems and Applications (ISA), 2nd International Workshop on, pp.1 - 4, 2010.

[7] Thomas, E., "Service Oriented Architecture: Concepts, Technology, and Design," Prentice Hall, PTR, 2005.

[8] Roy S., "The New Integration Scenario: Five Trends That Change How application Software Work," Gartner Application Integration and Web Service Summit, 2005.

[9] Xu, H., Hongqi, L., Qiaoyan, D. Zhuang, W., "The SOA Based Solution for Distributed Enterprise Application Integration," *Computer Science Technology and Applications, International Forum*, Vol.3, pp.330 - 336, 2009.

[10] Sha, Z. and Xie, Y., "Design of service oriented architecture for spatial data integration and its application in building web based GIS systems," *Geo Spatial Information Science*, Vol.13, No.1, pp.8 - 15, 2010.

[11] Haitao D., Bo Z. and Dingfang C., "Design and Actualization of SOA based Data Mining System," *Computer Aided Industrial Design and Conceptual Design, 9th International Conference*, pp.22 - 25, 2008.

[12] Awad M.M.I. and Abdullah M.S., "A framework for interoperable distributed ETL components based on SOA," *Software Technology and Engineering(ICSTE), 2nd international conference*, pp.67-70, 2010.

[13] Han S. and Kim J., "Rough set based decision tree using a core attribute," *Int. J. Inf. Technol. Decisi. Mak.*, Vol.7, No.2, pp.275 - 290, 2008.

[14] Hu Y. and Tseng F., "Mining simplified fuzzy if then rules for pattern classification," *Int. J. Inf. Technol. Decisi. Mak.*, Vol.8, No.3, pp.473 - 489, 2009.

[15] Peng Y., Kou G., Shi Y. and Chen Z., "A descriptive framework for the field of data mining and knowledge discovery," *Int. J. Inf. Technol. Decisi. Mak.*, Vol.7, No.4, pp.639 - 682, 2008.

[16] Witten, I. H., Frank, E. and Hall, M. A., "Data Mining: Practical Machine Learning Tools and Techniques," 3rd Ed., Morgan Kaufmann Publishers, 2011.

[17] L. Aijun, L. Yunhui and L. Siwei, Mapping a decision-tree for classification into a neural network, *Proc. 7th Int. Conf. on Computational Intelligence & Natural Computing*, pp.1528 - 1531, 2003.

[18] M. Kim, H. Na, K. Chae, H. Bang and J. Na, *A Combined Data Mining Approach for DDoS Attack Detection*, Lecture Notes in Computer Science (LNCS) Vol.3090, pp.943 - 950, 2004.

문 일 환

e-mail : mih80@naver.com
2006년 한경대학교 컴퓨터공학과(학사)
2008년 한경대학교 전자계산학과(석사)
2008년～현 재 한경대학교 컴퓨터공학과 박사과정 수료
관심분야 : GIS, Web Service, Data Mining, Mobile Computing, BI

허 환

e-mail : hurwhan0713@empal.com
2005년 충주대학교 전자계산학과(학사)
2007년 충주대학교 전자계산학과(석사)
2011년 한경대학교 컴퓨터공학과(박사)
1991년～현 재 경기동부과수농협 상무
관심분야 : DB, GIS, Spatial Data mining, SOA

김 삼 근

e-mail : skim@hknu.ac.kr
1985년 부산대학교 계산통계학과(학사)
1988년 숭실대학교 전자계산학과(석사)
1998년 숭실대학교 전자계산학과(박사)
1992년～현 재 한경대학교 컴퓨터공학과 교수
관심분야 : GIS, SOA, Data Mining, Mobile Computing, BI