시뮬레이션 데이터의 다양성을 고려한 데이터 전처리 프레임워크 설계

안선일*, 안부영*, 장지훈*, 이식*, 조금원*
*한국과학기술정보연구원 슈퍼컴퓨팅본부
e-mail:{siahn, ahnyoung, jangoq, siklee, ckw}@kisti.re.kr

Exploiting Data Diversity in a Simulation data Curation

Sunil Ahn*, Buyoung Ahn*, Jihoon Jang*, Sik Lee*, Kumwon Cho*
*Supercomputing Center, KISTI

요 약

계산과학 데이터를 공유하는 목적은 데이터의 분석을 통해 의미 있는 정보를 추출하기 위해서이다. 이를 위해서는 계산과학 데이터의 전처리 과정이 요구되며, 핵심 이슈는 계산과학 데이터의 다양성과 복잡성의 해결이다. 본 논문은 계산과학 데이터 저장소의 구축 과정에서 고려하였던 계산과학 데이터의 전처리에 대한 설계 이슈들과 해결 방안들에 대해 설명한다.

1. 서론

계산과학 데이터를 공유하는 것은 많은 시간이 소요되는 중복 계산을 회피하고, 분석을 통해 의미 있는 정보를 추출하기 위해서이다. 이를 위해서는 효율적으로 데이터를 저장하고 검색할 수 있는 저장소가 필수적이며, 데이터 저장소는 데이터의 수집, 전처리, 보존, 접근을 촉진하는 데이터 관리 시스템으로써 데이터의 재활용을 촉진한다.

계산과학 데이터로부터 의미 있는 정보를 추출하기 위해서는 데이터의 신뢰성을 검사하고 서술형 메타데이터를 자동으로 추출하는 등의 전처리 과정이 필수적이다. 본 논문은 계산과학 데이터의 다양성과 복잡성의 해결을 위한 설계 이슈들을 분석하고 해결 방안을 제시하였다.

2. 계산과학 데이터 전처리에 대한 요구사항

계산과학 데이터 전처리에 대한 요구사항은 다양하다. 첫 번째는 계산과학 데이터의 복잡성과 다양성 문제의 해결이다. 계산과학 데이터는 시뮬레이션 SW의 입력과 출력 파일들로 구성되며, 사용되는 시뮬레이션 SW와 그 버전에 따라 매우 다양한 형식을 가질 뿐 아니라 구조화되어 있지 않아 복잡하다. 동일한 목적을 위해서라도 여러 시뮬레이션 소프트웨어들이 활용될 수 있고, 각각 서술형 메타데이터 추출 방법이 상이하고, 데이터를 검증하는 방법은 물론 표현 방법들도 달라진다.

둘째, 데이터 생명주기에 대한 지원이 필요하다. 데이터 활용에 대한 사용자요구 및 환경변화에 대응하기 위해서는 계속적이고 지속적인 데이터 전처리와 큐레이션의 수행이 요구된다. 지속적 데이터 전처리는 여러 데이터를 상호 연계 분석하여 새로운 서술형 메타데이터를 추출하도록 돕는다는 점에서도 중요하다.

셋째, 데이터 전처리 모듈 개발의 생태계가 요구된다. 시뮬레이션 SW의 종류는 셀 수 없이 많으나, 계산과학 데이터 전처리 모듈의 개발은 해당 시뮬레이션 SW를 잘 알고 있는 계산과학 응용 연구자가 아니면 개발하기 어렵다. 이 때문에 시뮬레이션 SW를 잘 알고 있는 연구자들이 쉽게 개발에 기여하고 시험할 수 있는 방법의 제공이 필요하다. 특정 언어에 종속적이어서는 안 되며 계산과학 응용 연구자가 익숙히 알고 있는 언어나 기술을 활용하여 개발할 수 있어야 한다. 또한 데이터 전처리에 필요한 기능이 다른 SW를 통해 이미 개발되어 있다면, 이 SW를 설치하고 활용할 수 있도록 지원할 필요가 있다.

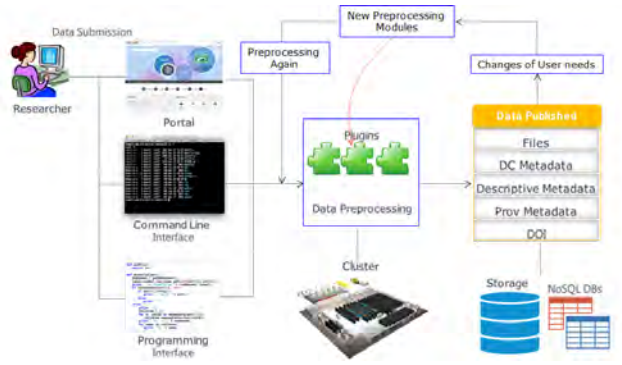
넷째, 데이터 전처리 모듈의 보안 문제의 해결이다. 데이터 전처리 모듈 개발의 생태계가 조성된 경우 어떤 연구자라도 모듈을 개발하여 기여할 수 있다. 이 경우 제3자가 제공한 데이터 전처리 모듈에서 보안 문제가 발생할 수 있으며, 이에 대응할 수 있는 방법이 요구된다.

다섯째, 서술형 메타데이터에 대한 시맨틱의 제공과 통합검색에 대한 요구이다. 사용된 전처리 모듈에 따라 다른 용어로 서술형 메타데이터가 기술된다면 통합검색이 어려워진다. 활용된 시뮬레이션 SW가 다르더라도 메타데이터 기술에 대한 동일한 용어를 사용한다면 다양한 유형의 데이터에 대한 통합 검색도 가능하다.

상기의 요구사항들은 모두 시뮬레이션 데이터의 다양성과 복잡성 때문에 유발되었다고 볼 수 있다.

3. 계산과학 데이터 전처리 프레임워크

계산과학 데이터의 전처리는 시뮬레이션 파일의 타당성 검증, 서술형 메타데이터 추출, 실행이력 메타데이터 추출, 파생 파일 생성, 메타데이터 타당성 검증 등으로 구성된다. 우리는 유연하고 확장성 있는 계산과학 데이터 전처리 기능을 제공하기 위해 플러그인 방식으로 전처리 기능을 설치하고 실행할 수 있는 프레임워크를 개발하였다. 이를 통해 데이터 유형별 특화된 전처리 기능을 개발하여 설치할 수 있고, 제3의 연구자들도 데이터 전처리 기능 개발에 기여할 수 있다. 플러그인 방식을 활용하여 유연한 전처리 기능을 제공하는 것은 기존 연구들[1-2]과 유사하지만 사용자요구 및 환경변화에 따라 지속적인 데이터 전처리와 큐레이션 방법을 제공한다는 부분에서 차별성이 크다.



(그림 1) 계산과학 데이터의 지속적 전처리

이용자들의 요구 변화나 환경 변화에 따라 데이터로부터 새로운 정보 추출이 필요한 경우 데이터 유형에 따른 신규 데이터 전처리 모듈들을 작성하여 플러그인 형태로 추가할 수 있다. 신규 전처리 기능이 추가되면, 관리자는 기존에 보존된 데이터를 대상으로 전처리를 재수행하여 데이터의 지속적 생명주기 관리를 지원할 수 있다. 수많은 데이터들에 대한 전처리가 다시 수행하기 위해서는 클러스터와 같은 고성능 시스템을 활용한다.

전처리 기능의 개발에 있어 기존 연구[1-2]들은 특정 언어에 종속적인 반면, 우리는 응용 연구자들이 익숙한 다양한 언어와 외부의 다양한 툴들을 활용하여 전처리 기능을 개발할 수 있도록 하였다. 이에 더하여 제3자가 개발한 전처리 기능의 보안 문제를 해결하기 위해 도커(docker) 컨테이너 내에서 전처리 기능이 실행될 수 있도록 하여 시스템 전체에 미치는 영향을 차단하였다.

전처리 기능의 개발은 데이터 유형별 도커 컨테이너 이미지를 다운로드하는 것으로부터 시작한다. 이미지에 데이터 유형에 따른 전처리에 필요한 외부 툴들을 설치하고, 이미지 내에서 전처리 기능들을 개발 및 시험한 후 저장한다. 개발이 완료된 전처리 기능과 도커 이미지는 커뮤니티를 통해 공유되고, 제3자에 의해 재활용이 가능하다.

4. 결론

계산과학 시뮬레이션 데이터의 공유는 데이터의 분석을 통한 새로운 지식창출을 촉진한다. 본 논문은 시뮬레이션 데이터의 다양성 문제 해결하기 위한 전처리에 대한 설계 이슈들을 분석하고 이를 해결하기 위한 방안들을 제시하였다. 향후의 연구 방향은 전처리 기능에 대한 개발 생태계를 확장시키기 위해 보다 쉬운 데이터 전처리 모듈의 개발 및 시험 환경을 확보하는 것이다. 한 방법은 기존의 ETL 툴들을 활용할 수 있는데, 현재까지 ETL 툴들은 정형화된 데이터의 처리에는 적합하지만, 비정형 데이터의 처리를 위해서는 보완이 필요하다.

사사

이 논문은 한국과학기술정보연구원(K-17-L01-C02)과 미래창조과학부, 한국연구재단의 지원을 받아 수행된 기초연구사업(No.NRF-2011-0020576)의 연구비 지원으로 작성되었습니다.

참고문헌

[1] Adewumi, A. O., and Omoregbe, N. A.: Institutional repositories: features, architecture, design and implementation technologies. Journal of Computing, vol. 2(8) (2011).

[2] Pizzi, G., et al.: AiiDA: automated interactive infrastructure and database for computational science. Computational Materials Science, vol. 111, pp.218−230 (2016).