식이데이터 분석 자동화 문제풀이환경 설계

최지은 안윤선 송지현 김윤희°
숙명여자대학교 컴퓨터과학과
jechoi1205@sm.ac.kr, ahnysun@sm.ac.kr, jihyun.song0829@gmail.com, yulan@sm.ac.kr

A Design of Problem Solving Environment for Automation of Dietary Data Analysis

Jieun Choi Younsun Ahn Jihyun song Yoonhee Kim°
Dept. of Computer Science, Sookmyung Women's University

요 약

웰니스는 개인 건강 및 삶의 패턴, 운동 또는 식이 패턴에 대한 많은 양의 측정된 데이터에 의해 결정된다. 특히 식이 패턴은 개인 건강과 밀접한 관련이 있어 개인 건강을 분석하기위해 식이 패턴을 분석하는 것은 중요하다. 식품 영양학 분야에서는 식이 분석을 위해 많은 양의 입력 데이터를 처리하는 것이 필요하고 그 단계는 시간이 많이 소모되는 지루하고 반복적인 과정이다. 한편, 과학자들에게 있어서 많은 양의 데이터를 다루고 반복적인 작업이 수행되는 실험의 경우 효율적인 실험과 실험의 편리성을 위해서 문제풀이환경을 구축하는 것이 필요하다. 본 논문에서는 식이데이터 분석 자동화를 위한 문제풀이환경을 제안한다. 제안된 문제풀이환경은 순차적이고 병렬적인 작업들을 적절하게 배치함으로써 전체 실험에 소요되는 시간을 단축할 수 있다.

1. 서 론

웰니스는 사람들이 개인의 건강과 삶의 질을 개선하고 향상시키기 위해 최근 관심을 갖는 가장 인기 있는 분야 중 하나이다. 신체 내외적인 상태를 나타내는 많은 양의 개인 건강 데이터는 물리적으로 측정되는 값뿐만 아니라 식이 패턴과 같이 일상에서 매일 반복적으로 기록되는 개인의 활동도 포함된다. 식이 패턴은 식품 영양학적으로 개인 건강과 밀접한 연관이 있다. 따라서 식품 영양학에서 식이 패턴을 분석하는 것은 개인이 가지고 있을 수 있는 잠재적인 질병을 분석하고 예방하기 위해서 필요하다.

축적된 개인 건강 데이터(life log, PHR (Physical Health Record), EMR(Electronic Medical Record), etc.)는 건강을 유지하기 위해 신체 상태와 특정 질병의 관계를 찾는데 유용하다. 이러한 개인 건강 데이터를 사용하여 통계적 분석을 수행하는 경우 빠르고 효과적으로 많은 양의 데이터를 처리하는 것이 필요하다. 그러나 대량의 축적된 개인 건강 데이터를 효율적으로 처리하고 관리 및 저장하는 것은 매우 어렵다.

본 논문에서는 많은 양의 식이데이터를 처리하고 분석하는 과정에서의 반복적인 작업을 자동화하기 하기 위한 분석 자동화 문제풀이환경을 제안한다. 제안된 문제풀이환경은 병렬 컴퓨팅 환경에서 식이 패턴 분석을 가속화하고 식품 영양학 적으로 식이 데이터 분석을 통한 개인 건강의 잠재적인 위협 요소를 파악한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고 3장에서 설계한 식이데이터 분석 자동화 문제풀이환경을 설명한다. 4장에서는 식품 영양학적 시나리오를 통해 설계한 시스템을 설명하고, 마지막 5장에서 결론을 맺는다.

2. 관련연구

다양한 분야에서 병렬 및 분산 클라우드와 그리드 컴퓨팅 환경에서 데이터 분석하는 것을 돕는 플랫폼에 관한 연구가 진행되고 있다. CAP(Collaborative Analytics Platform, 2013~2019)[1] 프로젝트의 경우 유럽에서 진행되는 클라우드 인프라 기반 표준 서비스 플랫폼 개발을 돕기 위한 프로젝트이다. 한편, 특정 분야를 대상으로 문제풀이환경을 제안하는 관련 연구가 활발하게 진행되고 있다. 기후 데이터 분석을 위한 문제풀이환경에 대한 연구[2]와 유체 역학 실험환경을 위한 문제풀이환경 연구[3]가 진행되었다. [2]에서는 클라우드 기반의 대량의 기후 데이터를 분석하기 위한 프레임워크를 제안한다. 클라우드 기반의 프레임워크는 효과적으로 클라우드 컴퓨팅과 데이터 스토리지, 유용한 워크플로우 관리, 스케줄러 엔진과 같은 컴포넌트를 활용한다. [3]에서 제안된 통합 과학적 실험 프레임워크는 사용자가 유체 역학에 대해 수치 해석적으로 분석 실험하는 것을 돕는다. [4]에서는 하둡을 기반으로 하여 건강 데이터에 관하여 분산 컴퓨팅 환경에서 데이터 관리 및 분석

° "본 연구는 미래창조과학부 및 정보통신산업진흥원의 ICT 융합고급인력과정지원사업의 연구결과로 수행되었음" (NIPA-2014-H0401-14-1022)

하는 것을 제안한다.

3. 식이데이터 분석 자동화 문제풀이환경 설계

식품 영양학에서 식이데이터 분석을 위한 문제풀이환경은 다음과 같은 4가지 과정을 거친다. 1. 가설 설정 및 계획, 2. 데이터 클리닝, 3. 데이터 분석, 4. 결과 해석. 그림1은 식이데이터 분석 자동화 문제풀이환경의 4단계를 나타낸다.

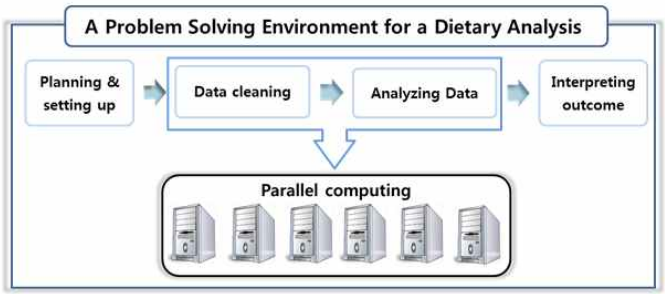

그림1. 식이데이터 분석 자동화 문제풀이환경의 4단계

먼저, 사용자는 데이터를 이용하여 증명하고자 하는 가설을 설정한다. 가설을 증명하기 위해 가공되지 않은 데이터를 준비한다. 가공되지 않은 수많은 독립된 데이터는 하나의 입력 데이터로 구성된다.

데이터 클리닝 단계에서는 데이터 분석을 위해 입력 데이터에 대한 데이터 정제가 필수적이다. 예를 들어 특정 값에 대해 데이터가 누락된 경우와 비정상적인 데이터를 포함한 경우는 해당 값을 정상적인 범위 안에서 수정하거나 제거하는 것이 필요하다. 또는 몇몇 연속적인 값들을 분석에 사용하기 위해 적절한 몇 개의 범주로 나누는 과정이 필요하다. 이 과정에서 사용자는 특정 조건을 사용하여 불필요한 데이터를 제거하거나 반복적인 작업을 통해 연속적인 데이터를 적절한 범주로 나누는 과정을 진행한다. 따라서 데이터 클리닝 단계는 전체 분석 과정에 있어서 시간이 많이 소요되고 사용자의 노동력이 요구된다. 제안된 시스템은 분석을 위한 데이터 클리닝 단계에서의 이러한 반복적인 과정을 자동화하여 사용자가 효율적으로 데이터 분석을 진행하도록 돕는다. 데이터 클리닝 단계가 완료된 후, 사용자는 정의한 가설을 분석하기위한 적절한 데이터를 얻게 된다.

분석 단계에서 사용자는 원하는 가설을 얻기 위해 다양한 데이터 분석 방법을 시도한다. 이 단계에서는 같은 데이터에 대해 여러 통계적 분석을 적용해야 한다. 따라서 많은 양의 데이터를 분석하는데 있어서 오랜 시간이 걸리고 각 분석에는 반복적인 작업이 포함된다. 제안된 시스템은 동일 데이터에 대해 다양한 통계적 분석 방법을 적용하는 것을 병렬 처리하여 보다 빠른 분석을 수행한다.

마지막으로 사용자는 가설을 입증하기 위해 실험 결과를 해석하여 다양한 표와 그래프를 도출해 낸다.

제안된 시스템은 대량의 데이터를 통계적으로 분석하기 위한 자동화된 문제 풀이 환경을 제공한다. 각 반복적인 작업은 분산되어 동시에 수행될 수 있다. 제안된 시스템은 식품 영양학에서 널리 사용되는 통계 분석 시스템 SAS[5]를 활용한다. 병렬 수행을 위해서 따라서 식이 데이터를 분석하는 과정에서 반복적인 작업과 병렬 수행이 가능한 작업이 분산되어 수행되기 때문에 전체 분석과정을 빠르고 효율적으로 수행한다.

4. 사례 연구

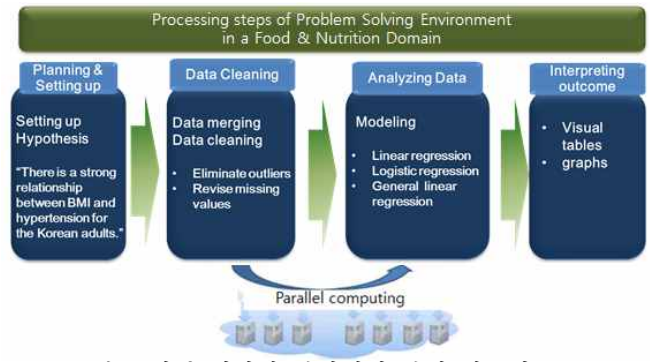

그림2. 식품 영양학 분야에서 분석 자동화 문제풀이환경의 처리 단계

그림 2는 식품 영양학 시나리오에 따라 식이데이터의 분석을 자동화하는 문제풀이환경의 4가지 처리 단계를 보여준다. 사용자는 가설 설정 및 계획 단계에서 "한국 성인의 BMI와 고혈압 사이에 관계가 있다"라는 가설을 세운다.

실험에는 국민 건강 영양 조사[6]의 2007년부터 2012년까지 각 연도별 미 가공된 데이터를 사용한다. 연도별 데이터 크기는 07년도(53.9MB), 08년도(114MB), 09년도(123MB), 10년도(70MB), 11년도(66.6MB), 12년도(63.0MB)이다. 분석에 사용되기 위해 연도별 데이터는 하나의 입력 데이터로 병합된다. 병합된 데이터는 50,405명의 측정 데이터를 갖는다. 입력 데이터는 데이터 클리닝 단계에서 가설 검증을 위해 불필요한 데이터를 제거 하는 단계를 거친다. 예를 들어 암에 대한 데이터는 증명하고자 하는 가설과 관련 없는 데이터이기 때문에 사용자는 특정 조건을 입력하여 불필요한 관련 데이터를 제거한다. 또한 알코올 섭취 및 혈압, 흡연량과 같은 연속적인 변수들에 관해서는 분석을 위해 적절한 그룹으로 나누는 것이 필요하다. 따라서 제안된 프레임워크는 많은 양의 데이터에서 불필요한 값을 처리하는 것과 특정 변수에 대해 파라미터 값을 나누어 절적한 범주를 나누는 것을 자동화 한다. 데이터 클리닝 과정을 마친 후, 가공된 18,220 명의 건강 데이터가 가설 분석에 사용된다.

제안된 프레임워크는 나이, 성별, 연도, BMI, 혈압과 같은 변수를 사용하여 적절한 모델을 찾기 위해, 사용자에게 선형 회귀분석, 로지스틱 회기분석, 일반 선형 모형과 같은 통계학적 분석 방법을 제공한다. 이와 같은 통계학적 모델은 다양한 데이터 타입에 따라 적용되는 것이 나뉜다. 예를 들어 선형 회기 분석은 오직 연속적인 변수에 대한 분석에서 적용된다. 통계학적 분석 결과 다음과 같은 선형 방정식이 정의된다. 이 방정식은 선형 회귀 분석을 사용하여 대부분의 데이터를 설명하는데

사용되는 가장 적합한 방정식이다.

age =−13.160934*(hp1) + 0*(hp2).

사용자는 각 모델의 p−values 값을 확인하여 한국 성인의 BMI와 고혈압 사이에 관계가 있다는 가설의 성립됨을 확인하였다. 마지막으로 사용자는 제안된 프레임워크를 통해 시각적 그래프 및 테이블을 통해 분석된 데이터를 확인한다.

위와 같이 일련의 순서로 실행되던 실험은 다음과 같은 부분을 병렬로 처리하여 진행할 수 있다. 데이터 클리닝 단계에서 3개의 머신은 각각 2년치의 데이터를 가지고 이상치(outlier)를 제거하거나 데이터의 필드에서 암에 대한 값과 같이 가설과 관련 없는 값을 제거하는 부분을 수행한다. 각 머신에서 데이터 클리닝을 통해 얻어진 데이터는 하둡 분산 파일시스템에 저장되고, 각 머신이 이를 가지고 데이터를 통합한 후 파라미터에 대한 실험을 진행한다. 파라미터 조절부분은 과거 흡연자와 현재 흡연자의 흡연 기간을 1~15 년으로 3머신이 다른 파라미터에 대해 병렬로 실행한다. 또한 음주를 섭취하는 빈도에 따라 구간을 4개로 나누는 부분에서 각 구간의 파라미터 값을 9가지로 다르게 수행하는 부분을 병렬로 실행한다. 데이터 분석 단계에서는 한 머신 안에서 순차적으로 실행되던 분석 모델 3가지를 각 머신에서 나누어 수행한다. 다음의 표 1은 제안된 병렬 실험에서 머신의 사양을 나타낸다.

	Machine1	Machine2	Machine3
OS	windows 7 64bit	windows 7 64bit	windows 7 64bit
RAM	8GB	4GB	4GB
CPU	i5, 3.40GHz, Quad Core	i5, 3.00GHz, Quad Core	i3, 3.07GHz, Quad Core

표 1. 제안된 병렬 실험 머신 사양

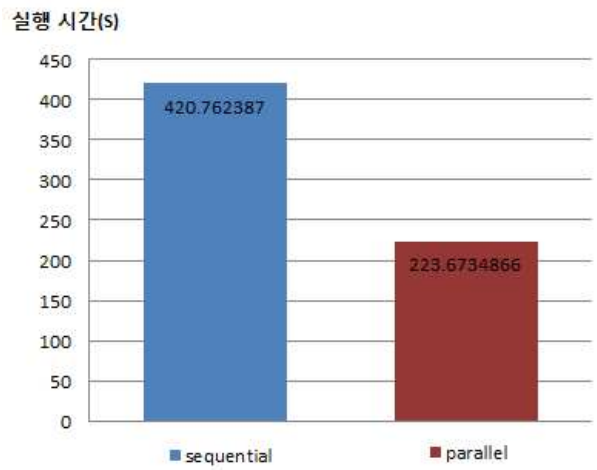

그림 3. 제안된 시스템의 성능 비교

그림 3은 순차적으로 Machine3을 사용하여 식이 데이터 분석의 4단계를 수행(sequential)했을 때의 실행 시간과 제안된 분석 자동화 문제 풀이 환경에서 3개의 머신을 이용하여 병렬 수행(parallel)했을 때의 실행 시간을 각 10번씩 수행하여 평균 수행 시간을 보여준다. sequential의 경우 420.76초, parallel은 223.67초가 걸려 제안된 분석 자동화 문제 풀이 환경의 경우 약 46.8%의 수행 시간을 단축하였다. 따라서 제안된 문제풀이 환경으로 식이 데이터 분석 자동화 문제풀이 환경을 구축하면 실험에 소요되는 시간을 단축하는 것이 가능하다.

5. 결론 및 향후 연구

본 논문에서는 식품 영양학 분야에서 식이데이터 분석을 위한 분석 자동화 문제풀이환경을 설계하였다. 또한 식이데이터 분석 자동화 프레임워크는 사용자가 병렬컴퓨팅 환경에서 효과적으로 통계학적 분석을 수행하는 것을 돕는다. 또한 실험 시나리오에 따라 입력 데이터를 생성하는 단계와 데이터 클리닝 단계 및 분석 단계에서 각 수행되는 작업에 있어서 반복적인 작업을 자동화하여 분석의 속도를 빠르게 한다.

참 고 문 헌

[1] CAP, https://itea3.org/project/cap.html

[2] Li, R.M., Tjhi, W.C., Kee Khoon Lee, Long Wang, Xiaorong Li, Di Ma, “A Framework for Cloud−based Large−Scale Data Analytics and Visualization: Case Study on Multiscale Climate Data”, Cloud Computing Technology and Science (CloudCom), pp. 618−622, 2011.

[3] S. Park, H. Kang, Y. Kim, C. Kim, Y. Hyun, “An Integrated Scientific Experiment Framework for Numerical Analysis in e−Science Environment”, Computation Tools, pp. 34−37, 2012.

[4] Hongyong Yu, Deshuai Wang, “Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop”, Computational and Information Sciences (ICCIS), pp. 514−517, 2012.

[5] SAS, http://www.sas.com/

[6] Korean National Nutrition Survey, http://knhanes.cdc.go.kr/