한국시뮬레이션학회 논문지
Vol. 25, No. 1, pp. 29-34 (2016. 3)
http://dx.doi.org/10.9709/JKSS.2016.25.1.029
ISSN 1225-5904

온톨로지 기반 대용량 코호트 DB 검색 시뮬레이션

송주형 · 황재민 · 최정석 · 강상길*

Ontology-based Cohort DB Search Simulation

Joo-Hyung Song · Jae-min Hwang · Jeongseok Choi · Sanggil Kang*

ABSTRACT

Many researchers have used cohort DB (database) to predict the occurrence of disease or to keep track of disease spread. Cohort DB is Big Data which has simply stored disease and health information as separated DB table sets. To measure the relations between health information, It is necessary to reconstruct cohort DB which follows research purpose. In this paper, XML descriptor, editor has been used to construct ontology-based Big Data cohort DB. Also, we have developed ontology based cohort DB search system to check results of relations between health information. XML editor has used 7 layered Ontology development 101 and OWL API to change cohort DB into ontology-based. Ontology-based cohort DB system can measure the relation of disease and health information and can be used effectively when semantic relations are found. We have developed ontology-based cohort DB search system which can measure the relations between disease and health information. And it is very effective when searched results are semantic relations.

Key words : Ontology, Big data, Cohort DB, OWL API

요 약

코호트 DB(DataBase)를 이용하여 질병 발생 예측 및 확산을 추적하는 많은 연구가 진행되고 있다. 코호트 DB는 대용량의 질병 및 건강정보가 단순한 개별적인 DB 테이블의 집합으로 구성되어있어 연관관계 검색을 위해서는 코호트 DB를 연구 목적에 맞게 재구성하는 작업이 필요하다. 본 논문에서는 대용량 코호트 DB를 온톨로지 기반으로 구축하기 위해 XML descriptor, editor를 이용하였다. 또한, 원활한 연관관계 검색결과 확인을 위해 온톨로지 기반의 코호트 DB 검색 시스템과 UI를 개발하였다. XML editor에서는 코호트 DB를 온톨로지로 구성하기 위해 7단계로 구성된 Ontology development 101 방법론과 OWL(Ontology Web Language) API를 이용하였다. 이와 같은 온톨로지 기반 코호트 DB 검색 시스템은 질병 및 건강정보의 연관성을 측정하고 의미적인 연관관계를 검색 시 효과적으로 활용 가능하다.

주요어 : 온톨로지, 빅데이터, 코호트DB, OWL API

1. 서 론

최근 코호트 DB를 이용하여 특정 질병 발생을 예측하거나 질병 확산을 추적 및 관찰하는 등의 연구가 활발하게 진행되고 있다[1]. 한국건강보험공단에서 공개한 코호트 DB는 재정, 질병, 나이, 소득 수준과 같은 대용량 자료를 바탕으로 구성되어 있으며 사용자의 ID나 지역과 같은 특정한 정보를 기준으로 정의되어 있는 단순한 개별적인 DB 테이블 구조이다. 이러한 구조로는 질병간의 의미적 연관관계를 찾기 위해서는 기존의 코호트 DB를 재정의하거나 수정하는 작업이 필수 불가결하다. 이러한 문제를 해결하기 위해 코호트 DB내의 의미적 연관관계를 용이하게 검색해 줄 수 있는 통합된 대용량 검색 시스템이 필요하다. 온톨로지 구조[2]를 이용하면 데이터간의 의미적 관

* 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2014R1A1A2056374).

Received: 16 January 2016, **Revised:** 22 February 2016, **Accepted:** 26 February 2016

***Corresponding Author:** Sanggil Kang
E-mail: sgkang@inha.ac.kr
Inha University Computer Science and Information Engineering

계를 검색할 수 있다. 본 논문에서는 원활한 의미적 연관 관계 구축을 위해 온톨로지 기반의 대용량 코호트 DB 검색 관리 시스템을 개발한다. 이러한 시스템을 구축하기 위하여 Ontology development 101 7단계 방법론을 따라 DB 시스템을 설계한 후 온톨로지 계층 구조를 정의하고, 이를 구축한다[3].

본 논문은 다음과 같이 구성되어있다. 2장에서는 코호트 DB의 구성과 기존에 연구되어 왔던 내용을 소개한다. 3장에서는 코호트 DB를 온톨로지 기반으로 변환하는 방법에 대해 자세히 설명하며 4장에서는 온톨로지 기반 대용량 코호트 DB 검색 시스템의 구현 환경과 실제 기동 화면을 설명한다. 마지막으로 5장에서는 구축한 시스템에 대한 결론과 향후 연구 진행 방향에 관해 고찰한다.

2. 관련 연구

특정 인구 집단으로부터 2002년부터 2013년까지의 사회적, 경제적 변수가 포함된 자격자료, 건강검진자료 및 진료내역자료를 수집하여 관리한 DB가 코호트 DB이다. 장기간의 건강관리 자료로 이뤄져있기 때문에 시간적 선후관계나 인과적 관계분석이 가능한 자료로써 여러 방면의 연구가 진행되고 있다.

기존의 코호트 DB에 관한 연구로는 데이터마이닝 기법을 이용한 당뇨병 치료분석[4], 어린이의 성장예측인자 식별연구[5], 건강보험 재정추계[6], 대기오염의 건강영양평가[7], MeSH tree를 이용한 건강 측정[8], 빅데이터를 이용하여 건강정보를 측정[9]하는 등 다양한 방면의 연구가 진행되어 왔다. Abdullah A. Aljumah et al.[1]는 코호트 DB를 사용하여 회귀 기반 데이터마이닝 기법으로 당뇨병 치료의 예측 분석을 연구하였다. 이 연구는 예측을 위해 소프트웨어 툴인 ODM(Oracle Data Miner)을 사용하여 당뇨병 치료 예측에 관한 알고리즘을 개발하였다. 연구결과 분석을 위해 SVM(Support Vector Machine)알고리즘[10]을 사용하고 WHO(World Health Organization)의 자료로부터 NCD(Non Communicable Disease) 위험 요소를 추출하여 다섯 개의 그룹을 두 개의 연령 그룹으로 나누어서 분류하고 이를 토대로 실제 발병과 예측 발병률을 비교하는 DB를 구축하였다.

Ken K L Ong et al.[5]는 2세에서 5세까지의 관계로부터 출생 후 회복 성장의 예측 인자를 식별하는 연구를 하였다. 영국의 산모와 어린이를 대상으로 2세부터 5세까지의 아이의 출생 체중과 산모의 임신 전 체중, 임신 체중 증가, 키, 흡연여부와 같은 정보를 수집하였다. 이를 통해 부모의 몸무게, 키, 흡연여부와 같은 정보가 0세에서 5세까지의 아기 몸무게와 성장지수와의 연관 관계가 있음을 증명하였다.

건강보험공단 코호트 DB를 이용한 재정추계[6]에서는 건강보험공단의 지출과 수입 두 가지 경제변수를 사용하여 2060년까지의 건강보험 재정추계를 하였다. 재정추계를 하기위해 질병별, 의료기관별, 남녀 유병률 자료를 이용하여 VECM-LC 모형[11]을 구축하고, 선형 회귀 모형[12]을 사용하여 건강보험공단부담 1인당 진료비를 추계하였다. 이를 위해 재정현황과, 재정수지, 가입자 현황, 건강검진비를 기준으로 DB를 구축하였다.

코호트 자료를 이용한 대기오염의 만선 건강영향 평가체계 구축[7]에서는 대기오염의 만성 건강영향을 시범적으로 평가하기 위해 코호트 DB를 이용하였다. 새로이 발생한 심혈관계 입원발생위험의 건강영향을 시계열분석하기 위해 2005년부터 2010년까지 대기오염 누적수치를 집계하였다. 코호트 DB 중 최종 표본 추출 대상 모집단 DB를 구성하고, 그 중 연령, 소득 수준으로 분류 하여 최종 표본 코호트 DB를 구성하여 대기오염의 만성 건강영향을 평가하였다.

위와 같이 코호트 DB는 질병 및 건강에 관한 대용량 자료를 바탕으로 다양한 연구가 진행되었다. 그러나 코호트DB의 개별적인 DB 테이블 구조상 연관관계를 검색하기 위해서는 연구목표에 맞게 재구성해야한다는 문제가 있다. 본 논문에서는 위와 같은 문제점을 해결하기 위해 질병이나 건강정보간의 의미적 연관성을 측정하고 검색할 수 있는 코호트 DB 검색 시스템을 개발했다.

3. 온톨로지 기반 코호트 DB 변환 작업

Fig. 1은 코호트 DB를 온톨로지 기반으로 변환하기 위한 프로세스를 그림으로 보여준다. 이 프로세스는 기존의 코호트 DB, XML descriptor, XML parser & editor, 온

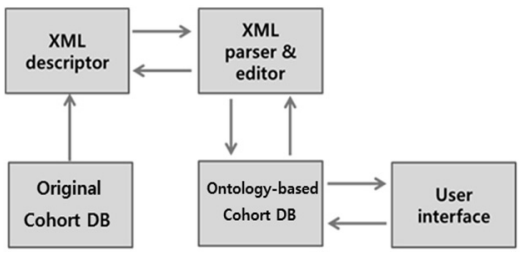

Fig. 1. Conversion process of cohort DB

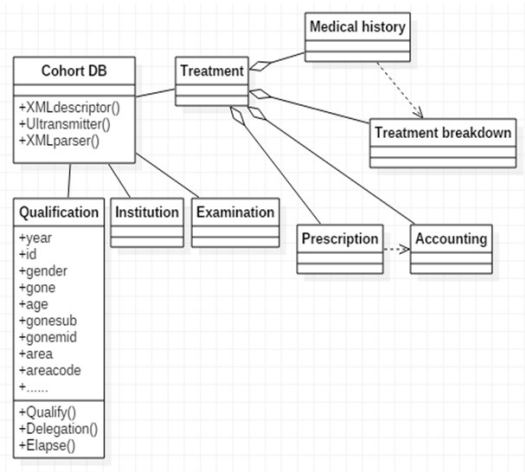

Fig. 2. Structure of Original Cohort DB

톨로지 기반 대용량 코호트 DB, 그리고 UI로 구성되어 있다. 기존의 코호트 DB는 Fig. 2와 같이 자격 DB(Qualification), 건강검진 DB(Examination), 진료 DB(Treatment), 요양기관 DB(Institution)로 총 4개의 DB 테이블이 있고 각 DB 테이블은 DB를 대표하는 속성과 해당 DB를 조작할 수 있는 오퍼레이션으로 구성되어있다. 예를 들어 자격 DB에는 year, id, gender 등과 같은 속성이 있고 이를 조작할 수 있는 Qualify(), Delegation()등과 같은 오퍼레이션이 있다. 진료 DB는 치료의 진행과정에 따른 처방 DB(Prescription), 처방에 대한 결제 DB(Accounting), 환자의 병력 DB(Medical history), 환자의 진료 기록이 저장된 진료내역 DB(Treatment breakdown) 와 같은 4개의 하위테이블을 포함한 집적테이블 구조로 구성되어 있다. 각 하위테이블은 위에서 설명한 자격 DB, 건강검진 DB와 같이 DB의 속성과 오퍼레이션으로 구성되어 있다.

이와 같이 개별적인 DB 테이블 구조에서 질병간의 의미적 연관관계를 찾기 위해서는 코호트 DB를 온톨로지 구조로 변환할 수 있는 XML descriptor와 XML parser & editor가 필요하다. XML descriptor는 코호트 DB의 XML 메타데이터를 분석하는 역할을 한다. XML parser & editor은 descriptor에서 분석한 메타데이터를 읽고 7단계의 Ontology development 101방법론에 의해[3] 온톨로지 계층에 따라 재구성한다.

Ontology development 방법론 7단계와 그에 따른 해결방법은 다음과 같다.

1) 온톨로지의 영역 및 목적에 관해 결정해야한다. 본 시스템의 영역은 코호트 DB 전체로 설정하며 온톨로지 기반 코호트 DB의 목적은 질병 및 건강정보의 연관성을 측정하기 위함이다.
2) 기존의 온톨로지 재사용을 위해 온톨로지 기반 대용량 코호트 DB의 계층구조에 맞는 OWL API를 이용한다.
3) 온톨로지에서 중요한 객체를 선택하기 위해서 본 시스템의 질병 및 건강정보의 연관관계를 최우선 기준으로 세운다.
4) 온톨로지를 구성하는 클래스와 계층을 정의하기 위해 Fig. 2의 DB 테이블을 주 클래스로 정의하고 각 계층은 DB 테이블과 하위 속성을 이용하여 정의한다. 이는 DB 테이블과 속성의 관계가 온톨로지 구조에서의 클래스와 객체의 관계와 흡사하기 때문이다.
5) 각 클래스의 특성을 정의하기 위해 먼저 3)에서 선택한 중요한 객체간의 의미적 연관관계를 고려한다. 그리고 클래스의 특성을 객체에 맞게 정의한다.
6) 온톨로지 기반 코호트 DB를 이루는 객체의 유효성 검증을 위해 값의 범위 및 종류를 제약한다. 이는 대용량 DB를 다루는 본 시스템의 특성상 무효 데이터를 삭제하기 위해 반드시 필요한 작업이다.
7) 코호트 DB의 질병 및 건강 정보를 해당 클래스의 인스턴스와 특성을 입력하는 것으로 인스턴스를 생성한다.

온톨로지 기반 코호트 DB는 위 방법론에 따라 OWL API[3]기반으로 기존의 코호트 DB에서 추출된 정보를 OWL온톨로지로 구축하였다. 이와 같은 작업을 통해 온톨로지 계층에 따라 변환된 XML메타데이터를 온톨로지 구축한 코호트 DB에 저장한다. Fig. 3는 온톨로지 기반 코호트 DB의 구조를 나타낸다. 온톨로지 기반 코호트 DB는 주 클래스 4개로 구성했다. 이는 진료, 기관, 자격, 검진이며 각 주 클래스의 하위에 서브클래스를 구성했다. 또한 각 서브 클래스 하위에는 환자의 질병 및 건강정보가 담긴 객체로 구성했다. 예를 들어 진료 클래스는 상병내역, 진료내역, 명세, 처방전의 4개의 하위클래스를 지니며 이는 진료의 사전, 사후 결과분석을 위해 사용될 것이다. 각 주 클래스와 서브 클래스는 클래스 명에 해당하는 객체속성을 지니며, 각 객체들은 일련번호를 통하여 구분되고, 분류할 수 있다. 또한 각 객체들은 온톨로지를 이용하여 서로 의미적으로 연결되어 있어 시스템에서 의미적인 검색을 하기에 유용하다. 온톨로지 기반 대용량 코호트 DB를 통하여 관리되는 의료정보는 UI를 통해 사용자가 쉽게 접근할 수 있다. 4장에서 보이게 되는 UI는 온톨로지 코호트 DB 내부데이터의 연관관계를 비교하고 연관

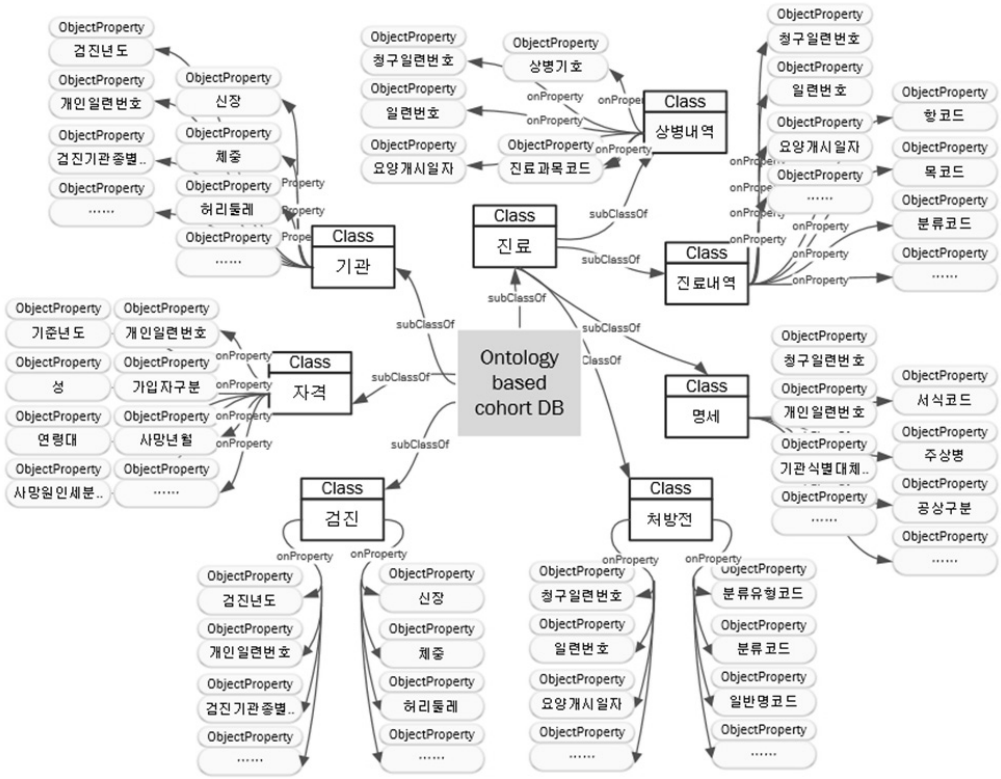

Fig. 3. Structure of ontology-based cohort DB

성을 측정하기 위해 변수들 간의 연관성을 비교할 때 사용되는 연관성 측도기법[13]을 이용했다. 유저가 검색한 정보의 연관관계 및 측정된 연관성을 화면에 출력한다.

4. 온톨로지 기반 코호트 DB 검색 시스템

온톨로지 기반 대용량 코호트 DB 검색 시스템 구현 환경은 Table 1과 같다. CPU는 Intel Core i7-3770, RAM은 8GB DDR3를 이용하였고 운영체제는 Windows 8을 사용하였다. 또한, 온톨로지 코호트 DB에 사용한 OWL API는 현재 가장 많이 쓰는 온톨로지 구축 시스템 중 가장 많이 쓰는 시맨틱 웹 프레임워크이다. Fig. 4는 코호트 DB에서 메타데이터를 가져오고 편집 및 저장하기 위한 시스템이다. DB에 저장된 메타데이터를 나열하는 리스트와 추가기능을 위한 버튼으로 구성되어 있다. 추가기능

Table 1. Experimental Environment

CPU	Intel Core i7-3770
RAM	8GB DDR3
OS	Windows 8
OWL	OWL API

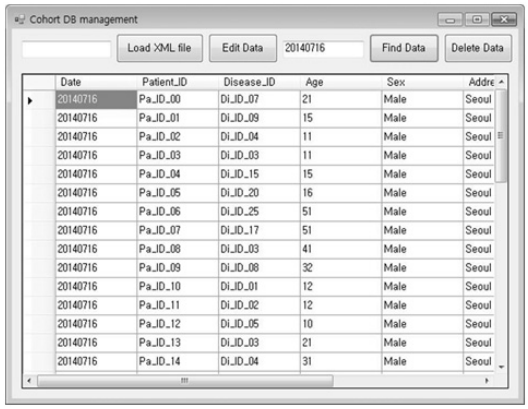

Fig. 4. Ontology-based Cohort DB

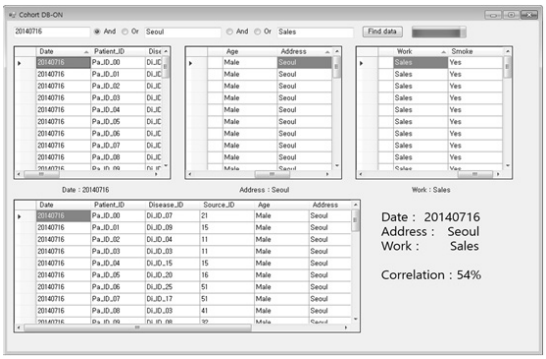

Fig. 5. Ontology-based cohort DB UI

버튼은 Load XML file, Edit data, Find data, Add data로 총 4개가 있다. Load XML file은 메타데이터를 편집하기 위한 XML파일을 불러오는 버튼이다. Edit data는 DB의 데이터를 수정할 수 있는 버튼이고, Delete data는 잘못된 데이터를 삭제하는 버튼이다. Find data는 온톨로지 기반 코호트 DB를 사용하기 위해 데이터를 검색하는 버튼이다. Find data를 실행하면 Fig. 5와 같이 온톨로지 기반 코호트 DB 검색 시뮬레이션을 수행하여 데이터간의 연관성 측정한다. Fig. 5는 코호트 DB의 온톨로지 기반 검색 시뮬레이션 결과를 확인할 수 있는 UI를 보인다. DB 검색 기능과 검색 시뮬레이션의 결과를 확인할 수 있는 리스트와 검색결과의 연관성 측정이 이루어져있다. 상단의 검색 창에 원하는 연관관계를 검색하면 해당관계에 맞는 데이터가 리스트에 나타난다. 이때, 리스트에서 원하는 조건을 기준으로 검색 시뮬레이션을 실행하여 해당 검색결과의 연관성을 측정한다. Fig. 5는 날짜 20140716의 진료데이터와 Seoul에 사는 사람의 진료데이터, 직업이 Sales인 사람의 진료데이터를 검색하는 결과 화면이다. 좌측 하단의 리스트에는 해당 3가지 검색이 모두 해당하는 데이터가 나타나고 있다. 또한 우측 하단에서는 3장에서 설명한 연관성 측도 기법을 이용하여 측정한 연관성을 수치로 나타낸다.

5. 결론 및 향후연구과제

본 논문은 질병 및 건강정보의 연관관계를 나타내기 위하여 온톨로지 기반의 대용량 코호트 DB를 구축하였으며, 구축한 온톨로지 코호트 DB의 연관성 측정 및 연관관계 구성을 위한 검색 시스템을 설계, 구현 및 시뮬레이션 하였다. 본 논문에서 개발한 온톨로지 기반 코호트 DB 검색 시스템은 기존의 코호트 DB와 비교하여 질병 및 건강정보 간의 의미적 연관관계 검색이 용이하다고 할 수 있다.

하지만, 온톨로지 기반 코호트 DB는 환자를 관리하는 클래스가 추가되는 경우 온톨로지 계층구조를 재구성해야한다는 오버헤드가 발생한다. 이는 온톨로지 계층구조를 설계하고 클래스와 특성에 인스턴스를 입력하는 절차상의 문제이므로 다음 후속연구로는 새로운 클래스가 추가하는 경우 설계된 온톨로지 계층구조에 해당 클래스 추가 작업만으로 해결할 수 있는 방법론 개발이 필요하다.

References

1. Wan, Joy, et al. “Risk of moderate to advanced kidney disease in patients with psoriasis: population based cohort study.” BMJ 347 (2013).
2. 노창현, 장성호, 김태영, and 이종식. “시멘틱 컴퓨팅 기반의 동적작업 스케줄링 모델 및 시뮬레이션.” 한국시뮬레이션학회논문지 18.2, (2009): 29-38.
3. 조대웅, 최지웅, and 김명호. “비정형 문서의 정보추출을 통한 OWL 온톨로지 구축 시스템의 설계 및 구현” 한국컴퓨터정보학회논문지 19.10 (2014): 23-33.
4. Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. “Application of data mining: Diabetes health care in young and old patients.” Journal of King Saud University-Computer and Information Sciences 25.2 (2013): 127-136.
5. Ong, Ken KL, et al. “Association between postnatal catch-up growth and obesity in childhood: prospective cohort study.” Bmj 320.7240 (2000): 967-971.
6. 박유성, 박혜민, and 권태연. “국민건강보험 표본코호트 DB를 이용한 건강보험 재정추계.” 응용통계연구 28.4 (2015): 663-683.
7. 배현주. “코호트 자료를 이용한 대기오염의 만성건강영향 평가체계 구축.” 기본연구보고서 2014.단일호 (2014): 1-103.
8. Yoo, Illhoi, et al. “Data mining in healthcare and biomedicine: a survey of the literature.” Journal of medical systems 36.4 (2012): 2431-2448.
9. Herland, Matthew, Taghi M. Khoshgoftaar, and Randall Wald. “Survey of Clinical Data Mining Applications on Big Data in Health Informatics.” Machine Learning and Applications (ICMLA), 2013 12th International Conference on. Vol. 2. IEEE, 2013.
10. 강윤정, 이재일, 배진호, and 이종현. “복소수 SVM을 이용한 목표물 식별 알고리즘.” 전자공학회논문지 50.4 (2013): 182-188.
11. 박유성, 장선화, and 김성용. “연구논문: 사망률 추계를 위한 오차수정 LC 모형.” 조사연구 14.2 (2013): 19-47.

12. 박주환, 김상구. “다중선형 회귀분석을 이용한 고속도로 터널구간의 교통사고 예측모형 개발.” 한국ITS학회논문지 11.6 (2012): 145-154.

13. 이승천, 허문열. “독립성검정에 의한 연관성의 측정.” 통계연구 10.0 (2002): 133-152.

송 주 형 (ringsilver@inha.edu)

2015 인하대학교 컴퓨터정보공학과 학사
2015～현재 인하대학교 컴퓨터정보공학과 석사과정

관심분야 : 빅데이터, 인공지능, 기계학습 등

황 재 민 (nulpis@inha.edu)

2013 인하대학교 컴퓨터정보공학과 학사
2013～현재 인하대학교 컴퓨터정보공학과 석사과정

관심분야 : 빅데이터, 인공지능 등

최 정 석 (jeongseokchoi.korea@gmail.com)

2015 인하대학교 컴퓨터정보공학과 학사
2015～현재 인하대학교 컴퓨터정보공학과 통합과정

관심분야 : Modeling & Simulation

강 상 길 (sgkang@inha.ac.kr)

1989 성균관대학교 전기공학과
1995 Columbia University 전기공학과 전자공학과 공학석사
2002 Syracuse University 전자공학과 공학박사
2004 한국정보통신대학교 연구교수
2006～현재 인하대학교 컴퓨터정보공학과 교수

관심분야 : 모바일 컴퓨팅, 인공지능 시스템, 정보 검색, 신경회로망, 멀티미디어 시스템, 유비쿼터스 시스템, 신호처리 등