하이퍼커넥트 2019년
[DEV] Data Scientist
포트폴리오

성명 : 최 동 근

Contents

WHO IS 최동근 ?

- 최 동 근
- chlehdrms3@gmail.com
- 홍익대학교 산업공학과 졸업 (2015)
- 한양대학교 일반대학원 산업공학과 석사 졸업 (2017)
  지능데이터 시스템 연구실 (이기천 교수)
- 솔트룩스 AI Labs (2017~ 현재)
- Java, Python, R

수행 프로젝트 요약

- **한양대학교 지능데이터시스템 연구실 (석사) - 데이터 마이닝 및 시계열 분석**

  ✓ APM의 성능정보 빅데이터를 이용한 Machine Learning 기반 장애예측 기술 개발 (2015.11. ~ 2016.11.)
  ("정보시스템에서 실시간 다변량 시계열을 위한 동적 전이 모델의 앙상블" -학위논문 / -특허 출원)

  ✓ SK하이닉스 Big Data 기반 대용량실시간분석 시스템 구축 (2015.08. ~ 2015.11.)

  ✓ 영상 빅데이터기반 기계학습을 통한 스마트 범죄예방 솔루션 개발 (2016. 08. ~ 2017. 07.)

  ✓ 서버접근통제 영상정보보안시스템 (2016. 11. ~ 2017. 08.)

- **솔트룩스 Saltlux AI Labs - QA시스템, 지식그래프 연구 개발**

  ✓ 지식그래프 임베딩 및 지식 추론 및 확장 (엑소브레인, 특허 출원 : 지식 임베딩 기반 지식 보강 시스템 및 방법)

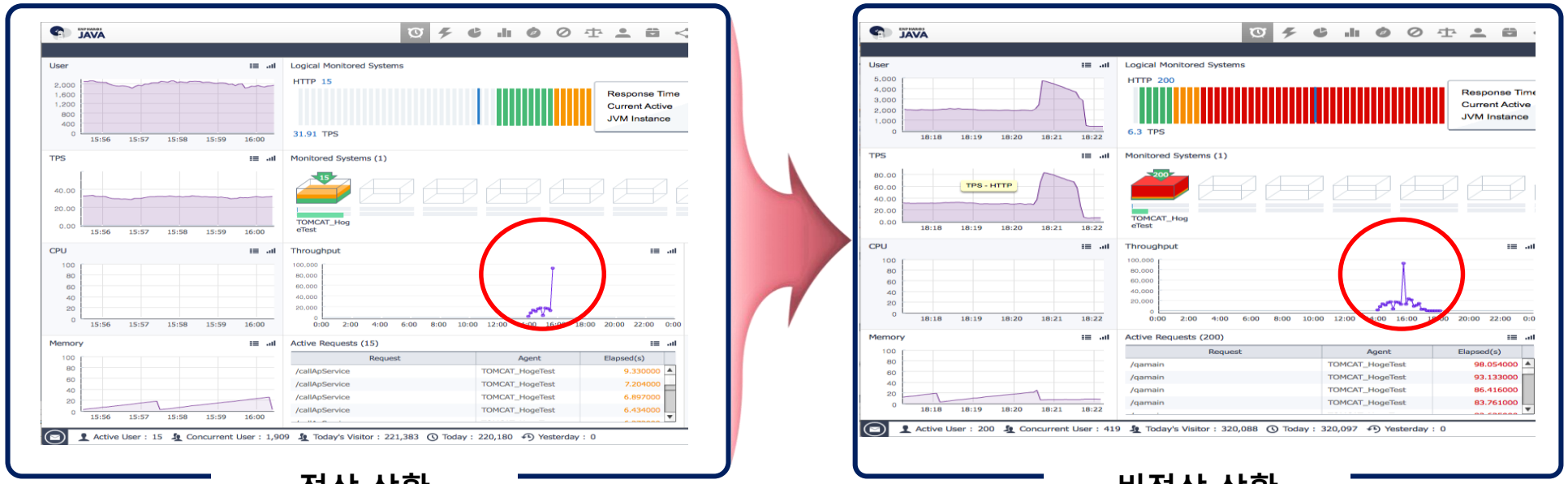
  ✓ 지식 구축 및 QA시스템 개발 (디지털동반자, 특허 출원 : 복수의 데이터 소스들 기반 지식 베이스 구축 시스템 및 방법)

  ✓ 질의 유형 분류기 개발 (*"질의 유형 분류기를 활용한 지식 베이스 기반의 복합 질의 응답 시스템" KIPS 2018*)

  ✓ 문장 임베딩을 활용한 검색 기반의 QA시스템

대표 수행 프로젝트 Ⅰ. APM의 성능정보 빅데이터를 이용한 Machine Learning 기반 장애예측 기술 개발

- 연구 주제
  - ✓ APM(Application Performance Management)에서 발생하는 다변량 시계열 데이터 분석을 통한 장애 예측 및 탐지
- 담당업무
  - ✓ 시뮬레이션을 통한 데이터 수집, 다변량 시계열 데이터 분석 및 예측 모델 개발
- 사용 기술
  - ✓ Jmeter를 활용한 시뮬레이션, R과 Java를 활용한 데이터 분석 및 모델 개발



정상 상황 비정상 상황

HYPERCONNECT

대표 수행 프로젝트 Ⅰ. APM의 성능정보 빅데이터를 이용한 Machine Learning 기반 장애예측 기술 개발

- 프로젝트 내용

데이터 수집	1. APM 테스트 환경 구축 및 장애 유발 시나리오 정의 2. J meter를 활용한 시뮬레이션 3. APM에서 수집되는 Memory, CPU. Active Thread, 동시접속자수, 응답시간 등의 데이터 수집
모델링 Ⅰ (시계열 예측)	수집된 APM 성능 데이터를 이용하여 응답시간에 대한 다변량 시계열 예측 알고리즘 개발 분포가 변하는 특성을 반영할 수 있는 앙상블 모델 개발 실시간 데이터 처리를 위해 RLS(Recursive Least Square) 적용을 통한 고도화
모델링 Ⅱ (시계열 분류)	장애 판단의 기준이 되는 반응변수인 서버의 응답시간 데이터에 대한 비정상 탐지 알고리즘 개발 일정 간격에서의 회귀 계수 추정치 분포에 대하여 G-ESD test를 통해 비정상 탐지 연속적인 비정상 분류에 대하여 LOESS(local regression)을 통해 예측

- 결과

Data		ARIMA	ANN	ENN	WNEF	VAR	W_ver	**RLS_ver**
APM_Ano	Time	301.4	223.6	796.8	1513	71.7	423.8	**218.4**
	RMSE	1.22	1.78	1.19	1.93	5.80	1.65	**0.817**

기존 예측 방법들에 비해 연구 결과의 성능이 우수

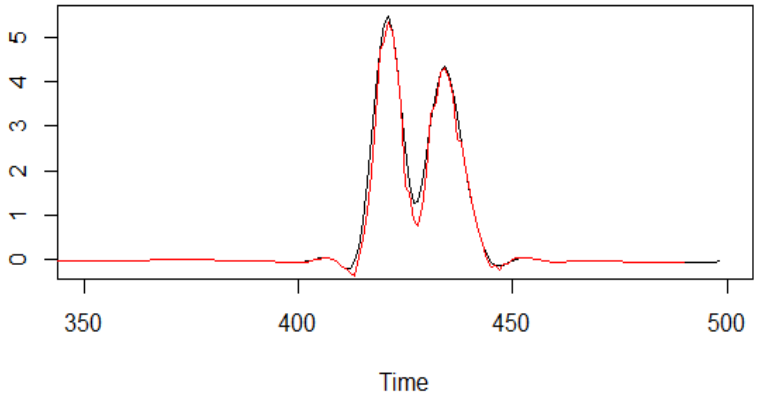
HYPERCONNECT

대표 수행 프로젝트 Ⅰ. APM의 성능정보 빅데이터를 이용한 Machine Learning 기반 장애예측 기술 개발

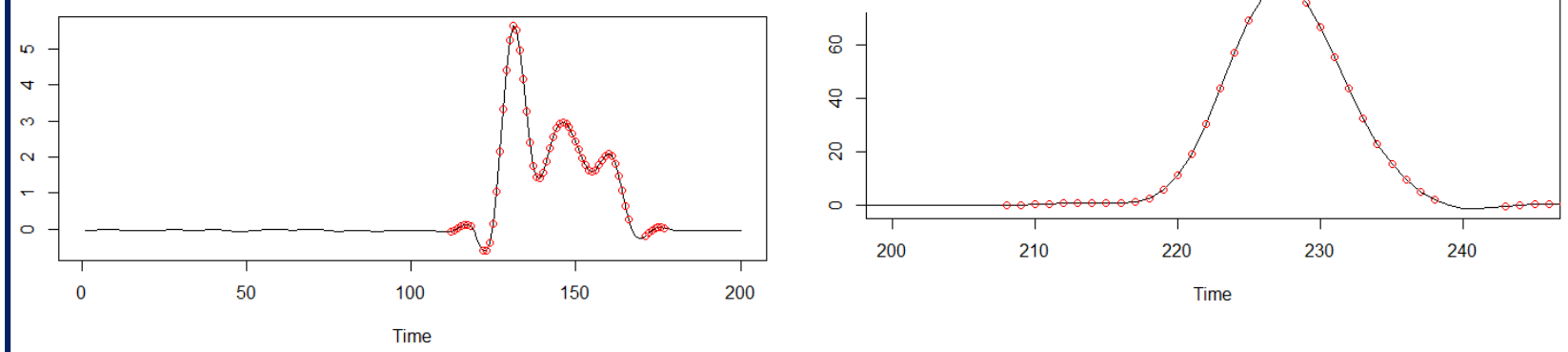
● 프로젝트 결론

- Offline 스텝과 Online 스텝으로 나누어 앙상블 모델을 학습하고 예측하여 분포가 다른 다양한 시나리오에서 좋은 성능을 보임
- RLS(Recursive Least Square)를 적용하여 실시간 예측이 가능
- 기존의 단순 탐지에서 보다 높은 정확도의 예측 모델과 분류 모델을 통해 조기 탐지 가능
- **학위 논문 : "정보시스템에서 실시간 다변량 시계열을 위한 동적 전이 모델의 앙상블"**

T+3 시점 예측 결과



장애 예측 결과

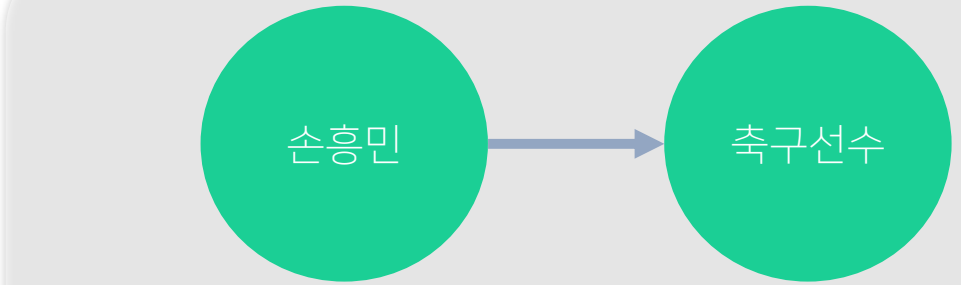


→ *예측 값이 Threshold를 넘는 경우 알람*

대표 수행 프로젝트 Ⅱ. 지식 그래프 임베딩을 활용한 지식 확장

● 프로젝트 개요

- 연구 주제
  - ✓ 지식그래프 임베딩을 통한 새로운 연결 정보 예측



손흥민과 축구선수 사이의 '직업' 이라는 관계를 예측 !

- 연구 배경
  - ✓ 지식그래프는 인공지능 시스템의 뇌 역할을 담당
  - ✓ 다양한 인공지능 분야에서 지식그래프를 활용
  - ✓ 인공지능 품질의 개선을 위해 필요한 지식그래프의 확장은 필수적
  - ✓ 기존의 지식그래프 확장은 사람이 일일이 개입하여 데이터를 구축
  - ✓ 보유하고 있는 지식그래프를 이용하여 연결이 없는 정보에 대한 예측이 필요

- 사용 기술
  - ✓ 지식그래프, SPARQL
  - ✓ Python을 활용한 데이터 정제
  - ✓ Keras (Tensorflow)를 활용한 DNN 모델

대표 수행 프로젝트 Ⅱ. 지식 그래프 임베딩을 활용한 지식 확장

- 프로젝트 내용

데이터 상황	다양한 소스로부터 자동으로 구축된 지식그래프 활용 지식그래프 전체를 사용하는 것은 코스트가 높아 개체 간의 연결 예측이 필요한 데이터 샘플링 예측이 필요한 데이터 샘플링
모델링	입력된 두 개체에 대한 관계를 분류하는 모델 정의 DNN(Deep Neural Network)를 활용한 관계 분류 네트워크 사용

- 프로젝트 결과

프로퍼티 추천

평가 데이터 업로드(엑셀) | 업로드 양식 다운로드

번호	인스턴스#1	인스턴스#2	추천#1	추천#2	추천#3	추천#4	추천#5	추천#6
1	adr:0000608360(손흥민)	adr:0000099551(마우리시오 포체티노)	adp:employer	adp:belongsTo	adp:memberOf	adp:participant	adp:artist	adp:composedOf
2	adr:0000251657(애플)	adr:0000523688(삼성전자)	adp:belongsTo	adp:employer	adp:composedOf	adp:memberOf	adp:opponent	adp:employ
3	adr:0000079756(구글)	adr:0000373738(소프트웨어)	adp:topic	adp:sp_touches	adp:job	adp:field	adp:type	adp:border
4	adr:0000079756(구글)	adr:0000104192(안드로이드)	adp:awarded	adp:ideology	adp:operatingSystem	adp:state	adp:sport	adp:degree
5	adr:0000339821(미녀와 야수)	adr:0000311716(월트 디즈니 컴퍼니)	adp:producer	adp:distributor	adp:participant	adp:productionCompany	adp:editor	adp:productionCrew
6	adr:0000061019(이선희)	adr:0000259471(조용필)	adp:colleague	adp:friend	adp:relative	adp:employer	adp:contribute	adp:opponent
7	adr:0000557137(김경수)	adr:0000552525(김해 김씨)	adp:belongsTo	adp:employer	adp:locatedIn	adp:manager	adp:education	adp:memberOf
8	adr:0000327173(문재인)	adr:0032276200(2018년 남북정상회담)	adp:work	adp:awarded	adp:majorWork	adp:prize	adp:debutWork	adp:write
9	adr:0000450792(유재석)	adr:0000587173(수유중학교)	adp:play	adp:belongsTo	adp:next	adp:bornIn	adp:previous	adp:education
10	adr:0011457761(트와이스)	adr:0000006082(K-pop)	adp:participate	adp:genre	adp:activity	adp:win	adp:diedBy	adp:sport

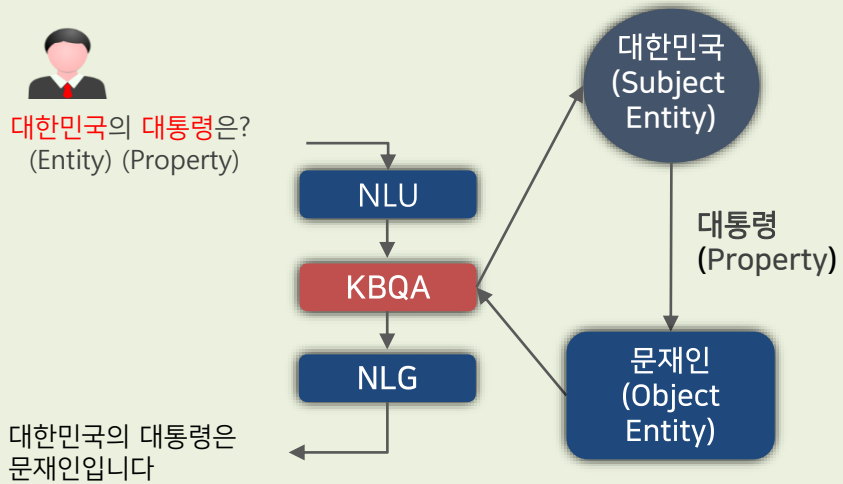
"기 구축 데이터 상세화 및 새로운 속성 추천 시스템"

HYPERCONNECT

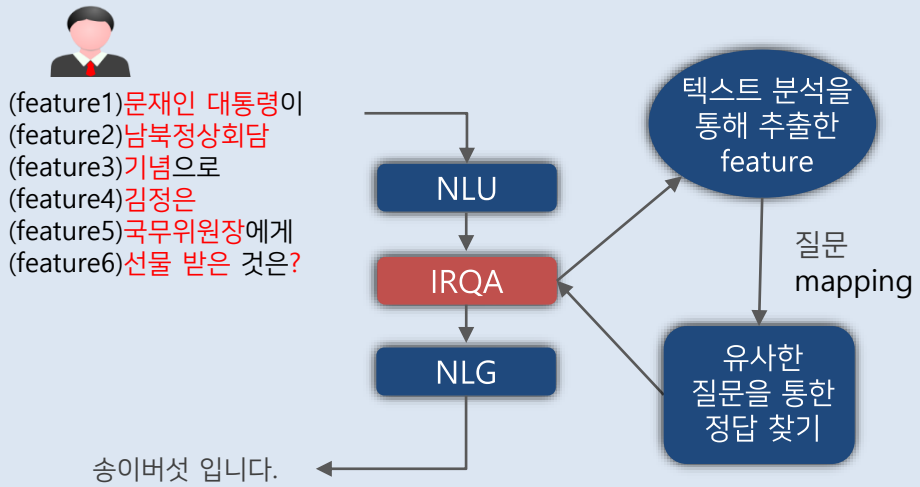
대표 수행 프로젝트 Ⅲ. 질의 응답 시스템

- 질의 응답 시스템
  - ✓ 사용자의 질의를 시스템이 처리해서 적절한 답을 주는 시스템
  - ✓ 질의 응답 시스템은 지식베이스 기반(KBQA)과 검색 기반(IRQA)으로 나뉨
- 사용 기술
  - ✓ SPARQL, 자연어처리, 정보검색, CNN

- KBQA(Knowledge Based Question Answering)
  - ✓ 지식베이스 기반 질의처리 모듈로써 구조화가 용이한 도메인에 적합한 QA



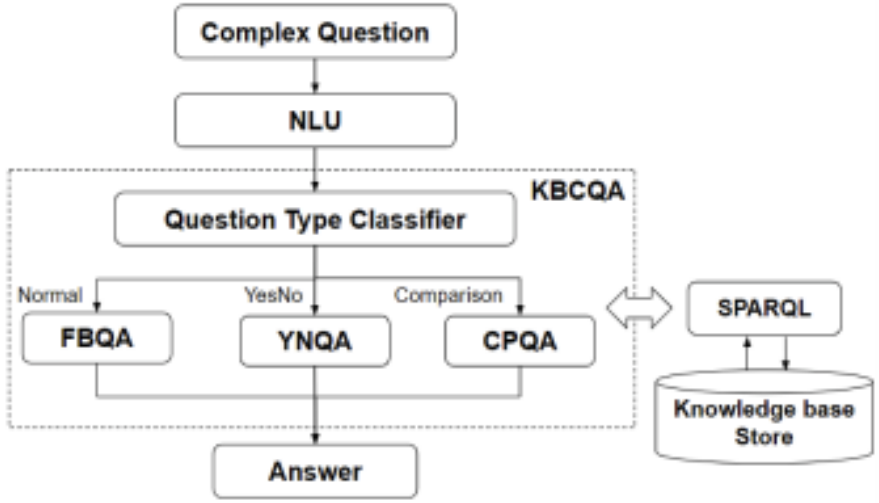
- IRQA(Information Retrieval Question Answering)
  - ✓ 질문/답 기반 질의처리 모듈로, 구조화가 용이하지 않은 도메인에 적합한 QA
  - ✓ 텍스트 분석을 통해 질문과 유사한 질문을 찾아 랭킹 된 답변 도출



대표 수행 프로젝트 Ⅲ. 질의 응답 시스템

● 프로젝트 내용

- 복합 질의 응답 시스템을 위한 질의 유형 분류기
  - ✓ 일반형, 비교형, 판정형 등의 질의 처리를 위한 시스템
  - ✓ CNN(Convolution Neural Network)기반 질의 유형 분류기를 통해 효율적인 시스템 구축



복합 질의 응답 시스템 구조도

● 프로젝트 결과

	학습 정확도	테스트 정확도
Morph+Tag	0.968	0.960
Tag	0.984	0.980

질의 유형 분류기 성능

	Recall	Precision	F1
KBCQA	0.76	0.78	0.76
CNN-KBCQA	**0.81**	**0.86**	**0.83**

복합 질의 응답 시스템 성능

"CNN기반 질의 유형 분류기를 통한 시스템 성능 향상"

- 높은 정확도의 질의 유형 분류기를 통해 질의 유형을 분류하여 질의 처리 최적화
- 보다 빠른 속도로 질의 처리가 가능하며 복합 질의 패턴이 다양해짐에 따른 시스템 확장 용이