해 외 출 장 복 명 서

<table>
<tr><td rowspan="2">출 장 자</td><td>소 속</td><td colspan="2">성인지정책연구실</td><td>직 위</td><td>선임연구위원
연구원(계약직)</td><td>성 명</td><td>문유경
동제연</td></tr>
<tr><td>당 초
변 경</td><td colspan="2" rowspan="2">6.26-7.3 (7박 8일)</td><td rowspan="2">출장지</td><td>당 초
변 경</td><td colspan="2" rowspan="2">미국 뉴욕</td></tr>
<tr><td>출장기간</td></tr>
<tr><td>출장목적</td><td colspan="7">- 2015 IEEE 4th International Congress on Big Data 참석을 통한 DB 시스템 운영 및 성인지 통계 연구의 최신 동향 파악
- Big data를 활용한 정책연구 전문가와의 네트워킹</td></tr>
<tr><td>경비부담</td><td colspan="7">우리원 부담(일반연구사업비): 12,546천원 (문유경 선임연구위원 외 1인)</td></tr>
<tr><td>주최기관</td><td colspan="7">The Institute of Electrical and Electronics Engineers (IEEE)</td></tr>
<tr><td colspan="2">방문기관</td><td colspan="2">면담자</td><td colspan="3">협의사항</td></tr>
<tr><td colspan="2">2015 IEEE 4th
International
Congress
on Big Data</td><td colspan="2">Ephraim Feig 외
학회 발표자 및 참석자</td><td colspan="3">-빅 데이터를 이용한 최신 통계 기법 및 동향 파악
-통계 분석을 활용한 빅 데이터 분석 영역 발굴
-빅 데이터를 이용한 연구의 최신 동향 파악 및 정책 적용 사례 검토
-빅 데이터를 활용한 정책연구 전문가와의 네트워킹</td></tr>
</table>

상세한 업무처리 및 세부내용은 별도 붙임

2015. 7. 15.

출 장 복 명 자 : 문 유 경

<표지>

해외출장 결과 보고서

2015년 제4회 빅 데이터 국제회의

(2015 IEEE 4th International Congress on Big Data)

보고자 : 문유경

성인지정책연구실 성별영향평가 · 통계센터 선임연구위원

Ⅰ. 출장개요

1. 배경 및 목적

가. 2015 IEEE 4th International Congress on Big Data(2015년 제4회 빅 데이터 국제회의)

- o 빅 데이터 분석을 위한 인프라 구축과 분석 기법의 도입은 다양한 분야에서의 서비스 수요가 증대하고 있는 영역임.
- o IEEE 빅 데이터 국제회의에서는 빅 데이터를 이용한 경제 분석, 정부 정책의 수요 파악 및 성과 분석, 의료 데이터 활용 등 다양한 연구의 최신 동향 파악이 가능함.
- o 특히, 클라우드 컴퓨팅, 웹 서비스, 모바일 서비스 등 기업 및 기관의 시스템 운영과 관련된 국제적인 최신 동향을 파악할 수 있는 종합 사이언스 학회로 본원의 성 인지 통계정보 시스템 운영에 도움이 됨.

나. 이에, 빅 데이터 국제회의 참석을 통해 다른 나라의 DB 시스템 운영 동향 및 통계 분석 연구 동향을 파악하고, 관련 연구자와의 네트워크를 구축하여, 본원 성인지 통계정보 시스템의 운영 및 빅 데이터를 이용한 성인지 통계 분석에 적용하고자 함.

1) DB 시스템 운영 동향 파악

- o 빅 데이터를 이용한 최신 통계 기법 및 트렌드 파악
- o 통계 분석을 활용한 빅 데이터 분석 영역 검토 및 발굴
- o 통계 DB 및 시스템 운영 활용 방안 검토

2) 통계 분석 연구 동향 파악

- o 빅 데이터를 이용한 통계 분석 연구의 최신 동향 파악
  - · 공공 데이터 활용
  - · 건강 데이터 활용
  - · 소셜 미디어
  - · 소셜 네트워크 분석
- o 빅 데이터를 이용한 성 인지 통계 분석 동향 파악
- o 빅 데이터를 이용한 정책 적용 사례 검토

3) 관련 연구자와의 네트워크 구축

- o 빅 데이터를 활용한 정책연구 전문가와의 네트워킹

2. 출장자 명단

	기관	부서	직위	이름
1	한국여성정책연구원	성인지정책연구실 성별영향평가 · 통계센터	선임연구위원	문유경
2	한국여성정책연구원	성인지정책연구실 성별영향평가 · 통계센터	연구원(계약직)	동제연

3. 일정

가. 출장일정

1) 기간 : 6월 26일(출국) - 7월 3일(입국) (7박 8일)

2) 장소 : 미국 뉴욕

나. 2015 IEEE 4th International Congress on Big Data 일정

1) 기간 : 6월 27일(토) - 7월 2일(목)

2) 장소 : 미국 뉴욕 (Millennium Broadway Hotel, 1st & 3rd Floor)

3) 세부일정

o 6월 26일(금) 출국: 한국(인천) → 미국(뉴욕), 오전 11시 20분 도착

o 6월 27일(토): Big Data Congress 등록 및 Short paper track 1-6 참석

시간	회의형식	발표주제
9:00-10:00	□ Short paper • Session1	A Real-Time Decision Support Tool for Disaster Response: A Mathematical Programming Approach
		Scalable Query Optimization for Efficient Data Processing using MapReduce
		Developing a Real-Time Data Analytics Framework using Hadoop
		Content Based Image Retrieval on Hadoop Framework
		Phishing URL detection using URL Ranking
10:15-11:15	□ Short paper • Session2	High-Order Tensor Decomposition for Large-Scale Data Analysis
		Optimal Feature Extraction and Classification of Tensors via Matrix Product State Decomposition
		A Workflow Model for Adaptive Analytics on Big Data
		Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty
		Analytics of industrial operational data inspired by natural language processing
		Online Nonlinear Classification for High-Dimensional Data
11:20-12:20	□ Short paper • Session3	Cloud Tree: A Library to Extend Cloud Services for Trees
		Matrix-based XML Stream Processing using a GPU
		A Flexible Data-driven Approach for Execution Trace Filtering
		Road traffic analytic query processing based on a timeline modeling
		Approximate querying in big heterogeneous data
13:30-14:30	□ Short paper • Session4	Queriosity: Automated Data Exploration
		Write Optimization using Asynchronous Update on Out-of-Core Column-Store Databases in Map-Reduce
		Social network analysis of developers and users mailing lists of some free open source software
		A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers: An Application Platform for Public Health Research
14:40-15:40	□ Short paper • Session5	How has Twitter changed the Event Discussion Scenario?:A Spatio-Temporal Diffusion Analysis
		Discovering Environmental Impacts on Public Health using Heterogeneous Big Sensory Data
		Detecting Corporate Social Media Crises on Facebook using Social Set Analysis
		Data Quality and Energy Management Tradeoffs in Sensor Service Clouds

o 6월 28일(일): Keynote 및 Plenary Panel 참석

시간	회의형식	발표주제
9:40-11:30	□ Keynote1	A Holistic View of Software Evolution toward Software as a Service(Carlo Ghezzi, Politecnico di Milano, Italy)
13:40-14:40	□ Plenary Panel	The Arc of Research Activity in Key Areas of Services Computing
16:15-17:25	□ Plenary Pane2	Big Data and IoT

o 6월 29일(월): Keynote, Plenary Panel 및 Big Data Congress Visionary Track 참석

시간	회의형식	발표주제
10:50-12:00	□ Keynote2	Big Data as a Service at NASA
14:10-15:20	□ Plenary Pane3	Service Economics on Cloud Computing
15:30-16:30	□ Visionary • Session 1	Research Directions for Big Data Graph Analytics
		MCD: Mutual Clustering across Multiple Social Networks
		Big SaaS: The Next Step Beyond Big Data

o 6월 30일(화): Big Data Congress Applications, Research Track 및 Plenary Panel 참석

시간	회의형식	발표주제
8:15-9:15	□ Applications • Session 1	Big Data Analytics Framework for System Health Monitoring
		Predictive Modeling for Comfortable Death Outcome using Electronic Health Records
		H-DRIVE: A Big Health Data Analytics Platform for Evidence-Based Decision Making
9:25-10:25	□ Applications • Session 2	Design and Realization of Cognized Routing Resource by Big Data Analysing in SDN
		Toa: A Web Based Network Flow Data Monitoring System at Scale
13:00-14:00	□ Research • Session 3	A Clustered Approach for Fast Computation of Betweenness Centrality in Social Networks
15:30-16:30	□ Research • Session 4	Can we Rank Emotions? A Brand Love Ranking System for Emotional Terms
		Deriving Topics in Twitter by Exploiting Tweet Interactions
		Matrix inter-joint Factorization - A New Approach for Topic Derivation in Twitter
16:50-18:00	□ Plenary Pane4	Convergence of Cloud Computing and Big Data: Making Big Social and Human Impact

o 7월 1일(수): Big Data Congress Research, Applications, Special Track(BDRH) 및 Plenary Panel 참석

시간	회의형식	발표주제
8:15-9:15	□ Research • Session 5	Privacy Preserving Data Analysis in Mental Health Research
		Enabling Privacy Mechanisms in Apache Storm
		Sensitive Disclosures under Differential Privacy Guarantees
	□ BDRH	Blood Pressure Management with Data Capturing in the Cloud among Hypertensive Patients: A Monitoring Platform for

시간	회의형식	발표주제
	• Session 1	Hypertensive Patients
		Indoor Air Monitoring Platform and Personal Health Reporting System: Big Data Analytics for Public Health Research
		Embracing Big Data for Simulation Modelling of Emergency Department Processes and Activities
9:25-10:25	□ Research • Session 6	Red-RF: Reduced Random Forest for big data using priority voting & dynamic data reduction
		Optimal and Efficient Distributed Online Learning for Big Data
		Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection
	□ BDRH • Session 2	Risk-adjusted Monitoring Method for Surgical Data: Methodology for Data Analytics (Work in Progress)
		Patient Flow Evaluation with System Dynamic Model in an Emergency Department: Data Analytics on Daily Hospital Records
		Two screening methods for genetic association study with application to psoriasis microarray data sets
13:00-14:00	□ Applications • Session 7	Study on Corporate Governance of Stock Market in Korea: Network Analysis with relationship of Major Shareholders
		Performance Evaluation of NoSQL Databases: A Cass Study
		Optigrow: People Analytics for Job Transfers
16:50-18:00	□ Plenary Pane5	Mobile and IoT Services

o 7월 2일(목): Big Data Congress Applications, Research Track 및 Closing Session 참석

시간	회의형식	발표주제
8:15-9:15	□ Applications • Session 9	Unsupervised Event Detection with an Infinite Poisson Mixture Model
		Reconstructability-aware Filtering and Forwarding of Time Series Data in Internet-of-Things Architectures
		NoSQL in practice: a write-heavy enterprise application
9:45-10:45	□ Applications • Session 10	Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander
		Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity
		Automation of the Validation, Anonymization and Augmentation of Big Data from a Multi-year Driving Study
10:55-11:55	□ Research • Session 11	Hybrid Traffic Speed Modeling and Prediction Using Real-world Data
13:00-14:00	□ Research • Session 12	Big Data Open Source Platforms
14:10-15:00	□ Closing Session	

o 7월 3일(금) 입국: 미국(뉴욕) → 한국(인천), 오후 2시 출발

Ⅱ. 주요내용

1. 기조발표

가. Keynote 1

o 주제: A Holistic View of Software Evolution toward Software as a Service

o 발표자: Carlo Ghezzi (Politecnico di Milano, Italy)

o 일시 및 장소: 6월 28일(일), 오전 9:40-11:30, Hudson Theatre

o 내용 요약

서비스로서의 소프트웨어라는 최근의 트렌드는 전례 없는 도전을 야기하고 있음. 이미 수십 년 간 요구되고 있지만 '더 나은 소프트웨어로의 진화'에 대한 요구 또한 끊임이 없는 실정임. 소프트웨어는 필수적인 요건들의 변화, 상호작용하는 환경의 변화, 사용 기반 플랫폼의 변화, 또는 소프트웨어에 사용되는 응용프로그램의 변화에 따라 지속적인 발전이 필요함.

지속적인 변화는 반복적이면서도 계속 성장해 나간다는 본질적인 특성을 지니고 있음. 특히, 소프트웨어의 진화는 프로그램이 실행되는 시간을 줄이는 측면에서 진행되어야 하며, 이를 통해 소프트웨어가 운영되면서 서비스를 제공하게 되는 것임. 더 나아가 변화에 스스로 적응할 수 있는 소프트웨어에 대한 요구 또한 높음.

소프트웨어 발전에 대한 기존의 접근은, 유동성과 의존성이 요구되는 이러한 도전적 상황에 호응할 수 있는 방법으로 재고되어야 함. 소프트웨어를 설계 시간(개발)과 실행 시간(운영)으로 구분하던 전통적인 분류방법은 모호하며, 이제는 둘 사이의 경계가 서서히 사라지고 있는 실정임. 끊임없이 지속되는 소프트웨어 개발과 진화에 대해 다시 생각해본다는 측면에서 모델링과 검증 시스템에 초점을 맞추어 소프트웨어의 진화적인 측면을 검토하였음. 또한 클라우드 컴퓨팅, 서비스 기반 컴퓨팅 등 가상 환경에서 정확성, 고성능, 에너지 소비 등의 비기능적인 필수요소에 대해 지속적으로 충족할 수 있도록 소프트웨어의 '자기 적응화' 필요성에 대해서도 언급함.

나. Keynote 2

o 주제: Big Data as a Service at NASA

o 발표자: Tsengdar J. Lee (NASA, USA)

o 일시 및 장소: 6월 29일(월), 오전 10:50-12:00, Hudson Theatre

o 내용 요약

빅 데이터가 직면한 문제는, 전통적인 저장 장치들을 이론적, 기술적으로 재배열 하는 문제로서, 사이즈가 큰 데이터를 어떻게 관리할 것인가에 대한 문제임. 이러한 접근법은 관심 사항 중 특정 부분을 발견하기 위해 많은 양의 구조화된 데이터를 저장 또는 관리하는 방법을 찾아내기 위한 것으로서 빅 데이터에 접근하는 방법임.

두 번째로 빅 데이터가 직면한 문제는, 분석 기법이나 기술을 재배열 하는 것으로서의 지식 경영의 문제임. 이 접근법은 관심 사항에 대한 깊이 있는 이해를 통해 많은 양의 비구조화된 데이터로부터 의미 있는 패턴을 찾아내기 위한 것으로서 빅 데이터에 접근하는 관점임.

'빅 데이터'를 다루면서 우리가 자연스럽게 갖게 되는 사이즈에 대한 고민을 해결

하기 위해서, 위에 제시한 두 개의 관점이 균형을 이루어야 함. 그동안 진행해 온 NASA의 연구 및 개발 실적들을 제시하면서, 이 두 관점을 어떻게 상호 운용함으로써 두 관점 사이의 균형을 맞추고 있는지 검토함.

2. 패널 토론

가. Plenary Panel 1

- o 주제: The Arc of Research Activity in Key Areas of Services Computing
- o 사회자: John Miller (University of Georgia, USA)
- o 패널: Hong Zhu (Oxford Brookes University, UK)
  Elisa Bertino (Purdue University, USA)
  Ling Liu (Georgia Tech, USA)
  Manish Parashar (Rutgers University, USA)
- o 일시 및 장소: 6월 28일(일), 오후 1:40-2:40, Hudson Theatre
- o 내용 요약

패널 토론의 목적은 연구 영역, 지표, 향후 연구 프로젝트를 검토 및 발굴하기 위함. 특히, (1) 신뢰와 액세스 제어 등과 같은 서비스 보안 및 개인정보 보호, (2) 클라우드 서비스 및 빅 데이터 등의 웹을 넘어선 서비스 응용, (3) 사용자 수 증대에 유연하게 대응할 수 있는 컴퓨터 확장성 및 역동성 등을 의미하는 서비스 컴포지션, (4) 설계부터 테스트까지의 서비스 지향 소프트웨어 공학 기술에 대한 영역을 다룸.

나. Plenary Panel 2

- o 주제: Big Data and IoT
- o 사회자: Tony Shan (Chief Technologist)
- o 패널: Cristian Sturek (Founder of IoT.do and InteliAthlete Corp)
  Dan Daogaru (General Manager of IoT at Hortonworks)
  Jason Kolb (CTO at Uptake)
  Tom Gilley (Founder & CTO of wot.io)
- o 일시 및 장소: 6월 28일(일), 오후 4:15-5:25, Hudson Theatre
- o 내용 요약

실시간 빅 데이터는 실생활에서 빅 데이터를 찾아서 수집, 변형, 저장, 처리, 분석하는 과학, 공학 기술, 첨단 기술 등의 지식 분야가 집적된 것이라고 할 수 있음. 수 시간, 수 일이 걸리던 분석 과정은 현재 1초도 안 되는 시간에 가능하도록 줄어들었음. 대게 전통적 분석 방법은 전통적인 자료 저장소로부터 검색된 구조화된 데이터를 가져와 기록하고 보고하는 과정임. 실시간 빅 데이터는 실생활의 구조화, 반구조화 또는 비구조화된 데이터의 다양한 근원적 차이를 이해하고, 각각의 장점을 찾아냄으로써 사업 가치를 증진시키므로, 앞서 제시한 전통적 분석 방법을 보다 발전된 단계로 끌어올릴 수 있음.

최신의 트렌드는 보다 큰 매출과 이윤을 가져올 수 있도록, 보다 빠르고, 지능적이면서도, 자동화되어 있고, 비용효율이 높고, 정확한 맥락에서 빠른 이해를 이끌어내면서도 구체적인 장점을 제안하는 양질의 데이터 처리 과정을 선호함.

컴퓨터 고성능 체계에서의 최근의 진보에도 불구하고, 온라인상에는 여전히 많은 문제, 장벽, 위험이나 어려움, 도전이 산재해 있음. 이에, 패널 토론에서 그 간의 경험을 공유하고 향후 빅 데이터와 사물인터넷 분야의 발전 발향을 전망함.

다. Plenary Panel 3

o 주제: Service Economics on Cloud Computing
o 사회자: Ajay Mohindra (IBM Research, USA)
o 패널: Bharat Bhargava (Purdue University, USA)
Dennis Gannon (Indiana University, USA)
Michael Goul (Arizona State University, USA)
Gueyoung Jung (AT&T Labs, USA)
Nianjun Zhou (IBM T.J. Watson Research, USA)
o 일시 및 장소: 6월 29일(월), 오후 2:10-3:20, Hudson Theatre
o 내용 요약

3차 패널 토론에서는 클라우드 컴퓨팅이 서비스 경제의 변화를 어떻게 이끌어낼 수 있는지를 중점적으로 토론함. 클라우드 컴퓨팅은 구독 모델(이용자가 이용에 대한 이용료를 정기적으로 내는 모델로서 콘텐츠 서비스, P2P 네트워킹 서비스, 신뢰서비스, 인터넷 서비스 제공자 등을 세부 모형으로 들 수 있음)을 활용하면서 IT 산업을 제품 기반의 경제에서, ‘필요에 의해 무엇인가를 사는’ 서비스 기반의 경제로 변화키시고 있음. 우리는 IaaS(Infrastructure-as-a-Service=host), PaaS(Platform-as-a-Service=build)와 SaaS(Software-as-a-Service=consume)가 출현할 때부터 이 과정을 접해 왔음.
그러나 이 변화의 과정은 아직 진행 중임. 각각의 서비스에 대한 가치평가에 기반한 서비스들 간의 상호작용 또는 특정한 사업의 운영 및 문제 해결을 지원하는 서비스 공급자의 능력 등을 포함하여 아직 해결해야 할 많은 기술적인 어려움이 남아 있음. 여전히 남아있는 과제 중 하나는 서비스 친화적인 시스템을 창조하는 것이고, 서비스 친화적인 시스템이란, 물적 서비스 세계와 유사하게 정교한 서비스 환경을 만들기 위해 다양한 각각의 서비스를 모으는 과정을 의미함.
사업-가치평가로 작동하는 구독 모델에 대한 한계를 인정하면서 클라우드 컴퓨팅에서 현재의 구독 모델의 역할을 살펴보고, 서비스 신뢰성과 속도, 서비스 자동화와 규모(크기), 그리고 API 중심 컴퓨터 시스템 구성을 통한 서비스 통합에 대해 살펴봄.

라. Plenary Panel 4

o 주제: Convergence of Cloud Computing and Big Data: Making Big Social and Human Impact
o 사회자: Calton Pu (Georgia Tech, USA)
o 패널: Henning Schulzrinne (Columbia University, USA)
Peter Chen (Carnegie Mellon University, USA)
Bhavani Thuraisingham (University of Texas at Dallas, USA)
o 일시 및 장소: 6월 30일(화), 오후 4:50-6:00, Hudson Theatre
o 내용 요약

클라우드 컴퓨팅은 잠재적으로 모든 사람이 사용 가능한 전례 없는 계산력, 저장 능

력을 축적하고 있음. 빅 데이터는 - 특히 인간 활동과 관련된 - 많은 영역에서 축적되고 있음. 클라우드 컴퓨팅과 빅 데이터는 개인 의료, 정밀 의학, 스마트 설비, 교통, 사이버 안보, 물리적 보안과 같은 광범위한 영역에서 우리 삶을 근본적으로 변화시킬 수 있음. 그러나 클라우드 컴퓨팅과 빅 데이터의 융합은 몇 가지 이슈 또한 동반함. 패널 토론에서는 다음과 같은 이슈에 대해 논의함. (1) 융합에 따른 잠재적 이익과 개인정보 손실과 같은 위험. (2) 이익을 달성하는데 있어서의 기술적 어려움과 윤리적, 법적, 사회적 영향력. (3) 선진국(가진 자들)의 잠재적 이익과 개발도상국(못 가진 자들)의 잠재적 이익.

마. Plenary Panel 5

o 주제: Mobile and IoT Services

o 사회자: Nimish Radia (Director of Research and Innovation, Ericsson Research, USA)

o 패널: Rong Chang (Member of IBM Academy of Technology)
Roberto S. Silva Filho (GE Global Research, USA)
Teresa Tung (Accenture Technology Labs, USA)
Manish Parashar (Rutgers University, USA)

o 일시 및 장소: 7월 1일(수), 오후 4:50-6:00, Hudson Theatre

o 내용 요약

2020년까지 500억 개 이상의 사물이 서로 연결될 것임. 모바일과 사물인터넷 서비스는 의료서비스, 금융, 자동차 네트워크, 교통/수송, 디지털 공장, 수도 · 전기 · 가스와 같은 공익사업 등 거의 모든 산업의 사이버스페이스와 물리적 연계라는 변화의 중심이 될 것임. 최근 이슈인 모바일 및 사물인터넷 서비스의 생산 및 활용과 관련된 주요 요소는 ‘크고, 빠르고, 작은’ 센서 데이터 기반의 상황 인지형 기기, API로 정의된 산업 플랫폼(사용 기반이 되는 컴퓨터 시스템이나 소프트웨어) 및 이동성 가능하도록 설계되고, 프로그램 작동이 가능한 서비스임.

5차 패널 토론에서는 산업의 디지털화를 위해 어떻게 이러한 서비스가 생산되고, 구성되는지에 대해 논의함. 주요 필수 기술 요소 및 트렌드 - 예컨대, 자동화 과정부터 새로운 혁신 기술에 이르기까지 더 나은 사업 결과를 도출하기 위해 플랫폼, 기반 시설 상부에서 모바일과 사물인터넷 서비스의 지속적인 개발과 운영을 가능케 하는 구체적인 기술 및 트렌드가 무엇인지에 대해 논의함. 또한 모바일과 사물인터넷 서비스의 연계 데이터, 산업 자산, 착용가능성, 시스템 등과 더불어 시각화 양식과 이동성으로 점철되는 고유한 특성에 대해 논의함(사물인터넷 서비스 사례: 스마트 홈, 웨어러블 기기(애플 워치, 신체 변화를 측정하는 신발), 스마트 자동차, 드론 등. 인터넷에 연결된 서비스+사물을 활용한 지능형 서비스)

3. 소논문

가. Session 1

o 일시 및 장소: 6월 27일(토), 오전 9:00-10:00, R3.11

- 사회자: Liqiang Wang (University of Wyoming, USA)

1) A Real-Time Decision Support Tool for Disaster Response: A Mathematical Programming Approach

- 발표자: Yong-Hong Kuo (The Chinese University of Hong Kong, China)
- 내용 요약

재난은 갑작스럽게 발생하는 사건으로 사회적으로 심각한 악영향을 줄 뿐 아니라 인적 손실까지 동반함. 정부와 인도주의 구호기관 등은 재난으로부터 기인한 부정적인 결과들을 줄이거나 피하기 위해 엄청난 노력을 기울임. 최근 몇 년, 재난 관리에 있어서 IT(정보통신기술)와 빅 데이터가 중요한 역할을 해오고 있음. 재난 정보 추출 및 전파에 있어서는 많은 역할을 하는 반면, 재해 대책에 대한 결정을 지원하기 위한 실시간 최적화는 빅 데이터 연구에서 좀처럼 다루어지지 않고 있음.

따라서 향후 긴급 보급품 전달과 관련한 재난 관련 결정에 활용하기 위해 실시간 재난 관련 정보에 대한 수리 계획적 접근법을 제안함. 이와 같이 개발된 결정 지원 도구는 재단 대책을 위한 신속하고 효과적인 해결책을 제공할 수 있음.

Figure 1. The flow of delivering emergency supplies to communities of affected areas.

- 활용 영역: 재난
- 활용 가능한 빅 데이터: (1) 자연재해 또는 인재를 관리 감독하기 위한 인공위성 원격 탐사 정보, (2) 재난 피해가 가능한 영역과 가까운 곳에 위치한 사람들에게 재난 경보를 전파하고, 피난처를 안내하기 위해 스마트폰과 모바일 애플리케이션을 활용한 정부 정보, (3) 정부와 구호단체에게 도로 상황 등 최신의 교통 상황을 제공함으로써 피해자들에게 긴급 구호 물품을 적절하게 전달할 수 있도록 지원한 GPS 시스템, (4) 재난 피해 사상자로부터 직접 수집된 신체 상황 정보를 통해 치료 받을 수 있는 곳에 대한 정보를 바로 제공, (5) 재난 발생 동안 사람들의 위치 정보를 공유함으로써 구조팀이 피해자들에게 쉽게 접근할 수 있도록 한 소셜 미디어, (6) 긴급한 상황이 발생한 지역, 규모 등 재난 대책 마련이 가능하도록 재난 관련 정보를 수신, 보관, 공유하는 클라우드 플랫폼, (7) 재난 복구 및 지역사회 회복 상황과 관련된 파생 정보, (8) 지진을 탐지하는 트윗 데이터

2) Scalable Query Optimization for Efficient Data Processing using MapReduce

- 발표자: Yi Shan (Arizona State University, USA)
- 내용 요약

맵리듀스(분산 컴퓨팅 지원을 위해 개발한 구글의 소프트웨어 프레임워크. 대용량 데이터의 신속, 안전한 처리를 위한 것. 맵 단계에서는 흩어져 있는 데이터를 연관성 있는 데이터끼리 분류하여 묶어주고, 리듀스 단계에서는 맵 작업 후 중복 데이터를 제거하고 원하는 데이터를 추출함. 대표적 맵리튜스 프레임워크 중 하둡(hadoop)이 가장 주목받고 있음)는 빅 데이터 쿼리 프로세싱(구조화된 형태로 데이터를 보관하고 있다가 쿼리를 통해 필요한 정보를 얻는 과정)을 위한 효율적인 프로그래밍 모델로서 산업과 학계에서 모두 널리 인정받음. 맵리듀스를 이용한 SQL(Structured Query Language: 데이터베이스를 직접 액세스할 수 있는 언어로, 데이터를 정의하고, 조작하며, 조작한 결과를 적용하거나 취소할 수 있고, 접근권한을 제어하는 처리들로 구성됨) 쿼리를 실행하기 위해 가장 효율적인 방법으로 최적화 프로그램을 설계하는 것이 중요함. 그러나 실제로 병렬 쿼리 프로세싱은 맵리듀스를 이용한 SQL 쿼리 최적화에 미치지 못하거나, 시간 복잡성을 기하급수적으로 늘어나게 함. 또한 HIVE, YSmart와 같은 산업 솔루션들은 SQL 쿼리의 배열 연계를 최적화하지 못함. 발표에서는 SOSQL이라는 명칭의, 맵리듀스 활용 SQL 쿼리를 위한 확장 가능한 최적화 프로그램을 구글 클라우드 플랫폼을 사용 · 개발하여 제안함.

3) Developing a Real-Time Data Analytics Framework using Hadoop

- 발표자: Sangwhan Cha (University of New Brunswick, Canada)
- 내용 요약

현재 활용되는 작업흐름 모델은 메타휴리스틱 방법을 기반으로 함. 메타휴리스틱 방법은 실제 상황에서 역동적으로 적용되고, 상황에 맞게 적합한 서비스를 제공하기 위한 워크플로우 작업을 발견하는 양질의 휴리스틱(부피가 큰 문제나 복잡한 문제를 풀기 위해 시행착오를 반복 평가해가면 자기 발견적으로 문제를 해결하는 방법)을 생산함. 그러나 메타휴리스틱 방법은 데이터 유형이나 실시간 처리과정을 고려하면서 분석적인 작업을 지원하는 능력은 부족함. 발표에서는 다양한 분석적인 작업의 실행, 데이터 흡수, 데이터 탐색 및 시각화를 처리하는데 필요한 비정형화 또는 정형화된 데이터에 대한 실시간 처리를 다루기 위한 실시간 데이터 분석 프레임워크를 개발함으로써 이러한 문제를 해결하는 방법에 대해 고려함. 본 연구에서는 정형화 및 비정형화된 데이터의 흡수, 탐색 처리, 스트리밍 시각화를 위해 Storm/YARN 프로젝트에 기반한 아키텍처를 제안함. 로컬 모드와 분산 모드를 위한 API 서버와 연결된 아파치 스톰을 활용하여 본 연구에서 제안한 아키텍처를 직접 실행해 봄. 아키텍처 시제품 완성을 위한 연구 과정을 설명하고, 각 구성요소들의 기능적인 필요성 및 실시간 업데이트, 구성요소들 간 데이터 흐름을 위해 필요한 시간 등 비 기능적인 테스트에 대해 평가함. 모든 구성요소들은 각각의 기능에 대해 적절하게 다루어졌고, 시스템 효율성을 입증하기 위해 비 기능적인 테스트의 주요 결과에 대해서도 제공함.

4) Content Based Image Retrieval on Hadoop Framework

- 발표자: U.S.N. Raju (National Institute of Technology, India)
- 내용 요약

하둡 맵리듀스 프레임워크에서의 내용기반검색(CBIR, Content Based Image Retrieval: 특정 이미지가 서버에 전달되면 데이터베이스에 저장되어 있는 색깔, 모양, 질감, 표정, 레이아웃 등 기본적인 수준의 특성들을 바탕으로 다른 이미지들과 비교해서 가장 유사한 이미지를 찾아주는 과정)에 대해 소개함. 병렬적으로 제공되는 양식이라면 어떤 알고리즘이라도 실제 검색에 활용될 수 있지만, 특히 LTrPs라는 접근법에 대해 논의함. 여기에 소개된 맵리듀스 실행 세부 사항은 특정 벡터 추출 및 거리 측정 등을 포함한 대부분의 내용기반검색 기술에 활용될 수 있음.

5) Phishing URL detection using URL Ranking

- 발표자: Mohammed Nazim Feroz (Texas Tech University, USA)
- 내용 요약

악성콘텐츠를 업로드 하는 범죄가 웹에 노출되고 있음. 사실, 광범위한 연구에도 불구하고, 스팸 메일을 걸러내는 이메일 기반 기술은 다른 웹 서비스를 보호하기에는 아직 한계가 있음. 따라서 호스트 URL을 피싱하는 범죄로부터 일반 웹 서비스 사용자를 보호하기 위한 대책 마련이 시급함. 발표는 호스트에 사용된 어휘나 호스트 특색을 검토함으로써 URL을 자동적으로 분류해내는 접근법에 대해 논의함. 클러스터링은 데이터세트 전반에 실행되고, 각각의 URL에는 분류 시스템에서 예측을 위해 차례대로 사용되는 클러스터 ID(라벨)가 부여됨. URL을 범주화하기 위해 온라인 URL 레퓨테이션 서비스가 사용됨. 또 URL 반송 내역도 URL을 평가하기 위한 추가 정보원으로 활용됨. 그 결과, 대다수의 피싱 호스트를 골라냄으로써 분류 효율성은 93~98%의 정확성과 통계적으로 무의미한 오류율을 보임. URL 클러스터링, URL 분류, URL 범주화 메카니즘이 결합하여 URL을 평가하는데 효과적으로 작용함을 밝혀냄.

나. Session 2

o 일시 및 장소: 6월 27일(토), 오전 10:15-11:15, R3.11

- 사회자: Kelvin K.F. Tsoi (The Chinese University of Hong Kong, China)

1) High-Order Tensor Decomposition for Large-Scale Data Analysis

- 발표자: Longzhuang Li (Texas A&M University-Corpus Christi, USA)
- 내용 요약

고차원 텐서 분해는 데이터 마이닝 작업에 있어서 매우 중요한 기본 과정임. 효율적인 대용량 텐서 분해 알고리즘은 클러스터링, 경향 탐색, 이상 탐지 등에 긍정적 영향을 미침. 발표에서는 하둡 맵리듀스 프레임워크를 사용한, Tucker(터커) 텐서 분해의 확장 가능하면서도 분해된 버전인 MR-T의 개발에 대해 소개함. 연구에서는 이중 행렬 곱셈 알고리즘을 피하는 대신, 중간 데이터와 연산을 최소화하기 위해 매개 행렬을 연속적으로 컴퓨팅하고 매개 텐서 벡터를 생산함으로서 대규모 데이터 세트를 분해함.

2) Optimal Feature Extraction and Classification of Tensors via Matrix Product State Decomposition

- 발표자: Johann A. Bengua (Centre for Health Technologies, University of Technology, Australia)
- 내용 요약

빅 데이터는 사이즈가 큰 여러 개의 다차원 데이터 세트로 구성되어 있기 때문에 원래의 텐서(3차원 이상의 고차원 배열. 데이터가 커질 때 확장성이 떨어지는 등 기존 알고리즘 분석에 한계가 있기 때문에 하둡 등의 분산 시스템을 도입하여 데이터를 분석하게 됨)를 이용해 분석할 경우 문제점이 발생할 수 있음. 따라서 최근 하나의 큰 차원의 특징 공간으로부터 필요한 특색을 추출하기 위해 텐서를 분해하는 것에 대한 관심이 증가하고 있음. 발표에서는 대규모 텐서의 특색 추출을 위해 MPS(matrix product state) 분해법을 사용함. 특히 특징 공간의 규모를 효과적으로 감소시킴은 물론 선택적 추출법의 적용 없이도 정확한 분류를 구현할 수 있는 MPS로부터 구축된 단핵구조 텐서를 소개함.

3) A Workflow Model for Adaptive Analytics on Big Data

- 발표자: Verena Kantere (University of Geneva, Switzerland)
- 내용 요약

빅 데이터 분석은 전통적이거나 현대적인 데이터 저장소, 도식이나 형식이 여러 가지 다른 종류로 이루어진 데이터 소스, 다양한 쿼리 엔진에 의해 이루어짐. 이러한 데이터 분석을 수행하는 사용자들은 사업 분석가, 기술자, 일반 소비자등과 같이 다양함. 따라서 빅 데이터 분석은 사용자와 시스템에 맞추어진 방식으로 표현되고, 이행되어야 함. 발표에서는 적응 분석의 기본적인 요소들에 대해 논의함. 또한 현재 진행 중인 워크플로우 모델 구축에 대해 설명하고 적응 분석의 구축과 실행을 가능케 하는 작업흐름 관리 시스템에 대해 설명함. 실제 텔레커뮤니케이션 도메인 활용 사례를 통해, 작업이 의존적이지 않도록 작업흐름 모델의 기능적인 측면을 강화하고 실행과정에서 응용 프로그램 논리를 분리시키는데 초점을 맞춤.

4) Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty

- 발표자: Muhammad Raza Khan (University of Washington, USA)
- 내용 요약

서비스 제공자를 바꾸는 고객 행동에 대한 예측, 또는 서비스 사용을 중단할 가능성이 큰 고객을 파악하기 위한 작업은 많은 다양한 사업 영역에서 중요한 일이며, 기업의 수익에 영향을 미치는 일임. 기업들은 고객의 행동 또는 성향에 대한 대규모의 다양한 데이터를 수집하고 있기 때문에, 고객의 변심을 예측할 수 있는 새로운 방안을 고안해낼 수 있는 환경임. 발표에서는, 고객 변심 징후를 사전 경고할 수 있는 통합 분석프레임워크를 제안함. 이것은 사전 정의된 시간 안에 변심할 가능성이 큰 특정 고객에게 '고객 변심 지수'를 부여하는 방법을 사용하고 있음. 이 프레임워크는 설비 공학적 접근법인 강제 기법(컴퓨터의 힘을 빌려 문제를 강제로 푸는)을 도용하여 적절한 속성을 기준으로 최종 특징 집합을 결정하기 전에 감독 학습 알고리즘에 적합

한 특색들을 선택적으로 걸러냄. 연구에서는 대용량 모바일 폰 네트워크로부터 수집된 테라바이트 크기의 여러 데이터를 활용하여, 직관적으로 고객 변심의 징후를 판단하여 사전 경고를 함. 최적의 모델로 실험한 결과, 고객 변심 징후를 보였던 고객의 89.4%가 실제로 변심을 한 것으로 나타남.

5) Analytics of industrial operational data inspired by natural language processing

- 발표자: Mariusz Kamola (NASK-Research and Academic Computer Network, Poland)
- 내용 요약

산업 정보 처리를 통해 다양한 기간 동안 운용된 많은 데이터들이 모이게 됨. 이러한 데이터들은 잘 구조화되어 있고, 일상적으로 통제되며, 긴 기간에 걸쳐 관리됨. 발표에서는 자연 언어 처리가 완료된 데이터의 예비적 처리 과정에 대해 제안함. 일반적으로 적용될 수 있는 방법을 찾아내기 위해, 처리 과정이 조직적으로 관리될 수 있도록 설계하고 오류를 사전 탐색함으로써 제조 과정에서 필요한 정보를 예측함. 연구에서는 가스 송전 운용 데이터 사례를 검토함으로서 데이터의 예비적 처리 과정을 제안함.

6) Online Nonlinear Classification for High-Dimensional Data

- 발표자: N. Denizcan Vanli (Bilkent University, Turkey)
- 내용 요약

온라인의 이진 분류 문제에 대해 논의함. 역동적인 계층 모델에 기반하여 특색 공간을 분리하는 신개념의 무작위 추출 알고리즘을 소개함. 이러한 접근법은 공통적, 연속적으로 특색 공간을 구분 가능하도록 하고, 누적 손실을 최소화하기 위한 최적의 분류기를 제공하도록 함. 손실 기능을 최소화하기 위해 전반적인 계층 모델을 조정하였더라도, 이 알고리즘의 계산 복잡성이 특색 공간의 차원 수에 비례하여 연속적으로 증가한다는 문제가 있음. 그럼에도 이 알고리즘은 사전 정보 없이도 스트리밍 데이터, 상황에 따른 즉각적인 데이터 처리 및 폐기에 적용 가능하다는 장점이 있음. 다양한 실제 데이터 세트에서 이 알고리즘의 적용 가능성에 대해 논의함.

다. Session 3

o 일시 및 장소: 6월 27일(토), 오전 11:20-12:20, R3.11

- 사회자: Bo Hu (Kingdee International Software Group, China)

1) Cloud Tree: A Library to Extend Cloud Services for Trees

- 발표자: Yun Tian (Eastern Washington University, USA)
- 내용 요약

클라우드 사용자가 트리자료 구조의 생성 및 관리가 가능하도록 하는 클라우드 저장소에 대해 논의함. 이러한 콘셉트를 입증하기 위해, 신개념 클라우드 서비스인 'CloudTree'를 실험함. CloudTree를 활용하여, 사용자들은 빅 데이터를 직접 선택한 트리자료 구조로 구조할 수 있고, 이것은 클라우드에 저장됨. 수행 능력 강화를 위해 CloudTree 설계 및 실험에 캐시(고속 기억 장치), 프리패치(진행 중인 처리와 병

행하여 필요하다고 생각되는 명령 또는 데이터를 사전에 판독하는 것), 데이터 종합 기술을 활용함. CloudTree의 구성 요소인 이진 검색 트리(BST)와 프리픽스 트리 서비스를 실험해 보고, 아마존 클라우드를 사용하여 이들의 수행과정을 벤치마킹함. BST와 프리픽스 트리의 설계와 실험 과정에서 사용한 아이디어와 기술은 포괄적이기 때문에 B-트리와 같은 다른 유형의 트리 및 연계 목록 및 그래프와 같은 링크 기반의 데이터 구조에도 적용될 수 있음. 사전 실험 결과, CloudTree는 다양한 빅 데이터 응용프로그램에 유용하며 효율적이었음.

2) Matrix-based XML Stream Processing using a GPU

- 발표자: Soo-Hyung Kim (KAIST, South Korea)
- 내용 요약

GPGPU(GPU 상의 범용 계산: 일반적으로 컴퓨터 그래픽스를 위한 계산만 맡았던 그래픽 처리 장치(GPU)를 전통적으로 중앙 처리 장치(CPU)가 맡았던 응용 프로그램들의 계산에 사용하는 기술로 이를 통해 그래픽이 아닌 데이터에 스트림 프로세싱(제한된 형태의 병렬 처리를 응용 프로그램들이 쉽게 이용할 수 있도록 하는 컴퓨터 프로그래밍 양식)을 사용할 수 있게 됨) 컴퓨팅의 출현으로, 병렬 처리로 전환되었던 주 패러다임이, 다시 저렴한 비용으로 대용량 병렬 처리가 가능하도록 해주고 있음. 그러나 기존의 XML 스트림 프로세싱 알고리즘은, 단일 스레드 실행을 위해 만들어졌으므로 GPU가 갖고 있는 장점을 활용하지 못함. 본 연구를 통해, 다량의 XPath 쿼리가 이진 값의 매트릭스로 전환되고, XML 스트림 또한 두 개의 매트릭스 지수로 전환된 형태의 매트릭스 기반 XML 스트림 프로세싱 방법론을 활용하여 GPU 가속지원에 대해 논의함. 이를 통해 다량의 쿼리 프로세싱이 단순한 Boolean(개별 오브젝트를 하나의 지오메트리로 만들어 주는 과정) 처리로 전환되는 과정을 살펴봄. Xmark 벤치마크 데이터 세트로 실험한 결과, 기존의 알고리즘에 비해 본 연구에서 제안한 알고리즘이 8배 정도의 더 나은 작업 처리를 할 수 있음을 입증함.

3) A Flexible Data-driven Approach for Execution Trace Filtering

- 발표자: Kadjo Kouame (Polytechnique Montreal, Canada)
- 내용 요약

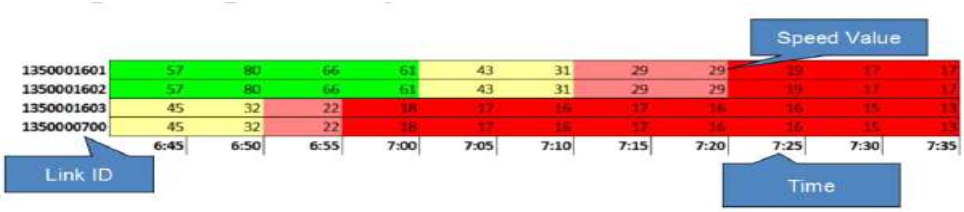
실행 추적은 흔히 시스템 실행 시간 양상을 파악하고 문제점을 파악하기 위해 사용됨. 그러나 실행 추적 시, 대용량 데이터는 이러한 분석을 복잡하게 할 수 있음. 또한 모든 사용자들이 항상 추적 과정의 모든 단계에 관심이 있는 것이 아님. 따라서 적절한 필터 접근법이 활용되어야 함. 필터링은 사이즈와 복잡성을 줄이면서 추적을 강화하기 위해 사용되며, 이를 통해 분석은 좀 더 쉬워짐. 발표에서 논의된 접근법은 선언형 XML에서의 일반적인 필터링 패턴을 정의하여, 가장 중요하면서도 관심이 있는 사건을 골라 분석할 수 있도록 해 주는 기법임. 필터링 시나리오에는 유한 상태 기계를 활용한 다양한 분석 패턴을 설명하기 위한 구문이 포함됨. 패턴은 아주 단순한 사건의 필터부터 복잡한 다차원 사건의 추출까지 실행 추적 데이터로부터 나타날 수 있는 다양한 유형의 종합적인 행동을 모두 포괄할 수 있도록 구성됨. 본 연구에서는 데이터 기반 필터링 접근법의 구체적인 방법에 대해 논의하고, LTTng 리눅스 핵심 추적장치를 사용해서 사례 연구를 실시함.

4) Road traffic analytic query processing based on a timeline modeling

- 발표자: Joonho Kwon (Pusan National University, South Korea)
- 내용 요약

수집되는 교통 데이터 증가로 인해 빅 데이터 기술을 이용해 도로 교통과 관련된 심층 분석이 가능한 상황임. 교통 관련 행동의 이해를 돕기 위해, 빅 데이터를 이용해 시간대별 교통 정보를 분석할 수 있음. 발표를 통해 교통 데이터의 시간대별 모델을 정립하고, 로데이터로부터 시간대별 데이터 구조를 설계하기 위한 알고리즘을 제안함. 또한 교통 분석을 위해 시간대별 정보를 활용한 쿼리 프로세싱 알고리즘을 제안함. 부산 컴퓨터 운송 시스템 센터로부터 수집한 실제 교통 데이터 세트에 알고리즘을 적용하여 효과성 분석을 시도함.

- 활용 영역: 교통
- 활용한 빅 데이터: 부산 ITS 센터로부터 획득한 도로 교통 데이터. 각 도로는 개별 ID가 부여되어 있고, 5분 단위로 시간대별 속도 정보를 수집함. 따라서 각 도로는 하루에 288개 셀의 속도 정보를 갖게 됨.



위의 그림에서 보면 도로 1350000700에 교통 혼잡이 시작된 시간은 7시 경으로 볼 수 있고, 이것이 얼마나 지속됐는지 그 시각 다른 도로의 상황은 어떠했는지 등의 정보를 얻을 수 있음. 또한 교통 혼잡도 단독으로 나타나는 현상인지, 다른 도로의 교통 혼잡과 연계되어 나타나는 것인지 등에 대한 분석이 가능함. 도로별 개별 아이디를 통해 교통 혼잡이 나타나는 도로의 총 길이(km) 파악도 가능함.

5) Approximate querying in big heterogeneous data

- 발표자: Verena Kantere (University of Geneva, Switzerland)
- 내용 요약

이질적인 환경에서 쿼리 재기록의 기본 가정은 재기록을 위해 사용하는 매핑은 복잡하다는 것임. 예컨대, 쿼리에서 언급되는 모든 관계와 속성은 매핑 전반에 관련되어 있고, 소스의 스키마 정보와의 연관성과 속성이 쿼리로 재기록 되는 것임. 실제로, 기존 소스들 간의 이러한 복잡한 매핑 구조는 드물게 나타나고 부분 매핑이 나타나는 게 일반적임. 그러므로 현실적으로 기존 쿼리 응답 알고리즘은 주요 사례에서 재기록을 생성하는데 실패함. 이러한 현상은 다양한 이질적인 데이터 소스로부터의 쿼리 응답이 꼭 필요한 새로운 빅 데이터 시대에 심각한 문제가 되고 있음. 따라서 어떻게 하면 불충분한 매핑 때문에 재기록을 할 수 없는 쿼리를 계산해 낼 수 있는지, 어떻게 관심이 있는 쿼리를 대략적으로라도 찾아낼 수 있는지와 관련되어 있음. 본 발표에서는 현재 연구 진행 중인 기술적인 방안으로, 입력 쿼리로부터, 이질적인 빅 데이터 소스를 자동적으로 통합하기 위한 환경을 평가하고, 재기록 할 수 있도록 쿼리 근

사치를 계산하는 방법에 대해 소개함. 쿼리 재기록의 전통적인 방법을 확장하여, 이러한 근사치를 계산하고 평가할 수 있는 휴리스틱 알고리즘을 소개함.

라. Session 4

o 일시 및 장소: 6월 27일(토), 오후 1:30-2:30, R3.11

- 사회자: Durga Toshniwal (Indian Institute of Technology Roorkee, India)

1) Queriosity: Automated Data Exploration

- 발표자: Abdul Wasay (Harvard University, USA)
- 내용 요약

점차 데이터가 중심이 되는 세상에서, 데이터 탐색은 과학, 사업, 개인 영역에서의 정보의 집약 과정과 같음. 그러나 데이터 탐색보다는 데이터 검색에 가깝게 설계된 현대 데이터 시스템은 데이터 검색만을 가능하게 함. 발표에서는 자동화, 개인 맞춤형 데이터 탐색 시스템인 Queriosity에 대해 제시함. 이 시스템은 데이터 탐색의 패러다임 이동을 구상하고 있으며, 단순한 데이터 검색 대신 데이터 세트 중 사용자가 관심있어 하는 것을 바로 제공할 수 있는 개인 맞춤형 '데이터 로봇'을 구축하는 것을 목표로 함. Queriosity는 트렌드, 통계적 특성, 사용자와의 상호작용에 기반하여 관심있는 정보를 자동적, 지속적으로 탐색할 수 있도록 함.

2) Write Optimization using Asynchronous Update on Out-of-Core Column-Store Databases in Map-Reduce

- 발표자: Feng Yu (Youngstown State University, USA)
- 내용 요약

칼럼 스토어(열 단위로 데이터를 저장하는 방법) 데이터베이스는 기존에 사용하던 행 단위 데이터베이스에 비해 데이터를 읽는 속도를 빠르게 해준다는 특징이 있음. 그러나 칼럼 스토어 데이터베이스의 쓰기 작업을 최적화 방법은 잘 알려진 문제점 중 하나임. 쓰기 성능 최적화와 관련된 대부분의 기존 작업들은 주기억 장치 칼럼 스토어 데이터베이스에 초점이 맞추어져 있음. 본 발표에서는 맵 리듀스 환경에서의 칼럼 스토어 데이터베이스로 연구를 확장하여, TBAT(Timestamped Binary Association Table)라고 부르는 데이터 저장 형식을 제안함. AMO 업데이트(Asynchronous Map-Only Update)라고 부르는 TBAT 기반의 새로운 업데이트 방법이 기존의 업데이트 방식을 대체할 수 있도록 설계됨. 기존의 업데이트 방법과 비교할 때 AMO 업데이트의 처리 속도가 많이 향상됨을 실험을 통해 입증함.

3) Social network analysis of developers and users mailing lists of some free open source software

- 발표자: Armel Jacques (Universite de Yaounde I, Cameroun)
- 내용 요약

Kevin Crowston 외의 최근 저술에 보고된 바와 같이, 오픈 소스 소프트웨어는 백만여 명의 프로그래머, 수많은 소프트웨어 개발사, 수백만 사용자가 관련된 매우 중요한 사회 현상이고, 또한 수백억 유로 가치로 추산되는 무료 소프트웨어 창조와 관련되는

등 그 재정적 영향력 또한 매우 큰 것임. 무류 오픈 소스 소프트웨어 프로젝트는 일반적으로 일련의 개발자와 사용자 메일링 리스트를 포함하고 있음. 이 많은 수의 메일링 리스트는 지속적으로 변화하고 있으며, 회원이나 관련 주제에 대한 다양성을 포괄하고 있음. 이러한 특성은 이들에 대한 관리를 어렵게 만드는 요소임. 이와 같은 빅 데이터 문제를 극복하기 위한 한 가지 방법은 쉽게 산출할 수 있는 글로벌 지수를 통해 발견하는 것임. 본 발표에서는, 4개 무료 오픈 소스 소프트웨어 프로젝트(CentOS, GnuPG, Mailman, Samba)의 개발자, 사용자 메일링 리스트를 비교하여 소셜 네트워크 분석을 시도함. 이들 메일링 리스트가 어떤 공통된 특성을 갖고 있음을 증명하고자 함. 메시지 수, 지속시간, 연결 시간 등에 대한 분석을 시도함. 이번 연구의 분석 경향을 검토함으로서 향후 모니터링 메일링 리스트의 구성 요소를 설계하는데 참고할 수 있을 것으로 보임.

4) A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers: An Application Platform for Public Health Research

- 발표자: Kelvin KF Tsoi (The Chinese University of Hong Kong, China)
- 내용 요약

보건 데이터 관리를 위한 클라우드 프레임워크 활용 기술과 분석방법은 공공 보건 연구 분야의 범위를 넓히고 있음. 흡연은 중독성 있는 행위이며, 심장마비, 폐암 등 다양한 질병으로 인한 사망 위험을 증가시킴. 최근 선진국에서 전자담배 사용이 증가하고 있고, 금연을 위한 효과적인 방법으로 추천되고 있음. 그러나 흡연 행위와 전자담배 간의 상관관계를 밝히는 연구는 충분치 않음. 본 연구에서는 독창적인 클라우드 기반의 장비를 사용하여 흡연 행위와 전자담배 사이의 관계를 파악하기 위한 데이터를 수집함. 즉, 흡연자의 전자담배 사용 데이터가 모바일 인터넷, 그리고 스마트폰과 전자담배를 연결하는 블루투스를 통해 클라우드에 매일 업로드 됨. 개인 정보 보호를 위해, 모든 개인 정보는 암호화되고 대신 연구 번호가 개개인에게 부여됨. 클라우드의 원격 플랫폼이 빠르게 축적되는 대규모 데이터를 효율적으로 분석 처리 할 수 있음. 흡연 행위와 관련된 데이터 마이닝(대규모 자료를 토대로 새로운 정보를 찾아내는 것)은 전자담배 사용 방식에 대한 이해를 높임. 이러한 데이터 장비는 공공 보건 영역에서 다른 종류의 역학 연구에 활용될 수 있음.

- 활용 영역: 흡연
- 활용한 빅 데이터: 참여한 흡연자들의 동의하에 질병, 흡연력, 니코틴 의존도 등의 정보를 클라우드에 기록. 참여자들에게 담배 대신 전자담배를 제공. 이 전자담배를 사용하는 빈도가 블루투스를 통해 자동적으로 스마트폰으로 전송됨.

<전자담배 설계 구조>

모든 정보는 주기적으로 클라우드 플랫폼(마이크로소프트 애저)으로 업로드 됨. 흡

연 습관 및 침여자의 기존 병력과의 관계, 금연 시도 등에 대해 파악할 수 있고, 다른 공공 보건 관련 임상 연구에도 이러한 플랫폼을 적용할 수 있음.

<클라우드 기반 데이터 수집 체계도>

마. Session 5

o 일시 및 장소: 6월 27일(토), 오후 2:40-3:40, R3.11
- 사회자: Yuhong Yan (Concordia University, Canada)

1) How has Twitter changed the Event Discussion Scenario?: A Spatio-Temporal Diffusion Analysis
- 발표자: Purva Pruthi (Indian Institute of Technology Roorkee, India)
- 내용 요약

초기 전통적인 인쇄 매체 시기 동안에는, 제한된 시간과 거리를 가지는 지리적 경계에 의에 한정될 수밖에 없는 일방적인 정보 전파가 이루어짐. 온라인 소셜 미디어가 출현하면서, 정보 전파 과정에도 큰 변화가 생김. 온라인 소셜 미디어는 큰 인기를 모으며 가장 빠른 의사소통 수단이 되었고, 페이스북, 트위터와 같은 온라인 소셜 네트워크는 지리학적인 한계를 넘어 세계적인 차원에서 자기 자신을 표현할 수 있도록 개개인에게 플랫폼을 제공함으로써 대인 커뮤니케이션 방법을 진화시킴. 이 영역의 주요 연구들은 일반적인 정보 전파 현상을 분석하는 것에 초점이 맞추어져 있음. 본 연구에서는 장소와 시간과 관련하여 트위터에서 논의된 특정한 실제 사건을 분석함으로써, 정보 파급의 역동성에 대해 연구함. 개별 사건을 다음과 같은 기준에 의해 범주화 함. 일시적(단기, 장기), 지리-공간적 분포(지역적, 세계적), 정보 전달 메카니즘(파급(퍼짐), 점진적), 영향력(인기 있는, 인기 없는), 원인(자연적인, 계획적인). 일시적 분석은 사건 전, 사건 중, 사건 후에 트윗의 빈도가 자연적인 사건과 관련해서 차이가 있음. 예컨대, 계획적인 사건 중 '델리 선거' 같은 것은, 사건 발생 중에만 활발한 논의가 진행되는 '오바마의 인도 방문'과 달리, 실제로 사건이 발생한 후에 논의가 활발함. 지리-공간적 분석을 통해서는, 지역적 차원에서 발생하는 사건들은 지

역적 경계를 넘어서 세계적 차원으로 논의가 확산됨. 또한 리트윗이나 답글에 의해 구조화된 사용자 간 상호 교류 그래프를 활용하여, 사건 전파 유형에 대해서도 연구함. 사용자들 간의 관계를 증폭시키는 실제 사건의 공간적-일시적 전파 역동성에 대해 3차원적 분석을 함.

2) Discovering Environmental Impacts on Public Health using Heterogeneous Big Sensory Data

- 발표자: Minh-Son Dao (National Institute of Information and Communications Technology, Japan)
- 내용 요약

이기종 센서(heterogeneous sensors)로부터의 데이터 스트리밍 트렌드를 요약하여 보건관련 사건 탐지 방법론에 대해 살펴봄. 이러한 방법론 저변의 주요한 아이디어 실시간 사건을 탐지하기 위함이며, 신체적, 사회적 센서 데이터들의 공간성-일시성-주제 간 상관관계를 활용하여 그 사건들에 대해 이해 가능하도록 설명하기 위함임. 방법론의 교육훈련 단계는 긍정, 부정적인 표본세트를 구성하기 위해 특성 벡터가 자동적으로 할당된 라벨을 포함한 논스톱 프로세스로 구성됨. 그런 다음, 사건 모델은 정확성을 안정적으로 극대화하기 위해 감독학습 접근법의 활용함으로서 발생됨. 천식 발작에 있어서 환경의 영향과 같은 문제는 제안된 방법론의 평가를 위해 활용됨. 실험 결과, 제안된 방법론이 특정한 공간-시간적 차원에서 천식의 전파 위험을 높은 정확성을 갖고 감지해낼 수 있음을 입증함.

3) Detecting Corporate Social Media Crises on Facebook using Social Set Analysis

- 발표자: Raghava Rao Mukkamala (Copenhagen Business School, Denmark)
- 내용 요약

소셜 미디어의 특징인 정보의 빠른 확산 속도로 인해 소셜 미디어 사용자들이나 소비자들에게 브랜드 이미지가 부정적으로 지속되는 현상은 기업이 직면한 큰 어려움 중 하나임. 발표에서는 사회 집합 분석(사회학과 집합 이론을 조합하여 컴퓨터 기반의 사회적 분석)을 통해 소셜 미디어로 인한 이러한 기업의 위기를 방지하기 위한 기술을 제안함. 소셜 데이터의 개념적, 형식적 모델에 근거하여, 4개의 덴마크 회사에 대한 페이스북 월 데이터에 대한 사회 집합 분석을 실시함. 조사 결과, 대용량이지만 일시적인 소셜 미디어의 특징을 살펴볼 수 있었고, 사용자 행동 패턴이 종합적으로 나타남을 보여줌. 상품 프로모션이나 선거 운동 등의 다른 종류의 사건을 분석하기 위해서도 본 연구에서 개발한 위기 대응 기술을 적용할 수 있을 것으로 봄.

4) Data Quality and Energy Management Tradeoffs in Sensor Service Clouds

- 발표자: Victor Lawson (Georgia Gwinnett College, USA)
- 내용 요약

클라우드 기반의 센서 데이터 수집 서비스는 사물인터넷(IoT)의 주요 분야로 자리 잡고 있음. 이러한 서비스 영역의 소비자 수요가 증가함에 따라, 스트림의 데이터 질(DQ)에 대한 문제 또한 그 중요성이 증가하고 있음. 데이터 질과 센서 사용에 따른 에너지 소비 간에 내재해 있는 상호 균형의 문제도 특별한 관심 영역임. 그동안 데이

터 사용자들로 하여금 에너지를 보존하면서 동시에 양질의 데이터를 받을 수 있도록 하는 이 둘 사이의 상호 균형의 관리와 관련된 연구는 충분치 못한 실정이었음. 본 연구에서는, 데이터 스트림 소비자를 위한 데이터 질 서비스를 데이터 공급 생산자를 위한 맞춤형 에너지 효율인 'EE' 조절 알고리즘과 조합함으로써 상호 균형의 관계를 보다 자세하게 탐색함. 이러한 에너지 관리 서비스는 데이터 질과 에너지 효율 중 어떤 것을 선택해야 할지 고민하는 소비자들에게 비용을 절감할 수 있도록 해줌. 연구를 통해 상호 균형을 관리하기 위한 클라우드 기반 서비스를 개발하였고, 공급자/수요자의 데이터 스트림과 데이터 질을 조절하는 최선의 매칭 클라우드 서비스 아키텍처를 개발함. 이러한 연구 결과가 향후 소비자가 에너지 효율을 위한 선택을 하는데 기여할 수 있음.

4. 심층조사 연구

가. Session 3 : Bigdata and Social Network

o 일시 및 장소: 6월 30일(화), 오후 1:00-2:00, R3.11
- 사회자: Vladimir Hahanov (Kharkov National University of Radioelectronics, Ukraine)

1) A Clustered Approach for Fast Computation of Betweenness Centrality in Social Networks
- 발표자: Eugenio Zimeo (University of Sannio, Italy)
- 내용 요약

지난 몇 년 간, 소셜 네트워킹 시스템을 통해 구축된 데이터는 지역 및 글로벌 사회 현상을 분석하는 데에 활용됨. 영향력 있는 사람이나 여론 주도자들을 식별하기 위한 측정방법은 매개 중심성 지수(betweenness centrality index)임. 비가중 그래프를 위한 O(nm) 시간 복잡도를 증명해야 하기 때문에 이 지수를 계산하는 것은 매우 힘든 작업임. 소셜 네트워크 데이터가 계속적으로 증가하고, 오늘날 관련 그래프의 노드가 수억 개 정도인 것을 고려해 본다면 이러한 복잡성은 계산 시간에 큰 영향을 미침. 이러한 문제를 해결하기 위해, 광범위한 네트워크에서 근사치의 중심성 지수를 빠른 시간 안에 계산할 수 있는 Brandes 알고리즘의 변형을 제안함. 구체적으로, 예비 단계에서 류벤(Louvain) 방법에 기초하여 확장 가능하고 효율적인 클러스터링 기술을 활용함. 이를 통해 제한된 수의 중추 노드의 선택을 가이드 하는 노드의 경계와 노드 커뮤니티를 구분할 수 있도록 해줌. 이번 실험을 통한 분석결과, (소셜 네트워크 그래프의 광범위한 등급을 나타내는) 제안 방법이 기존의 가장 효율적인 솔루션과 비교하더라도 계산 시간을 대폭 단축시킴을 보여줌. 반면, 이 근사치는 상단 k의 매개 값(top k betweenness values)에 집중할수록 정확해짐. 종합 무척도 그래프에 대한 확장성 분석을 통해서도, 연구에서 제안한 방법이 거대한 네트워크(이러한 경우에 추가적인 기계가 메모리 소비를 처리하는 데에 유용하더라도)에서 잘 작동한다는 것이 증명됨.

나. Session 4 : Bigdata and Social Network

o 일시 및 장소: 6월 30일(화), 오후 3:30-4:30, R3.11
- 사회자: Peter Chen (Carnegie Mellon University, USA)

1) Can we Rank Emotions? A Brand Love Ranking System for Emotional Terms
- 발표자: Eleanna Kafeza (Business School, Athens University of Economics and Business, Greece)
- 내용 요약

발표를 통해, 소셜 미디어에서 추출된 콘텐츠를 활용하여 브랜드에 대한 고객의 감정적인 애착을 검토함. 보다 구체적인 방법으로, 사용자의 트위터 게시물에서 용어의 형태로 나타나는 브랜드 선호에 관련된 감정을 분석해 냄. 기존의 브랜드 선호에 대한 일곱 개의 관점을 활용하여, 주제 식별 방법을 이용한 확률 네트워크 방안을 활용하여 브랜드 선호를 분석하는 방법을 제안함. 브랜드의 동의어와 사용자의 감정을 나타내는 용어를 찾아 순위를 매김. 이를 통해 사용자의 행동을 묘사하는 Twitter Behavior Metric을 소개하고 사용자 행동에 브랜드 선호와의 연관성을 분석함. 트위터를 통해 샘플링한 특정 브랜드명을 추론 네트워크 정보를 활용하여 사용자 행동 척도와 연계해 봄으로써 제안한 방법론에 대해 실험함.

2) Deriving Topics in Twitter by Exploiting Tweet Interactions
- 발표자: Robertus Nugroho (Macquarie University, Australia)
- 내용 요약

빅 데이터 소셜 네트워크로서의 트위터는 전 세계에서 일어나는 최신 유행의 사건을 수집하는 데에 가장 중요한 소스 중의 하나임. 트위터에서의 주제는 상황인식, 시장분석, 콘텐츠 필터링, 건의사항과 같은 다양한 응용 프로그램을 위한 중요한 요소임. 그러나 트위터의 주제는 짧은 메시지들로 구성됨. 현재의 빅 데이터 분석 방법은 트윗 콘텐츠의 다양한 의미 분석을 활용하고 있지만, 트윗 간의 상호작용 분석에 대해서는 간과하고 있음. 발표에서는 트윗을 보낸 사람들 상호간의 행동으로 정의한, 트윗 간의 상호작용, 행위, 트윗 콘텐츠를 모두 고려하는 새로운 주제어원 분석 방법에 대해 제안함. 구체적으로, 트윗에서의 주제는 트윗의 상호작용과 의미 분석에 2단계의 행렬 인수분해 알고리즘을 수행함으로써 도출함. 일정 기간 동안 수집된 트윗을 가지고 다수의 실험을 수행한 결과, 기존에 제안된 방법보다 나은 결과가 도출됨을 확인함. 또한 이 방법을 활용함으로써, 트윗 상호 연계 분석 시, 트위터가 가진 짧은 텍스트로 인한 한계가 완화된다는 것도 증명함.

3) Matrix inter-joint Factorization - A New Approach for Topic Derivation in Twitter
- 발표자: Robertus Nugroho (Macquarie University, Australia)
- 내용 요약

현재 주로 사용되는 모든 소셜 미디어 플랫폼 중, 트위터는 신속한 실시간 통신을 위한 주요 플랫폼으로서의 역할을 하고 있음. 특히, 트위터를 모니터링하고 대화의 주제를 이해하는 것에 관심이 높음. 그러나 기존의 방법이 대부분 내용의 특성만을 사

용하는 것과 달리 트윗의 짧은 내용은 주제 유도를 힘들게 만들기 때문에 제한된 상호작용 정보들을 통합시켜야 함. 발표에서는 NMijF(Non-negative Matrix inter-joint Factorization)이라는 새로운 방법을 제안함. 이 알고리즘은 단일한 반복갱신 과정 내에서 트위터 상호작용 특성과 내용의 결과물을 공동으로 인수분해 하는 방법임. 실제 트위터 데이터세트를 활용하여 포괄적인 실험을 통해, joint-NMF(Joint Non-negative Matrix Factorization), NMcF(Non-negative Matrix co-Factorization)를 비교함으로써, 제안한 방법의 효과를 평가함. 그 결과, 제안된 NMijF 방법이 주제 일관성, 순도, 정규상호 정보, 재현율 및 정확률 면에서 joint-NMF, NMcF, 그리고 여타 고급 주제 유도 방법을 능가하는 효과성이 있음을 입증함.

다. Session 5 : Privacy

o 일시 및 장소: 7월 1일(수), 오전 8:15-9:15, R3.11

- 사회자: HyeJung Moon (Seoul National University of Science and Technology, South Korea)

1) Privacy Preserving Data Analysis in Mental Health Research

- 발표자: Jingquan Li (Texas A&M University-San Antonio, USA)
- 내용 요약

정신건강 기록과 심리치료 노트의 디지털화는 환자, 정신과 의사, 연구원, 통계학자, 데이터 과학자 등 다양한 사용자들의 정신건강 데이터에 대한 접근을 용이하게 해줌. 그러나 매우 민감한 정신기록에 대한 접근 증가는 정신병 환자들의 프라이버시와 개인 정보를 위협하고 있음. 따라서 정신건강 연구 과정에서의 개인 프라이버시 침해에 대한 문제를 검토하고, 이러한 문제를 해결하기 위해 개인정보 보호 데이터 분석 방법을 개발하고자 함. 발표에서는 정신건강 기록과 심리치료 노트의 사용에 적용되고 있는 기존의 개인정보 보호 방법에 대한 부적절성을 검토하고, 현재 접근이 허락된 정신건강 기록을 활용하여, 정신질환을 가진 사람들의 개인정보를 보호할 수 있는 프라이버시 보호 데이터 분석 방법을 개발함. 나아가, 발표에서 제안한 접근법을 사용하여 시범 연구를 수행함.

2) Enabling Privacy Mechanisms in Apache Storm

- 발표자: Jingquan Li (Texas A&M University-San Antonio, USA)
- 내용 요약

실시간 계산 시스템을 전송하는 데이터를 분석하는 것은 통신 네트워크를 동적으로 최적화시키는 경우에 유용함. 빅 데이터와 같은 많은 양의 데이터의 분석을 위해서 이러한 분석 방법이 사용됨. 빅 데이터 분석은 데이터 분석에 필요한 핵심 응용프로그램뿐만 아니라 다른 응용프로그램들까지 사용하게 되므로 데이터에 포함된 개인정보를 포함한 사생활 침해 위험이 제기됨. 만약 데이터에 개인정보가 포함되어 있다면, 그 경우의 목표는 데이터 접근을 제어하는 것이며, 이는 데이터 접근에 조건들을 부과함으로써 이루어짐. 발표에서는 아파치 스톰과 같은 실시간 계산 시스템의 데이터 접근 과정에 제어 조건을 부과함으로써 개인 정보 보호정책 프레임워크에 기여함.

3) Sensitive Disclosures under Differential Privacy Guarantees

- 발표자: Chao Han (SFU, Canada)
- 내용 요약

NIR(Non-independent reasoning, 비독립적 사유)는, 데이터 내 레코드들이 동일한 기본 분포를 공유한다는 가정 하에, 다른 레코드들로부터 하나의 레코드 정보를 습득하는 것을 뜻함. 정확한 NIR은 개인의 사생활 정보가 드러나게 할 수 있음에도 이것이 개인정보 보호 위반이 아닌 것으로 간주된다는 문제가 있음. 본 연구에서는 개인정보 보호를 만족시키는 쿼리에 의해 NIR을 통해 공개될 가능성이 있는 개인 정보에는 어떤 것이 있는지 조사해 봄. 이 작업을 위해, 먼저, 무작위 쿼리에 의한 NIR에서 '공개'의 의미를 규정하였고, 차등적인 개인정보 쿼리에 의한 개인정보 공개의 과정을 시현함. 실제 데이터 세트를 통해 실험한 결과, 개인정보 보호를 위해 제한된 설정을 사용할 경우, NIR을 통한 개인정보 공개가 덜 발생하는 반면, 이것이 차등적인 개인정보 쿼리의 효율성에는 부정적인 영향을 줄 수 있음을 입증함. NIR을 통한 개인정보 공개가 사생활 침해라는 가정 하에, 분석 쿼리의 유용성과 개인정보 보호가 부정적인 상관관계를 갖기 때문에 차등적인 개인정보 쿼리를 사용하는 것이 유용하지 않음을 입증함.

라. Session 6 : Big data and Learning

o 일시 및 장소: 7월 1일(수), 오전 9:25-10:25, R3.11

- 사회자: Jingquan Li (Texas A&M University-San Antonio, USA)

1) Red-RF: Reduced Random Forest for big data using priority voting & dynamic data reduction

- 발표자: Hussein Mohsen (Indiana University, USA)
- 내용 요약

렌덤 포레스트(RF)는 분류를 위한 효과적인 모델로 사용되고 있음. 발표에서는 Red-RF라 불리는 랜덤 포레스트의 새로운 유형을 제안함. Red-RF는 새로운 동적 데이터 정리 원리와 Breiman의 RF 보다 정확성, 실행시간, AUC 값을 개선한 형태인 PV(Priority Vote Weighting)라 불리는 새로운 voting 매커니즘을 소개함. 또한, Red-RF는 트리들 간 눈에 띄는 상관관계 증가 없이도 랜덤 포레스트를 강화시키는 RF의 강점을 그대로 적용하고 있음. 8번의 실험을 통해, Red-RF와 Breiman의 RF의 수행 과정을 비교함으로써, 각기 다른 크기의 데이터 세트를 구분하는데 있어서의 문제점을 확인함. 또한, 각각 백만 점으로 구성된 빅 데이터를 사용하여 2번의 추가 실험을 수행함.

2) Optimal and Efficient Distributed Online Learning for Big Data

- 발표자: Muhammed O.Sayin (Bilkent University, Turkey)
- 내용 요약

빅 데이터 응용프로그램을 위한 최적의 효율적 분산 온라인 학습 전략을 제안함. 자

율적 데이터 소스의 분산 제어를 통한 학습률 향상 방안을 제안함. 이처럼 서비스 컴퓨팅에서의 기계를 통한 학습법은 학습률 향상과 관련하여 많은 주목을 받고 있으며, 광범위한 연구가 행해졌지만, 본 연구에서는 특히 오라클 알고리즘을 적용한 최적의 온라인 학습 전략을 제안하고 있음. 이에 더하여 통신 부하를 줄일 수 있는 최적의 효율적인 온라인 학습 알고리즘을 고안하였음. 연구에서 제안한 학습 전략이 현 기술적 수준에서 상당한 성능을 보여주고 있음을 입증함.

3) Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection

- 발표자: Hong Zhao (Xiamen University, China)
- 내용 요약

유전자 수준 데이터 추출 방법으로, 정상, 비정상 성분을 포함한 막대한 양의 유전자 발현 데이터가 존재함. 암 조직과 같이 빠르고 정확하게 대응 패턴을 검출하기 위한 방법으로서, 유전자 발현 데이터 마이닝은 시급한 연구 과제임. 유전자 기술 발전과 함께 유전자 발현 데이터 분류 문제가 광범위하게 연구되어 온 후로, 신경망과 관련된 훨씬 많은 방법들이 양질이면서 소량인 의료 데이터 분석 분야에서 개발되어 옴. 고도의 학습기계 등 클러스터링 접근법에 관한 많은 연구들이 수행됨. 이러한 연구들은 주로 저층 신경망 모델에 적용됨. 최근 심층적인 학습은 고차원 데이터 세트를 다루는데 최적의 효과와 수행능력을 보여줌. 최근의 일반적인 심층 신경망과 달리, 본 연구에서는 계속적으로 저층 신경망에 적용할 것이지만, 동시에 저층 신경망을 위한 혁신적인 알고리즘을 개발할 계획임. 관리 모델로서, 일정한도 내에서 저층 신경망 모델을 구현하고, 원활한 처리를 통해 최종적으로 더 나은 결과를 도출할 수 있도록 계산 과정을 좁혀나감. 실험 결과, 이러한 모델이 심층 신경망과 비교했을 때, 더 안정적이며 우수한 결과를 보여줌을 입증함. 알고리즘 분석 결과 또한 제시함.

마. Session 11 : Big data Platform/framework

o 일시 및 장소: 7월 2일(목), 오전 10:55-11:55, R3.11

- 사회자: Chen (Cherie) Ding (Ryerson University, Canada)

1) Hybrid Traffic Speed Modeling and Prediction Using Real-world Data

- 발표자: Rong Zhang (Zhejiang University, China)
- 내용 요약

차량 속도 모델링과 예측은 교통 정체와 이에 따른 사회 · 경제적 영향의 증대로 개인 및 정부 정책 차원에서 중요한 의미를 가짐. 그간 차량 속도를 예측하는 방법은 다양하게 제시되어옴. 그러나 장기 예측의 정확성은 만족스럽지 못한 수준임. 특히, 기상이변과 같은 사건이 발생할 때, 예측의 정확성은 많이 빗나감. 연구에서는 중국 항저우에서 수집된 택시 GPS 기록 880,000개와 기후조건, 휴일 데이터 세트를 바탕으로, 차량 속도의 multi-time-scale 상관관계 및 다양한 관련 사건의 결과를 도출함. 과거의 multi-time-scale 차량 속도 데이터뿐만 아니라, 입력되어 있는 관련 사건까지 모두 이용한 하이브리드 차량 속도 모델링과 예측의 틀을 제안함. 항저우의 주요 도로

의 모든 부분에 대하여, 반복되는 모델 식별 알고리즘을 통해 해당 차량의 속도 모델을 설정하고, 대규모의 추적 기반 시뮬레이션을 통해 다양한 조건에서 제안한 접근방법의 효과를 입증함.

바. Session 12 : Analsysis on BigData Research and Platforms

o 일시 및 장소: 7월 2일(목), 오후 1:00-2:00, R3.11
- 사회자: Sathish A.P. Kumar (Coastal Carolina University, USA)

1) Big Data Open Source Platforms
- 발표자: Pedro Daniel Coimbra de Almeida (ISEC - Coimbra Institute of Engineering, Portugal)
- 내용 요약

글로벌 시장에서 사용자 데이터를 찾아내 분석하는 능력은, 기업이 사용자의 요구에 시간과 정확도의 측면에서 가장 가까이 근접할 수 있는 한 가지 방법임. 빅 데이터 플랫폼은 이러한 측면에서 필요한 과제를 해결하고자 하는 기업에게 있어 중요한 해결책 중 하나임. 그러나 불행하게도, 다수의 다른 해결책들과 함께 해결이 필요했던 과제들은 하나의 플랫폼으로 확정하기도 어렵고 적절치도 않은 또 다른 다수의 플랫폼의 생성으로 이어지기도 함. 이 연구에서는 기업이나 단체가 그들의 요구에 가장 적절한 것을 선택하는 데에 도움이 되는 여섯 개의 가장 중요한 빅 데이터 오픈 소스 플랫폼을 비교함. 특히 Apache Mahout, MOA, R Project, Vowpal Wabbit, PEGASUS, GraphLab CreateTM와 같은 오픈 소스 플랫폼을 분석함.

5. 응용연구

가. Session 1 : Big Data & Health

o 일시 및 장소: 6월 30일(화), 오전 8:15-9:15, R3.02/3.03
- 사회자: Jeffrey Tsai (Asia University, Taiwan)

1) Big Data Analytics Framework for System Health Monitoring
- 발표자: Brian Xu (Honeywell Aerospace, USA)
- 내용 요약

본 연구에서는 보조 동력 장치(APU) 상태 모니터링 서비스의 질과 수행의 향상을 위해, 테스트를 거친 빅 데이터 분석 틀에 기반을 둔 Machine Learning(ML)을 제시함. 항공우주산업 응용을 위한 실용적이고 유용한 빅 데이터 분석 기술을 개발하고 지원하고자 하는 목적에서 시작됨. 실험을 통해 다양한 데이터 소스를 분석하고, 예측능력을 증가시키는 ML알고리즘 개발하고 사용하는데 기여함. (1) APU 39%에서 56% 손상 (2) APU 19%에서 60% 종료. 이러한 상태 모니터링 시스템은 현재 널리 사용되는 상태 기준 보수(CBM)와 통합 가능함. 사용자들은 이 클라우드 기반 분석 도구세트를 사용할 수 있으며, 언제 어디서나 어떠한 기기(PC, 태블릿, 스마트 폰)를 사용하더라

도 빅 테이터에 접속할 수 있도록 함.

2) Predictive Modeling for Comfortable Death Outcome using Electronic Health Records
- 발표자: Muhammad Kamran Lodhi (University of Illinois, USA)
- 내용 요약

전자 의무 기록(EHR)시스템은 환자의 치료 경과를 관찰하는 의료 산업 분야에서 사용됨. 데이터의 빠른 성장과 함께, EHR 데이터 분석은 빅 테이터 문제로 대두됨. 대부분의 EHR들은 엉성하고, 다차원적인 데이터 세트이며, 그것들을 수집한다는 것은 여러 가지 이유로 인해 힘든 작업임. 본 연구에서는 어떠한 요인이 죽어가는 환자의 죽음에 대한 공포와 같은 심리적인 문제에 영향을 미치는지를 결정하는 예측 모델을 구축하기 위해 간호 EHR 시스템을 사용함. 기존의 상이한 모델링 기술들은 환자 결과를 예측하기 위해 큰 단위의 모델뿐만 아니라 세밀한 모델을 개발하는 데에도 사용되어져 왔음. 대단위 모델은 각 입원기간 말의 결과를 예측하는데 도움이 되는 반면, 세밀한 모델은 각 교대시간 말에 결과를 예측하는 것을 도움으로서 예측 결과의 궤적을 제공함. 다른 모델링 기술을 바탕으로 비교적 오류가 없는 데이터를 구성함으로서, 연구 결과들이 상당히 정확한 예측을 보여줌. 이러한 모델은 효과적인 치료를 결정하고, 의료비용을 절감하며, end-of-life(EOL) 진료의 질을 향상시키는데 기여함.

3) H-DRIVE: A Big Health Data Analytics Platform for Evidence-Based Decision Making
- 발표자: Ashraf Abusharekh (Dalhousie University, Canada)
- 내용 요약:

의료 수술은 대량의 데이터를 생성함. 빅 데이터 분석 방법은 환자의 치료를 개선시키기 위해 '대량'의 의료 데이터로부터 결정 및 실행 가능한 '정보'를 유도하는데 필요함. 대량 의료 데이터 분석을 위한 기술적 도전과제들을 감안하여, 본 연구에서는 전문 의료 분석 플랫폼인 H-DRIVE(의료 데이터 조정 추론 및 시각화 환경)를 제시함. H-DRIVE는 분석실험을 설계하는 데이터 분석가들과 연구자들에게 권한을 부여하고, 그들이 의료 데이터에 대한 복잡한 분석을 수행하도록 설계된 통합적이고 철저한 의료 데이터 분석 서비스를 지향하는 워크 벤치임. H-DRIVE의 높은 기능 수준 및 기술 아키텍처를 제시함. 연구 사례로서, 주립 병리학 실험실의 작업을 최적화하고자 H-DRIVE의 애플리케이션을 활용함.

나. Session 2 : BigData & Network Management

o 일시 및 장소: 6월 30일(화), 오전 9:25-10:25, R3.02/3.03
- 사회자: Suzanne McIntosh (Cloudera Inc. and New York University, USA)

1) Design and Realization of Cognized Routing Resource by Big Data Analysing in SDN
- 발표자: Hongyan Cui (Beijing University of Posts and Telecommunications, China)
- 내용 요약

미래의 네트워크는 지적 능력의 성취를 위해 빅 데이터를 분석하여 사용자의 요구사항을 파악하는 것이 중요함. OpenFlow는 SDN 아키텍처의 제어 및 전달 층의 사이에

나타난 첫 번째 표준 통신 인터페이스임. 네트워크의 유연성과 유용성을 향상시키기 위해, 이 연구에서는 이것을 사용자 데이터 분석 클라우드 플랫폼에 연결함으로써 사용자의 기호인식능력을 갖춘 네트워크 자원 할당 방법을 제시함. 들어오는 정보 흐름의 속도와 유형을 예측하고, LARAC 알고리즘을 통해 가장 적은 부담을 주는 링크를 얻기 위해 Hadoop 플랫폼을 사용함. 이 시스템이 사전에 네트워크 부하를 예측하여 네트워크 자원 할당을 역동적으로 할 수 있다는 것은 BUPT SDN 실험을 통해 확인됨. 실험 결과, 연구에서 제안된 방법이 이전의 네트워크보다 부하 균형을 더 잘 해결함을 보여줌.

2) Toa: A Web Based Network Flow Data Monitoring System at Scale
- 발표자: Jose Ortiz-Ubarri (University of Puerto Rico, USA)
- 내용 요약

본 연구에서는 웹 기반 네트워크 흐름 데이터 모니터링 시스템(NMS) Toa를 제시함. Toa는 네트워크 시각화 분석을 위해 네트워크 흐름 데이터를 자동적으로 분석하고, 이 정보를 데이터베이스 시스템에 저장하며, 상호작용적 타임 라인 차트를 생성하는 스크립트의 모음들로 구성되어 있음. 시스템은 5분마다 생성되는 네트워크 흐름 데이터로부터 상호작용적 차트를 연속적으로 업데이트하는 실시간 시스템임. Toa는 더 심층적인 시각화와 분석을 위해 데이터베이스에 저장된 데이터로부터 맞춤형 차트를 생성하는 인터페이스 및 시각화 차트를 네트워크 흐름 데이터 파일과 연결하는 플러그인 또한 제공함. Toa web GUI는 (1) 네트워크 레이블(인터페이스, 자율 시스템 [AS], 또는 네트워크 블록) 당 트래픽, (2) 각각의 네트워크 레이블에 대한 포트 별 트래픽, (3) 네트워크 레이블 트래픽을 위한 네트워크 레이블, (4) 데이터베이스 데이터로부터의 맞춤차트, (5) 네트워크 흐름 데이터 파일의 심층 분석을 위한 플러그인 등의 네트워크 트래픽 시각화 옵션을 사용자에게 제시함.

다. Session 7 : Big Data Application

o 일시 및 장소: 7월 1일(수), 오후 1:00-2:00, R3.02/3.03
- 사회자: Tony Shan (Chief Technologist)

1) Study on Corporate Governance of Stock Market in Korea: Network Analysis with relationship of Major Shareholders
- 발표자: Hyejung Moon (Seoul Tech University, Korea)
- 내용 요약

이 발표의 목적은 복합적으로 얽힌 주요 주주의 관계를 분석함으로서 기업의 지속가능성을 측정하기 위함임. 연구 결과, 주요 주주 사이의 연결에 따른 클러스터링과 기업 집단별 네트워크를 밝혀냄. 첫 번째 유형은 소규모 네트워크와 기업집단의 비즈니스 스타일 및 의사결정 속도, 오너 리스크와 같은 형질들을 보여주고, 두 번째 유형은 주식시장에 고유의 법칙에 따른 무척도 네트워크와 투자선호도를 보여줌. 정보기술에 초점을 둔 주식 산업 분야별 클러스터링도 관찰 가능함.

2) Performance Evaluation of NoSQL Databases: A Case Study

- 발표자: Neil Ernst (Software Engineering Institute, USA)
- 내용 요약

빅 데이터 시스템 사용을 위한 특정 NoSQL 데이터베이스를 선택하는 것은, 현재의 기술 환경의 변화를 요구함에도 비싼 기술적 비용 때문에 망설이게 만들기도 함. 본 발표에서는 많은 의료 공급자에 의해 개발된 전자의료기록시스템의 사용을 위한 NoSQL 데이터베이스를 선택할 것을 제안함. 특정한 어플리케이션의 프로토타이핑, 데이터 모델 및 질문사용 사례에 맞는 NoSQL 제품의 식별 측정을 수행하고, 수행요건을 충족시킴. 사례 연구로, 10진법으로 변경된 데이터베이스 처리량을 알아내고, 5진법에 의해 변경된 연산지연 시간을 판독하고, 잠재시간을 기록함. 일반적인 빅 데이터 시스템과 NoSQL 데이터베이스의 세부적인 기술 평가를 수행할 때의 어려움을 비교함.

3) Optigrow: People Analytics for Job Transfers

- 발표자: Kush R.Varshney (IBM, USA)
- 내용 요약

IT 서비스 산업은 클라우드, 분석, 모바일, 소셜, 보안기술 등의 성장과 함께 급격한 변화를 겪고 있음. 이런 변화 환경에 맞추고자 하는 서비스 제공자들은, 직무역할적인 면에서의 변화가 요구되며, 초과비용 발생에 대해서도 고려해야 함. 본 발표에서는 성장분야와 기존분야에서 고용자의 적절한 내부 업무 이전을 통해 이러한 변화를 가능하게 하는 빅 데이터 접근방법에 대해 제안하고자 함. 성장분야의 직무에 필요한 수학적 프로필 기술세트를 개발하기 위해 직원의 전문지식 데이터를 사용하여 분석함. 그리고 성장분야 직무로 전환될 내부 후보자들의 우선순위를 매기기 위해 통계적 점수 알고리즘을 개발함. IBM Corporation의 IT서비스에 이 분석 과정을 어떻게 적용할 수 있는지를 실험한 결과를 제시함.

라. Session 9 : Evaluation

o 일시 및 장소: 7월 2일(목), 오전 8:15-9:15, R3.02/3.03

- 사회자: Barbara Carminati (University of Insubria, Italy)

1) Unsupervised Event Detection with an Infinite Poisson Mixture Model

- 발표자: Vinod Hegde (Insight Centre for Data Analytics, NUI Galway, Ireland)
- 내용 요약

센서와 웹 사용자들에 의해 생성된 대량의 시계열 데이터는 상황적인 정보를 제공하는 훌륭한 자원임. 본 발표에서는 가산 자료의 시계열 아웃라이어를 식별해냄으로서 사건을 탐지하는 무한 푸아송 혼합모델을 제시함. 연구에서 제시한 모델의 장점은 아웃라이어가 무한혼합모델에서 발견된 혼합요소를 매핑 한다는 것에 있음. 이러한 장점에 의해 추론된 아웃라이어 데이터의 규모를 기반으로 하여, 이벤트를 식별하고 분류하도록 함. 종합적인 현실세계 데이터를 사용하는 Markov Modulated Poisson Process(MMPP)에 기반을 둔, 잘 알려진 이벤트 검출 기술과 비교하여 본 연구에서 제

안한 모델의 수행도를 분석함. 분석 결과는 사건을 검출하는 본 연구의 접근방식이 MMPP에 비해서 주기적인 가산자료 분석에 더 적절하다는 것을 보여줌. 또한 제시된 모델이 아웃라이어 검출 분석에 있어 탄탄하고, 상세하며, 해석할 수 있는 결과를 제공한다는 것을 증명함.

2) Reconstructability-aware Filtering and Forwarding of Time Series Data in Internet-of-Things Architectures

- 발표자: Apostolos Papageorgiou (NEC Laboratories Europe, Germany)
- 내용 요약

사물인터넷 디바이스 모니터링에 기인한 시계열 스트리밍은 빅 데이터를 분석을 가능케 하는 요소 중 하나임. 사실상 거의 모든 개체가 연결된다면, 사물인터넷 플랫폼은 게이트웨이 단말기처럼 사물인터넷을 통해 수집되는 데이터 사이즈를 축소시키는 시스템을 필요로 하게 될 것임. 그렇지 않으면, 많은 시스템들은 저장비용, 대역폭, 에너지소비, 데이터베이스 I/O 처리량 제한과 같은 문제에 직면하게 됨. 이 발표에서는 사물인터넷 환경 재건과 관련하여, 시계열 데이터 축소를 위한 체계와 메커니즘을 제시함. 제시한 두 단계의 메커니즘은 원 데이터 재건을 위한 요구 정도를 유지하면서 부하를 줄이는 면에서 분석, 선택, 그리고 적절한 데이터 감소 조절에 대해 설명함. 이 접근 방식은 공적으로 이용 가능한 실제 시계열 데이터로 평가를 실시함. 복원된 시계열 데이터가 20% 이상이 되지 않음에도 불구하고 원 데이터 세트와 85%~99.9%의 유사성을 가질 정도로 재건되는데 불과 10~100초의 시간이 걸린다는 것을 입증함.

3) NoSQL in practice: a write-heavy enterprise application

- 발표자: Joao Ricardo Lourenco (CISUC. Centre of Informatics and Systems of the University of Coimbra, Portugal)
- 내용 요약

지속적인 정보의 성장은 데이터 저장과 관련하여 변화를 요구하고 있음. NoSQL과 같은 대체 기술은 기업계에서 계속 늘어나는 데이터 필요조건의 해결책으로 제시되었으나, 이러한 주장이 실제 연구에 의해 뒷받침 된 사례는 많지 않음. 현재 주로 사용되는 방법은, 합성데이터에 특정 쿼리를 수행함으로써 데이터베이스 성능을 평가하는 것임. 그러나 이러한 인공의 시나리오들은 현실 세계에서 적용 가능한 방법들을 제한하고, 실제 시스템에서의 데이터베이스의 기능을 특정하는 경우가 많음. 이에 대응하기 위해, 실제 기업 데이터 및 시스템을 사용하여 NoSQL 데이터베이스의 수행도를 평가하고, 그 결과를 SQL과 비교함. 특히, 기업 소프트웨어 및 빅 데이터를 사용한 첫 번째 write-heavy 평가 방법을 제시함. 이러한 방법으로 카산드라, 몽고DB, Couchbase Server, MS SQL Server를 테스트함으로써 각각의 성능을 비교함.

마. Session 10 : Big Data Framework

o 일시 및 장소: 7월 2일(목), 오전 9:45-10:45, R3.02/3.03
- 사회자: Tony Shan (Chief Technologist)

1) Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander
- 발표자: Salvatore Longo (NEC Laboratories Europe, Germany)
- 내용 요약

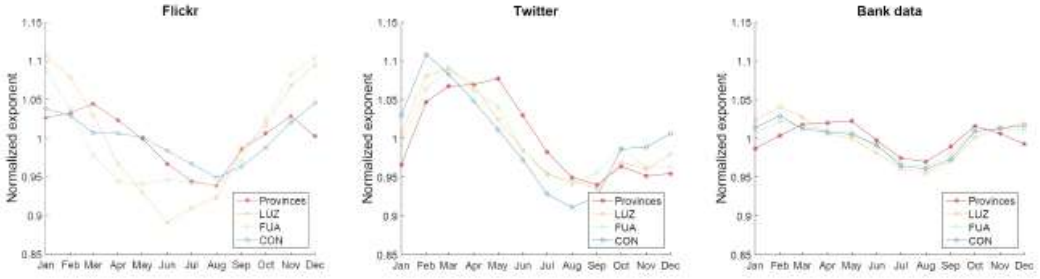
사물인터넷은 도시의 연결성, 편리성을 확대하고 있으며 동시에 지능화시키고 있음. 그러나 현재의 사물인터넷과 관련된 변화는 센서, 디바이스, 인간 활동을 연계함으로서 추출된 빅 데이터와 그에 대한 통찰력에 많이 의존하는 수준임. 기존의 많은 연구와 프로젝트는 우리가 살고 있는 도시를 스마트하게, 다양한 센서와 디바이스를 어떻게 배치할지에 대해 더 초점을 맞추어 데이터를 수집하도록 하고 있음. 그러나 이것은 스마트 도시를 향한 첫 걸음일 뿐이고, 다음 단계는 유동적인 빅 데이터 플랫폼을 통하여 모든 종류의 애플리케이션과 서비스에서 수집된 데이터와 상황인지 및 정보의 좋은 활용 사례를 만드는 것임. 이 발표에서는 시스템 아키텍처 및 살아있는 도시 데이터와 분석 플랫폼, 즉 CiDAP의 주요 설계에 대해 소개함. 더 중요하게는, 대규모의 스마트 시티 시험대인 SmartSantander라는 실제 시스템을 구축하면서 얻은 경험과 교훈을 공유함. 이 연구는 미래 스마트 도시 플랫폼 설계자에게 유용한 사례를 제공하여 향후 도시 설계자들이 실용적인 문제들을 예견하고, 스마트 도시 데이터 플랫폼을 설계할 때 적용할 수 있도록 함.

2) Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity
- 발표자: Iva Bojic (MIT, SENSEable City Lab, USA)
- 내용 요약

스페인의 은행카드 거래내역(스페인 전역에서 2011년 빌바오비스카야 은행의 국내외 사용자 은행카드 거래 내역), 지리적 위치 정보를 포함한 사진(일정 시기에 Flickr에 업로드 된 1억 개 이상의 지리적 사진들), 트위터 메시지(2012년 지리적 위치 정보를 포함한 트윗)에 지리로부터 수집한 빅 데이터를 분석하여 해외 방문객의 도시 매력지수를 측정함. 그 결과, 도시의 인구수와 도시 매력지수 간에 강한 초선형 관계가 나타남을 발견. 또한 도시 매력지수의 계절적 패턴을 분석하고, 도시 계획에 있어서 이산선택모형의 효과성에 대해 분석함.

- 활용 영역: 도시 계획 (관광)
- 활용한 빅 데이터: (1) 플리커(태그기반 인터넷 앨범 서비스) 데이터 두 개를 머징하여 1억 3천여 개의 사진/비디오 자료를 활용(2005~2014년 기간 동안, 각 데이터 당 1백만 명 이상의 사용자에 의해 찍힌 1억 개 이상의 사진/비디오 자료가 공유되고 있음). 사진/비디오별로 각각의 id가 부여되어 있으므로, 겹치는 부분은 제외함. 사용자가 해외방문객인지를 파악하기 위해, 가장 많은 사진을 올린 나라와 방문했을 것으로 생각되는 나라들을 구분함. 이러한 방법으로 사용자 50만 명의 나라를 구분해냄. 이들이 전체 데이터의 약 80%의 사진/비디오를 공유함. 나머지 50만 명의 경우는

간헐적으로 사진을 업로드하는 집단이기 때문에 출신국가를 구분하지 못함. 따라서 출신국가를 구분해 냄으로써 해외방문객으로 판정된 경우만 분석에 활용함. 스페인에서 찍힌 전체 350만 개의 사진 중 112개 국가의 약 1만 6천명의 해외방문객이 찍은 40만여 개의 사진을 분석에 활용함. (2) 2012년에 트위터로부터 지리적 정보를 포함한 메시지를 수집함. 1천 3백만 명의 사용자가 보낸 9억 4천 4백여 개의 트윗 중 2012년에 스페인에서 포스트 된 트윗은 64만 1천여 명의 트위터 사용자가 올린 3천 5백만여 개 메시지임. 이 중 2%의 트윗(약 1천 9백만여 개)이 180개 국가의 8만여 명의 해외방문객으로부터 보내진 트위터로 이것을 분석에 사용함. (3) 2011년 빌바오비스카야 은행(BBVA)카드 거래내역 데이터로, BBVA가 승인한 직불 또는 신용 카드 거래 내역과 3십만 여개 BBVA 카드 단말기를 통해 거래된 내역이 포함됨. 이 중 해외에서 발행된 카드 사용 내역을 구분해냄으로써, 175개 국가로부터 온 860만여 명의 해외방문객의 1천 7백만 거래내역을 분석에 활용함.



분석 결과, 여름 시즌에는 스페인 방문객이 보다 넓은 지역을 광범위하게 다니므로, 여름 동안에 대도시(Provinces: 스페인 52개 주, LUZ: 24개 대도시, FUA: 40개 기능적 도심 지구, CON: 211개 광역도시)의 도시 매력 지수는 떨어짐. 여름에는 해변 지역을 중심으로 방문객이 증가하고, 봄이나 가을에는 출장 등을 목적으로 한 해외 방문객의 대도시 집중이 증가함.

3) Automation of the Validation, Anonymization and Augmentation of Big Data from a Multi-year Driving Study

- 발표자: Bruce Wallace (Carleton University, Canada)
- 내용 요약

The Candrive/Ozcandrive 프로젝트는 고령 운전자들의 안전 증진에 초점을 맞춘 장기 연구 프로젝트로 올해 6년차에 접어듦. 연구 대상은 오타와 지역의 256명의 고령 운전자들이고 연구 대상자들의 자동차에 부착된 센서 정보를 빅 데이터로 구축해서 분석을 진행하는 패널 연구의 성격을 띰. 자동차가 시동을 건 후 초단위로 자동차 GPS와 온보드 진단기 II에 부착된 센서로부터 데이터를 수집함. 최종 산출된 데이터는 각각의 자동차로부터 수집된 수백, 수천 시간의 운행 정보로 구성되어 있음.

본 발표에서는 GPS로부터 구성된 대규모 데이터를 변환하는데 있어서의 어려움에 대해 언급하고, 이처럼 수집된 여타의 센서 데이터에 대해서도 살펴봄. 개별 센서 데이터로부터 문제점을 발견하고 교정하기 위해 쉽게 적용하여 활용할 수 있는 자동화 방법에 대해서도 함께 검토함. 연구 참여자들의 개인정보 보호를 위해 집 주소와 자동

차 위치 정보 및 집 주소와 센서 정보 등 개인이 드러날 수 있는 연계 정보에 대해서는 7개 매개변수를 기준으로 개인 민감 정보를 익명 처리하여 분석함. 데이터 세트에 날씨 정보, 도로 정보, 낮/밤길이 정보 등 센서 데이터 외에 외부로부터 가져온 데이터를 결합하는 방법을 제시함. 특히 클라우드 컴퓨팅 체계 안에서 데이터 세트를 구축하여 분석할 수 있는 방법을 제시함.

- 활용 영역: 교통(자동차 운행정보)
- 활용한 빅 데이터: 고령 운전자들의 자동차 운행정보. The Candrive/Ozcandrive 프로젝트는 캐나다, 호주, 뉴질랜드 전역의 1,000명 이상의 고령 운전자들 대상으로 하는 연구임. 현재 6년차인 이 연구는 개인별로 100시간 최대 1,000시간 넘는 자동차 운행정보를 담고 있는 데이터로 고령 운전자의 운전 습관을 파악하고 안전을 증진하기 위해 진행되고 있음. 이 연구의 결과는 단계적 운전면허 시험 기준 마련과 같이 고령 운전자 관련 정책이나, 의료 전문가들의 의사 결정을 위한 근거 자료로 활용됨. 본 발표에서는 오타와 사례만을 분석에 활용하고, 몇 개의 케이스만을 제시함. 개개인의 자동차의 GPS 및 온보드 진단기 등에 부착된 센서로부터 수집되는 자료의 종류는 아래와 같음.

<개인별 수집 자료>

자료원	측정된 데이터	수집 단위/형태
센서 수집 장치	센서 시리얼 넘버	정수
	운행 수	정수 - 자동차 시동 걸때마다 증가
GPS	날짜, 시간	(예) 18 January 2013 10:05:10
	위도, 경도	도
	GPS 수신 상태	글자 - (예)3D 수신
	GPS 정확도	부동 소수점
	속도	km/h
	제한 속도	km/h
	경고	글자
RFID	RFID 넘버 (한 개의 차량을 공유하는 두 명의 운전자에게 각각 부여)	16진법으로 변환된 일련번호
자동차 온보드 진단기 II	엔진 냉각기 온도	℃
	엔진 RPM	정수
	속도(자동차 계기판)	km/h
	공기 흡입구 온도	℃
	주변 온도	℃
	스로틀 포지션(절대값)	부동 소수점
	스로틀 포지션(상대값)	%

6. 기타 연구

가. Visionary Session 1

o 일시 및 장소: 6월 29일(월), 오후 3:30-4:30, R3.02/3.03

- 사회자: Latifur Khan (UT Dallas, USA)

1) Research Directions for Big Data Graph Analytics

- 발표자: John Miller (University of Georgia, USA)

- 내용 요약

현재의 빅 데이터 시대에는 방대한 데이터 그래프 정보를 분석하고 추출하는 것에 대한 관심이 빠르게 증가하고 있음. 이번 발표에서는 쿼리 프로세싱의 관점에서 그래프 분석에 대해 검토하고자 함. 최단 경로의 결정이든, 혹은 데이터 그래프 매칭에서의 패턴 찾기이든, 모든 문제는 그래프에서 관심 있는 특성이나 정보 콘텐츠를 찾는 것과 관련되어 있고, 대부분은 패턴 경로 문제로 요약될 수 있음. 이 문제를 해결하는 것이 쉽지 않고, 특히 반복되는 알고리즘은 병렬 맵리듀스를 통한 프로세싱 과정을 효율적으로 만들어야 한다는 과제를 안고 있음. 또한 규모가 큰 그래프에 대해서도 동일한 솔루션이 적용될 수 있도록 해야 함.

2) MCD: Mutual Clustering across Multiple Social Networks

- 발표자: Philip Yu (UIC, USA)
- 내용 요약

온라인 소셜 네트워크에서의 커뮤니티 탐색은 최근 몇 년간 뜨거운 연구주제가 되고 있으며, 현재의 사용자들은 공통의 정보를 공유하기 위해 다수의 온라인 소셜 네트워크에 동시적으로 참여하고 있음. 일반적인 사용자들이 포함된 네트워크는 다수의 '부분적으로 정렬된 네트워크'로 지정됨. 이 발표에서는 '상호 클러스터링'하는 다수의 불완전 정렬 네트워크 커뮤니티를 탐색하고자 함. '상호 클러스터링'의 특징이 있는 네트워크 커뮤니티는 두 가지의 중요한 문제를 가짐. (1) 상호 커뮤니티 인식 과정에서 어떻게 네트워크 특성을 보존할 것인가? (2) 공유된 사용자의 커뮤니티를 정제하고 그 차이를 명확하게 하기 위해 정렬 네트워크에서 정보를 활용하는 방법은 무엇인가? 이 두 가지 과제를 해결하기 위해, 새로운 커뮤니티 검출 방법인 MCD (Mutual Community Detector)에 대해 제시함. MCD는 (1) 각각의 네트워크 특성과 (2) 공유 사용자들의 정보를 동시에 고려하여 소셜 커뮤니티 구조를 검출하도록 함.

3) Big SaaS: The Next Step Beyond Big Data

- 발표자: Hong Zhu (Oxford Brookes University, UK)
- 내용 요약

SaaS(Software-as-a-Service)는 소프트웨어의 기능을 사용자에게 서비스로 전달해주는 클라우드 컴퓨팅 모델임. 가까운 미래에, SaaS 애플리케이션은 많은 인구에게 주문제작 가능한 지식 서비스를 제공하기 위하여 사물인터넷, 모바일컴퓨팅, 빅 데이터, 무선센서네트워크, 많은 다른 컴퓨팅 및 통신 기술을 통합할 것임. 이것은 전례 없는 복잡성과 규모를 가진 Big SaaS 시스템이라 부르는 시대의 도래를 초래할 것임. 그러나 이것은 큰 사회적 위험을 가져올 수도 있으며, 거기에는 다른 단점과 도전 과제도 있음. 예를 들어, 클라우드 소싱과 개념모델의 완전성을 유지하면서 데이터와 메타데이터의 질을 보장하는 것은 어려운 문제임. Big SaaS 애플리케이션이 계속적으로 진화해 가는 것 또한 필요함. 이 발표에서는 소프트웨어 수명주기의 모든 단계에서 어떻게 이런 문제들을 다룰 것인가에 대해 논의함.

나. BDRH (Big Data Research in Healthcare) Session 1

o 일시 및 장소: 7월 1일(수), 오전 8:15-9:15, R3.04/3.05
- 사회자: Kelvin KF Tsoi (The Chinese University of Hong Kong, China)

1) Blood Pressure Management with Data Capturing in the Cloud among Hypertensive Patients: A Monitoring Platform for Hypertensive Patients
- 발표자: Benjamin Yip (The Chinese University of Hong Kong, China)
- 내용 요약

고혈압은 심혈관과 신장질병의 위험요인이지만 치료가 가능한 질병임. 혈압(BP) 조절은 심혈관계 질환 관리에 매우 중요한 요소임. 최근의 홈 원격모니터링 BP는 서구에서 일반적으로 사용되는 방법으로, BP 조절에 효과적인 도구임. 이처럼 건강관리를 위한 기술 애플리케이션은 트렌드가 되고 있음. 의료 데이터는 보통 길고 규모가 아주 방대함. 이러한 데이터에 대한 효과적인 관리는 의료서비스의 질을 향상시켜줄 것임. 의료 데이터와 클라우드 기술의 접목은 새로운 지평을 열었음. 의료공급자 경보 기능과 자동 데이터 캡처링 클라우드 기술은 최근의 홈 원격모니터링 BP 시스템을 활성화하는 도구가 될 것임. 발표에서는 클라우드에 연결된 개인 BP 측정기가 연구기반 BP 데이터 캡처링 클라우드 플랫폼으로 변환되는 과정을 살펴보고, 이러한 BP 측정 및 업로드 데이터가 USB 허브 및 인터넷에 연결된 개인 컴퓨터를 통해 일상적으로 사용되는 과정에 대해서도 살펴봄. 전반적인 과정에서 모든 개인의 개인정보는 디코딩되며, 데이터 개인 정보 보호를 위해 연구식별 번호가 각 사용자에게 할당됨. 클라우드 플랫폼은 다른 단체, 고속 성능, 강력한 인프라 지원이라는 요소와 함께 어디에서는 활발한 데이터 분석을 위해 쉽게 사용될 수 있음.

2) Indoor Air Monitoring Platform and Personal Health Reporting System: Big Data Analytics for Public Health Research
- 발표자: Kin-Fai Ho (The Chinese University of Hong Kong, China)
- 내용 요약

대기오염은 호흡기 감염과 폐암의 위험도를 증가시킴. 외부환경에 대한 대기오염 모니터링 시스템은 일반적임. 발표에서는 가정 내 환경에서의 공기 모니터링 결과를 모바일 애플리케이션을 통해 개인 의료 보고 시스템에 연결하는 것과 관련 있음. 데이터는 계산효율과 데이터 저장용량 향상을 개선하기 위해 클라우드에 캡처되고 저장됨. 오염데이터는 연중 매시간 캡처될 수 있음. 이러한 이유로 클라우드에서 꽤 큰 데이터 스토리지가 필요함. 건강 상태는 자가 보고 시스템을 통해 업로드 될 수 있음. 이러한 데이터는 다른 의료연구 및 미래 도시계획 등에도 유용한 정보로 제공될 수 있음. 또한, 오염 데이터에 기반한 데이터 분석은 서로 다른 시점에서 고도의 오염 지역을 식별하는데 도움을 줄 수 있음. 이러한 데이터는 개인에게 예방책을 상기시키고 흡입오염원을 피하도록 경고하는 시스템의 개발에 유용함. 이 경고시스템은 가정, 상업용 건물, 공공장소에 적용 가능함. 이러한 클라우드 플랫폼의 누적데이터는 공기오염과 건강 결과 사이의 연계성을 찾아 데이터 마이닝을 지원함으로써 공중보건 연구를 지원하는데 기여함.

3) Embracing Big Data for Simulation Modelling of Emergency Department Processes and Activities

- 발표자: Yong-Hong Kuo (The Chinese University of Hong Kong, China)
- 내용 요약

시뮬레이션은 응급실에서의 프로세스 및 활동을 확인할 수 있는 방법임. 그러나 대부분의 애플리케이션은 부서의 직원에 의해 수동으로 입력된 데이터에 의존하고 있음. 첫째로, 이 방법은 응급실 프로세스에 대한 정보를 자동적으로 저장하지 않는 한, 시뮬레이션 모델을 구축하는데 필요한 데이터가 컴퓨터시스템에 캡처되도록 보장하지 않음. 둘째로, 수동적인 입력시스템에서 인간의 오류에 의한 데이터 누락은 일반적인 현상임. 응급실의 실제 시스템을 구현할 수 없는 시뮬레이션 모델은 병원 관리자에게 잘못된 결과를 제공하게 됨. 만약 의료진이 부정확한 시뮬레이션 결과를 신뢰한다면, 이로 인한 부정적인 결과를 초래하게 됨. 이 발표에서는 홍콩 응급실의 시뮬레이션 모델 개발 사례를 제시하고, 데이터와 관련한 도전 과제에 대해 논의함. 그런 후에 개발 사례와 관련 문제점을 고려한 높은 정확도를 가진 시뮬레이션 모델을 구축하기 위해, 응급실에서 환자의 활동에 관한 많은 양의 데이터를 자동적으로 캡처하는 RFID가 가능한 인프라를 제안함.

다. BDRH (Big Data Research in Healthcare) Session 2

o 일시 및 장소: 7월 1일(수), 오전 9:25-10:25, R3.04/3.05

- 사회자: Jeffrey Tsai (Asia University, Taiwan)

1) Risk-adjusted Monitoring Method for Surgical Data: Methodology for Data Analytics (Work in Progress)

- 발표자: Xin Lai (The Chinese University of Hong Kong, China)
- 내용 요약

홍콩 병원청은 홍콩 내 모든 공공병원의 수술실적을 감사하는데 사용되던 수술 결과 모니터링 및 개선 프로그램(SOMIP)을 시작함. 연간 SOMIP 보고서가 제공하는 가장 중요한 정보 중 하나는 30일내 사망률의 변화와, 각 병원에서 심각하게 악화된 상태가 있었는지에 대한 것임. 그러나 일상적인 모니터링 방법은 Variable Life-adjusted Display(VLAD)와 Cumulative Sum Charting(CUSUM)을 사용하므로 수술 성과의 변화를 효과적으로 감지하지 못할 수도 있고, 수술결과에 따른 개선 또는 악화를 예상하는 것이 현실과 정확히 일치하지 않을 수 있음. 발표에서는 수술실적의 변화를 검출하기 위해서 보다 효과적인 위험 조정 모니터링 방법에 대해 제안함. SOMIP에 이 방법을 적용하면 제안된 모니터링 절차로 환자에게 어떠한 수술을 해야 할 지 결정할 때 외과의, 마취의, 중환자 치료사들에게 판단 기준을 줄뿐만 아니라, 홍콩 병원청의 관리자들을 도와 수술실적과 질을 개선할 수 있을 것으로 예상됨.

2) Patient Flow Evaluation with System Dynamic Model in an Emergency Department: Data Analytics on Daily Hospital Records

- 발표자: Marc Chong (The Chinese University of Hong Kong, China)
- 내용 요약

병원에서 사고 및 응급 서비스에 대한 빅 데이터는 임상의에게 임상결과를 제공하고 환자에게 의료 정보를 제공할 목적으로 분석됨. 시스템 다이나믹 모델링은 조직이나 사회 시스템의 복잡한 행동양상을 모델링하는데 사용되는 기술임. 발표에서는 홍콩 내 병원에서의 사고나 응급상황에서 환자의 흐름을 모델링하는데 사용되는 시스템 다이나믹 접근방식에 대해 검토함. 이를 통해 응급실에서의 안전 및 다양한 행동상황에 대한 질 평가에 대해 검토함. 예를 들어, 응급실 초기 대기 시간(acute bed와 대기실)이나 홍콩의 사고 및 비상시스템이 얼마나 효율적으로 작동할 수 있는지에 대해 평가하기 위한 다양한 요인들을 조정(입원 병상 수 및 직원 수 등) 하여 적용해 봄.

3) Two screening methods for genetic association study with application to psoriasis microarray data sets

- 발표자: Maggie Haitian Wang (The Chinese University of Hong Kong, China)
- 내용 요약

원인이 다양한 질병의 근원을 식별하기 위해 게놈 데이터를 분석할 때, 실제 존재하는 변수들 간의 다양한 조합에 의해 이루어진 변수까지 검사해야 하므로 고려 대상이 되는 변수의 수가 실제의 변수 수보다 더 많아지게 됨. 이처럼 변수 간 상호작용 특성을 선택하기 위한 기존의 방법은 게놈의 양방향 조합을 철저하게 계산하거나 유전자 표지의 한계효과로 사전 스크리닝의 단계를 취하는 것이었음. 그러나 이러한 방법들은 추가 검사를 필요로 하는 작용 요인들을 발생시킬 수도 있음. 각각의 변수가 적당히 중요한 효과가 있지만, 이보다 더욱 강한 상호작용의 부분 집합을 형성하기 위해서 다른 유전자들을 필터링하는 경우도 발생함. 발표에서는 두 가지의 대안적인 검진 방법을 제안하는데, 하나는 가변 등장 빈도(VAF)를 사용하여 특성을 선택하는 것이고, 다른 하나는 비중복 기준을 사용하여 후보자 풀을 줄이는 것임.

Ⅲ. 시사점

1. 빅 데이터를 이용한 연구 동향 파악 및 정책 적용 사례 검토

- o 공공 데이터 활용: 교통, 재난, 기후 등 정보 공개가 필수적인 정부 데이터를 활용하여, 교통 혼잡도, 재난 대응 및 예측 등의 빅 데이터를 분석하여 관련 정책 수립을 위한 시사점을 도출
- o 건강 데이터 활용: 병원에서의 진료, 수술 기록이나 개인 흡연 습관 등의 건강행태 자료를 빅 데이터로 구축하여 건강 관련 정책 제안을 위한 자료로 활용함. 그러나 개인 정보가 민감한 자료이므로, 개인 정보를 보호할 수 있는 코드화 등의 작업이 선행되어야 함을 제안
- o 소셜 미디어 및 네트워크 분석: 트위터, 페이스 북 등 소셜 네트워크 사용자의 행태, 행동 패턴 등을 빅 데이터로 구축하여 분석함으로써 도시 개발 정책, 관광 산업 등의 정책 제안에 활용하고, 사용자들의 행동 패턴을 분석함으로써 사회 현상을 분석함.

다른 형태의 자료들과의 연계를 통해 다각도의 분석을 시도함.

2. 빅 데이터 통계 분석 영역 발굴

o 오픈 소스를 활용한 데이터 구축 및 분석

o 기존 빅 데이터 분석에 있어 성별 통계 기법을 적용할 수 있도록 제안

o 여성 관련 이슈에 대한 빅 데이터 구축 및 분석을 통한 시사점 도출

o 빅 데이터 구축 및 분석을 위한 정책 연구 제안

예) 여성 인구가 많이 분포하고 있는 취약 노인 계층에 대한 응급 상황 대처 방안 마련: 센서 데이터를 통해 대상자의 건강 상황을 정기적으로 체크. 클라우드 시스템 등을 활용하여 이를 빅 데이터로 구축. 데이터 분석을 통해 응급 상황에 대한 사전 대처 방안 등 마련

3. 정책연구 전문가와의 네트워크 구축

o 해외 연구자와의 네트워크 구축

o 포럼 기획

주제: 빅 데이터 사례와 연구 방법론

일시: 2015년 8월 중(예정)

발표: 문혜정(서울과학기술대학교 IT정책전문대학원)

4. 수집자료 목록

o 학회 일정 및 프로그램 설명서

o 논문 자료 (*대용량 자료이므로 별도 제출)