IT & Future Strategy

빅데이터 시대, 미래전략의 새로운 접근법

제14호(2015. 12. 16.)

목 차

NIA 한국정보화진흥원

「**IT & Future Strategy 보고서**」는 21세기 한국사회의 주요 패러다임 변화를 분석하고 이를 토대로 미래 초연결 사회의 주요 이슈를 전망, IT를 통한 해결 방안을 모색하기 위해 한국정보화진흥원(NIA)에서 기획, 발간하는 보고서입니다.

보고서 내용에 대한 문의나 제안은 아래 연락처로 해 주시기 바랍니다.

▶ 발행인 : 서 병 조

▶ 작 성
- 한국정보화진흥원(NIA) 정책본부 ICT미래전략팀
  이영주 수석 (053-230-1208, lyj@nia.or.kr)

▶ 보고서 온라인 서비스
- www.nia.or.kr

모두를 위한 미래, 인간중심의 초연결 창조사회

요약

◇ 빅데이터 시대, 미래예측 패러다임의 변화

- 미래연구는 수리적이고 단편적인 예측(Forecasting)에서 종합적 전망에 근거한 대응 방안을 포함하는 미래전략(Strategic Foresight)로 발전 중
- 빅데이터 처리 기술의 발달에 따라 텍스트마이닝, 소셜네트워크 분석, 계량정보학 등 대용량 데이터의 분석을 통한 통합적 미래연구가 주목받기 시작
- 주요 국가 정부와 산업계에서는 빅데이터를 각종 문제 해결 및 이슈 대응뿐 아니라 미래전략과 수반되는 전략적 의사경정의 중요한 도구로 활용 중

◇ 사례로 보는 빅데이터 기반 미래전략의 가능성

- 미래전략에 활용되는 데이터의 원천(Source)은 기존의 통계, 논문 뿐 아니라 소셜미디어, 공공데이터, 웹 활동데이터 등 다양한 빅데이터로 확대되는 추세
- 실시간 경기 예측, 사회적 위험 모니터링 등 단/중기 미래예측부터 장기적 미래전망까지 빅데이터를 활용한 다양한 미래전략 수립 시도 중

◇ 증거기반 미래전략의 시작

- 초연결 사회에서 발생할 훨씬 많은 분석 가능한 데이터들로부터 미래사회 대응에 필요한 유용한 정보와 지식화 역량 준비 필요
- 미래전략은 미래 예측과 전망 뿐 아니라 전략적 대응방안을 모색하는 실천(Action) 지향적인 활동으로, 지혜를 탐구하고 축적하는 과정
- 빅데이터 시대 증거와 객관적인 합의에 의해 형성된 지식을 기반으로 도덕이나 윤리적인 이슈까지 포함한 바람직한 미래에 대한 공론화 필요

◇ 데이터 기반 미래전략 프레임워크

- 빅데이터 시대의 미래전략은 데이터 특성과 전략 수립 목적에 따라 유형을 세분화하고 각각 특화된 방법론의 적용 필요

– 예측 시점(단기 vs 중장기)과 데이터 주제의 범위(특정 분야 vs 광범위)에 따라 4가지 미래전략 유형과 분석적 특성을 제안

* 문제해결형 / 예측·대응형 / 조기경보형 / 아젠다 발굴형

<table>
<tr><th>시점
범위</th><th>단기(1~3년) ⟷ 중장기(3~10년)
(프로젝트 단위 데이터 수집)</th><th>(지속적 데이터 축적)</th></tr>
<tr><td rowspan="2">특정
분야
데이터

↕

광범위
데이터</td><td>< 문제 해결형 >
‣ 현안이나 이슈의 세부적인 구조와 양상을 파악하고 해결안을 모색 (전통적인 예측적 분석)
‣ 다차원 분석이 가능한 상세하고 정확한 데이터 수집
‣ 예) 소셜미디어 기반 자살 위험 예측 모형</td><td>< 예측·대응형 >
‣ 축적된 데이터를 기반으로 추세분석을 통한 이슈(위기) 예측
‣ 변인 간 상호 역동성을 파악할 수 있는 종합적 분석역량 필요
‣ 대응전략과 실행방안 도출에 필요한 지식화가 중요
‣ 예) 기후변화/재난 예측</td></tr>
<tr><td>< 조기경보형 >
‣ 이상치 발견을 통한 비정상적인 이벤트, 사건들을 탐지
‣ 데이터의 정확도 보다 실시간 분석/탐지 가능성이 중요

‣ 예) 빅데이터 기반 물가 모니터링, 갈등 징후 포착</td><td>< 아젠다 발굴형 >
‣ 광범위 환경스캔을 통해 미래 트렌드와 이머징 이슈 전망
‣ 전통적인 미래연구와 유사
‣ 데이터의 해석과 전문가 집단 지성의 유기적 결합 필요
‣ 예) 환경스캔 기반 미래사회 유망기술 도출</td></tr>
</table>

◊ 증거기반 미래전략을 위한 빅데이터 활용 시 고려사항

– 미래의 변화를 읽어낼 수 있는 정확한 데이터의 소재 파악와 유효한 데이터의 선정, 선별

– 알고리즘에 의존한 예측 정확도보다 인간의 본성과 사회적 의미의 해석에 집중 (데이터 만능주의 또는 데이터 증거 과신의 오류 주의)

– 미래는 정해져 있지 않고 사회 구성원의 상호 행위 결과에 따라 바뀌어 가는 역동성(Dynamism)을 항상 고려, 미래의 이슈를 주도하고 만들어가는 개인과 집단의 목소리에 주목

Ⅰ 빅데이터 시대, 미래전략 패러다임의 변화

1. 미래연구 방법의 전환 시점 도래

□ 빅데이터 시대의 도래와 미래연구 환경의 변화

○ 미래연구는 수리적이고 단편적인 예측(Forecasting)에서 종합적 전망에 근거한 대응 방안을 포함하는 미래전략(Strategic Foresight)로 발전 중[1)]

– 빠른 기술발전과 세계화, 복잡성과 불확실성이 높은 상황에서 체계적 분석에 의한 미래예측과 대응 전략 수립은 필수

○ 미래학자들은 과학적 이론에 기초한 정량/정성적 예측 방법과 비과학적인 요소를 혼합한 방법을 활용해 옴

〈미래연구 방법론 동향에 따른 구분[2)]〉

과학적 원리 기반 미래연구 방법론	비과학적 요소가 혼합된 연구방법론
o 환경스캐닝 / 텍스트 마이닝 o 브레인스토밍 / 트렌드영향분석 o 교차영향분석 / 퓨쳐스 휠 o 기술예측 / 경제 통계학적 모델링 o 퓨쳐 로드맵 / 시스템 모델링 o 시나리오 / 시뮬레이션 게이밍 o 델파이 기법	o 소설, 일기, 신화, 공상과학, 예술작품 등에 나타난 미래 이미지 조사 o 개인의 통찰과 직관 분석 o 인과계층분석(Casual Layered Analysis) o 세대분석 (Age-Cohort Analysis) o 미래 비저닝 워크숍

– 정량적 데이터 분석에 의한 미래연구는 최근에 주목받기 시작하고 있으며, 빅데이터에 대한 관심이 증가함에 따라 점차 비중이 커지고 있음

1) 한국교육학술정보원, '테크놀로지와 미래교육에 대한 예측 방법과 해외 사례', 이슈리포트 2009.9
2) 한국정보화진흥원, '성공적 공공정책 수립을 위한 미래전략연구방법론', IT & Future Strategy, 2010.4.30., 내용 일부 수정 및 재정리

□ 미래연구과 빅데이터 접목의 시작

○ 대용량 데이터 처리기술의 발전에 따라 빅데이터는 미래전략의 새로운 정보 원천으로 부상

– 인터넷과 및 스마트폰의 확산에 따른 SNS 등 실시간성 데이터의 폭발로 사회적 현안이나 이슈, 환경변화 분석에 유용한 정보 제공

※ 기존의 언론, 잡지, 학술정보 또한 디지털로 축적되어 다양하게 활용 가능

– 미래를 전망할 수 있는 객관적인 정보들과 함께 디지털 사회에서 발생하는 대규모 데이터에서 통찰을 얻을 수 있는 장점이 부각

○ 글로벌 기업들은 각종 문제 해결 및 이슈 대응 뿐 아니라 미래전략과 수반되는 전략적 의사경정의 중요한 도구로 활용 중

– **(매출증가)** 아마존, 네플릭스 등은 수년간 축적된 데이터를 분석한 고객 추천서비스를 개발하여 수익 극대화

– **(품질개선)** 볼보와 GM은 자동차, 생산데이터, 운전자 데이터를 수집·분석하여 제품 품질 개선에 활용

– **(미래전략)** IBM은 사내에 200명 이상 수학자들이 분석해 도출한 핵심분야를 집중 연구함으로써 500개 이상의 관련 특허를 취득하고 미래 사업을 준비

○ 주요 선도 국가에서도 각종 현안 해결 뿐 아니라 미래예측 및 전략적 대응에도 빅데이터를 활용하기 시작

– 사회현안 및 미래이슈들에 대한 데이터 기반 분석과 선제적 대응 정책 마련을 위해 국가차원의 미래전략기구를 운영 중

※ 데이터분석을 통한 환경분석(호라이즌스캔)을 통해 미래 잠재적인 위협에 대한 징후 포착, 선제적 대응방안 모색 중

< 주요국의 빅데이터 기반 미래전략 추진 사례[3] >

▸ 영국의 HSC(Horizon Scanning Center)('05~)

- 내각 소속의 경영혁신기술부(DBIS) 소속 기구로 영국의 중장기 미래전략 수립을 위한 최신 과학이론과 데이터 증거기반의 정책분석 서비스 제공
- 기술변화와 혁신을 통한 미래역량 기법을 강화하고 전략적 미래예측 및 대응방안 수립, 미래예측 관련 문서 집적화
- 영국 비만대책 수립, 30-100년 내 위험관리대책 수립, 전염병 대응방안 마련 등 정량적/정성적 미래예측 기법을 적용하여 미래 이슈에 선제적 대응 지원

▸ 싱가포르의 RAHS(Risk Assessment and Horizon Scanning)('04~)

- 총리실 산하 기구로, 환경스캔을 통해 싱가포르의 미래에 영향을 미칠 수 있는 잠재적 위험요소와 불확실성 요소를 탐색, 이머징 이슈를 분석
  (해상안전, 테러, 조류독감 등에 대한 데이터수집 및 분석 절차 정립)
- 데이터 분석 실험센터를 운영하여 다양한 정량적 분석 기법을 연구하고 정부, 학계, 기업들과 공동 연구를 위한 플랫폼 제공

▸ 미국 행정부의 빅데이터 이니셔티브(Big Data Initiative)('12~)

- 오바마 행정부는 빅데이터 관련 연구개발에 2억 달러 이상을 투입하는 빅데이터 연구개발 이니셔티브를 발표[4]
- 유전자 연구 및 의료, 교육, 지구과학 및 국방분야 등 빅데이터 활용 효과가 뛰어난 분야의 기관들이 우선적으로 참여

※ 국립과학재단(NSF), 국립보건원(NIH), 국방부(DoD), 고등방위연구계획국(DARPA), 에너지부(DoE), 지질조사원(USGS)

▸ EU의 Future ICT 프로젝트와 iKnow 프로젝트

- Future ICT : 빅데이터를 활용하여 사회과학, 자연과학, 공학, 컴퓨터과학, 물리학(복잡계) 분야 간 협업을 통해 전 지구 차원의 지속가능성 확보를 위한 미래전략 플랫폼 개발
- iKnow(Interconnect Knowledge) : 유럽과 전세계의 과학, 기술 및 혁신을 위한 잠재적 지식 및 이슈 네트워크 구축, 전세계의 약신호(weak signal)과 와일드 카드(wild cards) 를 포착하기 위해 데이터 분석 기반의 horizon scanning을 활용

3) NIA, '선진국의 데이터기반 국가미래전략 추진현황과 시사점', IT&Future Strategy 2012-2호 (2012.4.6.)

2. 미래전략에 활용되는 빅데이터의 원천(Source)

□ 소셜미디어 데이터

○ 사회 구성원들이 자신의 의견을 표출하고 상호 소통하는 공간인 소셜미디어에서는 행동 패턴을 포착할 수 있는 데이터를 제공

– 구매패턴이나 취향 등 민간 영역의 활용 뿐 아니라 특정 시점의 사회적 이슈를 추적하여 정책 대안이나 미래 전망 도출에도 활용 가능

〈 전세계 인터넷 사용량을 실시간으로 보여주는 웹사이트 〉

※ 인터넷에서 생성되는 소셜미디어 데이터들은 이시간에도 1분에 약 1.3테라바이트(Terabite)이상 증가 중
(출처: http://pennystocks.la/internet-in-real-time/)

– 온라인 뉴스 또한 사회적 이슈에 대한 언론의 보도 비중과 보도 패턴을 파악하여 여론의 동향을 파악하는 데 귀중한 데이터를 제공

4) Executive Office of the President, Office of Science and Technology Policy, OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE, March 29.2012

○ 페이스북, 트위터 등 대표적인 SNS는 오픈 API 서비스를 통해 특정 유저(또는 단체)와 주제(이슈)에 대한 데이터 수집 기능을 제공

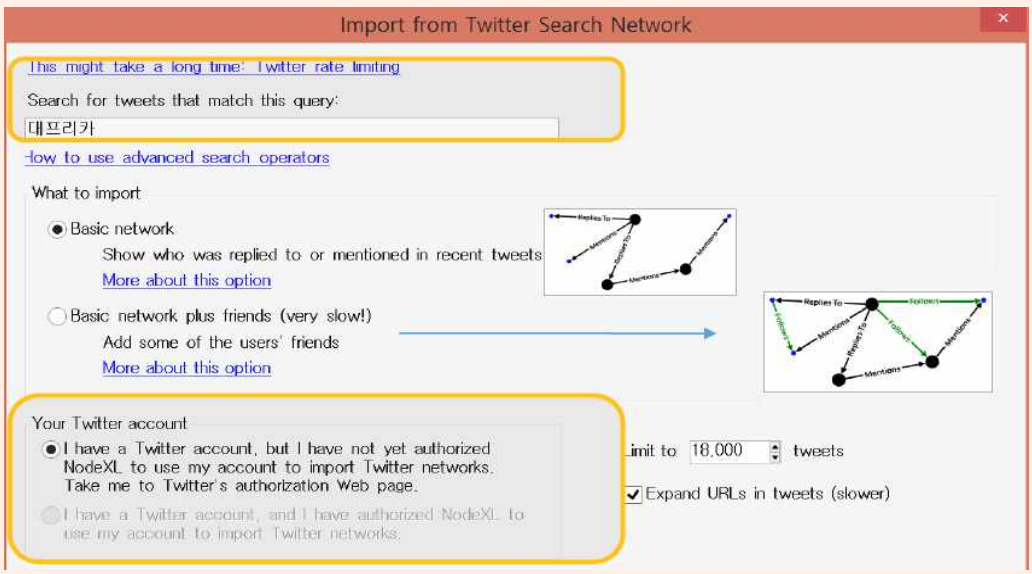
– 정성적 분석 뿐 아니라 네트워크 정보를 계량화하여 분석 활용 가능

○ 소셜 빅데이터는 사회적 현상과 이슈의 변화를 실시간성 데이터 증거로 파악할 수 있어 대한 객관성과 최신성이 높은 장점 보유

– 그러나 SNS 기업의 데이터 개방 정책 변화, SNS 서비스의 다양화로 미래전략에 활용 가능한 신뢰성 높은 데이터는 점차 제한되는 추세

* 최근에는 밴드, 인스타그램 등 소규모, 프라이빗 서비스로 사용자가 이동 중이어서 데이터 확보가 어려운 경우도 있음

< 노드엑셀(NodeXL)의 트위터 API를 활용한 데이터 수집 예시 >



▸ 노드엑셀은 별도의 프로그래밍 없이 엑셀의 형태로 소셜네트워크 분석을 할 수 있는 소프트웨어

▸ 트위터, 페이스북, 유튜브 등 SNS 및 소셜미디어에서 제공하는 Open API를 활용해야 해당 데이터를 임포트하여 분석할 수 있는 기능을 제공

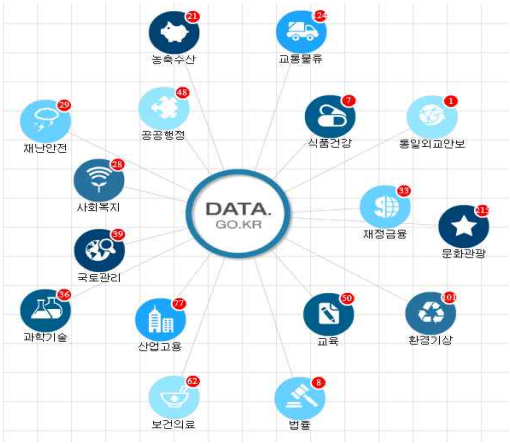
▸ 검색어 또는 특정 유저를 지정하여 트윗 기초정보와 리트윗(RT), 멘션한 유저들의 정보를 수집하여 관계망 데이터를 형성

□ 정부와 공공기관에서 제공하는 공공데이터

○ 행정기관 및 공공기관 등이 생성하고 관리하고 있는 데이터 중 공공의 목적으로 개방한 데이터로, 전통적인 통계데이터와 빅데이터를 포함

- 공공데이터포털(www.data.go.kr)에서는 각 기관에서 제공하는 공공데이터를 한 곳에서 통합하여 제공
- 1만4천여개의 파일데이터와 1,800여개의 오픈API 제공 중(지속적 확대 예정)

< 공공데이터포털 활용 분야 >



자료 : www.data.go.kr

□ 연구소 등에서 제공하는 통계 및 DB 자료

○ 분야별 전문연구기관에서 제공하는 통계나 DB 자료는 사회 변화와 기술의 발전을 파악할 수 있는 중요한 단서를 제공

- KISTI MIRIAN : 미래기술 지식베이스, 미래기술 디렉토리 등 미래 유망기술에 대하여 논문 통계, 기술전망, 관련 기사 동향 등을 제공[5)]
- KISTI KSCI : 국내 학술지에 대한 국내외에서의 인용정보 제공[6)]
- MediSys(EI, Joint Research Center)[7)] : 250개의 의학전문 사이트와 1,600개의 일반뉴스, 유럽 지역 20여개 상업 뉴스에서 의학 및 공공 건강 분야 미디어 데이터 제공
- EMM NewsExplorer(EU, FP-7) : 19개의 언어를 통해 일 단위로 뉴스를 요약해주며, 사람/기관/국가에 관한 정보를 군집화해서 연결, 시간상 변화하는 트렌드 분석 서비스 제공

5) http://mirian.kisti.re.kr/main.jsp
6) http://ksci.kisti.re.kr/main/main.ksci
7) 한국전자통신연구원, '소셜 빅데이터 이슈 탐지 및 예측분석 기술 동향', 전자통신동향분석 제28권 제1호, 2013년 2월

□ 검색포털에서 제공하는 웹 활동 데이터

○ 구글, 네이버는 자사의 검색엔진을 기반으로 웹 사용자의 검색 로그를 저장하여 Open API(Application Program Interface) 형태로 정보 추출 서비스를 제공

– 구글은 각 개인 관심사와 구글 검색 로그 기반 전 세계인 관심사와의 비교가 가능하며, 사용자의 관심 주제의 시간상의 변화 관찰, 지역별 사용자 관심 추이 관찰이 가능

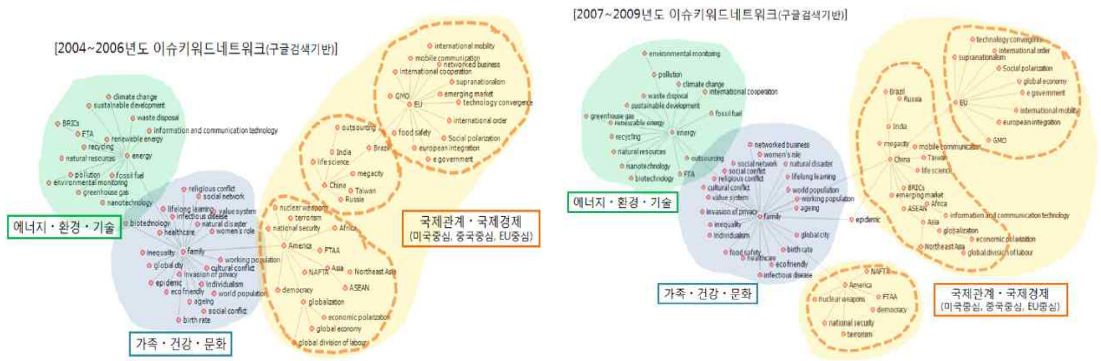
– 특정 주제를 형성하는 웹사이트 간의 네트워크 분석이 필요할 경우, 구글의 페이지랭크(Pagerank) 알고리즘을 이용하여 웹사이트 간의 연결(하이퍼링크) 정도를 수치화하여 네트워크 분석이 가능

※ 단어(키워드) 간 소셜네트워크분석에도 구글의 검색엔진이 반환하는 문서량(또는 검색량)의 숫자를 이용하여 노드 간 가중치 계산 가능

○ 최근 네이버에서도 10년여간 축적된 빅데이터를 공개하여 융합분석, 지역별 통계, 검색어 기반 트렌드 분석 기능을 제공 예정[8)]

참 고 **구글 검색엔진 기반 이슈 네트워크 분석 사례[9)]**

- 미래사회 이슈키워드 들의 동시 출현 결과(AND검색 결과 문서수)를 이용한 네트워크 도출 및 시각화 작업 수행
- 국제관계·국제정세 관련 키워드 그룹의 시간적 변화 양상을 파악h



[2007~2009년] [2004~2006년]

8) '데이터랩' 이라는 베타서비스로 제공 중(http://datalab.naver.com/)
9) 양혜영, '빅데이터를 활용한 기술기획 방법론', KISTEP 이슈페이퍼, 2012-14

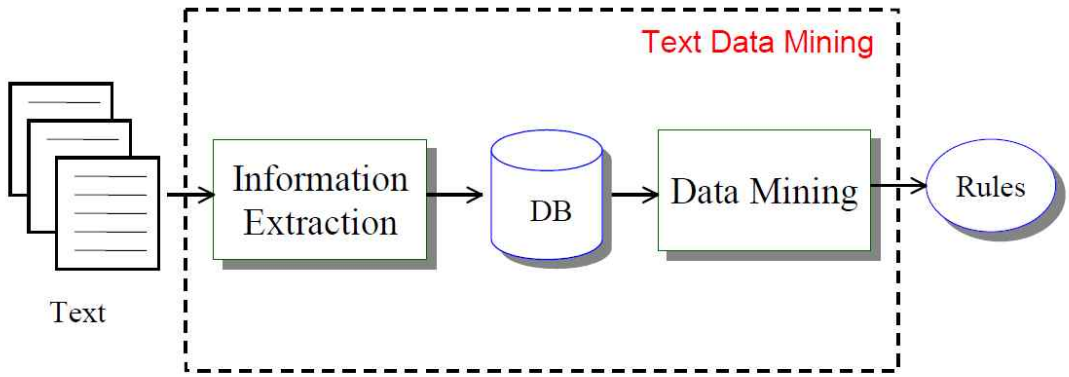
3. 미래전략에 활용되는 데이터 분석 기법

□ 텍스트 마이닝(Text Mining)

○ 대용량의 텍스트 데이터에서 숨겨진 맥락(Context)을 발견하고 의미를 찾아내는 기법

– 인터넷에 있는 빅데이터의 대부분(90%이상)은 텍스트, 동영상 형태의 비정형, 비구조화된 데이터의 형태로 존재

– 대용량 텍스트에서 의미있는 패턴을 발견하기 위해서 전처리 과정이 필요하며 데이터베이스 구조화 및 수학적 모델링(알고리즘)을 적용



〈 정보추출 기반의 텍스트 마이닝 프레임워크(Nahm & Mooney, 2002) 〉

○ 분석 목적과 데이터의 조건, 연구 환경에 따라 다양한 정보 추출 방법과 전처리 및 분석 알고리즘이 적용되고 있으며, 지속적으로 발전 중

– 텍스트의 시간정보를 이용한 종단적 분석으로 시계열적 변화를 추적하여 트렌드 분석, 기술 예측 등 미래예측과 전망에 활용

– 최근에는 단순히 텍스트만이 아니라 그림, 음성파일, 영상 등 멀티미디어의 비텍스트형 데이터도 구조화된 처리를 통해 활용 중

< 빅데이터 기반 텍스트 마이닝의 정보추출 방법 >

▸ 키워드 추출(Keywoed/Keyphrase Extraction)

- 텍스트마이닝의 다양한 분석에 활용되는 기본적인 출발점으로 문서의 핵심어를 추출하거나 문서 사이의 비슷한 정도를 구하여 핵심 키워드를 도출하는 과정
- 다양한 키워드 추출 알고리즘(Keyphrase Extraction Algorithm, KEA)을 사용
- 도출된 키워드는 상호 연관성을 고려하여 군집화(클러스터링) 하여 네트워크 분석을 하거나 등장 빈도 등을 고려한 가중치를 산정하여 워드 클라우드 등의 형태로 시각화 하여 직관적으로 표현

▸ 역문헌 빈도수(TF-IDF, Term Frequency-Inverse Document Frequency)[10)]

- 어떤 단어가 수집된 여러 문서에 동시에 출현할 경우 중요한 의미를 가질 것이라는 전제로 단어들의 중요도(가중치)를 계산하는 기법
- 분석 대상 문서의 핵심어(키워드)를 추출하거나 검색 엔진에서 검색어 순위, 문서 사이의 유사도 검증 등에 사용됨
- TF(Term Frequency)는 특정 단어가 분석 대상 문서 집합 내에서의 등장한 빈도수이며, DF(Document Frequency)는 특정 단어가 등장한 문서의 숫자를 의미, IDF(Inverse Document Frequency)는 DF의 역수로 TF와 IDF의 곱이 특정 단어의 전체적인 중요도를 계산하는 가충치(중요도 지수)를 산출

▸ 토픽 모델링(Topic Modeling)

- 단어들 간의 동시출현 빈도와 거리 등을 이용하여 유사한 의미를 가진 단어들을 클러스터링 하는 방식으로 분석 대상 텍스트의 주제(맥락)를 파악하는 기법(정다미, 2013)
- 빅데이터 처리기술의 발전에 따라 대량의 텍스트에서 형성된 주제와 키워드의 그룹을 효율적으로 추출할 수 있게 되면서 자주 사용되고 있음
- 텍스트 내에 잠재된(Latent) 의미를 추출하는 방법으로 확률 모형이 주로 사용되며 LSA(Latent Semantic Analysis), PLSA(Probabilistic Latent Semantic Analysis), LDA(Latent Dirichlet Analysis) 등의 알고리즘이 개발되어 있음
- 어떠한 알고리즘이 가장 우수한지는 학계에서 여전히 논쟁 중이며, 최근까지도 보완, 발전되고 있음

□ 소셜 네트워크 분석(Social Network Analysis)

○ 사회를 구성하는 객체들의 '관계' (사람과 사람, 정보와 정보 등)와 '상호작용' 을 계량적으로 분석하여 의미를 찾아내는 기법

– 네트워크를 구성하는 노드(점)과 링크(선)를 데이터로 구조화하여 관계의 형성, 구조, 변화를 추적

○ 네트워크를 형성하는 모든 유형의 사회적 현상이 분석 대상이 되며 행위자 들의 상호작용 분석을 통해 특이점과 의미를 발견

– 조직 내 부서·업무 간 상호작용 네트워크, 문헌 내 키워드 간 상호작용 네트워크, 소셜 미디어의 사용자(인물), 금융 거래, 문헌 등

※ 미국 워싱턴주 노동산업부는 산재보험 사기행각 적발, LA카운티는 보육서비스업체의 사기 탐지에 소셜네트워크 분석 기법을 적용

○ 특히 빅데이터분석 기술의 발달과 함께 대용량 문헌에서 키워드 간 관계를 계량적으로 분석하여 의미구조를 파악하는 방법인 사회의미망 분석(Semantic Network Analysis) 기법으로 발달

– 특정 사회적 이슈에 대하여 SNS, 뉴스 등에 나타나는 다양한 의견들을 수집하여 객관적인 방법으로 여론을 파악할 수 있음

– 키워드 추출 방식의 텍스트 마이닝 기법과 결합하여 미래 이슈들의 상호 연관관계와 관계의 시간상의 변화 추적이 가능

○ 빅데이터 기반 네트워크 분석은 객관적인 방법으로 이슈의 내용과 구조를 파악할 수 있어 기존의 정성적인 미래예측 방법을 보완하고 객관성을 부여

– 그러나 방대하고 비구조화된 텍스트데이터를 네트워크 구조로 파악하기 위한 데이터 수집 · 처리 인프라와 분석 알고리즘에 대한 투자가 필요

10) Salton G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill

□ 계량적 분석 방법론

○ 계량경제학(Econometrics) 기반 예측 기법

– 경제현상을 모델링하기 위한 정량적 분석기법에서 출발, 분석대상(종속변수)의 미래 상태(또는 가치)를 예측하거나 영향을 미치는 요인(독립변수)의 크기를 비교하는 경우 주로 사용

* 주로 회귀모델(Regression model)을 기반으로 경제학 뿐 아니라 고객 분석, 제조, 기후변화 등 정량적 예측이 필요한 모든 분야에 적용 가능

– 미래예측에는 시계열(time-series)자료를 이용한 확률적 추정방법이 사용되며 예측 뿐 아니라 외부 충격에 의한 반응을 장기적으로 파악할 수 있음

* 단순 예측 뿐 아니라 정책평가, 거시경제 및 외부 환경 변화에 따른 영향이나 독립변수들의 민감도 분석에도 유용

– 다양한 시나리오가 예상되는 불확실한 상황에서 개별 시나리오에 대한 영향을 비교하여 대응전략을 수립하는 데도 활용

* 예) 글로벌 경제위기로 인한 우리나라 경제영향의 방향이 명확하지 않은 경우 몇가지 시나리오를 설정, 각각의 영향을 선제적으로 파악하여 대응방안 마련

참 고 **계량경제학 모델링을 이용한 예측적 분석 사례**

◗ 기상기후 데이터 분석을 통한 과학적 농업경영 지원

- 한국정보화진흥원, 기상청, 농촌진흥청, 농촌경제연구원 등이 2014년 데이터 기반 미래전략 컨설팅 과제로 수행
- 과거 14년간의 기상기후 데이터와 같은기간 양파, 고추, 마늘 생산량과의 연관성을 분석하여 생산량 예측의 핵심 변수 도출

※ 이 모델을 활용하여 '14년도의 양파 생산량을 예측 한 결과, 실제 생산량과 6.48% 오차 범위로 신뢰할만한 결과를 도출

〈 (예시)양파 생산량 예측 모형 〉

$$Y=7303 - 12.12*RAIN\_DAY11 - 16.35*RAIN\_DAY1 - 78.30*ID1 + 116.22*txq90p5 + 127.71*\text{지역더미(경북)} - 38.12*\text{지역더미(전남)} - 675.69*\text{지역더미(전북)} - 852.19*\text{지역더미(충남)} - 988.47*\text{지역더미(충북)}$$

○ 계량정보학(Informetrics)[11] 기반 분석 방법론

– 정보(Information)에 대한 정량적 분석 기법을 통칭, (모든 형태의) 정보 생성, 사용, 유통, 흐름 과정을 계량화하여 분석하는 방법론

– 분석 대상 정보의 특성에 따라 세부 분야가 생성

▸ 계량서지학(Biblioometrics) : 학술문헌의 정량적 분석을 통한 지식구조분석, 효율적인 문헌관리 등

▸ 계량과학정보학(Scientometrics) : 특허나 논문정보를 활용하여 미래 연구분야 탐색, 유망기술 예측 등

▸ 웹계량정보학(Webometrics) : 웹에 있는 정보(콘텐츠)들의 속성을 이용한 기술, 사회, 경제 등 각 분야별 정보구조의 변화 및 추세 파악(Thelwall, Vaughan & Björneborn, 2005).

– 데이터베이스화 된 빅데이터에 대한 모델링 과정이 필요하며 네트워크분석 등 다른 정량적 분석과 결합하여 미래예측에 활용

참 고 **미래유망기술 예측에 사용된 계량정보학 분석 기법**

- 사례: KISTI는 계량정보분석 기법과 원내외 전문가 검증기법을 융합한 Hybrid형 발굴 프로세스를 통해 매년 미래 유망기술을 발굴
- 내용분석 : 국제표준분류체계(IPC)를 통한 특허 분류 및 융합 동향 분석
- 인용분석 : 논문 및 특허의 인용관계 분석(Citation Analysis)을 통해 10년간 분야별 피인용 상위 논문 검색 및 유망기술 후보영역 발굴
- 텍스트마이닝 : 키워드 클러스터링을 통해 미래유망 후보기술을 도출(100여개)하고, 전문가 검증을 통해 30~40개로 압축, 동향분석을 통해 최종 10개의 유망기술 선정
- 기술 선정과정에서 전문가의 주관에만 의존하지 않고 특정 특허나 기술의 유망성지수, 융합지수 등 데이터를 정량화하여 객관적으로 비교한 점이 전통적인 기술예측 방법과 차별화됨

11) 최초 용어는 다음 문헌에서 제창 : Bar-Ilan, Judit (2008). "Informetrics at the beginning of the 21st century: A review". Journal of Informetrics 2 (1): 1-52. doi:10.1016/j.joi.2007.11.001.

Ⅱ 사례로 보는 빅데이터 기반 미래전략의 가능성
- STEEP 각 분야를 중심으로 -

1. Society : SNS 분석을 통한 사회 위험요인 예측

□ 소셜 빅데이터를 활용한 청소년 자살 요인 예측[12)]

○ 우리나라 자살율은 OECD 국가중 최고 수준이며, 특히 청소년의 자살 문제가 사회적 이슈로 대두되며 정부의 적극적인 대책이 시급한 상황

○ 대책 마련의 실마리를 찾가 위해 SNS에 나타나는 감정이나 심리상태를 분석

- 청소년들이 SNS에서 표현하는 우울한 감정이나 스트레스, 고민 관련 글들을 수집하여 행태를 분석하여 위험 징후의 패턴을 감지
- 자살과 관련된 사회적 인식(buzz: 입소문)을 통해 개인의 극단적 선택과 사회적 현상과 상호 연관성을 통해 사회적 위험도를 파악

○ SNS 데이터에 대한 텍스트마이닝을통해 청소년 자살위험 예측모형을 제시

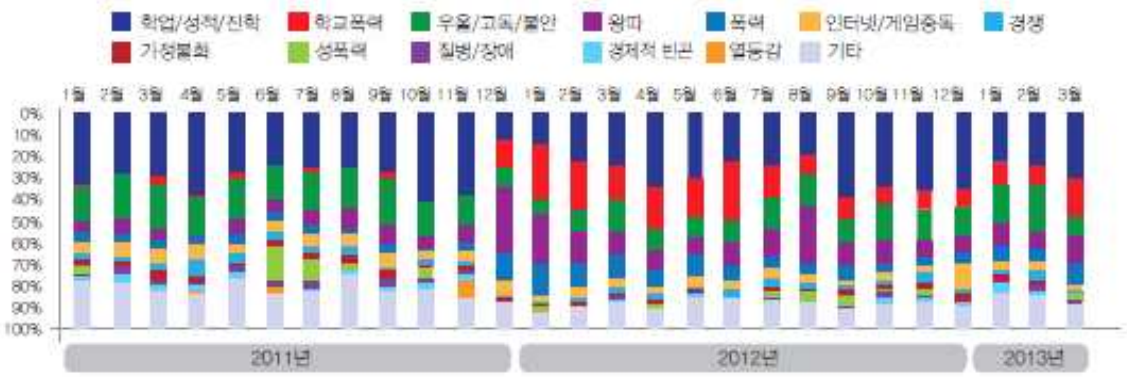
- 2011.1.1.~2013.3.31.(821일)동안 SNS에 나타난 자살 관련 소셜 빅데이터를 수집, 청소년 자살의 심리적 요인을 탐색

□ SNS에 나타난 청소년 자살 버즈를 통해 자살의 다양한 원인을 규명

○ '학업/성적/진학' 이 청소년 자살과 가장 자주 함께 언급

- 2011년 12월 이후 '학교폭력' 과 '왕따' 가 주요 청소년 자살 버즈 원인으로 지속 등장
- 2012년 통계청 사회조사에서도 13~19세 청소년 기준 '학교성적/진학문제' 가 39.2%로 자살충동 이유 1위로 조사됨
- 자살로 이어지는 심리적 요인은 '우울/고독/불안' 으로 지목

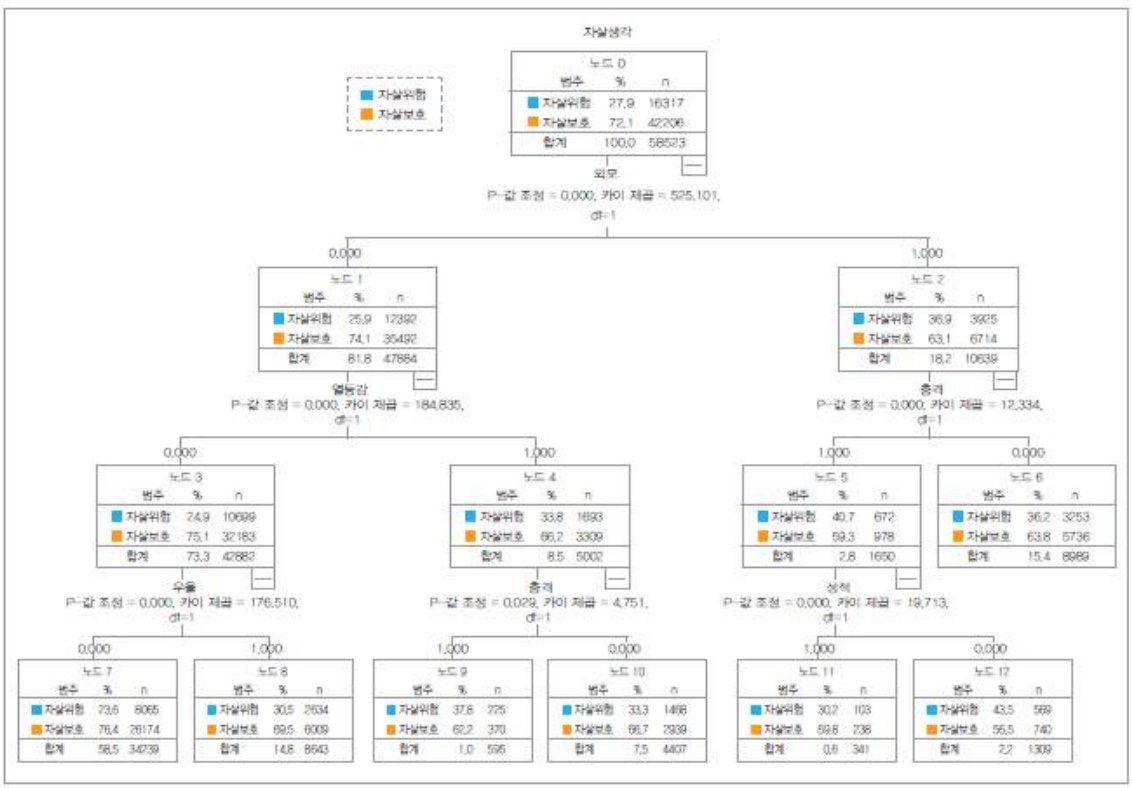
12) 송태민, '소셜 빅데이터를 활용한 사회위험 요인 예측: 청소년 자살과 사이버따돌림을 중심으로', 한국보건사회연구원 보건·복지 Issud & Focus, 제 238호 (2014-17)



< 자살 충동 요인의 시계열적 변화 >

○ 청소년자살 위험 예측에 영향력이 높은 요인으로 '외모요인', '열등감 요인', '충격요인'을 도출, 의사결정나무(Decision Tree)[13]로 시각화

* 예) '외모요인'의 위험이 높은 경우 자살 위험은 27.9%→36.9%로, '외모요인'이 높고 '충격요인'이 높으면 36.7%→40.7%로 증가

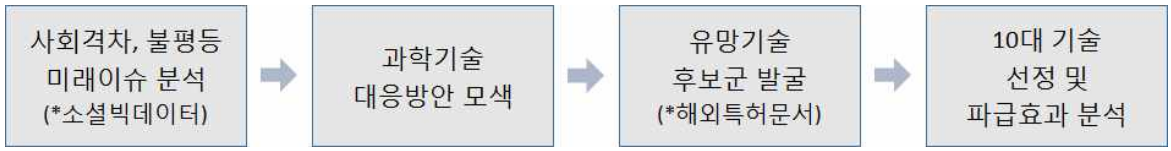


< 자살 충동 요인이 자살 위험도에 미치는 영향 분석 결과 >

13) 불확실한 상황에서 어떤 대안이 선택될 것인지를 확률적 사건으로 정의하고 나뭇가지처럼 그려 놓고 분석하는 기법으로, 최근에는 데이터마이닝을 통한 통계적 자료 분석에 활용됨

2. Technology : 미래사회 현안 해결을 위한 유망기술 도출

□ 미래사회 현안과 과학기술의 대응방안을 데이터 증거에 기반하여 탐색[14)]



< 유망기술 도출 절차 >

○ 소셜 데이터 및 뉴스 키워드를 분석, 문헌과 전문가 검토를 통해 우리 사회의 격차와 불평등과 관련된 미래 이슈를 도출

– 2005년부터 2014년까지 블로그/카페 문서, 트위터, 뉴스를 수집, 형태소 분석과정을 거쳐 정제한 후 토픽모델링 기법을 적용, 연관키워드 도출

– 키워드 대상 문헌 조사를 통해 미래이슈를 정리, 인문·사회 분야 전문가의 의견을 거쳐 향후 이슈화될 세부 내용 정리

< 사회적 격차와 불평등 관련 미래 이슈 도출 결과 >

격차 분야	미래 이슈
의료 격차	‣ 소득계층별로 의료서비스 이용격차 증가 ‣ 의료기관 및 인력의 지역간 불균등 분포 심화 ‣ 건강불평등의 세대간 대물림 현상 발생
정보 격차	‣ 모바일 기반의 새로운 정보격차(digital divide) 증가 ‣ 장노년층의 정보소외 현상(digital exclusion 증가 ‣ 참여(participation) 격차, 활용 격차 등의 증가
에너지 격차	‣ 에너지비용으로 인한 에너지 빈곤층 발생 ‣ 이상기후로 인한 에너지격차 심화
문화/교육 격차	‣ 소득계층 간 문화생활의 격차 발생 ‣ 도시지역과 농촌지역 간 문화예술 관람 격차 뚜렷 ‣ 소득계층 간 교육환경의 격차 발생

14) 한국과학기술기획평가원, “2015년 KISTEP 10대 미래유망기술 선정에 관한 연구”, 2015.2

○ 도출된 격차 및 불평등 관련 미래 이슈에 대응하는 유망기술 도출

– 특허문서 텍스트마이닝 : 과학기술 니즈 키워드를 활용하여 특허 검색, 특허 문서에서 키워드 추출 및 클러스터링 수행

– 미래유망기술 DB를 활용 : 국내외 기관에서 선정·발표한 미래기술과 KISTEP 보유 미래기술 DB에서 후보기술 도출

– 기술 전문가의 신규아이디어 제시 : 전문가 자문을 통해 앞서 도출된 유망기술 후보군검토와 더불어 격차, 불평등 증가의 이슈와 니즈 충족을 위한 미래기술 추가 선정

○ 분석 결과를 종합, 미래사회 격차·불평등 증가 이슈에 대응하여 모두가 혜택을 누릴 수 있는 포용적인 미래 유망기술 도출

< KISTEP 10대 미래유망기술 선정 결과(안) >

번 호	미래 유망기술
1	스마트폰 이용 진단기기
2	의료 빅데이터 기술
3	바이오스탬프(신체부착 센서)
4	Li-Fi 기술
5	가상촉감 기술
6	비콘 기술
7	진공단열물질 기술
8	에너지하베스팅 나노소재기술
9	개인맞춤형 스마트러닝
10	실감공간 구현 기술

3. Economy : 소셜데이터를 활용한 경기 예측 및 모니터링

□ 사례 1 : 소셜미디어 분석을 통한 식량 가격 단기 예측[15]

○ UN Global Pulse 에서는 세계식량계획(UN World Food Programme, WFP), 인도네시아 통계국과의 협력으로 트위터에서 추출된 식품가격 데이터와 공식 가격 자료를 비교하는 연구를 수행

– 시장에서 실시간으로 변화하는 식품들의 가격을 추적하여 정책에 활용하기에는 기존 방법으로 데이터의 취합과 분석이 어려운 상황

– 트위터리안의 다양한 버즈 중 쇼핑에 대한 트윗 메시지에서 가격 정보를 얻을 수 있음에 착안[16]

○ 공개된 트윗 데이터를 이용한 식품 가격 단기 예측(Nowcasting)[17] 모델 개발

– 2012년 6월부터 2013년 9월까지의 공개된 트윗 메시지를 수집

– 4대 주요 품목(소고기, 닭고기, 양파, 칠리)에 대해 키워드 기반 가격 추출 알고리즘을 개발

※ 언급량이 많은 가격정보는 신뢰도가 높도록 가중치 반영, 과거 공식 가격데이터를 통한 이상치 검증 절차 등을 반영하여 통계 처리

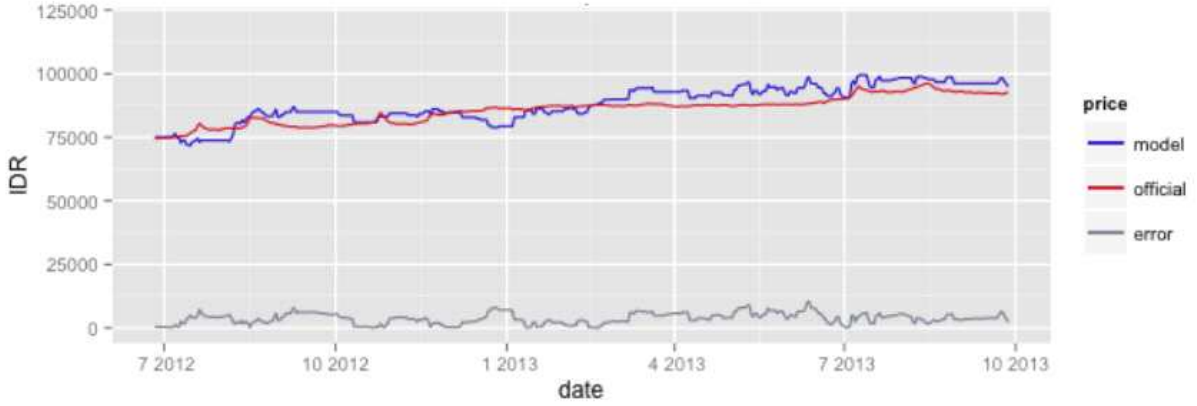
– 산출된 각 품목별 추정 가격은 동일기간 공식 물가통계자료와 비교를 통해 모델 검증

– 일부 구간 데이터가 부족한 양파를 제외하고는 세 개 품목 모두 통계적으로 유의한 상관관계를 보이고 있는 것으로 나타남

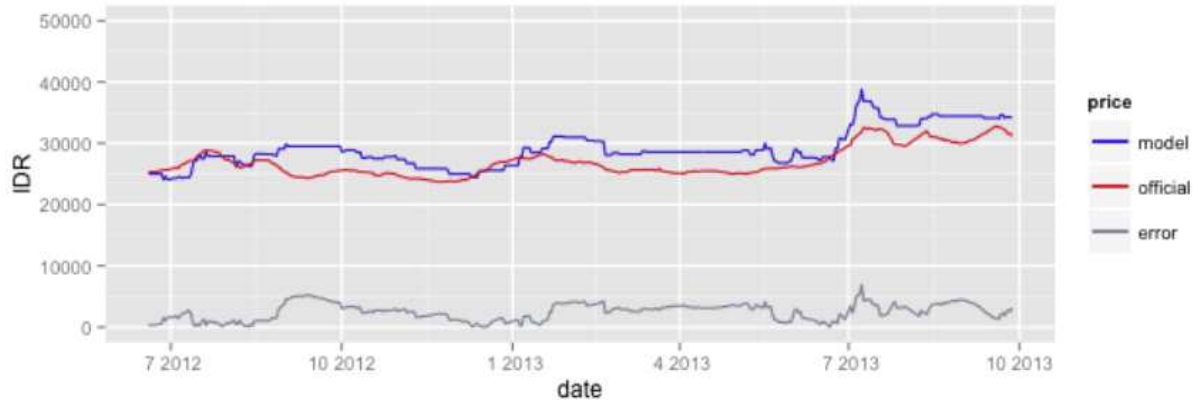
15) UN Global Pulse, 'Nowcasting Food prices in Indonesia using Social Media Signals', Global Pulse Project Series no. 1, 2014

16) 관련 원문 예시: “Pemerintah lagi sibuk! RT @promoasyik: Capai Rp 100 Ribuper Kg, Harga Cabai Rawit Merah Setara Daging http://bit.ly/1pUmUmo” (인도네시아어), “The government was busy! RT @promoasyik: Reaching Rp 100 thousand per kg, red chili is equivalent to the price of meat http://bit.ly/1pUmUmo”(영어)

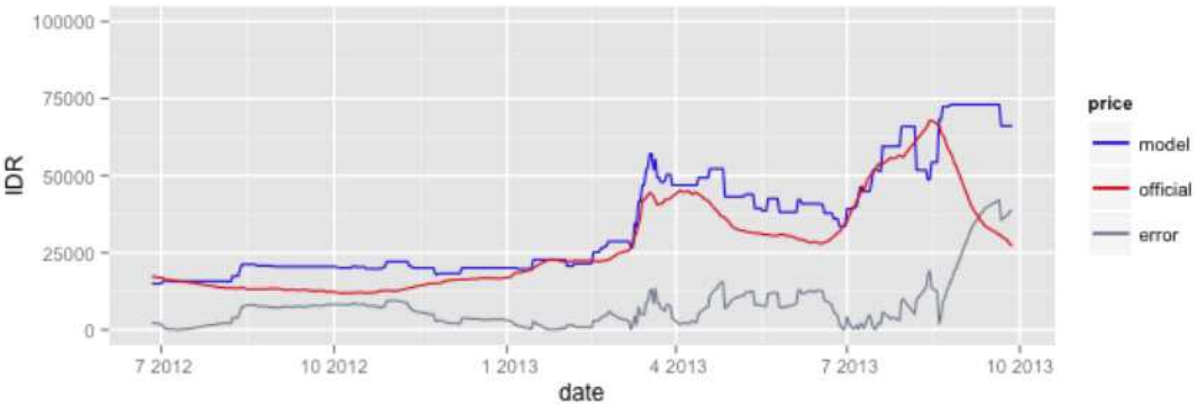
17) 특정 지역 대상 2~6시간 내외의 초단기 일기예보에서 유래한 용어로, 경제, 의료 분야 등에서 실시간성 빅데이터를 이용한 초단기예측 사례가 발표되면서 주목받기 시작한 개념



< 소고기 가격 예측 결과(14,473개 트윗에서 가격 언급) >



< 닭고기 가격 예측 결과(5,223개 트윗에서 가격 언급) >



< 양파 가격 예측 결과(1,954개 트윗에서 가격 언급) >

○ 도출된 모델과 수집체계를 시스템화 할 경우 실시간성으로 식품가격 지수를 추적할 수 있으며, 경제적 위험 관리나 정책 의사결정에 활용 가능

- 가격과 함께 언급된 내용에 대한 의미 분석까지 함께 할 경우 갑작스러운 가격변동의 원인이나 대책 마련에 필요한 정보도 파악 가능
- 모델이 정교화될수록 가격 급등이나 폭락에 대한 조기경보까지 가능

□ 사례 2 : 모바일폰 신용구매 데이터 기반 식량안보 예측[18]

○ 중앙아프리카 지역에서 쇼핑에서 일상적으로 사용되는 모바일폰 기반 신용구매(Airtime Credit Purchase, 일명 Top-up) 데이터에 주목

- 한 국가[19]를 대상으로 6개월 동안의 익명화된 모바일폰 사용이력(CDR, Call Detail Records)과 신용구매 이력 데이터를 취합
- 비교자료로는 UN의 조사에 기반한 식량소비지수(FCS, Food Consumption Score)와 식량부족 대응지수(CSI, Coping Strategy Index), 빈곤지수(Poverty Indext)를 활용

○ 분석 결과 모바일폰 사용데이터와 신용구매액에 기반한 지수가 식량부족 현황에 대한 대리지표(Proxy Indidcator)로 활용 가능함을 확인

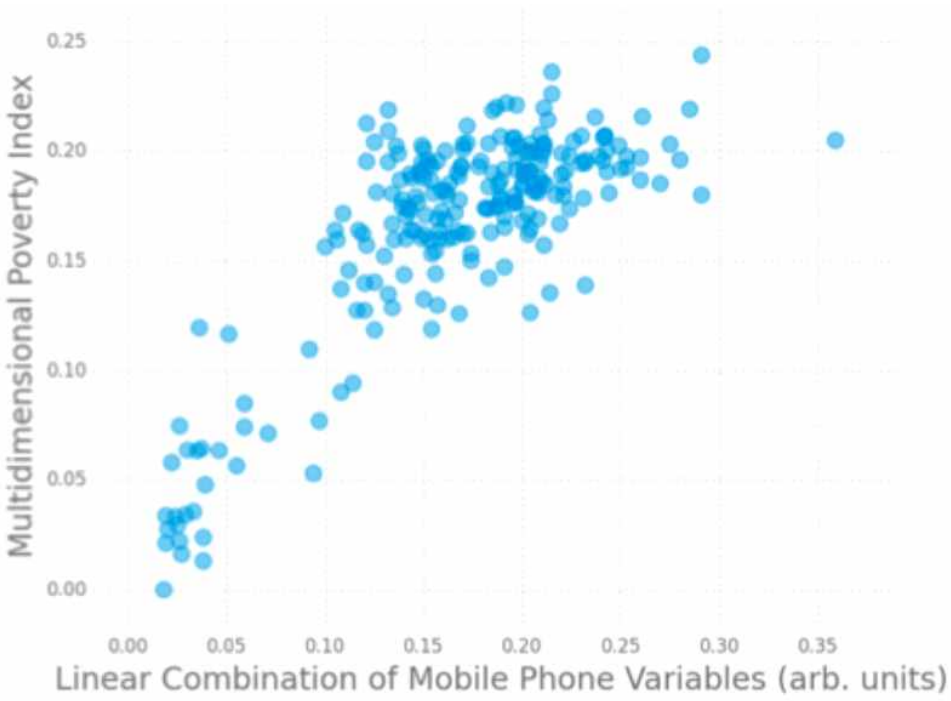
- 비타민 함유 야채, 쌀, 빵, 설탕, 고기 등이 기존의 지수 데이터와 모바일 데이터 간 가장 높은 상관관계(0.7~0.8)를 보임
- 감자, 수수, 땅콩, 생선, 우유 등 가정에서 직접 재배하거나 수확하는 항목은 낮은 상관관계를 보임

○ 설문조사가 어려운 지역에 대하여 비재무적 빈곤(Non-monetary poverty) 상황을 파악하고 실시간으로 모니터링할 수 있는 초기 타당성을 확인

18) UN Global Pulse, 'Using Mobile Phone Data and Airtime Credit Purchases to Estimate Food Security', Global Pulse Project Series no. 14, 2015.
19) 국가 실명은 보고서에서 미공개

FOOD ITEM (VARIABLE)	CORRELATION RANGE
Vitamin-rich vegetables (carrot, orange, sweet potato), rice, wheat, bread, sugar, meat	[0.7–0.8]
Eggs, oil, milk, butter, organ meat	[0.5–0.6]
Sorghum, ground nuts, seeds, fish, fruits, cooking banana, green leafy vegetables, beans, peas, maize, white roots, tubers, pumpkin, squash, cassava	[0.0–0.4]
White sweet potato	-0.4

< 식품 소비량(설문조사)과 모바일 신용구매액 간 상관관계 >



< 다차원 빈곤지수(UN)와 모바일 빈곤지수 간 상관관계 >

4. Environment : 복합적 미래예측 방법론 기반 재난 예측

□ 과학적 방법과 통계적 방법을 적용한 미래 재난 예측[20]

○ 국립재난안전연구원에서는 재난과 관련한 대응 전략을 도모하고 위험성을 완화시키기 위해 '미래재난예측 분석시스템' 프로토타입을 개발

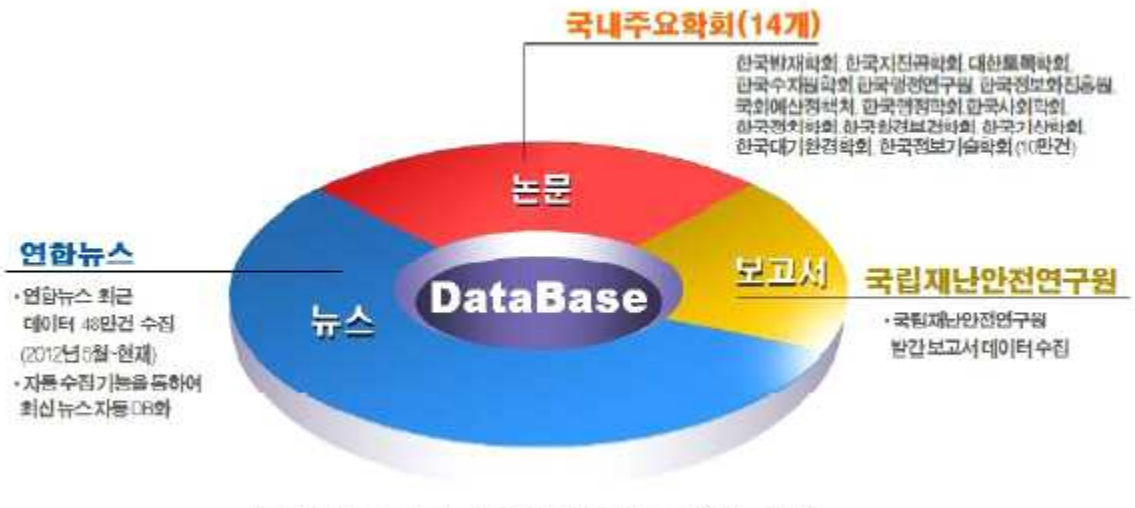
○ 단기적 예측과 중장기적 예측을 구분하여 분석 프레임워크를 제시

– 단기적 관점 : 트윗, 다음아고라 등 수시간~수일 동안의 자료를 이용하여 특정 재난 상황에 대한 신속한 분석을 실시, 재난 전조를 감지

* 재난예측 경보, 재난 진행상태 파악, 피해자 및 주변인 심리 파악 등

– 장기적 관점 : 수년 ~ 수십년 동안의 뉴스기사, 논문, 보고서 등의 자료를 정제하여 미래 재난 예측에 활용

* 재난 원인 기제 파악 및 What-if 시나리오 기법을 통해 재난의 진행 경로와 영향력을 파악



< 중장기적 미래 재난 예측을 위한 데이터 수집 현황 >

20) 국립재난안전연구원, "복합적 미래예측방법론 분석을 통한 미래 재난 예측기법 개발", 2013.12

□ 장기적 관점의 미래재난 예측 분석 시스템 프로토타입 개발

○ 뉴스자료와 논문, 주요 재난관련 보고서를 수집하여 전처리 및 통계 DB 생성을 자동화하는 분석시스템을 시범 구축

- STEEP(사회,기술,경제,환경,정치) 분류체계 별 환경스캐닝을 통해 미시환경에 영향을 미치는 광범위한 사회적 요인 파악

  * 세계경제포럼에서 발표한 「2012 글로벌 리스크 보고서」에서 제시한 50개의 리스크를 기준으로 글로벌 리스크가 국내 재난 환경과 어떠한 연관성일 가지고 있는지와 공통된 토픽이나 이슈를 가지고 있는지에 대한 분석을 실시

- STEEP 분야별 환경스캔은 자연재해 뿐 아니라 정치적, 경제적 리스크에 의한 재난 가능성까지 포괄적으로 포착하기 위함

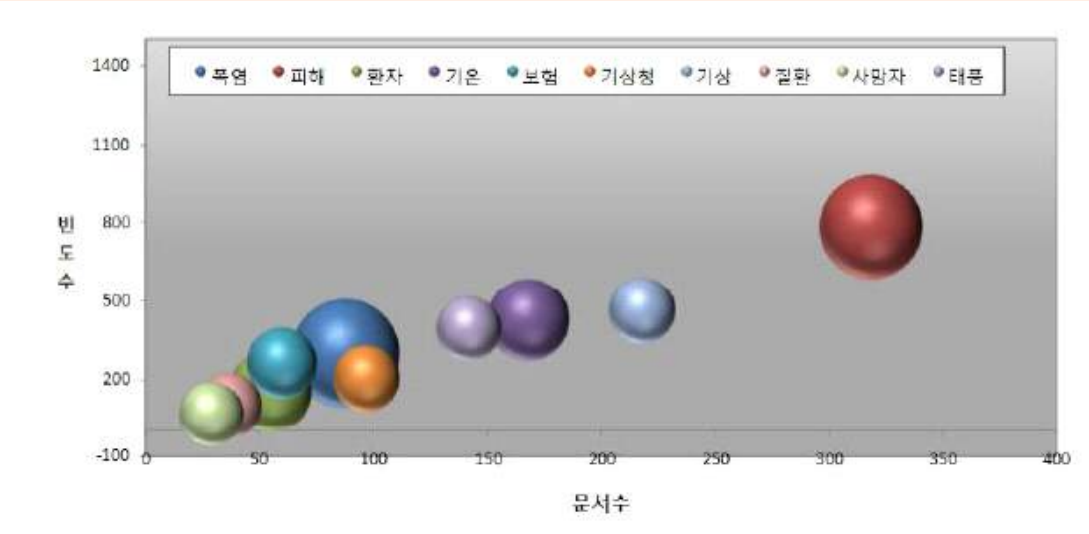
○ 환경분석 데이터 원천에 대한 분류체계(Taxonomy)를 체계화하여 재난재해 시나리오 선정을 위한 후보 이슈를 증거기반으로 찾아낸 것이 특징

< 발생가능성 및 파급효과를 고려한 핵심리스크 >

범 주	리스크	주요 내용
경제	안정적 재정 불균형	정부 부채 문제 해결 실패
	에너지/농산물 가격 변동성	극심한 가격변동으로 인한 경제위기, 대중의 시위를 촉발하거나 지정학적 긴장관계 조성
	소득불균형	최상위 계층과 최하위 계층 간 격차 확대
환경	기후변화 적응 실패	정부와 기업이 기후변화의 영향으로부터 시민을 보호하는 방안의 실패
	토양/수로 관리 실패	삼림벌채, 수로변경, 광물채굴 등이 생태계와 자연환경을 황폐화하는 수준으로 진전
지정학	글로벌 거버넌스 실패	취약하거나 부적절한 국제기관, 국제적 합의 혹은 네트워크가 각국의 이해관계와 충돌하여 협력이 어려운 상태
	테러	테러로 인한 대규모 인적/물적 피해 유발
사회	식량부족	양질의 식량 및 영양분의 불충분하고 불확실한 공급 상태
	수자원 공급 위기	식량 및 에너지 생산과 같이 자원 집약적 시스템 간의 수자원에 대한 경쟁 증가
기술	사이버 공격	범죄자, 테러리스트, 정부주도 또는 정부와 연계된 그룹이 자행하는 사이버 공격

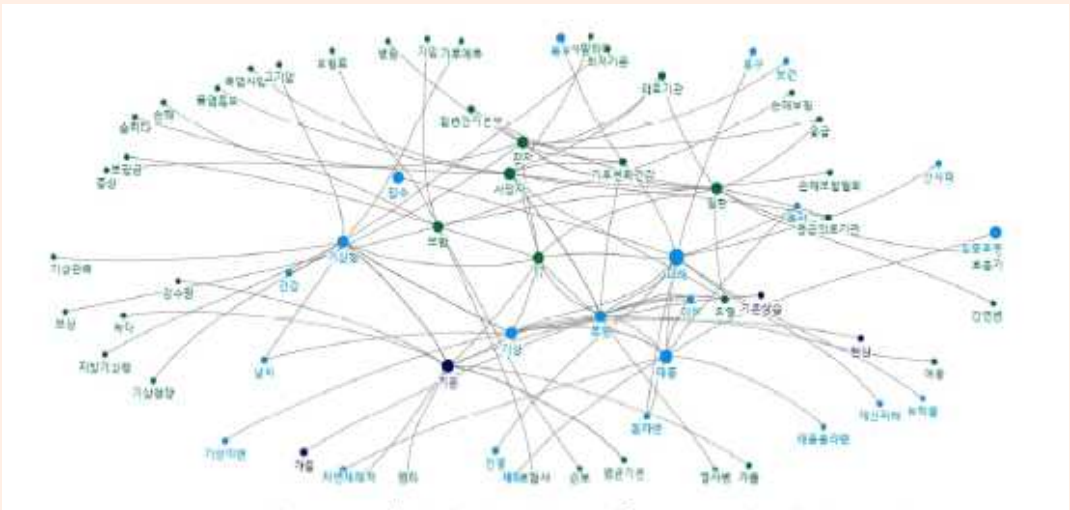
< 이상기후와 폭염 상세 예측 사례 >

▸ 구축된 DB를 이용하여 키워드 맵을 그리고 토픽을 추출, 기후변화 적응 실패와 수자원 공급위기 두 개의 리스크를 선정하고 각 리스크 별로 미래 발생가능성이 높은 이슈를 도출, DB로 검증



< '이상기후와 폭염' 토픽의 키워드 분포 시각화 >

▸ 폭염으로 인한 직접적인 인명피해와 가뭄, 전염병과 같은 2차 피해 가능성을 데이터 기반으로 파악



< '이상기후와 폭염' 토픽의 키워드 맵을 통한 인과분석 >

5. Politics : 뉴스미디어를 통한 정치 갈등 징후 탐지

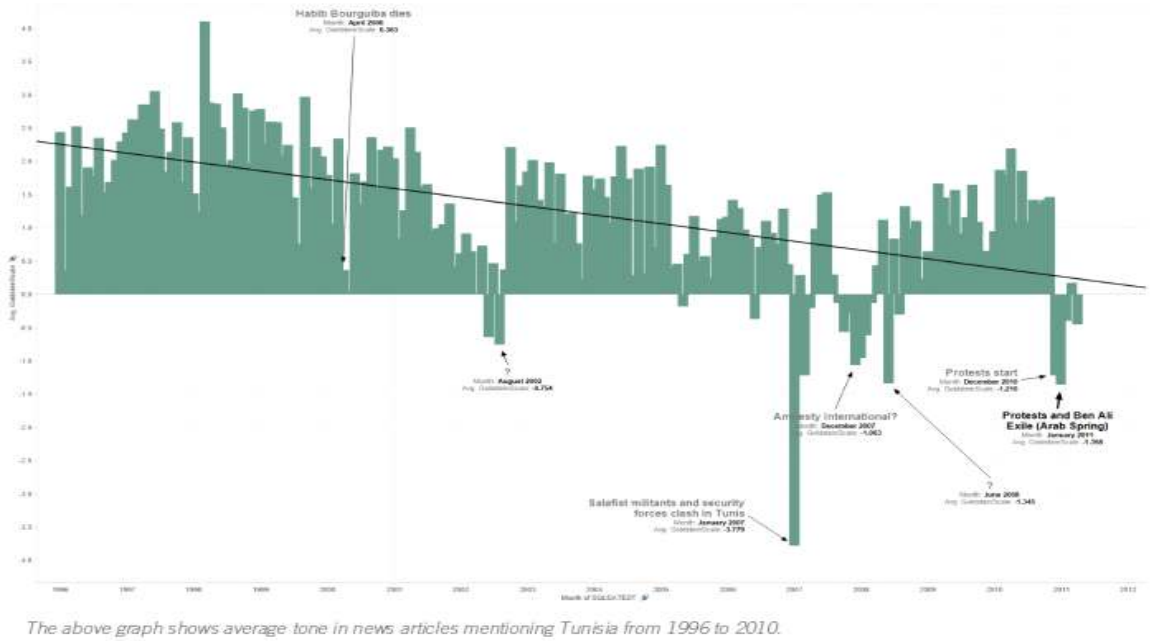
□ 소셜미디어의 감성(Sentiment) 변화를 추적, 정치적 갈등의 전개 양상 파악[21]

○ 현대사회에서 정치적 갈등은 비대칭성, 초지역성(Inter-communal), 범죄와의 연관성이 높아지고 있는 양상을 보임

– 특정한 사회적 이벤트에 대해 다양한 행위자들의 관계를 이해하고 패턴을 분석하는 것은 중요하지만 어려운 일

○ UN Global Pulse 는 UN 개발계획(UNDP, UN Development Programme)과 함께 뉴스 미디어 데이터 분석을 통해 갈등관리를 보조할 수 있는지 타당성 연구를 수행

– 1979년부터 2011년까지 구글이 수집한 250만 여개의 방송, 신문, 온라인 뉴스데이터 중 튀니지의 2011년 재스민 혁명[22]을 회고적으로 분석



The above graph shows average tone in news articles mentioning Tunisia from 1996 to 2010.

< 튀니지의 정치갈등지수[23]와 뉴스미디어 감성지수와의 트렌드 비교 >

21) UN Global Pulse, 'Feasibility Study: Analysing Large-Scale News Media Content for Early Warning of Conflict', Global Pulse Project Series, no.3, 2015.

22) 노점상을 하던 청년의 분신자살을 계기로 시위가 전국적으로 확산되어 당시 23년간 독재를 해오던 벤 알리 대통령을 하야시킨 민주화 혁명. 쿠데타가 아닌 민중봉기로 독재정권을 무너뜨린 첫 사례로 이후 이집트, 알제리, 등 독재정권에 시달리던 인근 아프리카 및 아랍 지역 국가로 확산되는 계기를 제공

23) Goldstein Scale: 군사위협 등 체제 간 갈등 상황을 -10 ~ + 10까지의 단계로 지수화 한 체계

○ 분석 결과 온라인 뉴스 데이터가 기존의 갈등 분석 및 조기경보 방법을 보완할 수 있음을 증명

– 분석결과 뉴스미디어에 나타난 부정적 감성의 급격한 증가는 이후 정치적 갈등의 기폭제(예. 분신자살, 대규모 시위)로 이어지는 패턴을 발견

* 몇가지 이상치(Outliers)를 제외하고는 전체 기간감성의 가장 큰 하락이 일어난 지 1개월 후 정권 교체 등 이벤트가 이어짐

*** 사례분석 시사점 : 빅데이터를 활용한 미래전략 수립 트렌드 변화**

▸ **소셜미디어(SNS, 온라인뉴스) 데이터의 활용도 부상**

- 초단기 예측부터 중장기 전망에 이르기까지 다양한 분석 시점 별로 미래예측에 활용되고 있음
- 비정형 분석 뿐 아니라 계량화를 통한 예측적 분석에도 활용 가능성이 검증되는 추세

▸ **융복합 데이터 분석**

- 미래예측과 전망에 필요한 데이터의 원천과 형태가 다양화, 대용량화되고 있으며, 활용 가능한 데이터의 범위가 점점 증가 중
- 하나의 원천에서만 데이터를 수집하지 않고, 상호 검증 가능한 여러 데이터를 교차 분석하여 분석 결과의 신뢰성을 높이는 방향으로 발전

▸ **통합적 방법론으로 진화**

- 기존의 조사방법론 기반 미래연구에 대한 패러다임 변화가 진행

* 정량분석의 경우 사실상 모집단 전체의 데이터 분석이 가능해져, 변수 간 상관성만 입증되면 미래예측에 바로 활용, 기존의 표본을 통한 모수추정, 인과관계 검증이 불필요

- 하나의 분석방법론이나 기법에 의지하지 않고 있으며 다양한 이론, 방법론, 데이터를 결합하는 삼각화(Triangulation)* 기반 분석이 확산 중

* 연구결과의 신뢰도와 타당도를 높이기 위해 연구 대상을 다각도에서 관측하고 검증하려는 사회과학 접근법

** 기존에는 자원의 제약으로 장기간 다수의 연구 프로젝트를 통해 삼각화 과정이 일어났다면, 빅데이터의 처리와 분석 효율성이 개선됨에 따라 하나의 연구프로젝트에서도 적용이 가능해짐

Ⅲ 증거기반 미래전략과 빅데이터 활용 방안

1. 증거기반 미래전략의 시작

□ 빅데이터, 분석을 넘어 미래를 준비하는 출발점

○ 미래에 펼쳐질 초연결 사회에서는 지금보다도 훨씬 많은 분석 가능한 데이터들이 매일매일 쏟아질 것

– 복잡하고 급변하는 정책 환경의 변화에서 기존의 경험과 직관에만 의존한 미래전략과 정책 대응에는 한계 존재[24)]

– 특히 다양한 데이터 원천으로부터 증거에 기반한 미래 이슈를 발굴하는 호라이즌 스캐닝에 핵심적인 역할을 수행할 수 있음

□ 데이터 기반 미래전략의 특징

○ 미래를 예측하고 전망하는 활동에 데이터분석을 접목하여 객관적, 과학적으로 문제 해결을 위한 대안과 전략을 마련하는 것

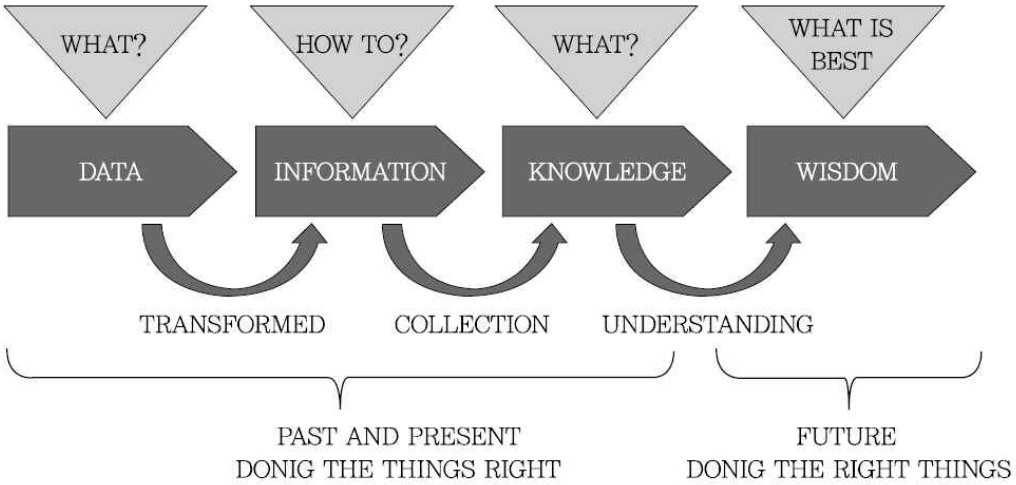
– 여기서 '데이터' 는 객관적 분석이 가능한 정형, 비정형의 모든 데이터를 의미하며, 빅데이터에 한정하지 않음

– 데이터분석의 목적은 현상에 대한 정확한 이해 뿐 아니라 미래 이슈의 동인까지 파악하여 미래 변화를 예측하고 대응방안을 도출하는 과정을 포괄

○ 미래전략은 미래 예측과 전망 뿐 아니라 전략적 대응방안을 모색하는 실천(Action) 지향적인 활동

24) 한국정보화진흥원, "데이터 증거기반의 과학적 정책 수립 방안", IT & Future Strategy 2015-6호 (2015.9)

○ 데이터 분석은 단순한 수치적 예측(Forecast) 뿐 아니라, 정보의 축적과 지식의 발견을 통해 사회를 정확히 읽고 바람직한 미래의 모습을 그리는 지혜(Wisdom)의 추구 과정



< 데이터에서 지혜로의 진화과정[25] >

– 지혜는 비결정론적(non-deterministic)이고 비확율적(nonprobablistic)이라는 점에서 지식보다 미래지향적인 개념

* '내일의 비올 확율이 20%' 라는 결정론적인 예측(지식)보다는 갈매기가 낮게 날면 비가 온다 "라는 속담(지혜)가 미래전략에 더 중요한 역할을 수행

– 빅데이터 시대 증거와 객관적인 합의에 의해 형성된 지식을 출발점으로 하여 도덕이나 윤리적인 이슈까지 포함한 바람직한 미래에 대한 공론화 필요

25) "From Data to Wisdom", (http://blogs.southworks.net/vfugante/2008/09/06/from-data-to-wisdom)

2. 데이터 기반 미래전략 프레임워크

□ 데이터 기반 미래전략의 유형

○ 데이터기반 미래전략은 개념적 정의 뿐 아니라 일관된 방법론과 절차가 정립되어 있지 않는 상황

– 빅데이터의 다양성 및 실시간성과 분석의 목적을 함께 고려한 미래전략의 유형 세분화와 각 유형별 특화된 방법론의 개발 필요

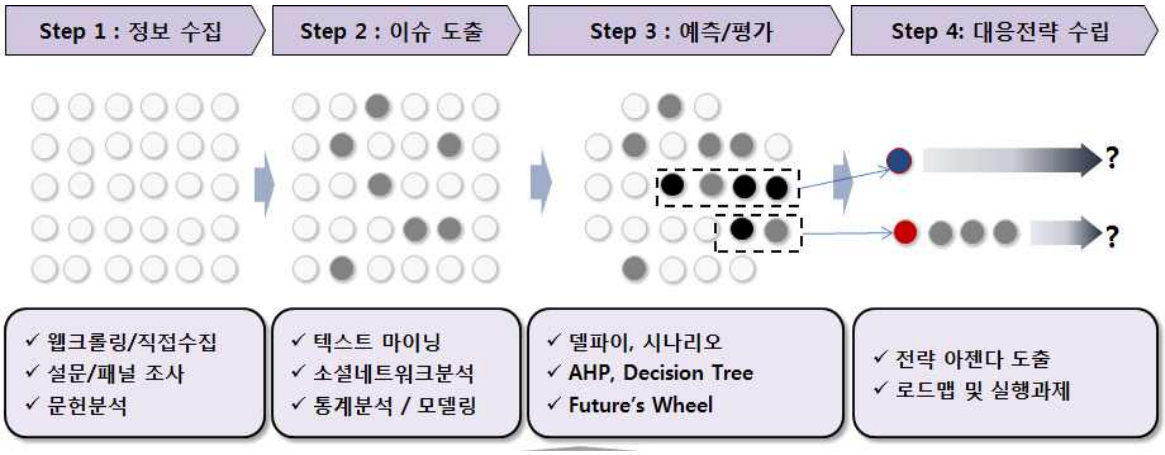
○ 본 보고서에서는 예측 시점(단기 vs 중장기)과 데이터 주제의 범위(특정 분야 vs 광범위)에 따라 4가지 미래전략의 유형을 제안

시점 / 범위	단기(1~3년) (프로젝트 단위 데이터 수집)	중장기(3~10년) (지속적 데이터 축적)
특정 분야 데이터	**< 문제 해결형 >** ‣ 현안이나 이슈의 세부적인 구조와 양상을 파악하고 해결안을 모색 (전통적인 예측적 분석) ‣ 다차원 분석이 가능한 상세하고 정확한 데이터 수집 ‣ 예) 소셜미디어 기반 자살 위험 예측 모형	**< 예측·대응형 >** ‣ 축적된 데이터를 기반으로 추세분석을 통한 이슈(위기) 예측 ‣ 변인 간 상호 역동성을 파악할 수 있는 종합적 분석역량 필요 ‣ 대응전략과 실행방안 도출에 필요한 지식화가 중요 ‣ 예) 기후변화/재난 예측
광범위 데이터	**< 조기경보형 >** ‣ 이상치 발견을 통한 비정상적인 이벤트, 사건들을 탐지 ‣ 데이터의 정확도 보다 실시간 분석/탐지 가능성이 중요 ‣ 예) 빅데이터 기반 물가 모니터링, 갈등 징후 포착	**< 아젠다 발굴형 >** ‣ 광범위 환경스캔을 통해 미래 트렌드와 이머징 이슈 전망 ‣ 전통적인 미래연구와 유사 ‣ 데이터의 해석과 전문가 집단 지성의 유기적 결합 필요 ‣ 예) 환경스캔 기반 미래사회 유망기술 도출

< 데이터기반 미래전략의 유형 >

○ 각 유형별로 수집 데이터의 특성과 분석 방법, 결과의 활용법이 달

라지며, 전략수립의 목적에 따라 복수의 유형을 통합·적용도 가능



< '아젠다 발굴형' 미래전략 수립 절차 예시 >

3. 데이터 중거기반 미래전략 수립 시 고려사항

□ 미래를 읽을 수 있는 정확한 데이터 원천의 확보

○ 인터넷 등 수많은 자료 속에 미래의 변화를 읽어낼 수 있는 정확한 데이터의 소재 파악과 유효한 데이터의 선정, 선별이 중요

– 목적의식이 없는 무분별한 데이터의 확보는 자원의 낭비를 초래

○ 분석에 방해되거나 그릇된 결론을 유도할 수 있는 노이즈 데이터의 식별과 정제에도 상당한 노력이 필요

– 이는 최신 데이터 처리기술의 지속적 활용과 더불어 분석가의 편의(Bias)를 배제할 수 있는 기술적, 방법론적 보완장치의 마련

○ 데이터의 신뢰성(Reliablility)과 타당성(Validity)은 여전히 중요한 가치[26]

☞ 국가 차원에서 신뢰성 있는 데이터가 개방 · 공유될 수 있도록 소재 정보의 통합과 표준화(국가 데이터 관리체계) 확대

26) 특히 소셜미디어를 분석에 활용할 경우 SNS를 사용하는 연령층은 여전히 20~30대가 다수인 점을 감안, 사용자 집단과 미디어의 특성이 분석 목적과 부합하고 대표성을 보장하는지에 주의해야 함

□ 데이터 만능주의 또는 데이터 증거 과신의 오류 주의

○ 데이터에만 충실한 일차원적인 분석은 때로는 사용자를 기만하는 결과를 도출할 수 있음

– 데이터분석에만 의지한 사실과 숫자에 집착할 경우 잘못된 데이터를 진실이라고 믿을 가능성이 높아짐

* 2004년 미국의 상원 의원 테드 케네디가 단순히 테러리스트 후보 명단과 이름이 같다는 이유로 탑승을 제지당하고 심문을 받은 일화가 대표적[27)]

– 빅데이터에만 의존하지 않고 분석 목적에 따라 인터뷰, 설문조사 등 기존의 데이터 수집방법을 적절히 결합도 필요

○ 알고리즘 보다는 인간의 본성과 사회적 의미의 해석에 주목

– 미래전략의 궁극적 목적은 정확한 알고리즘에 의한 예측의 정확도를 높이는 것이 아님을 유의

– 단순히 데이터가 제공할 수 없는 지혜와 미래 혜안의 영역은 전문가가 반드시 참여하는 집단지성이 필요

– 혁신적인 해결책(아이디어)는 데이터에 의존하지 않고 때로는 직관과 창의적 발상에서 나오는 경우도 많음에 유의

☞ 데이터분석 결과에 대하여 전문가 참여는 물론, 공공과 민간, 계층과 세대를 아우르는 대중의 참여를 미래전략 수립 과정에 반영

27) http://www.factcheck.org/2015/12/ted-kennedy-and-the-no-fly-list-myth/

참고 자료

[1] 국립재난안전연구원, 복합적 미래예측방법론 분석을 통한 미래 재난 예측기법 개발, 2013.12

[2] 김주환, '특허정보: 특허를 활용한 기술예측 방법론 (SNA 를 중심으로)' 고무기술 16.1 (2015): 23-38.

[3] 삼성경제연구소, '효과적 수요 예측 방법과 사례', SERI 이슈페이퍼, 2012.3

[4] 송태민, '소셜 빅데이터를 활용한 사회위험 요인 예측: 청소년 자살과 사이버따돌림을 중심으로', 한국보건사회연구원 보건·복지 Issud & Focus, 제 238호 (2014-17)

[5] 양혜영, '빅데이터를 활용한 기술기획 방법론', KISTEP 이슈페이퍼, 2012-14

[6] 정다미· 김재석· 김기남· 허종욱· 온병원· 강미정, '사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안' , 지능정보연구, 제19권 제3호 2013년 9월

[7] 한국과학기술기획평가원, '2015년 KISTEP 10대 미래유망기술 선정에 관한 연구', 2015.2

[8] 한국교육학술정보원, '테크놀로지와 미래교육에 대한 예측 방법과 해외 사례', 이슈리포트 2009.9

[9] 한국전자통신연구원, '소셜 빅데이터 이슈 탐지 및 예측분석 기술 동향 ' , 전자통신동향분석 제28권 제1호 , 2013년 2월

[10] 한국정보화진흥원, '성공적 공공정책 수립을 위한 미래전략연구방법론', IT & Future Strategy, 2010.4.30

[11] 한국정보화진흥원, '이슈 스캐닝(Horizon Scanning) 기반 국가 미래전략 수립방향', IT&Future Strategy, 2013. 5.

[12] 한국정보화진흥원, '선진국의 데이터기반 국가미래전략 추진현황과 시사점', IT&Future Strategy, 2012.4.6.

[13] 한국정보화진흥원, '데이터 증거기반의 과학적 정책 수립 방안', IT & Future Strategy 2015-6호(2015.9)

[14] Bar-Ilan, Judit (2008), 'Informetrics at the beginning of the 21st century: A review', Journal of Informetrics 2 (1): 1-52. doi:10.1016/j.joi.2007.11.001.

[15] Executive Office of the President, Office of Science and Technology Policy, OBAMA

ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE, March 29.2012

[16] Joshua S. Goldstein, ‘A Conflict-Cooperation Scale for WEIS Events Data’, Journal of Conflict Resolution, Vol. 36, No. 2 (Jun., 1992), pp. 369-385

[17] Nahm, Un Yong, and Raymond J. Mooney. ‘Text mining with information extraction’, AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Vol. 1. 2002.

[18] Salton G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.

[19] Thelwall, Mike, Vaughan, Liwen and Björneborn, Lennart, ‘Webometrics’, Annual Review of Information Science and Technology, 39-1

[20] UN Global Pulse, 'Using Mobile Phone Data and Airtime Credit Purchases to Estimate Food Security', Global Pulse Project Series no. 14, 2015.

[21] UN Global Pulse, ‘Feasibility Study: Analysing Large-Scale News Media Content for Early Warning of Conflict', Global Pulse Project Series, no.3, 2015.

[22] UN Global Pulse, 'Nowcasting Food prices in Indonesia using Social Media Signals', Global Pulse Project Series no. 1, 2014.

[23] 관련 웹사이트(Accessed : 2015-10~12)

- Effective Pattern Discovery for Text Mining IEEE Transactions on, Knowledge and Data Engineering, Volume: 24, Issue: 1
(http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5611523)
- A Mixture Model for Contextual Text Mining University of Illinois at Urbana-Champaign (http://sifaka.cs.uiuc.edu/czhai/pub/kdd06-mix.pdf)
- Text Mining with Information Extraction University of Texas, Austin (http://www.cs.utexas.edu/~ml/papers/discotex-melm-03.pdf)
- http://blogs.southworks.net/vfugante/2008/09/06/from-data-to-wisdom
- http://www.factcheck.org/2015/12/ted-kennedy-and-the-no-fly-list-myth/
- http://mirian.kisti.re.kr/main.jsp
- http://ksci.kisti.re.kr/main/main.ksci

IT & Future Strategy 보고서

- 제1호(2015. 3. 20), 「초연결 사회를 견인할 IoT 데이터화(Datafication) 전략」
- 제2호(2015. 4. 29), 「초연결 지능형 공장(CSF)을 통한 제조업 혁신 방안」
- 제3호(2015. 5. 29), 「IoT 융합 신산업 발전방향 및 정책 대응 방향」
- 제4호(2015. 6. 30), 「IoT 공통플랫폼 구축 및 활용 전략」
- 제5호(2015. 7. 15), 「新융합의 가능성과 저력, '인터넷 융합 경제(IConomy)'」
- 제6호(2015. 9. 15), 「데이터 증거기반(Evidence-Based)의 과학적 정책 수립 방안」
- 제7호(2015. 10. 15), 「사이버물리시스템(CPS) 기반의 사회시스템 최적화 전략」
- 제8호(2015. 10. 23), 「디지털 시대의 글로벌 키워드 : 사람·혁신·공존」
- 제9호(2015. 10. 26), 「시니어 경제활동 참여 확대를 위한 사회적 지식공유 모델 수립」
- 제10호(2015. 11. 20), 「초연결 기술, 날개를 달다 : 드론의 성장과 대응방향」
- 제11호(2015. 12. #), 「ICT 기반 Healthcare 서비스의 변화와 대응방향」
- 제12호(2015. 12. 11), 「2015 해외 신간도서로 읽는 미래사회」
- 제13호(2015. 12. 14), 「운송수단의 변화동인과 이슈분석」
- **제14호(2015.12.16.), 「빅데이터시대, 미래전략의 새로운 접근법」**

ISBN 978-89-8483-218-3