

빅데이터 분석 방법론

(youngsik.park@bsl-lausanne.ch)

강의에 들어가기 전에

데이터 분석은 기술만이 아닙니다.

수 많은 통계이론과 분석기술(R & Python)이 필요한 건 사실이지만, 이것이 전부는 아닙니다.

누구는 이상한 소리라고 할 수 있으나,

우리는 과학적기술(IT, Math)과 아티스트적 인문학 소양을 같이 길러야 합니다.

그래서 이 과정에서는 R과 기초적인 통계지식 및 사회과학적 접근법을,

여러분들께 말씀드릴까 합니다. 요즘과 같은 데이터 홍수 속에서 남다른 가치를 창출하고 그에 따른 인사이트(Insight)를 도출하는 사람을 이제 저는 '데이터 사이언티스트'가 아닌 '데이터 아티스트라' 부릅니다.

INDEX

- 1) 데이터란?
- 2) 데이터의 가치와 미래
- 3) 우리에게 데이터 분석이란?
- 4) 데이터 분석 기획 –이론과 사례

1) 데이터란?

◎ 데이터라는 용어

『데이터』: “ 추론과 추정의 근거를 이루는 사실 ” <OED, vol. IV, 264>

데이터는 객관적 사실 이라는 존재적 특성을 갖는 동시에 추론, 예측, 전망, 추정을 위한 근거로 기능하는 당위적 특성 / 수요조사, 실험, 검사, 측정, 마케팅 리포트, 경영전략, 정책의 기초

- “데이터”라는 용어는 1646년 영국문헌에 처음 등장
- 라틴어인 Dare(주다)의 과거분사형, 주어진 것
- 1940년대 이후, 경영학, 통계학 등 구체화 <한국데이터진흥원, 데이터의 이해 中>

정량적 데이터

정형 데이터
통계 분석
객관적 결론

VS

정성적 데이터

비정형 데이터
요약
주관적 결론

모델		구성	내용
	암묵지	사회화	경험을 통한 지식습득
		외부화	지식을 말이나 글로 표현
	형식지	종합화	새로운 형식지 창출, 조합
		내면화	형식지 이해, 습득

중요

정량 데이터 : 그 형태별로 언어, 문자 등으로 기술되는 데이터
정성 데이터: 수치, 기호, 도형으로 표시되는 데이터

1) 데이터란?

◎ 데이터 ?? 정보??: DIKW에 대한 이해

데이터(Data)는 개별 데이터 자체로는 ‘의미가 중요하지 않은 객관적인 사실’

정보(Information)는 **데이터**의 가공처리와 데이터간 연관관계에서 ‘도출된 의미’

지식(Knowledge)은 **데이터**를 통해 도출된 유의미한 정보를 분류 + 개인적 경험 =
‘고유지식’

지혜(Wisdom)은 **데이터**를 통해 만들어진 이러한 지식의 축적+아이디어= ‘창의적 산물’

<한국데이터진흥원, 데이터의 이



중요

[데이터]는 자체로는 의미가 없으나
모든 [정보], [지식], [지혜]를 생성하는데 핵심!

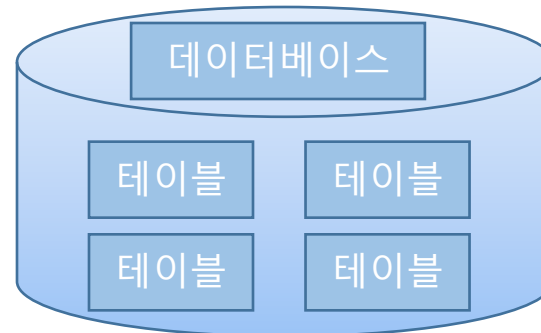
1) 데이터란?

◎ 데이터 관리의 필요성!!!

『 **데이터 베이스(Data Base)** 』: “체계적이거나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물”
<EU, 데이터 베이스의 법적 보호에 관한 지침.>

『 **데이터 베이스(Data Base)** 』: “관련된 레코드의 집합, 소프트웨어로는(DBMS: Database Management System)을 의미” <Wikipedia>

- 데이터 베이스는 통합된 데이터 이다.
- 데이터 베이스는 저장된 데이터 이다.
- 데이터 베이스는 공용 데이터이다.
- 데이터 베이스는 변화하는 데이터이다.



데이터 베이스 개념도식화

1) 데이터란?

◎ 기업내부에서 활용되는 데이터베이스

부문	DB구축형태
제조부분	ERP, SCM, DW, CRM,BI(Business Intelligence)
금융부분	EAI(Enterprise Applications Integration), ERP, e-CRM, EDW
유통부분	CRM, SCM,KMS,BSC,KPI

1) 데이터란?

◎ 사회기반구조로서의 데이터베이스

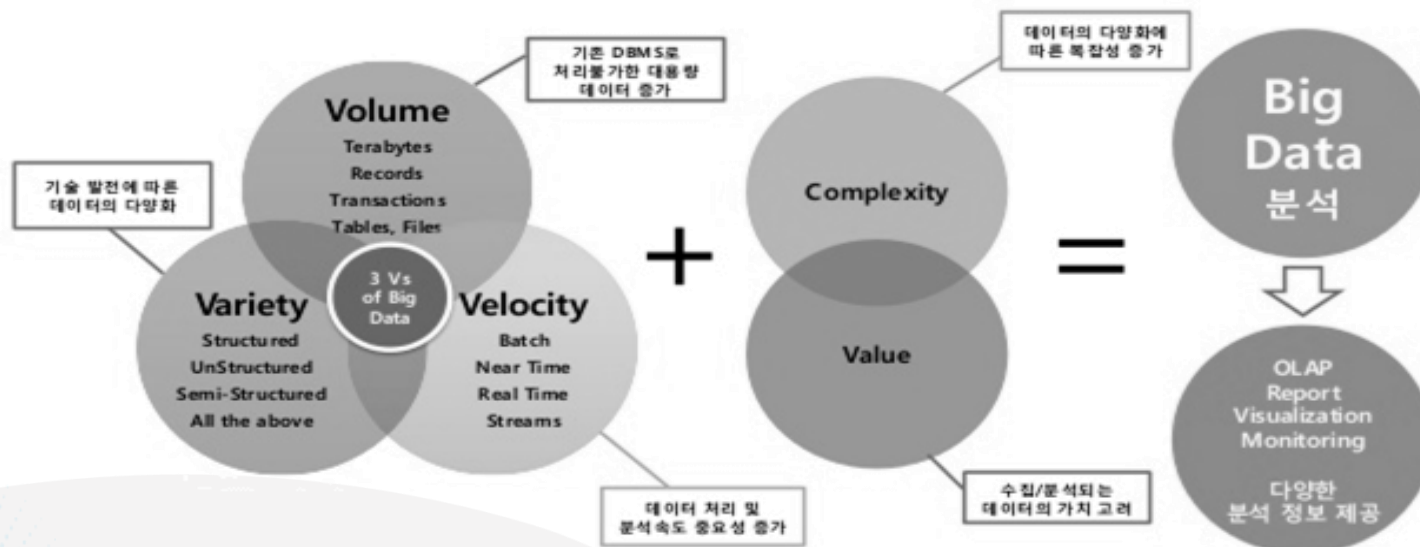
부문	DB구축형태
물류부분	EDI, CALS, CVO, 항만운영정보시스템(PORT-MIS), 철도운영정보시스템(KROIS), 민간기업 물류 VAN 등
지리부분	국가지리정보체계(NGIS), 토지종합정보망(LMIS), GIS, RS, GPS, ITS, LBS, SIM, 공간DBMS
교통부분	지능형교통시스템(ITS)
의료부분	처방전달시스템, 전자의무기록, 영상처리시스템(PACS)
교육부분	교육행정시스템(NEIS)

2) 데이터의 가치와 미래?

◎ 빅데이터의 이해

<McKinsey>

빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석 할 수 있는 범위를 초과하는 규모의 데이터.

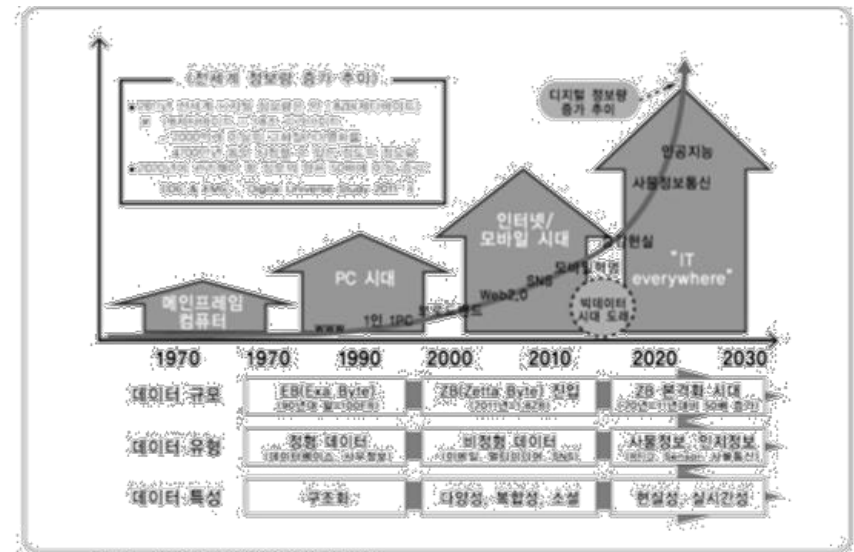
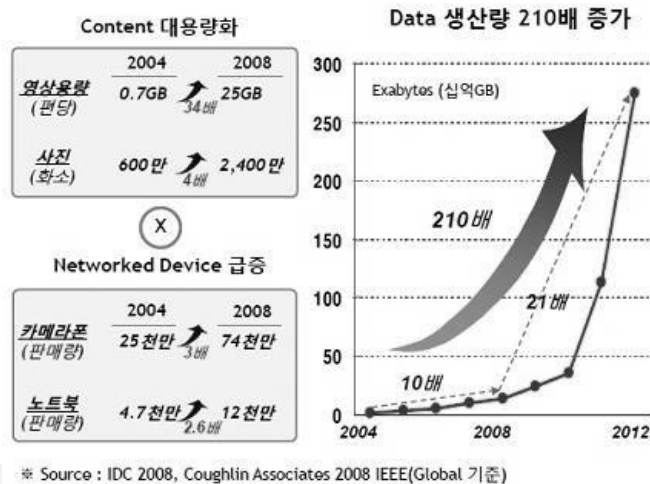


데이터의 크기, 다양성 및 속도에 복잡성이 더해지면서 Big Data에 대한 개념도 변화하고 있음

2) 데이터의 가치와 미래?

◎ 빅데이터의 출현 배경

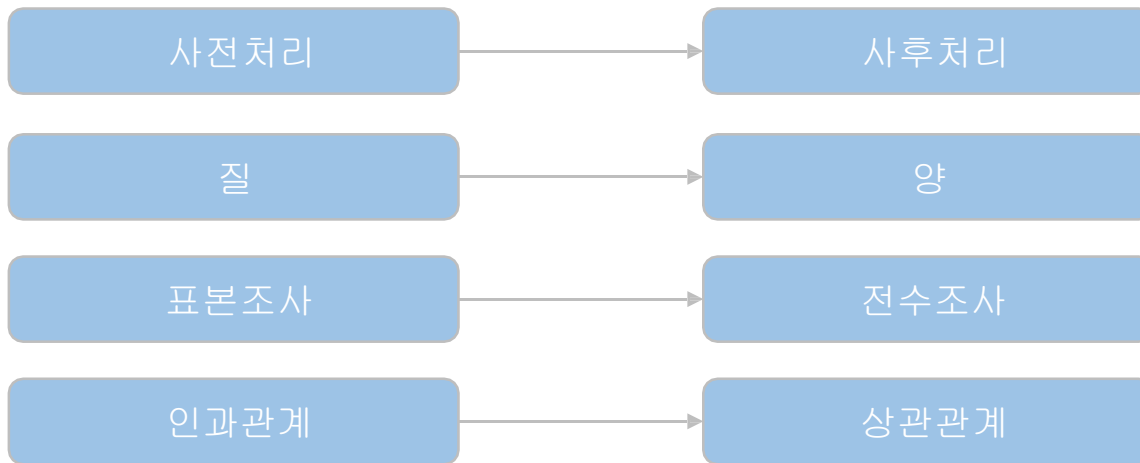
- ① 산업계-고객데이터 축적
- ② 학계-거대 데이터 활용 과학 확산
- ③ 관련 기술 발전(디지털화, 저장기술, 인터넷 보급, 모바일, 클라우드 혁명)



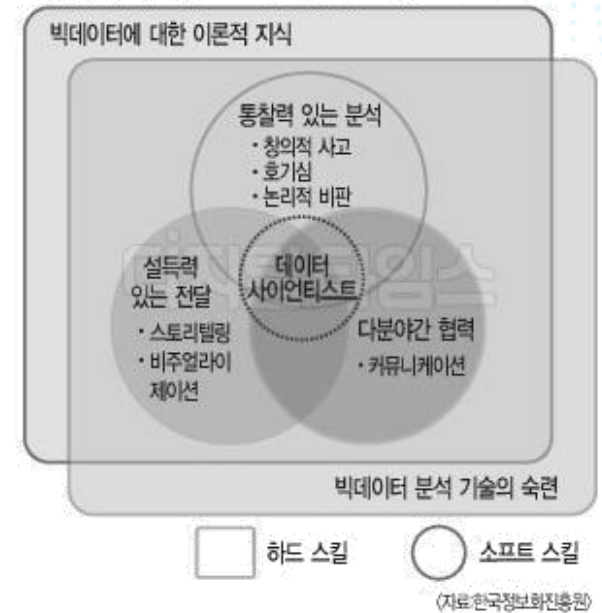
출처: NIA(한국정보화진흥원) 「새로운 미래-빅데이터 시대」(2019)

2) 데이터의 가치와 미래?

◎ 빅데이터가 만들어낸 본질적 변화



데이터 사이언티스트의 역량과 조건

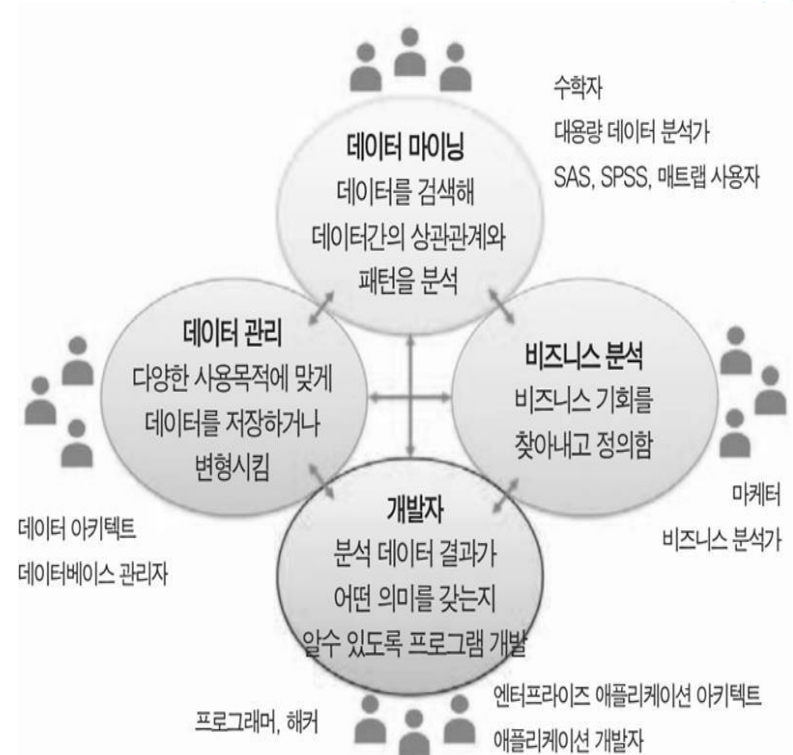
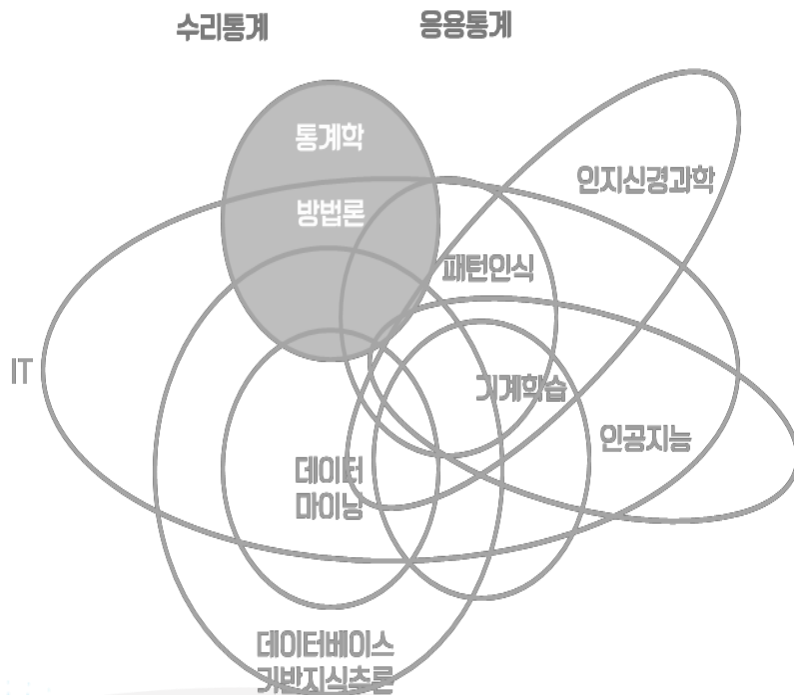


<빅데이터 활용 기본 테크닉>

- ① 연관 규칙 학습(Association Rule) ② 유형분석 (Classification Tree Analysis) ③ 유전알고리즘 (Genetic Algorithms) ④ 기계학습(Machine Learning) ⑤ 회귀분석 (Regression Analysis) ⑥ 감정분석 (Sentiment Analysis) ⑦ 소셜네트워크 분석

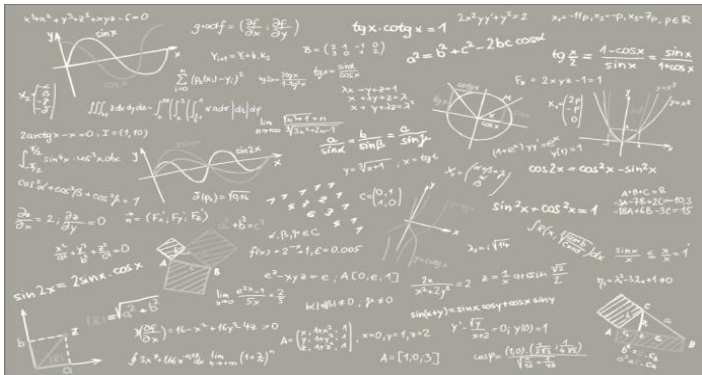
3) 우리에게 데이터 분석이란?

◎ 데이터를 가지고 노는 사람들



3) 우리에게 데이터 분석이란?

◎ 데이터를 가지고 노는 사람들



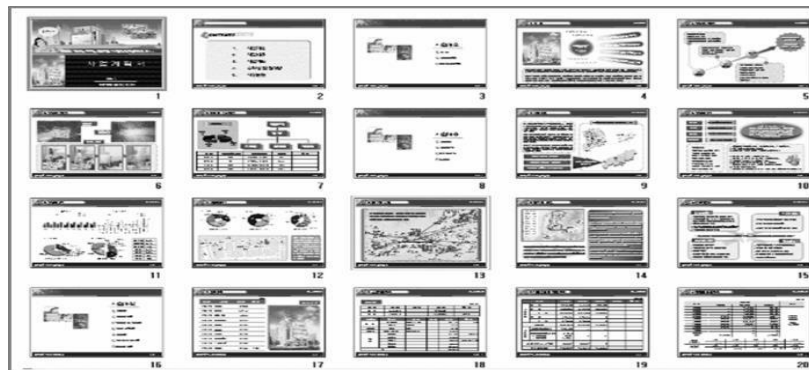
빠곡한 숫자와 잘 꾸며진 공식, 선형대수학, 벡터, 기하학



돈이 만들어지는 방식, 주식과 도박



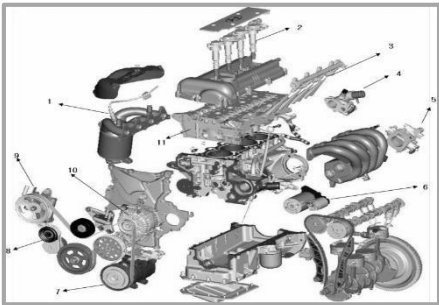
명확히 데이터를 보여주는 것. 시각화, 인포그래픽



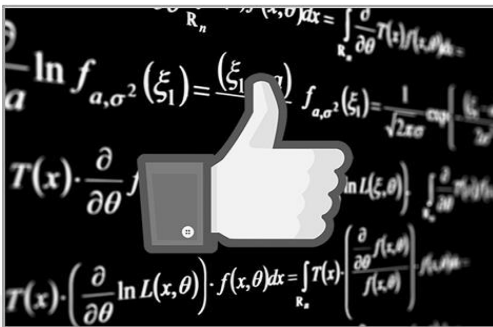
각종 보고서

3) 우리에게 데이터 분석이란?

◎ 데이터를 가지고 노는 사람들



V
S



V
S





데이터 속에서 이야기를 찾아내는 역량을 키우는

빅 데이터 분석 실무

- 경영/마케팅 실무 중심 -

EXAMPLE 1.

EXAMPLE 1.

EXAMPLE

서울 시민, 말라간다!?

아모레퍼시픽 아리따움 스킨터치 100만 데이터 분석.
09~15년, 7년간 스킨수분 감소추세
특히, 강서구, 용산구에서 활동하는
2030대 여성이 위험하다.



EXAMPLE

아모레퍼시픽 아리따움 스킨터치 100만 데이터 분석.
촉촉한 속대, 건조한 서울대



아리따움 매장명	수분지수
아리따움 숙대입구역점	66.1
아리따움 건대교운점	54.7
아리따움 이대직영점	53.7
아리따움 외대역점	50.6
아리따움 신촌세브란스직영점	48.4
아리따움 경희대점	43.8
아리따움 교대직영점	40.2
아리따움 서울대역점	34.6



EXAMPLE 1.

EXAMPLE

노는 물이 다르다.

아모레퍼시픽 아리따움 스킨터치 100만 데이터 분석.
촉촉한 건대입구 / 메마른 이태원

아리따움 건대플러스	70.9
아리따움 강남대로직영점	68.4
아리따움 홍대클럽직영점	63.0
아리따움 노랑진직영점	59.4
아리따움 건대고운점	54.7
아리따움 압구정직영점	51.4
아리따움 이태원점	48.0



EXAMPLE 1.

1. 100만 데이터 분석

2-2. Deep Dive

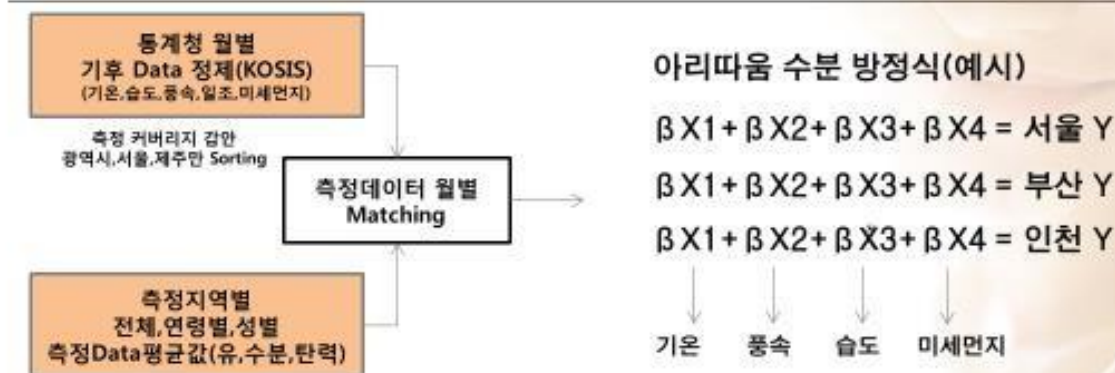
#Issue3.외부변수 Join

외부변수와의 관계성 및 영향도

- 탐색분석 : 지역별 측정값의 연도별 편차가 심함.
- Data Filter : 서울,제주,5대광역시 중심 외부 기후변수와 측정값 JOINT
- 가설구조 : 기후변수와 측정변수 관계로 측정변수를 예측 가능할 것이다.
- 분석 Scheme : 기후변수에 따른 측정변수 관계성 검증 > 회귀분석

⇒ Offer Message : 아모레의 수분 방정식

#분석 Scheme



EXAMPLE 1.

 1. 100만 데이터 분석

2-2. Deep Dive

#Issue3. 외부변수 Join

- 지역별 회귀방정식 모델에선 각 지역별 유의계수가 상이함. 우선 전체모형으로 수분 방정식을 활용하고
- Data 추가속적 및 기상 Data 정교화(일별, 측정지역) 를 통해 세부분류별 모형을 추가하는 것을 제안함.

④모델선택



아리따움 수분 방정식

$$= 45.7 + 0.51 * \text{기온} - 0.027 * \text{일조} - 0.018 * \text{강수} - 1.27 * \text{풍속} \\ + 0.08 * \text{습도} + 0.0002 * \text{미세} * \text{기온} * \text{강수}$$

EXAMPLE 1.





데이터 속에서 이야기를 찾아내는 역량을 키우는

빅 데이터 분석 실무

- 경영/마케팅 실무 중심 -

EXAMPLE 2.

EXAMPLE 2.

4) 추천 시스템이란?

◎ 무한한 정보들에서 적시에 사용자가 흥미를 갖거나 구매하기 원하는 상품을 쉽게 찾도록 도와주는 시스템


Your recently viewed items and featured recommendations




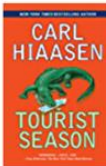

Inspired by your browsing history



 Apple 13" MacBook Pro, Retina Display, 2.3GHz Intel Core i5 Dual Core, 8GB RAM, 128GB... ★★★★☆ 94 \$1,249.00	 Apple 13.3" MacBook Air, 128GB SSD, MQD32LL/A Starters Bundles [Mid- 2017 - Newest... ★★★★☆ 75 \$897.00	 Google Pixelbook (i5, 8 GB RAM, 128GB) ★★★★☆ 63 \$899.00 ✓prime	 Apple iPhone 8 Plus 5.5", 64 GB, Fully Unlocked, Gold ★★★★☆ 59 \$942.98	 Apple iPhone 8 4.7", 64 GB, Fully Unlocked, Gold ★★★★☆ 51 \$829.00 ✓prime	 Apple iPhone 8 Plus 5.5", 64 GB, GSM Unlocked, Space Gray ★★★★☆ 8 \$891.01	 iPhone X Screen Protector, Maxboost (Clear, 3 Packs) iPhone X Tempered Glass... ★★★★☆ 5,196 \$9.95 ✓prime
---	--	---	---	---	---	---

Best Sellers



 Say You're Sorry (Morgan Dane Book 1) › Melinda Leigh ★★★★☆ 1,627 Kindle Edition \$2.49	 The Man from St. Petersburg › Ken Follett ★★★★☆ 635 Kindle Edition \$1.99	 Russian Roulette: The Inside Story of... › Michael Isikoff ★★★★☆ 46 Kindle Edition \$15.99	 Bones Don't Lie (Morgan Dane Book 3) › Melinda Leigh ★★★★☆ 34 Kindle Edition \$5.99	 True Fiction (Ian Ludlow Thrillers Book 1) › Lee Goldberg ★★★★☆ 403 Kindle Edition \$4.99	 Tourist Season › Carl Hiaasen ★★★★☆ 284 Kindle Edition \$1.99	 Bone Music (The Burning Girl Series Book 1) › Christopher Rice ★★★★☆ 564 Kindle Edition \$4.99
---	--	---	--	--	---	--

EXAMPLE 2.

4) 추천 시스템이란?

◎ 하지만 선택할 수 있는 상품이 너무 많다면...



EXAMPLE 2.

4) 추천 시스템이란?

THE JAM STUDY

A grocery store conducted 2 tasting sessions. In one session shoppers were allowed to sample 24 flavors of jams, and in the other session they were allowed to sample 6 flavors



24 Choices of Jam vs **6 Choices of Jam**

Attracted **60%**
of Shoppers

Shoppers sampled **2**
flavours on average

3% of shoppers
bought jam

Attracted **40%** of
Shoppers

Shoppers sampled **2**
flavours on average

30% of shoppers
bought jam

EXAMPLE 2.

4) 추천 시스템 목표

◎ 사용자 기록 정보, 거래 상세 정보, 상호작용 로그와 같은 **사용 가능한 사용자의 인터넷 활용 정보**와 제품 사양, 사용자 후기, 다른 제품과의 비교 등을 아우르는 **제품 정보**를 고려함으로써 **좀 더 개인화된 추천**을 하는 것

EXAMPLE 2.

4) 관점에 따른 추천 시스템의 정의

사용자 관점

- 결정을 내리는데 신뢰할 수 있는 데이터로부터 관심있을 만한 아이템을 추천을 받는 것이 중요

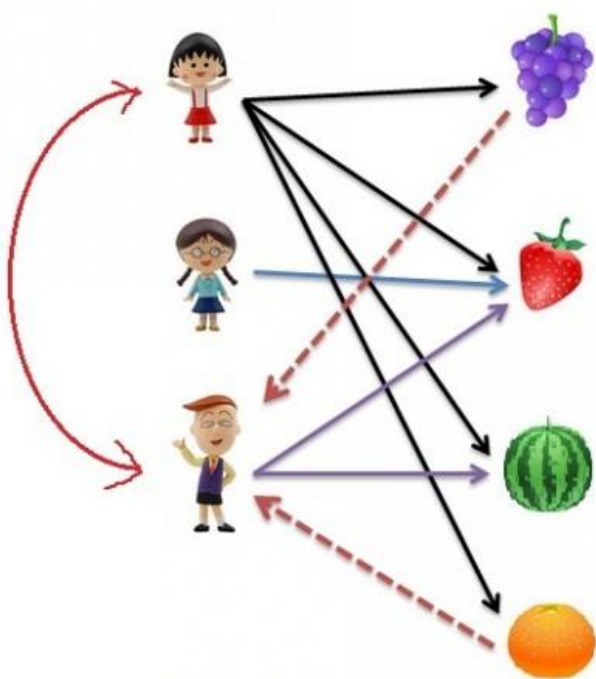
제공자 관점

- 사용자에게 개인 맞춤 수준으로 추천을 하는 것이 중요. 그를 통해 제공자의 목표(매출, 조회수,이용자수 등)를 달성하기 위한 수단

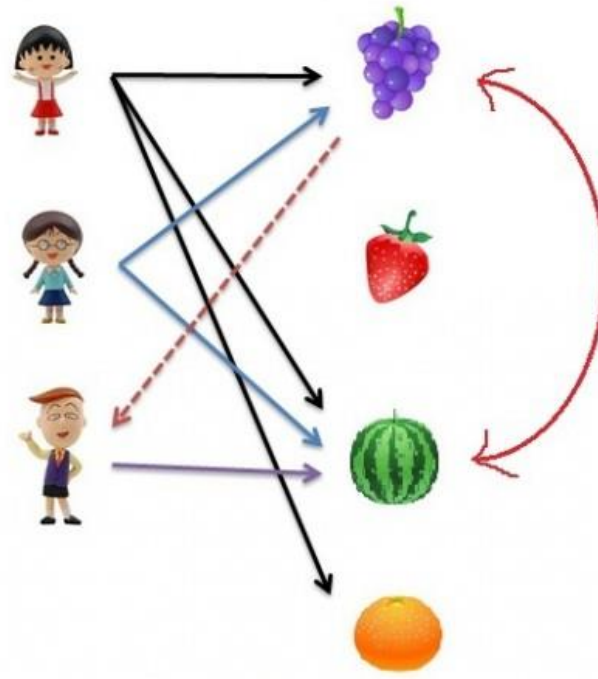
EXAMPLE 2.

© Collaborative filtering Model

<http://www.salemmarafi.com/code/collaborative-filtering-with-python/>



User-based filtering



Item-based filtering

사용자 기반 협업 필터링

- 단계

- 영화의 평가 정보를 이용해서 사용자 간의 유사도 계산
- 각각의 추천 대상 사용자에게 대해 해당 사용자는 안 보고 다른 사용자들이 본 영화 검토
- 대상 사용자가 아직 평가하지 않은 영화의 평점을 예측

- 예제

- Toby에게 새 영화 추천하기

Movie/User	Claudia Puig	Gene Seymour	Jack Matthews	Lisa Rose	Mick LaSalle	Toby
Just My Luck	3	1.5		3	2	
Lady in the Water		3	3	2.5	3	
Snakes on a Plane	3.5	3.5	4	3.5	4	4.5
Superman Returns	4	5	5	3.5	3	4
The Night Listener	4.5	3	3	3	3	
You Me and Dupree	2.5	3.5	3.5	2.5	2	1

사용자 기반 협업 필터링

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \times \sum (y - \bar{y})^2}}$$

● 단계

- 먼저 Toby와 유사한 사용자 찾는다. Claudia Puig, Lisa Rose, Mick가 비슷하다
- 사용자 간의 유사도는 사용자들이 제공한 평점을 기반으로 계산한다
- 유사도를 계산 하는데 가장 흔히 사용되는 방법은 유클리드 거리와 피어슨 상관관계수다.
- 지금은 피어슨 상관관계수 방식을 선택해 주어진 다음 식을 이용해서 유사도를 계산한다.

	Claudia Puig	Gene Seymour	Jack Matthews	Lisa Rose	Mick LaSalle	Toby
Claudia Puig	1	0.7559289	0.9285714	0.9449112	0.6546537	0.8934051
Gene Seymour	0.7559289	1	0.9449112	0.5	0	0.3812464
Jack Matthews	0.9285714	0.9449112	1	0.7559289	0.3273268	0.662849
Lisa Rose	0.9449112	0.5	0.7559289	1	0.8660254	0.9912407
Mick LaSalle	0.6546537	0	0.3273268	0.8660254	1	0.9244735
Toby	0.8934051	0.3812464	0.662849	0.9912407	0.9244735	1

사용자 기반 협업 필터링

- Toby가 아직 평가하지 않은 Just My Luck 영화에 대한 예상 평점 계산

$$(3 * 0.8934051 + 1.5 * 0.3812464 + 3 * 0.9912407 + 2 * 0.9244735) / (0.8934051 + 0.3812464 + 0.9912407 + 0.9244735) = 2.53$$

Movie/User	Claudia Puig	Gene Seymour	Jack Matthews	Lisa Rose	Mick LaSalle	Toby
Just My Luck	3	1.5		3	2	

의 추천

	Toby
Claudia Puig	0.8934051
Gene Seymour	0.3812464
Jack Matthews	0.662849
Lisa Rose	0.9912407
Mick LaSalle	0.9244735
Toby	1

Quiz : Toby에게 The Night Listener 영화를 추천해야 할까요?

아이템 기반 협업 필터링

- 아이템 간의 유사도 이용
- 사용자가 과거에 아이템 A를 좋아했다면 A와 유사한 아이템 B도 좋아할 것이라는 가정이 전제됨
- 사용자 기반 협업 필터링의 단점
 - 사용자 평점 정보가 매우 부족할 경우 성능 저하 발생함. 사용자들이 광범위한 카탈로그에서 일부 아이템만 평가하는 실세계에서는 매우 흔한 경우임
 - 데이터 규모가 매우 큰 경우 모든 사용자에게 대한 유사도 값을 구하기 위한 계산 비용이 매우 큼
 - 사용자 프로필 또는 사용자 입력이 빠르게 변하면 유사도 값을 다시 계산해야 하며, 이로 인한 계산 비용이 높아짐
- 아이템 기반 추천 엔진은 아이템 간의 유사도를 계산하기 때문에 계산 비용을 줄임

아이템 기반 협업 필터링

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

- 단계
 - 아이템 간의 유사도 계산
 - 다른 유사한 아이템에 대해 주어진 기존 평가를 이용해서 활성 사용자가 평가하지 않은 아이템에 대한 평점 예측
- 이 접근법에서는 코사인 유사도를 사용함
- 코사인 유사도는 벡터 공간에서 두 개의 n 차원 벡터 간의 각도를 이용해서 유사도를 계산함
- 코사인 유사도 적용 시, 아이템 열을 n 차원의 벡터로 여기고 두 아이템 간의 유사도를 두 벡터 사이의 각도로 여김
- 각도 크기가 작을수록 아이템은 유사함

아이템 기반 협업 필터링

- 영화 Lady in the Water에 대한 Toby의 등급 예측
 - Lady in the Water와 유사한 영화 확인
 - You, me and Dupree가 Lady in the Water와 유사하다는 것 확인(0.8897565)
 - Toby가 영화 Lady in the Water와 유사한 영화에 제공한 등급의 가중치의 합을 계산해서
Lady in the Water에 대한 Toby의 등급 예측
 - $(0.795 * 4.5 + 0.814 * 4 + 0.889 * 1) / (0.795 + 0.814 + 0.889) = 3.09$

	Just My Luck	Lady in the water	Snakes on a Plane	Superman Returns	The Night Listener	You Me and Dupree
Just My Luck	1.000000	0.6339001	0.7372414	0.7194516	0.8935046	0.7598559
Lady in the water	0.6339001	1.000000	0.7950515	0.8149529	0.7977412	0.8897565
Snakes on a Plane	0.7372414	0.7950515	1.000000	0.9779829	0.8585983	0.9200319
Superman Returns	0.7194516	0.8149529	0.9779829	1.000000	0.8857221	0.9680784
The Night Listener	0.8935046	0.7977412	0.8585983	0.8857221	1.000000	0.9412504
You Me and Dupree	0.7598559	0.8897565	0.9200319	0.9680784	0.9412504	1.000000

Movie/User	Toby
Just My Luck	
Lady in the Water	
Snakes on a Plane	4.5
Superman Returns	4
The Night Listener	
You Me and Dupree	1

아이템 기반 협업 필터링

● 장점

- 구현하기 쉽다
- 추천 생성 시에 아이템의 콘텐츠 정보 또는 사용자 프로필 정보는 필요하지 않음
- 사용자가 생각하지 못할 새로운 아이템을 추천

● 단점

- 유사도 계산을 위해 모든 사용자, 제품, 평가 정보가 메모리에 로드되기 때문에 계산 비용이 비쌈
- 사용자에 대한 정보가 전혀 없는 경우에는 적합하지 않음.(i.e. cold start problem)
- 데이터가 거의 없는 경우 성능이 저하됨
- 사용자 또는 아이템에 대한 콘텐츠 정보가 없기 때문에 평가 정보만으로는 정확한 추천을 생성할 수 없음



주요 이력

現) (주)RTMC 전략기획실장
前) (주)B사 웹로그분석 및 DP사업 完
前) (주)H금속사 회계팀
前) (주)B건설사 회계팀
前) K문고 CRM VIP 군집전략 CRM프로젝트 보조연구원
前) L백화점 CRM Alert 전략 CRM프로젝트 보조연구원

BSL(스위스 로잔 비즈니스 스쿨) MBA
ASSIST 빅데이터경영통계 MBA

국가공인 ADSP(빅데이터 준전문가)

現 코리아IT아카데미 빅데이터 R 강사
現 코리아IT아카데미 빅데이터 기초 파이썬 강사
現 코리아IT아카데미 빅데이터 기초통계 전담강사

[박영식] [완성에 이르기까지](#)