



# 지식기반 사회를 위한 시도별 도서관 데이터 분석과



도서 이용 증진방안 탐구







# 목차



# 02 데이터 분석

- 2-1. 활용 데이터 선정
- 2-2. 분석 범위 선정
- 2-3. 데이터 전처리
- 2-4. 활용 알고리즘
- 2-5. 개인화추천시스템



# 01 배경

- 1-1. 분석 필요성
- 1-2. 주제 선정



# 03 분석 활용전략

- 3-1. 도서관 내 활용방안
- 3-2. 기대효과
- 3-3. 향후 연구방향





## 시작에 앞서

# "국가의 미래 비전은 국민들의 독서력에 달려있다"

지식정보를 기반으로 한 산업사회에서 새로운 부의 창출은 지식 역량에 좌우되고, 그런 역량을 키우는 단단한 기반이 바로 독서에 있기 때문이다.



# 01 배경

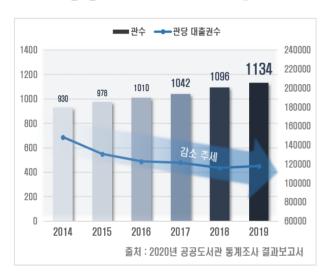
1-1. 분석 필요성

1-2. 주제 선정



# 1-1. 분석 필요성

#### ○ 전국 공공도서관 수 및 관 당 대출건수



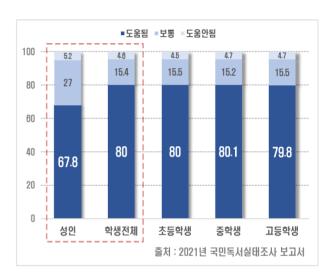
- 전국 공공도서관 수는 매년 증가( '19년, 1134관)
- 공공도서관의 도서대출건수는 연간 약 120,000회로 매년 감소추세
- 늘어나는 도서관 수에 대비하여 실제 독서대출건수는 <mark>상반적인 현상</mark>을 보임

#### ○ 연평균 독서인구 비율 및 평균독서량



- <mark>매년 독서인구는 감소</mark>하고 있으며, 15년 대비 2021년의 경우 약 11% 감소
- 독서인구 1인 평균 독서량은 2021년 9.4권으로 연간 약 10권 미만
- 전체 인구대비 1인 평균 독서량은 2021년 4권으로 국민독서실태조사 진행 이후 최저 수치

#### ○ 연령별 독서의 유용성 정도



- 성인의 약 70%, 학생의 80%가 독서가 사회 또는 학교생활에 도움이 된다고 응답
- 독서의 중요성과 필요성은 인지하지만 실제 독서를 이용하지 않는 추세

## 매년 공공도서관 수는 증가하는 반면 연평균 독서 인구 비율 및 평균 독서량 감소





## 1-2. 주제 선정

공공도서관은 문화 소외계층을 포함한 다양한 사람들이 이용하므로 도서대여 관련 서비스 질 향상은 독서 인구 저변확대에 있어 매우 중요

다양한 업종과 분야에서 **개인화추천시스템을 활용해 이용자에게 맞춤형 서비스를** 제공, 도서와 관련된 개인화추천시스템은 공공도서관에 적용된 사례가 전무

빅데이터 기반 개인맞춤형 추천서비스를 제공하여 이용자들의 독서 생활 영위에 도움을 주고자 함, 나아가 개인과 더불어 공공도서관 차원에서 활용할 수 있는 방안에 대해 제안하려 함

공공도서관에 방문하여 책을 읽는 사람들에게 독서에 대한 편의를 제공을 통한 도서이용 증진방안을 탐구



# 02 데이터 분석

- 2-1. 활용 데이터 선정
- 2-2. 분석 범위 선정
- 2-3. 데이터 전처리
- 2-4. 활용 알고리즘
- 2-5. 개인화추천시스템



# 2-1. 활용 데이터 선정

#### ○ 활용 데이터 정보

출처	사용 데이터	기준년도	사용목적	비고
도서관 정보나루	OPEN API (도서관/지역별 인기대출 도서 조회)	2020년	- 개인화추천시스템 적용 - 결과 활용방안 제시	-
	국민독서실태조사 (성인/학생)	2019년, 2021년	- 독서인구관련 현황 집계 - 결과 활용방안 제시	보고서 및 마이크로데이터 동시 활용
	국민여가활동조사	2019년, 2020년	- 독서인구관련 현황 집계	마이크로데이터 활용
통계청 MDIS	통계청 사회조사 (복지/사회참여/문화와여가/소득과소비/노동)	2011년, 2013년, 2015년, 2017년, 2019년, 2021년	- 독서인구관련 현황 집계	보고서 및 마이크로데이터 동시 활용
	전국도서관통계 (문화체육관광부 공공도서관통계조사)	2020년	- 개인화추천시스템 활용 - 결과 활용방안 제시	보고서 및 마이크로데이터 동시 활용

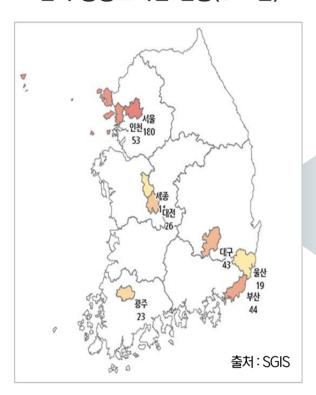
- 도서 개인화추천시스템: 도서관 정보나루(OPEN API), 전국도서관통계 2020년 [기준일자: 19.01.01~19.12.31]
- 분석결과 활용방안 제시: 도서관 정보나루(OPEN API), 국민독서실태조사 2019년, 전국도서관통계 2020년 [기준일자: 19.01.01~19.12.31]

## R, Python, SPSS 을 통한 데이터 추출 및 전처리 진행



## 2-2. 분석 범위 선정

○ 전국 공공도서관 현황(399관)



### ○ 분석에 활용한 인구통계학적 변수 및 코드정보(도서관 정보나루)

	지역(region)		성별(gender)	연령(age)			
코드	코드명	코드	코드명	코드	코드명		
11	서울특별시(180관)	0	남성	0	영유아(0~5세)		
21	부산광역시(44관)	1	여성	6	유아(6~7세)		
22	대구광역시(43관)			8	초등(8~13세)		
23	인천광역시(53관)			14	청소년(14~19세)		
24	광주광역시(23관)			20	20대		
25	대전광역시(26관)			30	30대		
26	울산광역시(19관)			40	40대		
29	세종특별자치시(11관)			50	50대		
				60	60세 이상		

- 공공도서관은 세종특별자치시를 포함한 총 8개의 특·광역시를 대상으로 하며 수도권에 약 56%가 분포한다.
- 코드 정보는 도서관 정보나루에 제시 된 자료를 이용하였으며, 추후에 분석 할 '2019년 국민독서실태조사'에 포함되는 3가지 인구통계학적 변수를 선정하였다.

## 전국 공공도서관 현황 및 활용하는 변수의 조건 기준 설정

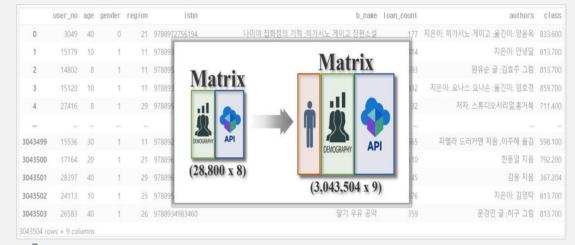


## 2-3. 데이터 전처리

- 도서정보 데이터 생성 (28,800 × 8 Matrix)
  - □ API 호출을 통한 5개 도서정보 변수 추출
  - □ 인구통계학적 3가지 조건 변수 추가
  - □ 지역(8) × 성별(2) × 연령(9), 총 144개의 그룹 생성
  - ✓ 도서명 (Bookname)
  - ✓ TATB (Authors)
  - ◀ 국제표준도서정보 (ISBN13)
  - ✓ 도서주제분류 (Class)
  - ✓ 대출건수 (Loan\_count)
  - 시 지역 (Region)
  - ✓ 성별 (Gender)
  - ✓ 연령 (Age)
- 이용자정보 데이터 생성 (3,043,504×9 Matrix)
  - □ 각각의 도서를 도서별 대출건수 만큼 확장하여

[도서정보 데이터]를 [이용자정보 데이터]로 변환하여 분석에 활용

- + Add1) 기상의 이용자 난수 생성 (User\_no) 후 부여
- + Add2) 도서와 관련된 개인성향 반영을 위해 평점(Rating score) 부여
- + Add3) 최종데이터에서 무응답, 오타 등 값(16,066건) 제외



🍑 최종 생성된 데이터를 <u>1. 이용자-도서</u> / <u>2. 도서-도서정보</u> 데이터로 분할하여 시스템 구현에 활용







	isbn	b_name	authors	class	loan_count
0	9788979196856	변신	히가시노 게이고 지음;이선희 옮김	833.60	16
1	9788953582750	구름빵 그림자가 사라졌어	글·그림: GIMC,김향수	813.70	13
2	9788930202299	Why? 로켓과 탐사선	글: 황근기 ;만화: 김성래	408.00	9
3	9788972081555	수수꽃다리	임선영 지음	813.60	15
4	9788997357659	삐악삐악 크리스마스	쿠도 노리코 글 그림 ;한라경 옮김	833.60	8
5	9788958282631	임꺽정 :벽초 홍명희 소설	[저자]: 홍명희 ;[그림]: 박재동	813.60	28
6	9788932907871	나무	베르나르 베르베르 지음 ;이세욱 옮김	863.00	9
7	9788960243446	(빈대 가족의) 생활의 달인	글: 이봉기 ;그림: 류수형	320.00	106
8	9791158360535	돌돌 말아 김밥	최지미 글 그림	813.70	10
9	9791189159320	오! 한강	글: 김세영 ;그림: 허영만	911.07	19

활용 데이터 전처리(R, Python)를 통한 도서정보 데이터의 이용자정보 데이터 변환 진행





## 2-4. 활용 알고리즘

• 본 개인화 추천시스템에는 잠재요인 협업 필터링(Latent Factor Collaborative Filtering)의

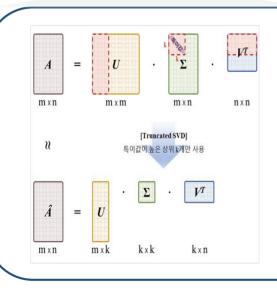
행렬 분해 MF-알고리즘(Matrix-Factorization Algorithm) 사용 (개인화 추천시스템을 활용하는 넷플릭스(Netfilx) 알고리즘 경연대회 수상 알고리즘)

- 사용자들의 기본 정보와 고객들의 선호도(도서관의 경우 상위 대출 성향으로 변환 가능)를 바탕으로 유사한 패턴을 보이는 이용자를 분류하는 기법
- 행렬 처리에서 정보의 유사도에 기반한 SVD(Singular Value Decomposition) 방법론 중

특이값이 높은 상위 K개의 요인(속성)을 사용하는 Truncated SVD를 활용



## **Procedure**



#### Step 1. 원 행렬(Original Matirx)을 세 가지 행렬로 분해(Decomposition)

- //: '이용자정보'와 '잠재요인' 사이의 관계를 내포한 행렬

- 🛽 : '이용자정보'와 '도서정보'사이의 관계를 내포한 잠재요인 정사행렬

- 1/: '잠재요인' 과 '도서정보' 사이의 관계를 내포한 행렬

Step 2. 분해 된 세가지 행렬 중 잠재요인(Latent Factor)을 포함한 정시행렬(♡)에서 특이값이 높은 상위 K개의 잠재요인을 선택하여 행렬 축소

Step 3. 축소된 행렬을 기반으로 재 내적(Inner Product)을 진행하여 원 행렬 형태로 복원

## 본 탐구의 개인화추천시스템에 활용하는 빅데이터 분석기법 (머신러닝, 알고리즘)

## 2-5. 개인화추천시스템(1)

- 도서정보 기반 추천시스템
  - 도서정보(ISBN13)만 활용하여 유사한 도서목록의 추천이 가능한지 확인
  - 단순한 도서 간 상관성을 통해 추천도서 확인이 가능하며 후에 적용하는 시스템의 추천 정확도(Accuracy) 평가를 위해 진행

## Result

1 isbn\_title = user\_isbn\_rating.columns (1) Input 2 isbn\_title\_list = list(isbn\_title) 3 coffey\_hands = isbn\_title\_list.index("라플라스의 마녀 :히가시노 게이고 장편소설 4 corr\_coffey\_hands = corr[coffey\_hands] 5 list(isbn\_title[(corr\_coffey\_hands >= 0.9)])[:15] 2 Output '(간단함, 병맛, 솔직함으로 기업의 흥망성쇠를 좌우하는) 90년생이 온다' '(지적이고 아름다운 삶을 위한) 라틴어 수업' '7년의 밤 :정유정 장편소설', '가면산장 살인사건', '개인주의자 선언 :판사 문유석의 일상유감', '걷는 사람, 하정우', '고양이 :베르나르 베르베르 장편소설', '고요할수록 밝아지는 것들 :혜민 스님과 함께 지혜와 평온으로 가는 길' '골든아워 :생과 사의 경계, 중증외상센터의 기록', '그대 눈동자에 건배 :히가시노 게이고 소설집', '나는 나로 살기로 했다 :냉담한 현실에서 어른살이를 위한 to do list' '나미야 잡화점의 기적 :히가시노 게이고 장편소설' '내게 무해한 사람 :최은영 소설', '당신이 옳다 :정혜신의 적정심리학'

☑ 피어슨 상관계수를 기반으로 도서간의 상관성 측도 계산

$$r_{xy} = \frac{\varSigma(X \! - \overline{X})(\, Y \! - \overline{Y})}{\sqrt{\varSigma(X \! - \overline{X}^2)} \, \sqrt{\varSigma(\, Y \! - \overline{Y}^2)}}$$

① Input: "라플라스의 마녀: 히가시노 게이고 장편소설" 을 통해 추천 된 결과

② Output: 다양한 소설 분류의 도서들이 추천되는 것이 확인 가능하며,

이 중 동일한 작가의 소설이 추천되는 것을 확인 가능

도서 정보를 기반한 추천시스템을 통해 사용하려는 데이터 유효성 확인





## 2-5. 개인화추천시스템(2)

### ○ 이용자-도서정보 행렬

isbn	9788925550633	9788936446819	9788936454951	9788936455286	9788937838750
user_no					
1	0.0	0.0	0.0	0.0	7.0
2	1.0	1.0	6.0	0.0	0.0
3	0.0	1.0	0.0	0.0	7.0
4	0.0	0.0	0.0	0.0	5.0
5	2.0	9.5	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0
7	0.0	6.0	10.0	0.0	0.0
8	0.0	6.0	0.0	0.0	0.0
9	0.0	8.0	1.0	9.0	0.0
10	6.0	0.0	4.0	1.0	0.0
11	0.0	0.0	10.0	4.0	0.0
12	0.0	8.0	8.0	0.0	1.0
13	0.0	4.0	8.0	6.0	3.0
14	0.0	0.0	0.0	9.0	0.0
15	8.0	0.0	0.0	2.0	4.0

• 시스템 적용을 위해 데이터 선정과정에서 추천 된 144개 그룹의 인기도서 200선 목록 28,800건 중 중복 도서명을 제외하여

#### 총 4,936건의 도서정보(ISBN13)를 바탕으로

이용자(User\_no)의 도서 대출여부를 행렬의 형태로 표현

(해당 도서를 대출하지 않음: 0 / 해당 도서를 대출 함: n)

### ○ 이용자-도서정보 평점 행렬

isbn	2788963290874	8809362342450	8809362345505	8809448445143	8809544940054
0	0.096080	0.044661	0.047109	0.051150	0.043868
1	0.084448	0.040232	0.042384	0.043999	0.039536
2	0.062296	0.030351	0.031906	0.035779	0.029886
3	0.084844	0.041446	0.043605	0.045801	0.040820
4	0.083564	0.048757	0.050501	0.052075	0.048211
5	0.072203	0.030032		0.033944	0.029413
6	0.103332	太昌	§ Data S	eL <sub>0.053673</sub>	0.047672
7	0.077417	0.	0.050768	0.052156	0.048873
8	0.083821	0.043094	0.045118	0.046806	0.042466
9	0.080119	0.038706	0.040682	0.044100	0.038094
10	0.083319	0.043974	0.045837	0.049340	0.043347
11	0.067850	0.031255	0.033092	0.035555	0.030688
12	0.061922	0.022831	0.024822	0.026083	0.022227
13	0.088158	0.042940	0.045138	0.047123	0.042235
14	0.070499	0.040169	0.041753	0.043101	0.039722

- 앞서 생성한 이용자-도서정보 행렬에 총 4,936건의 도서별 평점 평균과 개인-도서별 평점 사이의 차이를 계산한 후 Truncated SVD에 적용하여 1차 결과(평점 편차 행렬)를 산출
- 산출 된 1차 결과(평점 편차 행렬)에 4,936건의 도서별 평점 평균을 재 적용함(각 cell에 덧셈)으로서 도서별 평점 예측 행렬 생성

## 이용자정보-도서정보 기반 추천시스템 설계를 위한 데이터 가공 및 분석



## 2-5. 개인화추천시스템(2)

- 이용자정보-도서정보 기반 추천시스템
  - 앞서 생성한 (1) <u>이용자-도서정보 행렬</u>에서 이용자(user\_no), 도서정보(isbn), 평점 정보(Rating score)를 추출하고, Truncated SVD의 결과로 도출된 (3)<u>이용자-도서정보 평점 행렬</u>에서 도서(해당 이용자가 대출하지 않은 도서만 선택)의 예측 평점(Predicted Rating)이 가장 높은 상위 10권의 도서를 사용자 지정 함수(Function)를 구현하여 추출 진행

## Result

t	ount	loan_c	class	authors		b_name		gende	age	region	value	isbn	_no	user		
2	702		833.6	노 게이고 ;옮긴이: 양윤옥	지은이: 히가시노	스 :히가시노 게이고 장편소설	위험한 비너		20	11	10	8972758242	213 978	<b>1882248</b> 15		
ass loar	clas	uthors	ā	í	b_name					ender	ige ge	region a	value	isbn	user_no	
33.6	833	기: 양윤옥	고 ;옮긴	지은이: 히가시노 게(	게이고 장편소설	라플라스의 마녀 :히가시노 기				1	20	11	10	788972757573	15213	126
14.7	814	기: 신형철	지은	:	급금 :신형철 산문	슬픔을 공부하는 슬				1	20	11	10	791160401967	15213	235
13.7	813	기: 최은영	지은		소 :최은영 소설	쇼코의 미				1	20	11	10	788954641630	15213	25
33.6	833	석희 옮김	ㅏ지음 ;김	무라타 사야?	편의점 인간					1	20	11	10	788952235268	15213	311
32.6	332	영목 옮김	보통 ;정	지은이: 알랭 !	불안					1	20	11	10	788956605593	15213	298
33.6	833	윤옥 옮김	[지음] ;양	히가시노 게이고	게이고 장편소설	악의 :히가시노 7				1	20	11	10	788972754190	15213	214
91.1	591	아연 지음	공		남의 필수 아이템	인 가구 돈 관리 :잘사는 혼섩				1	20	11	10	791162203897	15213	292
48.0	848	종인 옮김	니 지음 ;이	폴 칼라니!	나의 마지막 순간	람 될 때 :서른여섯 젊은 의시	숨결이 바			1	20	11	10	788965961956	15213	290
37.2	337	지양 옮김	시 지음 ;노	록산 게(	못하는 당신에게	나서 페미니스트라고 말하지 !	불편하고 두려워	니스트 :	+쁜 페미	1 L	20	11	10	791195716937	15213	325
14.7	814	기: 신형철	지은		금:신형철 산문	슬픔을 공부하는 슬				1	20	11	10	791160401967	15213	236
				ication)	Classifi	ean Decimal	C, Ko	(KD	子亞	신분후	십	한국				
900		300	8	700	600	500	100		)	300		200		100	000	
역기		근학	1	언어	예술	기술과학	수과학	순	한	사회고	1	종교		철학	총류	

- 서울(Region : 11)소재 도서관을 이용하는 20대(Age : 20)의 여성(Gender : 1) 이용자(User\_no : 15,213)의 추천 도서목록
- ✓ 기존에 "위험한 비너스: 히가시노 게이고 장편소설"을 대출 한 기록과 인구통계학적 유사도를 반영한 결과로,
  상위 10권의 도서 중 1, 6순위에 동일한 작가의 도서가 추천됨과 앞서 대출하였던 '문학'과 관련된 7권의 도서가 추천됨

도서개인화추천시스템의 도출 된 결과를 통해 추천시스템의 적합성 확인



# 03 분석 활용전략

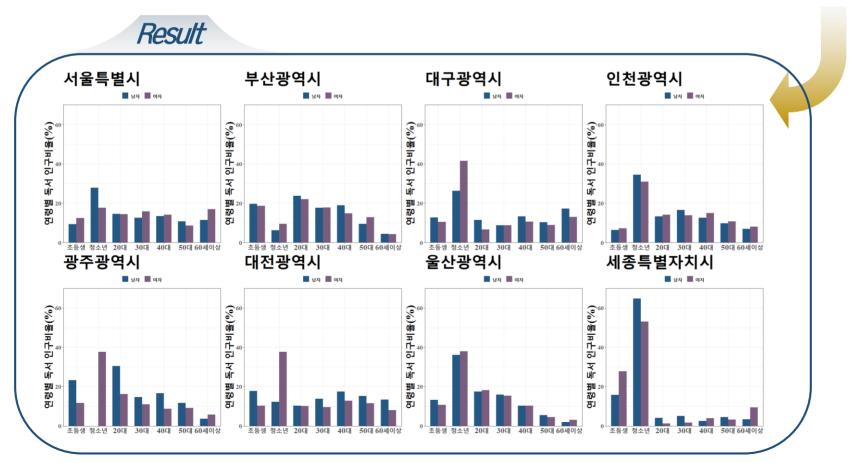
- 3-1. 도서관 내 활용방안
- 3-2. 기대효과
- 3-3. 향후 연구방향





## 3-1. 도서관 내 활용방안

- 이용자 특성 맞춤형 도서배치 전략
  - 앞서 제안한 개인화 추천서비스의 기반인 '추천시스템'을 활용하여 공공도서관의 도서관리 방안에 활용 목적
  - 도서관에 운영 될 추천시스템을 통해 예측 된 다수의 이용자 도서추천 정보를 분석하여 해당 도서관의 보유도서현황과 비교하는 전략 Step 1. 2019년 국민독서실태조사(2020년 전국도서관통계와 조사실시년도 동일) 마이크로데이터를 활용하여 인구통계학 조건에 기반한 '도서인구 분석'진행







## 3-1. 도서관 내 활용방안

○ 이용자 특성 맞춤형 도서배치 전략

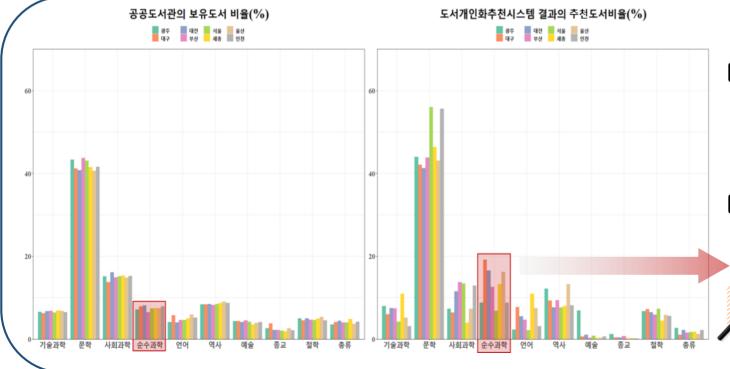
Step 2. 인구통계학적 조건에 기반한 도서 인구 비율을 토대로, 전처리 된 3,027,461건(무응답 제외 개인 대출목록) 중 5,000명의 개인을 임의 선정하여 추천 시스템에 적용

Step 3. 5,000명의 개인을 통해 추천 된 50,000권의 추천도서를 한국십진분류표(KDC)를 기준하여 동일한 조건하에 '주 이용자의 추천도서' 분석

Step 4. 2020년 전국도서관통계 마이크로데이터 분석을 통해 Step 3.의 조건과 동일하게 '도서관의 보유도서' 산출 및 결과비교

## Result

	한국십진분류표(KDC, Korean Decimal Classification)													
000 100 200 300 400 500 600 700 800 900														
총류	철학	종교	사회과학	순수과학	기술과학	예술	언어	문학	역사					



- ☑ 다양한 분야 중 '순수과학' 과 관련된 도서의추천 비율이 실제 공공도서관 내 보유도서비율보다 상대적으로 높음
  - (보유도서보다 적은 추천 비율을 보이는 분류는 제외)
- ✓ 해당 시도의 공공도서관 이용 계층이 '순수과학'분류의 도서가 많이 추천되는 특성을 가진이용자로 해석 가능
- **(1)** '순수고

'순수과학' 분류의 도서를 도서관에 추기배치



3-2. 기대효과



# 도서 개인화 추천서비스



### 이용자 측면

- 다양한 특성 계층의 이용자에게 맞춤형 도서 추천
- 이용자의 도서 선택과 관련한 신속한 의사결정 지원
- 다양한 주제의 도서 접근성 향상



## 공공도서관 측면

- 방문자의 니즈(Needs)를 예측하여 효율적인 도서 자원 운영방안으로 활용 가능
- 도서관의 이용자에 대한 'Cold-Start' 문제 완화

(Cold-Start : 새로운 또는 어떤 유저의 충분한 정보가 없으므로 해당 유저에게 적절한 상품을 추천하지 못하는 상황)

지식기반 사회를 위한 국민의 도서이용 증진에 한 발 다가설 수 있는 시작점이 될 것이다.





- **생 비대면 서비스의 확대로 인한 독서 수단 변화** 
  - 🧼 e-Book 전자 대여, 종이책 DB화, 다양한 비대면 대출 서비스 활용 필요
- ✓ 다양하고 고도화 된 추천 시스템 알고리즘 적용
  - SVD++, MI Algorithm, 딥 러닝 등을 활용
- ✓ 다양한 매체별 빅데이터 분석 적용
  - 각 도서관에서 운영하는 홈페이지 검색 키워드 분석(Text-Mining)어플리케이션 이용 로그분석자료의 활용

빅데이터 시대에 맞는 다양한 매체와 독서 수단을 적용한 차세대 도서개인화추천서비스로의 발전 필요





#### 참고문헌

## 참고문헌

- [1] 강현철, 한상태, 정병철, 신연주, (2003), 개인화를 위한 추천시스템 알고리즘에 관한 연구, 한국통계학회 학술발표논문집, 307-311
- [2] 손지은, 김성범, 김현중, 조성준. (2015). 추천 시스템 기법 연구동향 분석. 대한산업공학회지, 41(2), 185-208
- [3] 이가배, 이효맹, 이현창, 신성윤. (2017). 협업 필터링 알고리즘에 관한 비교연구. 한국정보통신학회:학술대회논문집, 151-153
- [4] 류영호. (2021). 2021년 국민독서 실태조사와 독서문화 진흥방향. 국회도서관
- [5] 문화체육관광부. (2013). 독서문화진흥기본계획 (2014~2018), 11-1371000-00054-2-13
- [6] 문화체육관광부. (2009). 독서진흥에 관한 연차보고서, 11-1370000-000023-10
- [7] 문화체육관광부. (2020). 2020년('19년 기준) 공공도서관 통계조사 결과보고서, 11-1371000-000824-10
- [8] 문화체육관광부. (2021). 2021년 국민 독서실태 조사. 11-1371000-000154-10
- [9] 통계청. (2021). 2021년 사회조사 결과 (복지·사회참여·여가·소득과 소비·노동)
- [10] 박상현 연합뉴스기자. (2022-06-29). 1~5월 공공도서관 대출량 1위 책은 '달러구트 꿈 백화점 1'.https://n.news.naver.com/article/001/0013276434?sid=103
- [11] 신정호 뉴시스기자. (2022-07-22). 용인시. 도서관 고정관념 허물다…멀티문화공간 호응 🕇. https://n.news.naver.com/article/003/0011316135?sid=102





# 감사합니다

