PART. II
연구과제

서울대학교 중앙도서관 도서 대출 빅데이터 분석

연구과제
연구내용
연구결과
향후계획

서울대학교 중앙도서관 도서 대출 빅데이터 분석

서울대학교 대학혁신센터 데이터통합관리부
사회학과 교수 **이준환**

연구과제

연구의 필요성 및 목적

- 도서관 대출 빅데이터를 이용하여 효율적인 도서구입을 위한 계획 수립을 지원하여 실제 이용자들의 만족도를 제고할 필요가 있음
- 빅데이터를 이용하여 도서 주제별 장서 현황과 이용도를 분석하고 도서관의 주 이용자인 학생 이용자들의 주제별, 도서연령별 장서 이용 현황을 분석하여 서울대학교 중앙도서관의 효율적인 도서구매계획 수립을 지원하기 위한 데이터 활용방안 모색
- 이를 통해 궁극적으로는 장서구성의 질적 수준을 높이고 이용자들의 만족도를 제고할 수 있을 것으로 기대함

연구의 필요성 및 목적

- 서울대학교 중앙도서관 소장자료 및 대출이력 데이터 이용
- 서울대학교 도서관 분류체계 기준 (대분류) (v_book_refine 데이터 이용)
- 분석대상: 학생이용자(학부생 및 대학원생 전체)
  - 28,645명 (2016-2021년)
- 전공계열: 학교 전공계열 구분 기준에 따라 인문사회계열, 자연과학계열, 공학계열, 예체능계열, 의학계열 5개 그룹으로 구분 (진행과정, 단과대, 학과 정보 활용)
  (계열 구분에 있어 간호대학은 이용도서 특성상 의학계열로 분류함)
  - 인문사회계열 (14,001명), 자연과학계열(6,050명), 공학계열(5,847명), 예체능계열(1,463명), 의학계열(1,284명)

분석 내용

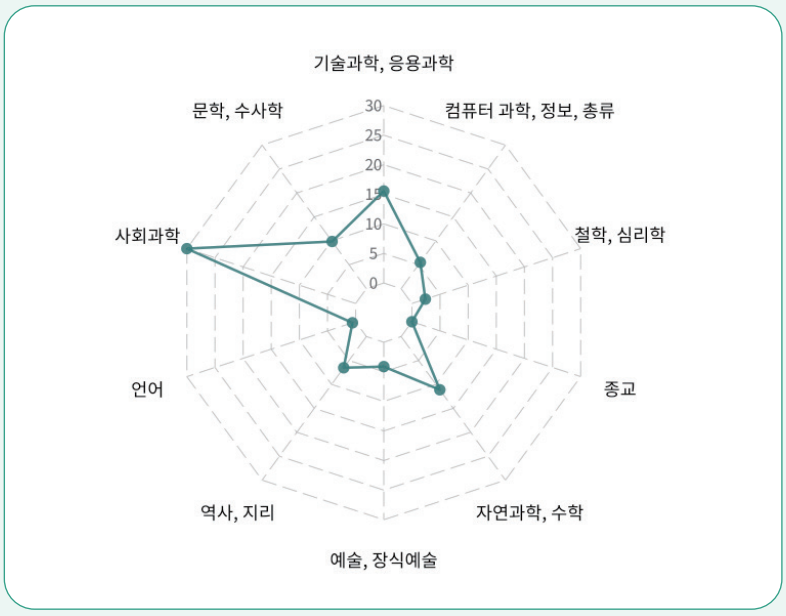
- 중앙도서관 주제별 장서 이용도 분석
- 학생 전공계열에 따른 주제별 이용도 분석
- 학생 전공계열에 따른 도서연령별 자료대출 현황 분석

연구내용

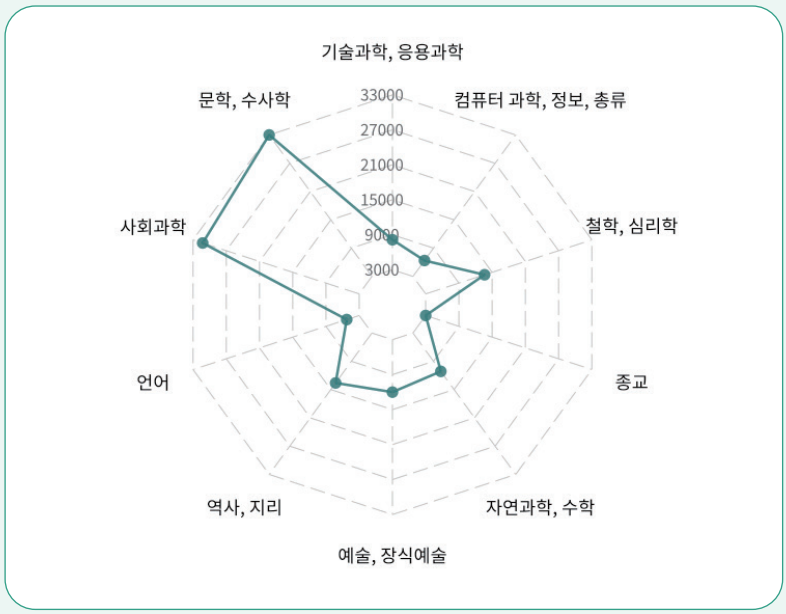
주제별 장서 이용도 분석

- 주제별 장서보유현황

- 2016-2021년 기간 주제별 장서보유현황(6년 평균)은 사회과학(28.50%)과 기술과학 및 응용과학(16.03%)이 가장 많았으며, 언어(3.10%)와 종교(2.61%)가 가장 낮은 구성을 보임
- 한편 학생이용자들의 주제별 대출분포도와 비교하여 살펴보면 사회과학 외에도 문학 · 수사학, 철학 · 심리학 등에 대한 대출 수요가 높은 것으로 드러나 보유 중인 장서와 학생이용자들이 자주 이용하는 장서 간에는 다소 차이가 있는 것으로 보임



[그림 1] 주제별 장서분포도(6년 평균)



[그림 2] 주제별 대출분포도(6년 평균)

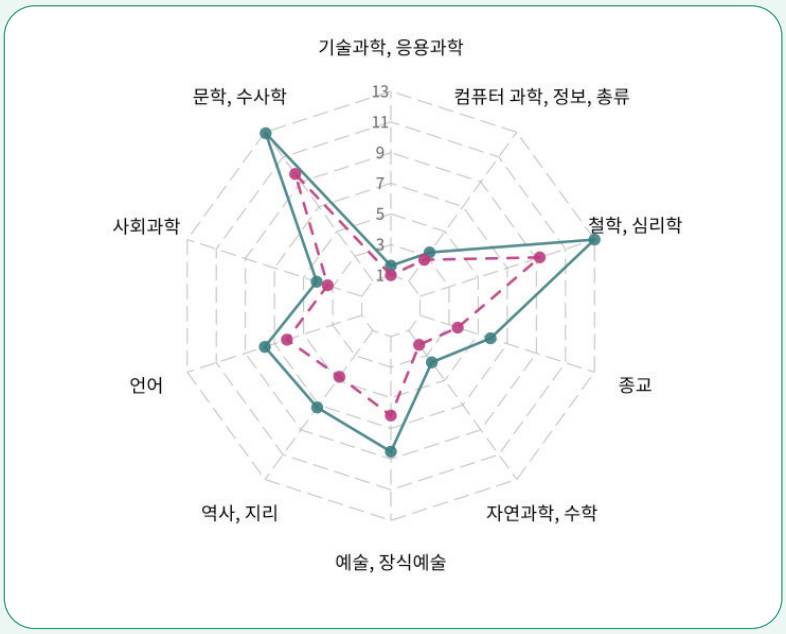
〈표 1〉 중앙도서관 보유 장서 현황 (2016-2021년, 각 연도별 8월 기준)

구분	2016년	2017년	2018년	2019년	2020년	2021년
	N(%)	N(%)	N(%)	N(%)	N(%)	N(%)
컴퓨터 과학, 정보, 총류	177,563 (7.45)	179,515 (7.40)	182,544 (7.41)	184,342 (7.37)	184,969 (7.32)	186,978 (7.23)
철학, 심리학	109,954 (4.62)	112,453 (4.64)	114,791 (4.66)	117,188 (4.69)	118,995 (4.71)	121,739 (4.71)
종교	109,954 (4.62)	62,650 (2.58)	64,158 (2.60)	65,699 (2.63)	66,688 (2.64)	68,067 (2.63)
사회과학	674,509 (28.31)	690,551 (28.47)	703,457 (28.54)	715,359 (28.61)	721,817 (28.57)	736,707 (28.49)
언어	73,510 (3.09)	74,828 (3.08)	75,821 (3.08)	77,851 (3.11)	78,833 (3.12)	79,926 (3.09)
자연과학, 수학	295,841 (12.42)	299,420 (12.34)	302,058 (12.25)	303,688 (12.14)	305,830 (12.11)	310,351 (12.00)
기술과학, 응용과학	388,576 (16.31)	391,797 (16.15)	394,239 (15.99)	396,930 (15.87)	398,532 (15.78)	415,401 (16.07)
예술, 장식예술	144,611 (6.07)	148,680 (6.13)	151,733 (6.16)	154,345 (6.17)	157,716 (6.24)	161,486 (6.25)
문학, 수사학	263,486 (11.06)	268,795 (11.08)	273,827 (11.11)	278,507 (11.14)	282,380 (11.18)	287,833 (11.13)
역사, 지리	193,199 (8.11)	197,123 (8.13)	202,450 (8.21)	206,734 (8.27)	210,308 (8.33)	217,010 (8.39)
합계	2,382,331 (100)	2,425,812 (100.00)	2,465,078 (100.00)	2,500,643 (100.00)	2,526,068 (100.00)	2,585,498 (100.00)

주) N: 장서수(권), %: 비율

주제별 장서회전율

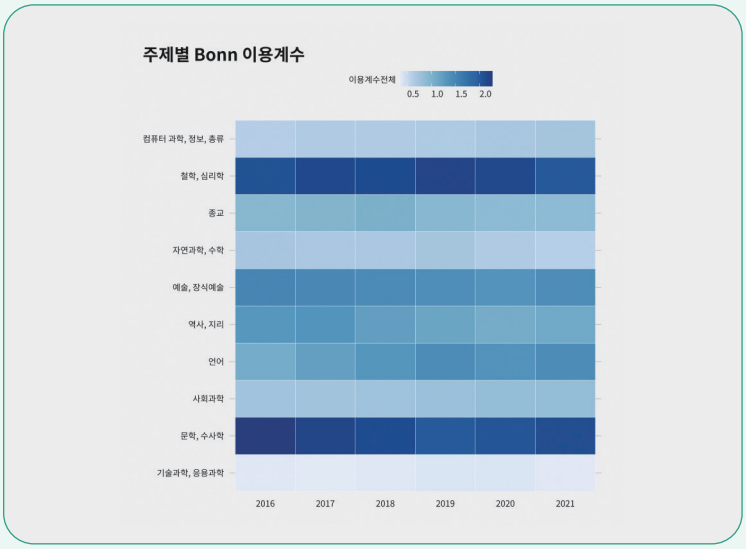
- 장서구성의 적합성을 평가하는 방법으로 장서회전율이 사용됨(이순영·이수상, 2021)
- 장서회전율 = 주제별 도서 대출건수/주제별 보유 도서 수(값이 클수록 도서 대출수가 많음을 의미)
- 코로나19 이전과 이후 시점인 2018년도와 2021년을 대상으로 주제별 장서회전율을 비교분석한 결과 주제별 장서회전 비율은 비슷한 경향을 보이는 반편 2021년의 경우 회전율이 증가한 것으로 나타남



[그림 3] 주제별 장서회전율 분포도
(자주색: 2018년, 청록색: 2021년)

주제별 이용계수

- 이용계수에 의한 주제별 이용정도를 살펴보기 위해 Bonn의 이용계수(Use Factor)를 적용하여 이용정도를 환산(양지안·남영준, 2016)
  - 이용계수(Use Factor) = [(특정 주제자료의 총 대출수 ÷ 도서관 총 대출수) × 100] ÷ [(특정 주제 분야의 장서수 ÷ 도서관의 총 장서수) × 100]
  - 이용계수가 1보다 크면 활발히 이용되는 장서로 평가(양지안·남영주, 2016)
- 이용계수에 의한 주제별 이용정도를 살펴보면 문학 및 수사학(2.13), 철학 및 심리학(2.12)으로 가장 많은 대출이 발생, 반면 가장 낮은 수치를 보인 주제는 컴퓨터 과학, 정보 및 총류(0.66)와 기술과학 및 응용과학(0.38)으로 나타남

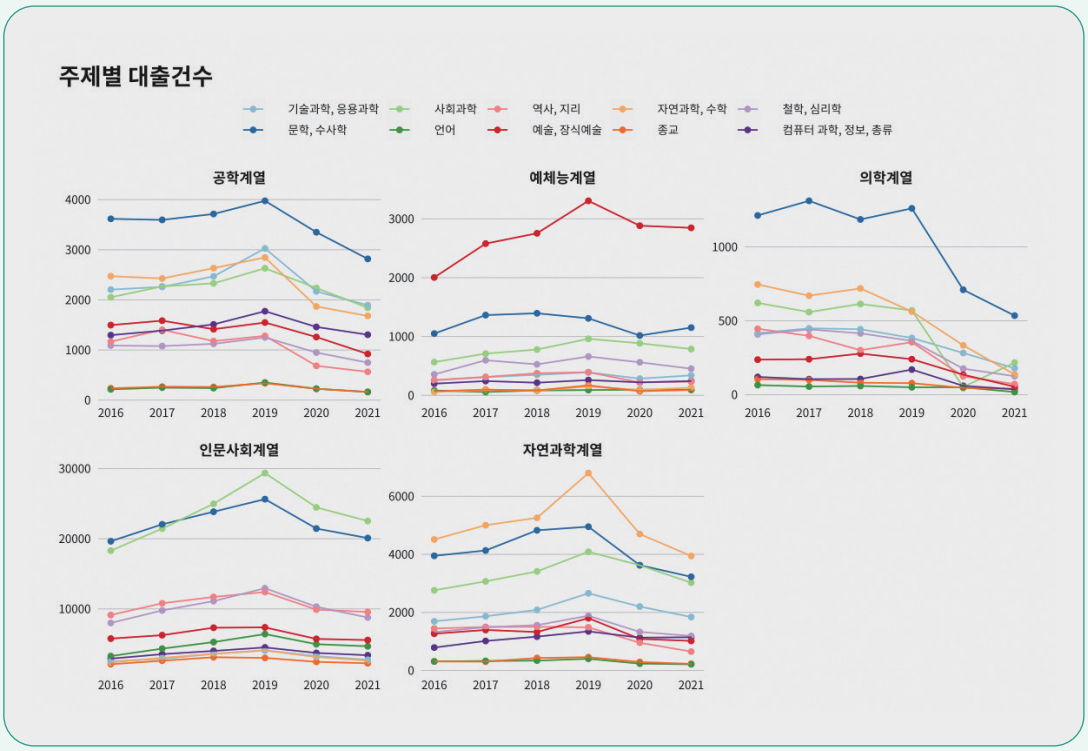


[그림 4] 연도별 주제별 이용계수

학생 전공계열에 따른 주제별 이용도 분석

- 주제별 대출건수

○ 예체능계열, 인문사회계열, 자연과학계열의 경우 전공 관련 주제 도서 대출 건수가 높게 나타나는 반면 공학계열과 의학계열의 경우 문학 및 수사학 도서를 가장 많이 대출하는 경향이 있는 것으로 보임



[그림 5] 전공계열별 주제별 대출건수

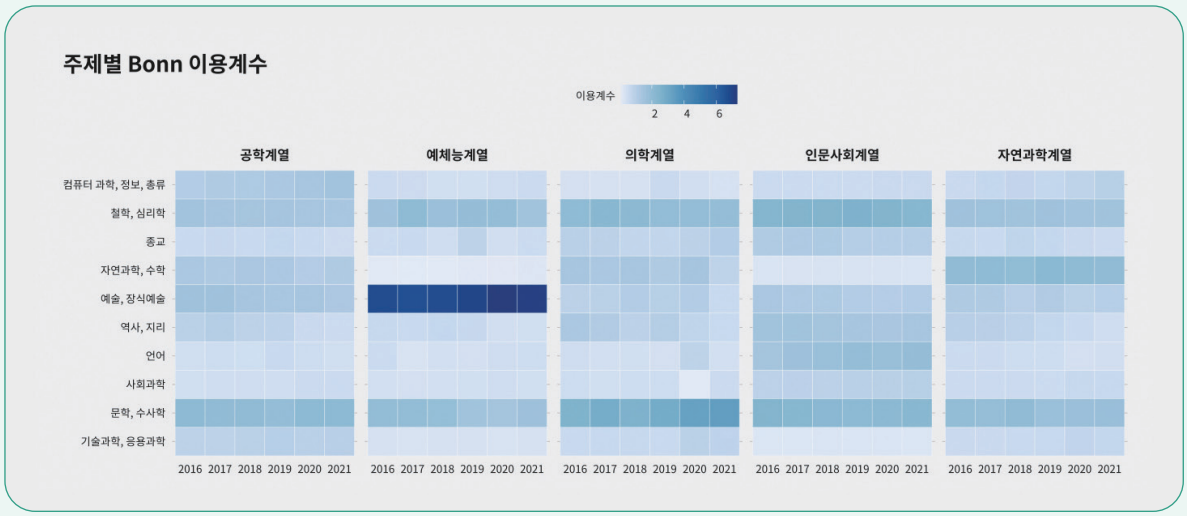
- 주제별이용계수

○ 이용계수 분석 결과 가장 이용정도가 활발한 주제로 인문사회계열은 철학 및 심리학(2.40), 자연과학계열은 자연과학 및 수학(2.02), 공학계열은 문학 및 수사학(2.01), 예체능계열은 예술 및 장식예술 (6.92), 의학계열은 문학 및 수사학(2.87)인 것으로 확인됨

○ 예체능계열의 경우 전공 관련 주제인 예술 및 장식예술 장서 이용도가 매우 활발한 것으로 나타남

〈표 2〉 주제별 계열별 이용계수 평균 (2016-2021년)

구분	인문사회계열	자연과학계열	공학계열	예체능계열	의학계열
컴퓨터 과학, 정보, 총류	0.5544	0.7416	1.2631	0.4935	0.3960
철학, 심리학	2.4038	1.5410	1.4066	1.7546	2.0073
종교	1.1027	0.6296	0.5989	0.5743	0.8777
사회과학	0.9139	0.5785	0.4973	0.4282	0.4320
언어	1.7147	0.4822	0.4871	0.4341	0.4956
자연과학, 수학	0.2836	2.0224	1.1963	0.1273	1.2300
기술과학, 응용과학	0.2197	0.6343	0.9239	0.3143	0.6966
예술, 장식예술	1.1435	1.0460	1.4030	6.9165	0.9241
문학, 수사학	2.2183	1.8252	2.0091	1.7252	2.8690
역사, 지리	1.4302	0.7480	0.7908	0.5697	0.9479



[그림 4] 연도별 주제별 이용계수

- 계열별 1인당 장서 대출현황

○ 계열별 1인당 대출건수 산출

- 계열별 1인당 대출건수 = 계열별 총 대출건수 ÷ 계열별 재적생 수(재학생 및 휴학생)
- 계열별 인원수의 경우 연도별 4월 기준 고등교육통계자료를 이용하였음

○ 계열별 전체 재적생 수를 고려했을 때 1인당 대출건수가 가장 많은 계열은 인문사회계열인 반면 가장 적은 계열은 의학계열인 것으로 확인됨

〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열	자연과학계열	공학계열	예체능계열	의학계열
2016	5.66	2.23	2.03	2.39	1.74
2017	6.66	2.46	2.12	3.09	1.75
2018	7.56	2.70	2.19	3.19	1.67
2019	8.52	3.18	2.50	3.78	1.63
2020	7.00	2.33	1.87	3.05	0.79
2021	6.44	1.97	1.52	3.04	0.57

- 계열별 주제편중도 분석

○ 계열별 주제 편중도 산출

- 최대 주제편중도 = (계열별 최대 대출 주제 분야 대출건수 ÷ 계열별 총 대출건수) × 100
- 최소 주제편중도 = (계열별 최소 대출 주제 분야 대출건수 ÷ 계열별 총 대출건수) × 100
- 주제 편차정도 = 최대 주제편중도 ÷ 최소 주제편중도
- 주제 편차정도가 크면 주제 분야와 비주제 분야 사이의 이용정도의 차이가 크다는 의미인 반면, 해당 값의 수치가 작으면 다양한 주제를 균형적으로 이용한다는 것을 의미함(양지안·남영준, 2016)

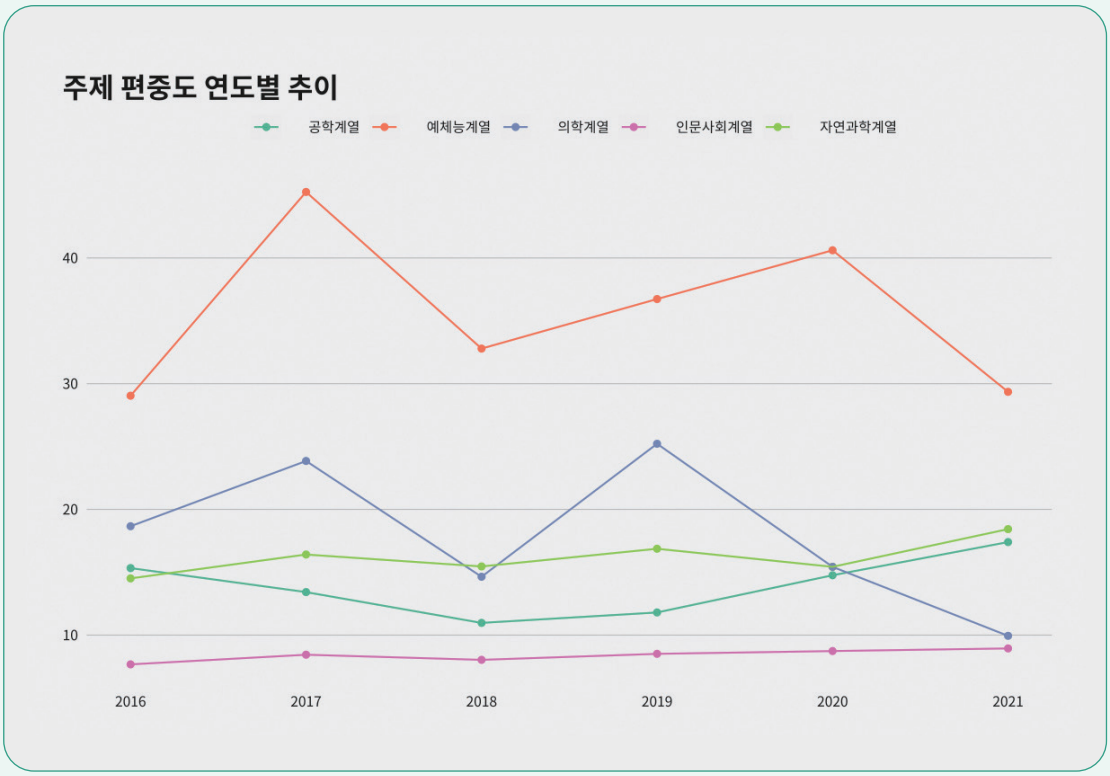
〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열		자연과학계열		공학계열		예체능계열		의학계열	
2016	문학, 수사학	24.72	자연과학, 수학	24.58	문학, 수사학	22.82	예술, 장식예술	40.91	문학, 수사학	27.71
2017	문학, 수사학	25.42	자연과학, 수학	24.89	문학, 수사학	21.76	예술, 장식예술	40.56	문학, 수사학	30.30
2018	사회과학	25.35	자연과학, 수학	23.99	자연과학, 수학	15.60	예술, 장식예술	41.45	문학, 수사학	28.22
2019	문학, 수사학	23.35	자연과학, 수학	26.28	문학, 수사학	20.90	예술, 장식예술	42.97	문학, 수사학	31.25
2020	문학, 수사학	23.99	문학, 수사학	18.91	문학, 수사학	23.21	예술, 장식예술	45.54	문학, 수사학	36.06
2021	문학, 수사학	24.44	자연과학, 수학	23.93	문학, 수사학	23.29	예술, 장식예술	44.67	기술과학, 응용과학	12.67

〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열		자연과학계열		공학계열		예체능계열		의학계열	
2016	문학, 수사학	3.22	언어	1.69	종교	1.49	종교	1.41	사회과학	1.49
2017	종교	3.01	종교	1.52	종교	1.62	언어	0.90	언어	1.27
2018	종교	3.16	언어	1.55	언어	1.42	종교	1.26	종교	1.93
2019	종교	2.74	언어	1.56	종교	1.77	언어	1.17	언어	1.24
2020	종교	2.75	언어	1.23	언어 & 종교	1.57	종교	1.12	종교	2.34
2021	종교	2.73	언어	1.30	언어	1.34	언어	1.52	언어	1.27

○ 도서 이용행태에 있어 인문사회계열이 이용편차가 가장 작게 나타나 주제별로 균형 있게 이용하는 것으로 보임. 한편 주제 편차정도가 가장 크게 나타난 계열은 예체능계열로 타 주제 대비 전공 관련 주제인 예술 및 장식 예술 도서에 대한 선호도가 매우 높은 것으로 드러남



[그림 7] 전공계열별 주제 편중도 변화 추이 (2016-2021년)

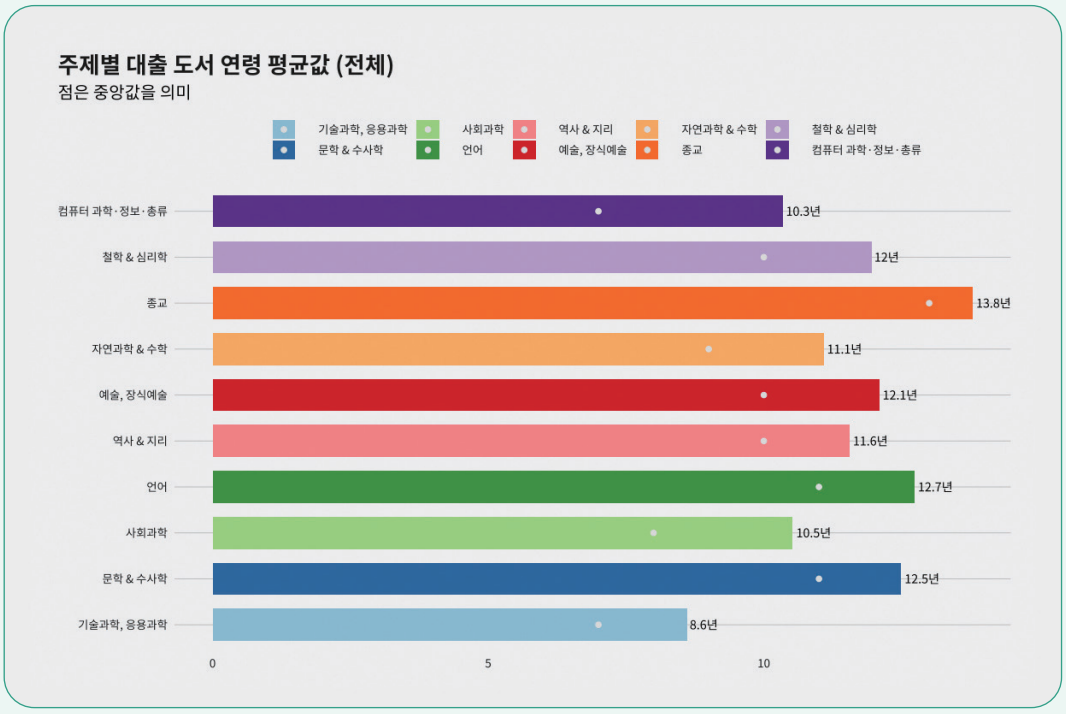
학생 전공계열에 따른 도서연령별 자료대출 현황 분석

- 학생 전공계열에 따른 도서연령별 자료대출 현황 분석

○ 도서연령 = 대출연도 - 출판연도

○ 주제에 따라 이용되는 자료의 최신성에 차이가 있는 것으로 나타남

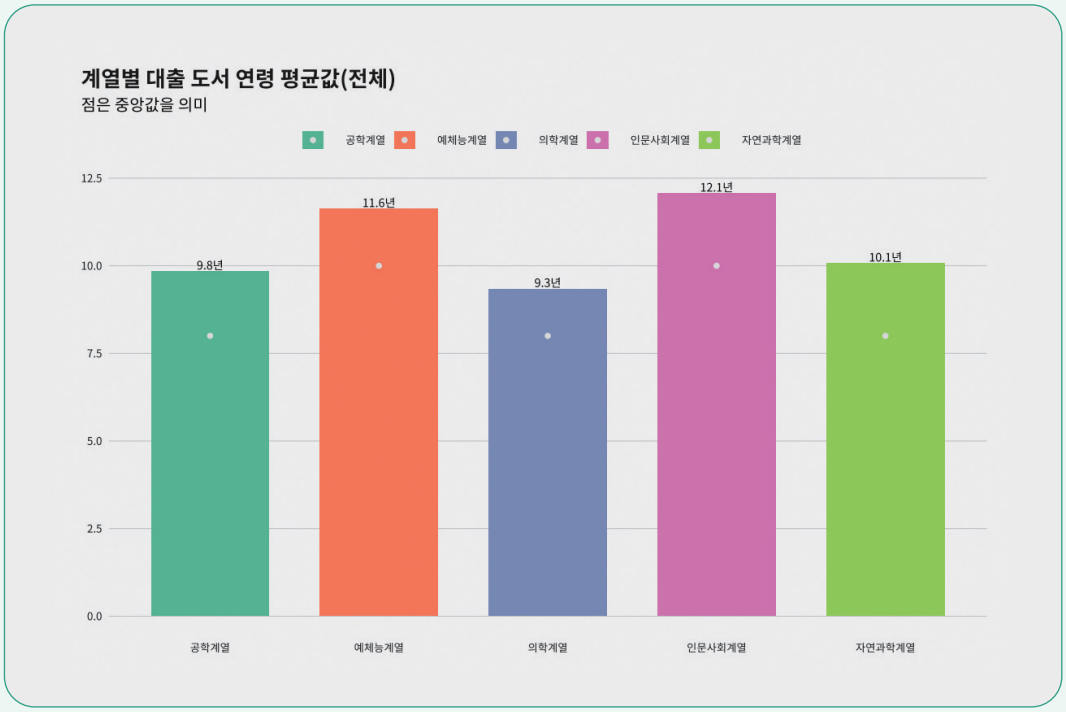
○ 평균보다 작은 값을 가진 분야에 해당하는 주제는 컴퓨터과학·정보·총류, 자연과학·수학, 사회과학, 기술과학·응용과학로 확인됨



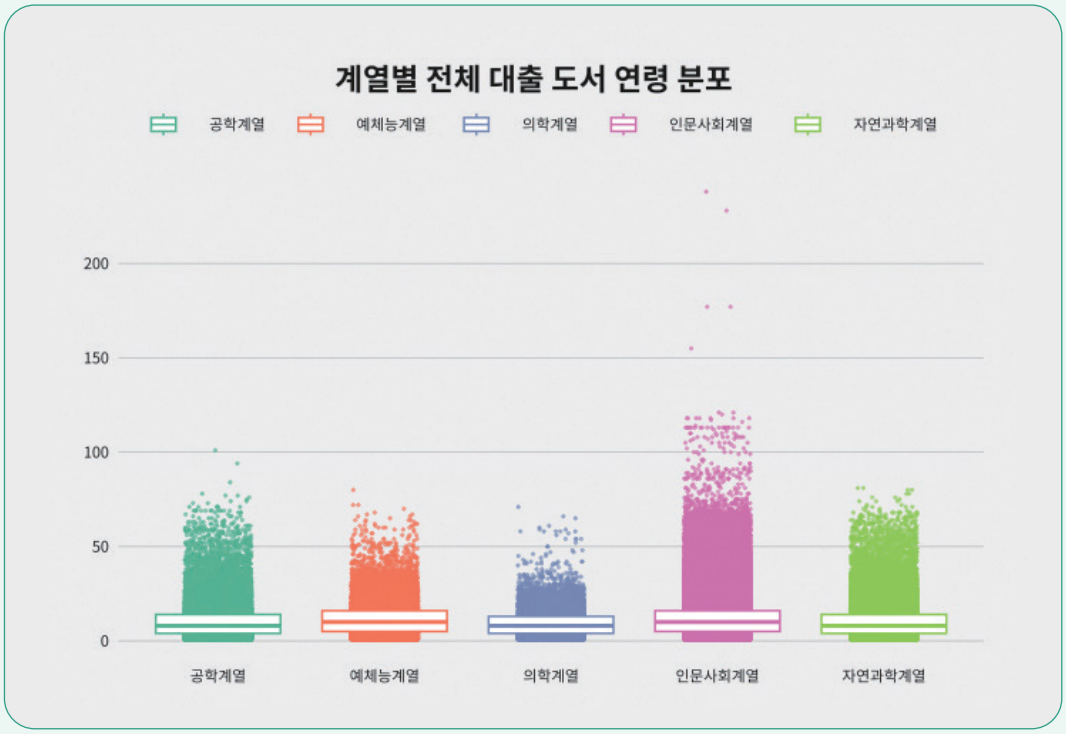
[그림 8] 주제별 대출 도서 연령 평균값

- 도서연령에 따른 자료대출 현황 분석

○ 의학계열과 공학계열은 비교적 최신 도서를 대출하는 경향을 보이는 반면 인문사회계열의 경우 그러한 경향이 가장 약한 것으로 나타남



[그림 9] 계열별 대출 도서 연령 평균값

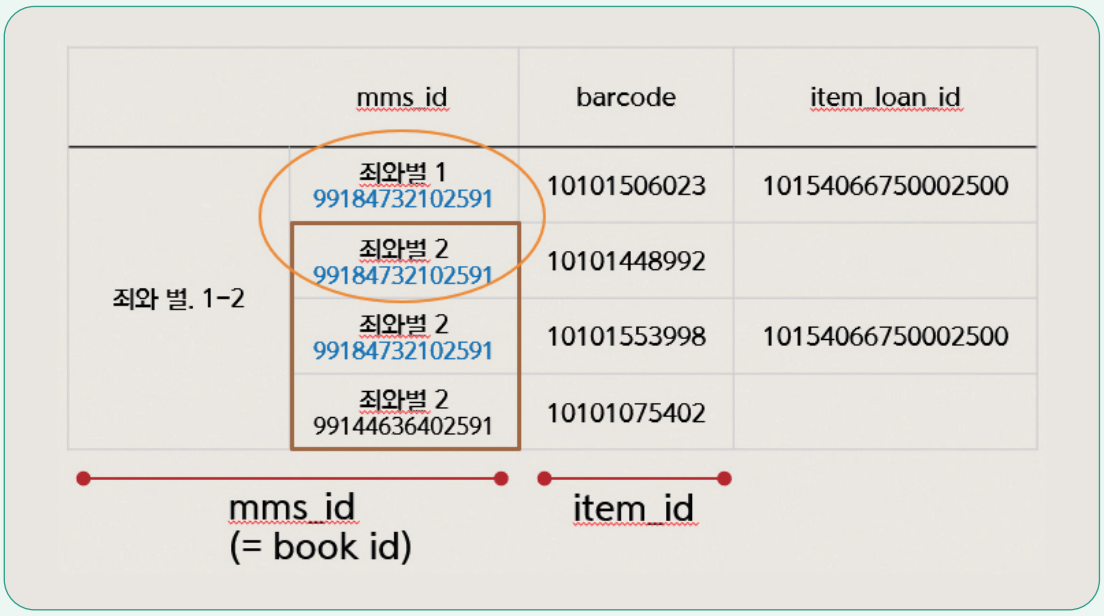


[그림 10] 전공계열별 대출 도서 연령 분포

연구결과

대출데이터 분류체계 개선 필요

- 현재 데이터는 여러 권수로 구성된 단행본의 경우 권수가 다른 동일 제목의 단행본의 경우 동일한 id(mms_id)를 갖는 방식으로 설정되어 있음
- 한편 바코드 정보(item_id, barcode)를 이용할 경우 각각의 단행본 서적별로 고유의 값을 가지고 있어 동일한 제목, 동일한 권수의 서적 여부를 구분할 수 없는 구조임
- 단행본 권수(volume number)별 별도 고유 번호를 부여하거나 단행본 제목에 책 권수를 반영하여 동일한 제목, 동일한 권수의 개별 서적들을 동일한 단행본으로 분류할 수 있도록 새로운 분류체계 마련이 필요할 것으로 보임
- 이를 이용하여 향후 "핵심장서" 분석에 활용하여 핵심장서만을 대상으로 대출 패턴을 분석하여 이용이 활발한 자료와 이용되지 않는 자료를 구별할 필요가 있음 (트루스웰 80/20 법칙에 따른 핵심장서 분류)
- 핵심장서 대출 패턴 분석을 통해 대출이 빈번한 핵심장서에 대한 구매를 통해 이용자 만족도 제고 방안을 모색할 수 있을 것으로 기대됨



[그림 11] 현재 데이터 문제 사례

- 또한 동일 단행본에 대해 도서 주제에 따른 분류가 다른 경우가 있어 개선이 필요할 것으로 보임. 가령 (그림 12)와 같이 동일한 도서임에도 불구하고 상이하게 분류가 되는 한계가 존재함

publishe	title	author_y	call_numbe	section_desc	division_	categor
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서

[그림 12] 현재 데이터 주제별 분류 문제 사례

출판연도 데이터 관리 개선 필요

- 도서 연령 산출 시 출판년도와 대출연도에 대한 정보를 이용하였음. 그러나 출판연도에 대한 정보가 없거나 잘못 기록된 경우가 있어 개선이 필요할 것으로 보임. 출판연도에 대한 정보가 없을 경우 출판일자에 대한 정보를 활용하였으나, 출판일자 데이터를 정제하는데 추가적인 노력이 요구되었음
- 출판경과년수에 따른 대출 패턴을 보다 정확히 분석하기 위해 대출연도에 대한 데이터 관리 개선이 필요할 것으로 판단됨

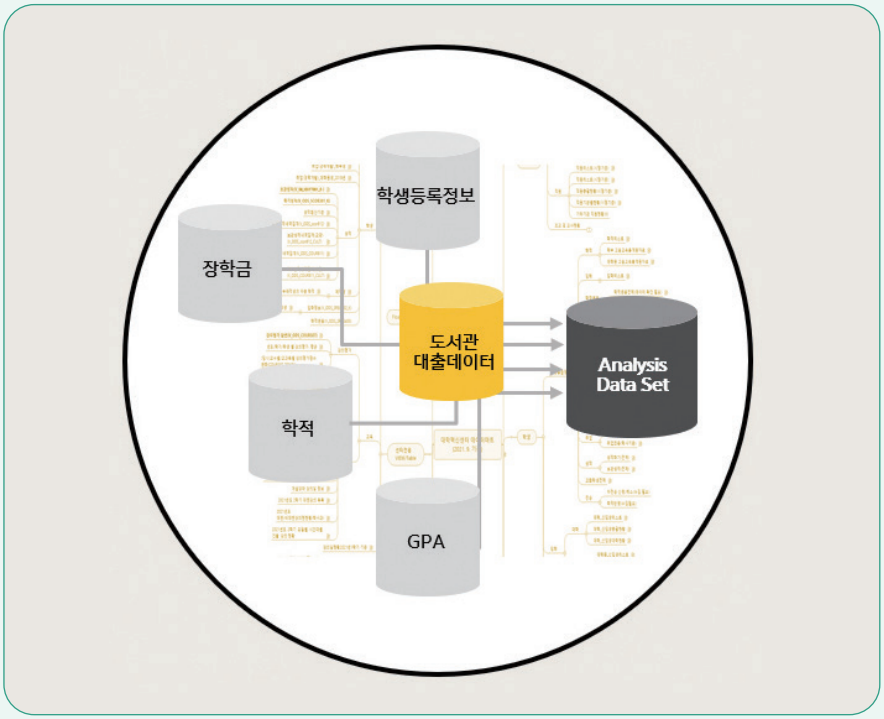
정보화본부가 관리하는 학내 데이터와의 연계방안 모색 필요

- 정보화본부는 학내DB에 축적되는 다양한 데이터를 관리하고 있으며 데이터마트를 통해 원본데이터를 제공하고 있음. 데이터마트에는 학적, 교원, 직원, 졸업, 등록, 장학, 수업, 강의평가, 성적, 국제협력 외에도 건물 단위 고정IP 및 무선랜(WIFI) 접속정보 등에 관한 데이터가 관리됨
- 대학혁신센터는 데이터마트의 데이터를 활용하여 학내 정책 수립을 지원하기 위한 다양한 분석을 수행하고 있는 바, 향후 도서관과의 긴밀한 협력을 통해 도서관에서 관리하는 다양한 빅데이터 분석을 통해 의미있는 분석을 수행할 수 있을 것으로 기대됨

※ 대학혁신센터는 2021년 데이터분석과제로 장학복지과와의 협업을 통해 '학내 밀집도 분석을 통한 효율적인 코로나19 방역대책 마련'을 주제로 고정IP 및 와이파이 빅데이터를 이용하여 서울대 건물 단위 유동인구를 분석하였음. 해당 분석의 일환으로 관정관과 중앙도서관의 시간대별 유동인구 분석을 수행한 바 있음

○ 당 분석을 수행는데 있어 정보화본부가 관리하는 데이터마트(DM)로부터 학생 전공계열에 대한 정보를 담고 있는 학부 및 대학원 고등교육통계원자료를 활용하였음. 이를 vuser_refine 데이터에 포함된 신분, 소속대학 및 소속학과 정보와 매칭시켜 학생이용자들의 전공계열에 관한 정보를 이용하여 분석을 수행하였음

○ 향후 도서관 빅데이터를 정보화본부의 데이터마트와 연계할 수 있는 구체적인 방안에 대한 논의가 필요할 것으로 보임. 구성원별 ID의 비식별화 방식을 통일할 경우 데이터 연계를 보다 용이하게 할 수 있을 것으로 판단되며, 이와 같은 노력을 통해 도서관 빅데이터를 이용한 다양한 분석 수행이 가능할 것으로 기대됨



[그림 13] 도서관 빅데이터와 정보화본부 DM 데이터 간의 연계 예시

향후계획

핵심장서 분석

○ 도서관 단행본의 이용률 분석을 통해 핵심장서를 산출할 수 있으며, 가장 대표적인 방법으로 Trueswell(1969)의 80/20법칙(약 20%의 장서가 전체 대출의 약 80%를 차지)이 있음

○ Trueswell은 단행본의 대출횟수가 높은 장서는 장서활용도가 상대적으로 높다는 것을 복수의 연구를 통해 밝혀내었음(양지안, 2017)

○ 서울대학교의 경우에도 대출횟수를 이용하여 전체 대출건수의 과반(50%) 또는 그 이상(80%)을 차지하는 대출

도서 법칙을 확인하여 연도별 대출 패턴을 상세히 살펴보면 보다 의미있는 시사점 도출이 가능할 것으로 보임

- 이와 같은 분석을 수행하기 위해서는 상기에서 제안한 바와 같이 장서 권수별 고유 번호를 부여하는 등의 대출 데이터 분류체계 개선이 선행될 필요가 있음
- 향후 핵심장서 대출 패턴 분석을 통해 이용이 활발한 자료와 이용되지 않는 자료를 중심으로 효율적 장서 구성 및 구매 방안 마련이 가능할 것으로 사료됨

장서 대출 패턴 관련 대시보드 구축

- 빅데이터에 가치를 부여하고 이를 통해 사용자들에게 인사이트를 제공하는 데 있어 데이터 시각화의 중요성이 높아지고 있음
- 빅데이터 분석의 핵심은 얼마나 많은 양의 데이터를 사용했는지가 아니라, 수집된 빅데이터의 분석 결과를 쉽게 이해할 수 있도록 하는 시각화를 통한 유용성에 있음
- 도서관 빅데이터의 다양한 활용이 가능할 수 있도록 기반과 체계를 강화하고, 이를 토대로 다수의 시각화 차트를 종합하여 실시간으로 보여줄 수 있는 데이터 시각화 대시보드의 구축을 통해 적시에 인사이트를 도출할 수 있을 것으로 기대됨

참고문헌

양지안, 남영준(2016). 대학도서관 단행본 대출이력통계를 통한 집중장서에 관한 연구. 한국문헌정보학회지, 50(3): 429-435.

양지안 (2017). 대학도서관 대출데이터분석을 통한 장서 이용행태 및 특성에 관한 연구. 한국도서관·정보학회지, 48(2), 263-293.

이순영, 이수상(2021). 부산지역 공공도서관의 빅데이터 분석 연구: 도서관 정보나루 장서/대출데이터를 중심으로. 한국문헌정보학회지, 55(4): 89-114.