95-b-7

온라인 데이타베이스정보의 분포특성 분석

The Distribution of Citations in Online Databases

이 효 숙
(한남대학교 문헌정보학과 강사)

Hyo Sook Lee
Dept. of Library and Information Science, Hannam Univ.

Twenty six databases relevant to 'trade' have been searched to test for a Bradford's law of scatter. Citations in the databases adhere to the linearity of Bradford's distribution, however, they show that the concentration of citations in core databases is not as great as being expected.

I. 서론

현재 생산되고 있는 데이타베이스들은 그 수와 종류에 있어서 매우 다양하고 급속한 증가 추세를 보이고 있다. 그러나, 특정 주제에 관한 탐색시 단일의 데이타베이스 탐색으로는 만족할만한 수준의 검색결과를 얻지 못하는 경우가 적지 않다.

본 논문에서는 무역분야에서 이용가능한 데이타베이스를 조사하고 데이타베이스 정보의 분포특성을 분석한다. 연구목적은 이 분야 정보탐색에 있어 원하는 재현수준에 이르기 위해 탐색되어야 할 데이타베이스 수를 파악하고, 데이타베이스에 수록된 정보의 분포패턴 및 정보의 집중현상을 조사하기 위한 것이다.

II. 정보의 분포특성과 수량적 측정

정보의 분포특성에 대한 수량적 측정은 정보처리 과정의 효율성을 증가시킨다. 이것은 또한 현 서비스의 결함을 분석하고, 정보생산 경향에 대한 예측을 통해 효과적인 정보이용으로 이어지게 한다.

정보의 분산 현상에 대해 설명한 브래드포드 법칙은 이미 많은 연구문헌을 통해 소개되고, 논의된 바 있다. 브래드포드 법칙에 대한 연구는 초기의 학술잡지 논문에 대한 분석에서부터 비롯하여 이 법칙에 대한 수학적 해석과 이론의 발전으로 이어졌고 최근에 와서는 특히 정보전문가들의 세부적 관심영역에 적용하여 이 법칙의 예측적 능력을 확장하고 있다.

본 논문에서는 우리말 온라인 데이타베이스 정보의 분포특성에 대한 분석을 위해 브래드포드 분산법칙을 적용한다.

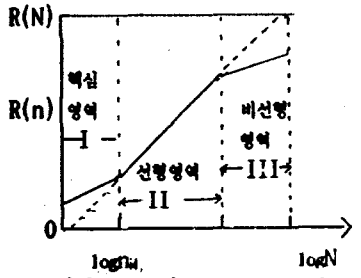
1. 브래드포드 분산법칙

브래드포드 분산법칙은 학술잡지와 생산된 논문

간에 공식 (1)과 같은 규칙적인 관계가 있음을 설명하고 있다.

$$R(n) = k \log(n) \quad \text{.....(1)}$$

R(n)은 적합한 논문의 총수, n은 논문을 생산한 잡지의 총수를 나타낸다. k는 상수로서 주제에 따라 다른 값을 가질 수 있다. 이것을 그래프로 표현한 것이 <그림-1 >과 같다. < 그림-1 >은 세부분으로 구성됨을 볼 수 있다. 첫번째 부분은 상승곡선의 형태를 갖는 고밀도 분포영역이다. 두번째 부분은 n과 R(n)의 관계가 선형적인 관계를 나타낸다. 세번째 부분은 n과 R(n)의 관계가 선형성으로부터 이탈되는 현상을 나타낸 것으로 이에 대한 해석은 연구자 간에 일치된 견해를 보이지 않는다.



< 그림-1 > 브래드포드 분포의 일반적 형태

온라인 데이타베이스 정보의 분포에 대한 테스트를 위해서 공식 (1)을 적용하고, < 그림-1 >과 같은 분포패턴을 보이는지를 조사하였다.

III. 데이타 수집

1. 탐색된 주제분야

탐색이 실시된 분야는 무역분야의 한 세부주제로서 '무역제도 및 동향'에 관한 분야이다. 일반적으로 무역분야에는 다양한 많은 주제들이 관련되는 학제적 성격을 갖는다. 실제 무역업무에 종사하고 있는 실무자들은 절차 및 실무요령, 업체 및 상품정보 등과 같은 직접적이고 사실적인 정보 이외에도 대외무역 활동에 중요한 영향이 있는 법령이나 기술수준 등을 포함하여 한 국가의 최근 변화에 관한 여러 측면에서의 정보를 필요로 하고 있다.

질문을 수집할 목적과 주제분야에 대한 보다 정확한 이해를 위해서 국내의 무역 종사자들이 주로 이용하고 있는 한국무역협회, 대한무역진흥공사, 산업기술정보원 산업무역실의 상담자와 면담하였다. 그리고 이 세기관으로부터 그동안 상담을 통하여 의뢰된 이용자의 질문내용을 수집하였다.

2. 질문

수집된 질문은 209건으로 이 가운데 조사된 세기관에서 가장 많은 비율로 의뢰된 질문들을 기초로 다음과 같은 6개의 탐색질문이 선택되었다. 이것은 특정 상품이나 국가에 한정하지 않는 질문으로 작성되었다. 탐색식은 조사대상 데이타베이스들로부터 재현율이 높은 검색결과를 가져올 수 있도록 본 연구자가 작성하였다.

i) 수출입동향 및 수출입 업무에서 고려하여야 할 절차 및 요령
ii) 관세율 및 관세대상에 관한 정보와 통관절차
iii) 무역 금융과 신용장 및 신용도 조사
iv) 대외무역에 관련된 법규, 법령,공시
v) 국내외 시장조사, 산업시장 동향
vi) 전시회, 박람회 및 기타 홍보활동의 일정, 개최장소, 품목, 특징

탐색시 탐색전략으로 빌딩블럭 방식을 적용하였다. 이를 위해 각 시스템에서 제공하는 색인어에 관한 정보를 탐색한 후 탐색식 작성에서 이를 사용하였다.

3. 데이타베이스

탐색대상 데이타베이스로 26종이 선택되었고, 데이타베이스 선정을 위해 사용된 자료는 '데이타베이스 총람'과 '국가전산 총람'을 참고로 하였다.

26개 데이타베이스를 제공하는 정보시스템은 KOTIS(8), KOTRA-NET(1), KINITHR(3), ETLARS(3), 천리안(10), 행정종합정보(1)등이다.

동일한 탐색어가 각 데이타베이스에 대해 탐색되었으나, 탐색에 사용되는 검색시스템이 다름에 따라 탐색에 사용된 명령어나 탐색기법에서는 다소 차이가 있다.

탐색된 데이타베이스의 정보는 다음과 같은 다양한 정보원들로부터 색인작성된 것이다.

- 정부간행물, 학술잡지 및 일반잡지,단행본, 편람, 보도자료, 법령, 회의록, 보고서, 신문, 해설기사.

검색된 정보의 적합성을 판정하기 위해 각 데이타베이스로부터 30건의 정보를 다운로드하여 총 780건의 표본문헌을 수집하였다. 적합성 판정은 무역업무에 종사하는 1명의 실무자가 하였고, 탐색된 질문이 특정성이 높은 질문이 아니므로 적합성 판정 척도로는 적합, 부적합 여부만을 사용하여 보다 상세한 적합성 정의는 적용되지 않았다. 표본문헌에 대한 적합성 판정결과는 각 데이타베이스에 대해 0.95의 신뢰수준에서 부적합문헌의 비율 추정에 사용되었다.

IV. 데이타 분석 및 논의

1. 데이타베이스 수록 건수와 데이타베이스의 선택

탐색 및 적합성 판정 후에 각 데이타베이스별로 정보의 수록건수를 조사하였다. < 표-1 >은 탐색질문에 의해 검색된 정보 건수에 따라 내림차순으로 작성된 것이다. 이 순위는 '무역제도 및 동향'에 관한 데이타베이스 탐색에서 데이타베이스선택시에 참고될 수 있을 것이다.

반면에 < 표-1 >에서 수록건수 간에 미미한 차이를 보이는 데이타베이스 간의 순위는 다소 변화될 수 있다. 그 이유는 데이타베이스 상의 부적합문헌의 비율은 표본문헌에 의한 추정치이기 때문이다.

2. 정보의 분포

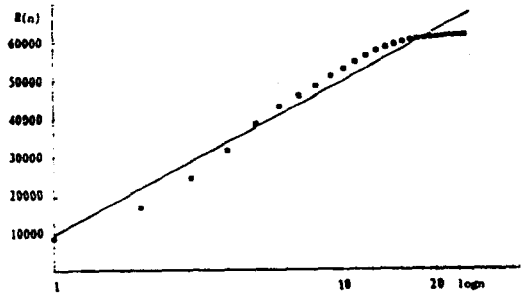
각 데이타베이스에서 검색된 정보건수를 모두 합하여 전체 집단을 구성하였다. 데이타베이스에 수록된 정보의 분포현상이 브래드포드 분산을 따르는지 테스트하기 위해 공식 (1)을 적용한 결과,< 그림-2 >와 같이 작성되었다. < 그림-2 >에서 점선에 의한 그래프는 실제값, 직선은 이에 대한 예측치 산출에 의한 최적선이다.

작성된 그래프를 통해 볼 때 데이타베이스간에 분포하는 정보는 브래드포드 분산법칙의 선형적 부분에 일치됨을 보여준다. 반면에, 핵심영역에 해당하는 부분은 나타나 있지 않다. 즉, 조사된 "무역제도 및 동향"분야의 탐색 결과, 핵심적 데이타

<표-1> 데이타베이스별 검색된 정보건수

데이타베이스명	검색정보건수	신뢰구간
JOINS속보	8393	-+ 0.0946
행정종합정보	7996	-+ 0.1395
무역관련법규	7783	-+ 0.0566
서울경제	7183	-+ 0.0946
일간무역기사	7181	-+ 0.1073
BIST	4385	-+ 0.148
국내외산업동향	2841	-+ 0.1073
BIIT	2637	-+ 0.0946
수출입요령	2439	-+ 0.0946
농수산해외시장속보	1959	-+ 0.0566
무역조사보고서	1683	-+ 0.1073
운송정보	1612	-+ 0.1487
주간기술동향	1556	-+ 0.0946
EPIC	751	-+ 0.1262
하이테크정보	686	-+ 0.1446
통상정보	636	-+ 0.0566
ITCH	506	-+ 0.1520
수출입절차	227	-+ 0.0566
농수산공사지원사업	218	-+ 0.1073
세계전기통신동향	170	-+ 0.1073
국제협약	162	-+ 0.0566
기술연구보고서	120	-+ 0.0947
농수산사업통합실시요령	94	-+ 0.1487
세계농업정책동향	64	모집단
외국인투자정보	59	-+ 0.0786
표준화소식	34	-+ 0.0566

베이스에 포함된 정보건수는 예상한 것 만큼 크게 집중적이지 않으며, 조사된 데이타베이스 수와 수록된 정보건수 간에는 전반적인 선형 관계를 이룬다.



< 그림-2 > 데이타베이스정보의 분포도

< 표-2 >는 누적 집계에 의한 정보건수와 누적비율을 나타낸 것이다. 이 결과는 이 분야에 관한 정보를 얻기 위해 탐색해야 할 데이타베이스 수에 대한 예측을 가능하게 한다.

'무역제도 및 동향'에 관한 데이타베이스 탐색결과를 1: a: a^2 에 근사화 한 결과, 3종의 데이타베이스 탐색시에 전체 건수 중 약 1/3, 그리고 6 종 (3×2^1)에서 다음 1/3 까지, 그리고 적어도 17종의 데이타베이스를 탐색하여야 거의 나머지 1/3에 해당하는 정보를 검색하는 결과가 된다.

< 표-2 > 데이타베이스정보의 누적율

데이타베이스명	누적건수	누적율(%)
JOINS속보	8393	13.6
행정종합정보	16389	26.7
무역관련법규	24172	39.3
서울경제	31355	51.0
일간무역기사	38536	62.7
BIST	42921	69.9
국내외산업동향	45762	74.5
BIIT	48399	78.8
수출입요령	50838	82.8
농수산해외시장속보	52797	86.0
무역조사보고서	54480	88.7
운송정보	56092	91.3
주간기술동향	57648	93.9
EPIC	58399	95.1
하이테크정보	59085	96.2
통상정보	59721	97.3
ITCH	60227	98.1
수출입절차	60454	98.4
농수산공사지원사업	60672	98.8
세계전기통신동향	60842	99.1
국제협약	61004	99.3
기술연구보고서	61124	99.5
농수산사업통합실시요령	61218	99.7
세계농업정책동향	61282	99.8
외국인투자정보	61341	99.9
표준화소식	61375	100.0

이분야 탐색에서 어떤 데이타베이스도 전체 정보건수의 13.6% 이상을 제공하지 않는다. 즉, 특정 데이타베이스에 한정한 정보 탐색은 매우 제한적 검색결과를 가져올 수 있음을 나타낸다. 따라서 탐색자가 원하는 재현 수준에 따라 탐색할 데이타베이스의 수가 결정되어야 할 것이다. 70% 이상의 재현 수준을 원하는 탐색에서는 7종 이상의 데이타베이스가 탐색되어야 한다. 반면에, 매우 높은 재현 수준을 요구할 경우 탐색할 데이타베이스 수는 매우 증가한다.

v. 결론

'무역제도 및 동향'에 관한 우리말 데이타베이스 정보의 분포는 브래드포드 분산법칙에 의한 선형적 특성에 일치한다. 반면에 소수 데이타베이스에 정보가 집중적으로 분포되어 있는 현상은 보이지 않는다.

데이타베이스들은 이 분야의 정보 수록 건수에 따라 3: 6: 17의 세 개의 영역으로 구분된다. 각 영역은 거의 동일한 수의 정보를 가지며 첫 영역에서 세번째 영역으로 감에 따라 첫영역과 동일한 수의 정보를 검색하기 위해 탐색할 데이타베이스의 수는 매우 증가한다.

이상과 같은 분석결과들은 다중데이타베이스(multidatabase) 탐색기능을 제공하는 우리말 정보검색시스템 설계시에 실제로 다음과 같이 적용되어야 할 것이다.

탐색자가 데이타베이스 추가 및 삭제시에 가장 적절할 후보데이타베이스에 대해 간략한 안내정보를 주고, 재현수준과 관련하여 시스템에서 탐색가능한 데이타베이스 수와 순위 정보를 제공한다.

본 연구에서는 데이타베이스 간에 포함될 수 있는 중복정보 수준은 고려되지 않았다. 따라서 후속 연구에서는 이를 분석하여 각 데이타베이스가 갖는 고유한 정보에 기초한 분포특성이 연구되어야 할 것이다.

참고문헌

Bookstein, A. (1976) " The Bibliometric Distributions, " Library Quarterly 46 : 416-423.

Brookes, B.C. (1977) " Numerical Methods of Bibliographic Analysis," Library Trends 22 : 18-23.

Brookes, B.C. (1977) " Theory of the Bradford Law," Journal of Documentation 33 :180-209,

Cline, Gloria S. (1981) " Application of Bradford's Law to Citation Data," College & Research Libraries 42 : 53-61.

Evans, J.E. (1980) " Database Selection in an Academic Library," Online 4(2): 35-43.

Epstein, B.A. and J.J. Angier (1980) "Multi-database Searching in the Behavioural Sciences, Pt.1" Database. 3(3) :9-15.

Lancaster, F.W. & J. Lee (1985) " Bibliometric Techniques applied to issues management : a case study " JASIS 36 : 389-397.

Meyer, D.E. & D. Ruiz (1990) " End-user Selection of Databases : Business/Law," Database 13(4): 35-42.

Tenopir, C. (1989) Issues in Online Databases Searching. Englewood; Libraries Unlimited. 111-121.

Yerkey, N. & M. Glogowski (1990) "Scatter of Library and Information Science Topics among Bibliographic Databases," *JASIS* 41(4) :245-53.