과학기술 데이터베이스 품질 평가에 관한 연구

Evaluative Study on the Informaion Quality of Science & Technology Database

김선형*,유사라**서울여자대학교 문헌정보학과

Kim, Seon-Hyung, Yoo, Sarah. Seoul Women's Univ. Dept of Information & Library Science

본 연구는 국내에서 제작한 데이터베이스의 품질을 평가하고자 4개의 문헌 데이터베이스를 선정하여 각 데이터베이스를 정확성, 일관성, 완전성, 현행성 등 네가지 기준에 따라 분석하였다. 실험결과는 각 평가기준별로 정리되었으며 그에 대한 최우선 대책을 제안했다. 데이터베이스의 품질은 정보 검색과 서비스에서 매우 중요한 것이므로 시스템 설계자나 이용자 모두 데이터베이스의 품질에 대해 넓은 안목을 갖추고 체계적인 분석을 수행하여야 하며 이와 함께 제도적인 뒷받침도 이루어져야 할 것이다.

1. 연구목적

정보시스템이 효과적이기 위해서는 정보가 도출되는 데이터의 질이 우선 전제되어야 한다. 정보화 사회에서 이용자들은 정보시스템에 더 의존해감에도 불구하고 여전히 최신의 정확한 자료들을 제공받지 못하고 있으며 불량 데이터로 인한 피해와 분쟁은 더욱 증가하고 있는 실정이다.

본 연구는 데이터베이스 품질에 대한 체계적인 연구와 분석의 필요성과 의의를 제기하면서 국내에서 제작한 과학기술정보 데이터베이스를 대상으로 정확성, 일관성, 완전성, 현행성 등 4가지 평가기준에 의해 온라인 전자정보의 품질을 평가하려 한다.

2. 연구범위: 데이터베이스 평가기준

(1) 정확성(Accurarcy)

정확성이란 실세계에서 객체(objects)들이 가지고 있는 속성 값(value-1)과 데이터베이스내 데이터 값(value-2)이 서로 일치하는지를 의미한다. 즉 원정보의 형태에 충실하도록 지정되는 것이 데이터의 정확성이다. 반대로 부정확성은 잘못된 데이터나 틀린 정보를 데이터로 표현한 경우의 오류정도라고 할 수 있다.

(2)일관성(consistency)

일관성이란 데이터베이스내 둘 이상의 데이터가 서로 상충되지 않고 일관된 상태를 이루는 정도로 정의된다. 데이터베이스 생산자나 시스템 설계자들이 어휘 통제 용어들을 적용하는데 얼마나 일관성 있고 신뢰할만한 방식과 기준을 적용하고 있는가를 평가한다.

(3) 완전성(Completeness)

완전성은 데이터베이스가 의도하고 있는 분야를 얼마나 폭넓고 깊이있게 망라하고 있는지에 관한 것과 레코드 각각이 개별적으로 충분한 정보를 갖고 있는지에 관한 측정이다. 구체적으로 크게 범위(scope), 구조(structure), 접근성(accessibility) 등 3가지 요소로 평가될 수 있다.

---

* 서울여자대학교 대학원 문헌정보학과

**서울여자대학교 문헌정보학과 부교수

(4) 현행성(Currentness)

현행성이란 데이터베이스의 데이터가 얼마나 최신의 데이터로 갱신되고 있는지를 의미한다.

이것은 전체 데이터 중 얼마나 많은 데이터가 낙후되었는지를 통해 측정될 수 있고 이와는 달리 시간적으로 얼마나 최근에 갱신되었는지로 표현되기도 한다.

3. 자료수집과 분석방법

본 연구는 국내에서 제작한 과학기술정보 데이터베이스를 연구 대상으로 선정하였다. 그중 과학기술처 산하 한국과학기술원 부설 연구개발정보센터(KORDIC : Korea R&D Information Center)에서 제작한 데이터베이스 중에 문헌 데이터베이스인 UNION, SATURN, KRIST, TREND 등을 선정하였다.

(1) 정확성 측정

첫 번째 방법은 같은 뜻이지만 외래어를 다양하게 표기함으로써 검색결과가 달라지는 오류 정도가 어느 정도인가를 측정한다. 외래어로는 'DIGITAL'와 'NETWORK'를 선정하였다. 이 단어는 '디지탈' 또는 '디지털' '네트웍' 또는 '네트워크'로 국문 표기할 수 있다. 정확한 검색이라면 여러 가지 표기방법이 사용되고 있는 용어일지라도 이를 통제할 수 있는 것이 이상적인 데이터베이스라 할 수 있다. 데이터베이스의 부정확성의 양을 체계적으로 평가할 수 있는 방법은 색인어를 브라우징하여 검토해 보는 것이다. 색인어를 브라우징하면 오류로 기입된 데이터의 양을 측정할 수 있다. 마지막으로 한 레코드내에 잘못된 데이터만이 탐색의 유일한 접근점이 되어 검색에 심각한 오류를 갖는 경우를 검토하였다. 예를 들어, 탐색자가 저자명만을 알고 있기에 저자명 필드로 탐색을 시도했지만 기입된 저자명이 잘못된 데이터라면 이 탐색은 이루어지지 못한다.

(2) 일관성 측정

데이터의 비일관성을 나타내는 예는 약어나 국가명과 같은 고유명사의 색인에서 쉽게 찾아볼 수 있다. 일관성이 있는 데이터베이스이라면 고유명사에 대한 표기 방법이 같아야 할 것이며 띄어쓰기 인식을 하고 있는지를 살펴본다. 동일한 의미의 용어이지만 띄어쓰기 함으로써 각각 다른 용어로 인식하는가를 실험해 본다. 선정된 용어는 'database', 'data base'이다. 마지막으로 대소문자 구별을 하고 있는지를 검토하였다. 선정된 용어는 'union', 'UNION', 'cobol', 'COBOL' 이다.

(3) 완전성 측정

범위란 여러 측면을 포함할 수 있지만 데이터베이스의 주제 범위의 완전성에 대해 평가했다. 레코드의 주제 범위를 분명히 정의하고 있는가를 판단하고 불완전성에 대해 이용자 매뉴얼에 미리 제시하고 있는가를 살펴본다. 두 번째 측정은 데이터베이스 내에 기록되는 데이터가 원문(원시 데이터)에 담긴 정보를 완전하게 담고 있는지를 살펴보는 것이다. 그 방법은 레코드의 필드 값이 비어있는 경우를 검토해 보았다. 마지막으로 데이터베이스의 구조에 관한 것으로 필드 구성의 완전성을 검토하고 실제로 접근이 가능한 필드의 다양성과 필드 검색의 제한 여부를 측정하였다.

(4) 현행성 측정

이것은 전체 데이터 중 얼마나 많은 데이터가 시간적으로 낙후되었는지를 측정하는 것이다. 예컨대 출판년도 검색을 통해 최신 데이터로 갱신되었는지의 여부를 파악할 수 있다. 그러나 본 연구대상의 데이터베이스는 전체적으로 출판년도 검색이 가능하지 않으므로 이 측정방법 대신에 갱신주기를 살펴봄으로써 시간적으로 얼마나 최근에 갱신되었는지를 측정하였다.

4. 자료분석 결과

4.1. 정확성

정확성에 관한 평가는 외래어 표기의 정확성 검증, 색인어 브라우징의 여부, 오류 데이터의 여부로서 정확성 측정 결과를 도표화하면 <표 1>과 같다.

분석한 결과, 동일한 의미의 외래어를 다르게 처리하여 검색결과의 총건수가 다르게 나타났고, 한 레코드내에 두 용어가 함께 혼합 표기된 레코드는 없었다. 이것은 질의어가 통제되지 않아 검색 결과가 달라짐을 말해주고 있다. 색인어 브라우징의 경우를 보면 외래어 표기가 정확하지 않았다. 브라우징을 한 결과, 'government' 색인어가 'goverment', 'governement'로 잘못 표기된 레코드들이 발견되었다. 오류 데이터의 여부를 검증하기 위해

'Michael'이라는 성(姓)을 선정했을 때, 저자명 필드에서는 'Michael'이 바르게 표기되었지만 서명저자필드에서는 'Michel'으로 잘못 표기하였다. 이런 경우에는 탐색자가 검색을 저자명 필드에 제한시킨다면 별 문제가 없지만 서명저자필드에 제한시킬 경우에는 탐색에 실패하게 된다. 다른 경우에는 모든 필드에서 'Michel'으로 잘못 표기하여 이 레코드는 읽혀지지 않게 됨을 알 수 있었다.

<표1> 정확성 측정결과

측정 방법	검색 용어	검색 건수	AND 연산자 이용 건수	비 고
외래어 표기의 정확성 검증	디지탈 디지털	463 107 합: 570	0	두가지 용어를 혼합 표기한 레코드는 없음
	네트웍 네트워크	47 285 합: 332	0	〃
색인어 브라우징의 여부	government			색인어가 부정확하게 표기되었음
오류 데이터의 여부	Michael			필드에 따라 부정확하게 표기되었음

4.2. 일관성

일관성에 관한 평가는 고유명사 표기의 일관성 검증, 띄어쓰기 인식 여부, 대소문자 구별 여부를 통해 분석하였다. <표2>는 일관성 측정결과를 도표화한 것이다.

고유명사 표기의 일관성 검증을 위해 국가명 'The Netherlands'를 선정하여 검색한 결과, 'Tne Netherlands'가 'Netherlands'으로 혼합 표기되고 있었다. 다음은 동일한 의미의 단어이지만 띄어쓰기를 함으로써 검색 결과가 달라지는가를 보았는데 결과는 예상대로 전부 달랐으며 한 레코드내에 두 용어가 혼합 표기된 건수가 244건으로 나타났다. 대소문자 구별 실험에서는 동일한 검색 결과가 나왔으므로 이는 대소문자 구별을 하지 않는다고 판단된다. 결국 UNION에서는 띄어쓰기에 상관없이 동일한 의미의 용어를 검색하고자 할 때 이를 자동적으로 통제하지 못한다는 것과 대소문자를 구별하지 못하므로 이용자가 원하는 문헌은 이용자가 직접 하나씩 체크해서 선별해야만 한다는 것을 알 수 있다.

4.3. 완전성

완전성에 관한 실험은 레코드의 주제범위에 대한 사전언급의 여부, 누락된 필드, 접근 필드의 다양성 여부 등으로 평가하였다.

먼저 주제범위에 대한 정의는 초기화면에서 'UNION DB 소개'란을 통해 언급되고 있었다. 그러나 누락된 데이터가 있을 수 있다는 사전언급이나 그 해결안이 전혀 없었으며 데이터베이스상에 아예 기록되지 않아 데이터가 누람됨으로써 생기는 오류는 불가피한 결과였다. 다음은 레코드의 필드 값이 누락된 경우를 검토해 보았는데 '형태사항' 필드에 데이터 값이 누락된 사례가 발견되었다. 필드별로 접근할 수 있는 필드는 저자명, 서명, 단체명, 개인명, 회의명 등이다. 따라서 출판년도나 자료형태별 해당 레코드의 총수를 알고자 할 때는 검색의 제한을 받아야 했다. 전체 문헌에서 검색한 레코드의 총수와 필드별 해당 레코드의 총수를 비교할 수가 없었으며 모든 검색이 1회로 종결되었다. 완전성의 측면에서 볼 때, 데이터베이스는 모든 검색 단계에서 원하는 필드로 접근하여 해당 레코드의 결과를 얻을 수 있어야 하지만 UNION는 검색이 다양하지 못하고 제한점

이 컸다.

4.4. 현행성

현행성 측정은 수록정보의 갱신주기를 점검하여 평가하였다. 갱신주기를 점검하기 위해 한 색인어에 의해 검색된 레코드의 총수를 정기적으로 비교해 보았다. 결과는 평균 월 1회 갱신으로 나타났다.

<표2>일관성 측정결과

측정 방법	검색 용어	검색 건수	AND 연산자 이용 건수	비 고
고유명사 표기의 일관성 검증	The Netherlands			비일관적으로 표기하였음
띄어쓰기 인식 여부	databse data base	799 688	244	한 레코드내에 두 용어를 혼합 표기함
대소문자 구별 여부	UNION union	1370 1370	0	대소문자 구별을 하지않음
	COBOL cobol	346 346	0	〃

5. 결론

과학기술정보 데이터베이스를 평가기준에 따라 검색한 결과 각 평가기준별로 오류가 적지 않게 나타났다. 이것은 데이터베이스의 품질이 신뢰할 수 있는 수준을 갖추어 정확한 정보와 적합한 정보 검색이 이용자에게 제공된다고는 단언할 수 없는 결과라 할 수 있다.

본 연구자는 데이터베이스 생산자나 제작자들이 준거할 수 있는 품질 평가기준안이 공식적으로 마련되어야 하며 데이터베이스 품질을 심사하는 권위있는 전문기관이 반드시 있어야 한다고 여긴다. 두 번째로 일단 생산된 데이터베이스 품질 등급을 공식적으로 명시하는 제도를 제안한다. 품질이 낮은 데이터베이스에 대한 사용자의 실망과 피해는 어제 오늘의 일만은 아니다. 많은 데이터베이스가 급속히 만들어지고 있지만 데이터베이스의 현행화 등 유지 보수에 대한 대책은 전무한 실정이다. 이에 일정한 기준에 의해 평가된 데이터베이스를 품질 수준에 따라 차별화 함으로써 데이터베이스 품질에 대한 사용자 인식을 높이고 개발자 노력을 유도할 필요가 있다고 판단한다.

<참고문헌>

유사라.(1997). 하이퍼미디어 도서관 정보시스템. 서울 : 한국도서관협회.

유혜영.(1996). 국내제작 데이터베이스 평가에 관한 연구. 서울여대석사학위논문.

이응봉.(1996). 데이터베이스 품질에 관한 비평적 평가. 국회도서관보. 33(4).

한국데이터베이스진흥센터.(1996). 데이터베이스 표준화 연구보고서.

Basch,Reva.(ed).(1995). Electronic Information Delivery:Ensuring Qualityand value. Englnad. Gower.

Brodie, M.L.(1980). Data quality in information systems. Information and Management. v.3.

Date, C.J.(1986). An Introduction to Database Systems. v.1.

Dolan, D.R.(1992). Quality control at system level. Online. v.16. n.2. March. p.30-35

Fox, Christopher. et. al.(1994). The Notion of Data and its Quality Dimensions. Information Processing & Management v.30. n.1.

O'Neill, E.T. & Vizine-goetz, Diane.(1988). Quality Control in Onlin Database. ARIST v.23

Tenpoir, Carol.(1987). Online Databse:Quality Control. Library Journal. Feb.15;112(3).