데이터(셋) 검색 시스템의 구축 및 중요성에 대하여

정보검색론 박진호 교수님
문헌정보학과 2019311821 강성하

정보사회·인터넷의 발달로 인해 거의 모든 분야에서 데이터가 수집되고, 데이터 베이스 관리 시스템에 의해 데이터(정보)에 대한 조작과 관계 맺기가 가능해졌다. 인공 지능의 등장과 함께 범용적인 사용을 목적으로 한 지식베이스의 구축과 활용의 필요성이 대두되고 있다. 초기 정보 검색은 도서관에서 서지 정보를 검색하는 것이 주 업무였으며, 대학과 연구소에서 학문에 관한 연구 논문과 자료를 중심으로 하는 텍스트 검색으로 시작되었다. 구조적 자료 중심의 검색부터 문서 검색, 웹 검색, 다매체 검색 등의 비구조적 자료 중심의 검색으로 정보 검색 시스템은 점점 확장되고 있다. 최근에는 웹에 무분별하고 방대하게 퍼져있는 수많은 분야의 데이터들 중 원하는 정보를 효율적으로 찾고자, 데이터를 가이드라인에 따라 표준화된 형태로 분류하고 효율적으로 검색할 수 있도록 검색 시스템이 발전 중이며, 이러한 시스템이 바로 데이터(셋) 검색 시스템이라고 할 수 있다.[1] 데이터셋, 그리고 데이터(셋) 검색 시스템의 구축과 중요성에 대해 구체적으로 살펴보겠다.

우선 데이터셋이란 무엇인가. 자료의 모임, 다시 말해 주제별 데이터의 집합이자 관련된 데이터를 충분히 모아 즉각적 활용이 가능하도록 정리하여 공개하는 데이터가 바로 데이터셋이다.[2] 문서 검색과 데이터(셋) 검색은 어떻게 다른가. 문서 검색은 문서 집합에서 사용자가 원하는 용어가 포함된 문서들을 검색하는 것이다. 여기서 문서란 자연어 문장의 나열이며, 색인은 그 문서에서의 의미를 가지는 키워드 또는 키워드 무리로 이루어진다. 문헌에서 색인어·검색어·주제어 키워드로 사용되는 것은 주로 명사형인데 이는 의미 전달 능력을 위해서다. 데이터(셋) 검색과의 근본적 차이로는 문서 검색 시스템에서의 '문서'는 '사람'이 읽을 수 있는, '사람'이 이해하기 쉬운 자료이지만, 데이터는 '기계'가 이해할 수 있는 차원의 자료임을 들 수 있다. 이처럼 시스템에서 다루는 자료의 차이를 시작으로, 정보 검색이 이루어지는 환경(검색자의 질의, 요구, 데이터의 종류 및 양), 색인 구축 등 많은 차이가 있다.

데이터 활용의 핵심은 데이터 간 상호 연계와 공유를 통해서 새로운 가치를 창출하는 것이다. 고립된 형태로 존재하는 데이터는 자원으로서 활용이 어렵고 데이터 간의 시너지 효과도 미미하다. 따라서 데이터는 공동의 자원을 창조하기 위해 연계된 데이터를 오픈 방식으로 공유하는 방향으로 나아가야 한다.[3] 데이터 연계(data linkage) 또는 데이터 매칭(data matching)이란[4],

---

[1] "과학자를 위한 구글의 데이터 검색 엔진 '데이터셋 서치(Dataset Search)'", 테크플러스, 2018.09.13, https://blog.naver.com/tech-plus/221358295150(2020.06.19).
[2] "데이터셋이란?", Morphometry Open AI Innovation,http://www.wonmoai.org/learningdata/dataset(2020.06.19).
[3] 박다정, "보건의료 빅데이터의 연계 데이터셋 추출 기법", 충북대학교 대학원 박사 학위 논문, 2017, 1쪽.
[4] 오미애 외 4명, "보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안", 한국보건사회연구원, 2014, 23쪽.

서로 다른 복수의 데이터 파일을 결합하여 보다 풍부한 정보를 제공해 줄 수 있는 하나의 완전한 통합데이터를 만드는 방법으로 정의될 수 있다. 데이터셋 시스템 구축에서 중요한 것으로 메타데이터를 언급하고자 한다. 메타데이터는[5] 네트워크 상 데이터의 속성을 구조화하여 그 구조를 통해 내용, 형식, 관리 및 위치 식별 사항을 기술하는 데이터인데, 전통적 서지 데이터보다도 현대의 데이터셋에서의 메타데이터로 한정시키는 게 일반적이다. 다양한 정보 시스템 간의 정보 공유에 대한 요구를 위해서는 정보자원의 의미와 표현에 대한 공통의 약속이 필요하며, 검색과 관리의 용이성을 위해서는 필수적이다. 하나의 메타데이터로 통합하고, Linked Open Data(데이터 간의 링크를 통해 데이터를 의미적으로 연결, 웹에 개방해서 생성되는 데이터 웹이라는 하나의 거대한 데이터 베이스. 그리고 이를 자유롭게 이용할 수 있도록 오픈.) 등을 통해 메타데이터의 상호 운용성을 확보하여 데이터셋 검색 시스템의 효율성을 높이는 것이 현재 우리에게 주어진 과제(시스템 구축 관련)라고 할 수 있다.

데이터셋 검색 시스템이 왜 새로운 정보 자원으로 등장하고 있으며, 그 중요성이 증대하고 있을까? 정보 탐색 환경이 변화하고 있다. 사람들은 방대한 데이터베이스에서 이미 존재하는 정보를 단순하게 가져오는 정보검색보다도 '존재하지 않는 새로운 정보'를 얻고자 지식검색을 한다. 과거 Hayes-Roth, Waterman, Lenat(1984)은 정적인 데이터나 정보를 단순히 수집하는 것이 아니라 인공지능, 전문가 시스템의 일부로써 학습할 수 있는 동적인 정보 자원인 지식 베이스를 강조하였는데 이 때는 웹이 없던 시기로 데이터가 존재하지 않았다. 우리가 현재 창출하고 필요로 하는 '지식'은 데이터, 그리고 정보가 있어야 창출될 수 있다. 현대 사회의 우리는 이 요건을 충족하며 더불어 이 데이터의 양이 증가하고 있는 시대에 살고 있다. 전문적·특수 상황에 검색을 할 수 있던 시대가 아니며 일반인들도 쉽게 데이터와 정보에 접근할 수 있다. 이와 맞물려 데이터셋 검색 시스템은 그 중요성이 나날이 증대하고 있다. 앞서 이야기한 문서 검색과는 상이한 데이터(셋) 검색 시스템으로서의 특징들, 그리고 이를 구축하기 위한 필요 요건들을 잘 마련하여 변화하는 정보 환경 속에서 새로움을 창출해야 할 것이며, 나아가 계속해서 변화하는 환경에 적응해야 할 것이다.

[5] 데이터셋 구축에서 메타데이터의 중요성에 대한 내용은 '정보검색론' 수업과 고영만 교수님의 '메타데이터론' 수업 내용을 기반으로 기술하였다.