데이터의 홍수에서 가치있는 정보를 건져올리는 핵심단어 시각화

김원배 정보통신용어표준화위원회(WORDSTD) 위원, 전자신문 부장

현대는 '데이터 홍수 시대'다. IBM의 추정에 따르면 세계 곳곳에서 하루에 약 25억 기가바이트의 정보가 생성된다고 한다. 게다가 기하급수적으로 생성되는 정보의 증가속도가 빨라지고 있다. 영국에서는 앞으로 150~350년 사이에 지구에 존재하는 원자 수보다 디지털 비트의 수가 더 많아질 것이라는 계산까지 나왔다. '디지털 재앙'이라는 표현이 어색하지 않을 정도다.

정보가 많아질수록 정확한 정보를 찾기는 어렵다. 에드가 앨런 포의 대표작, '도둑맞은 편지'에는 중요한 편지를 흔해 빠진 편지꽂이에 평범한 편지처럼 보이게 해서 감춰두는 이야기가 나온다. 평범한 정보가 널려 있어서 역설적으로 정작 중요한 편지가 보이지 않았던 것이다. 정보가 지나치게 많아진 현대 사회에서 중요한 것은 '노하우(know-how)'가 아니라 '노웨어(know-where)'라는 이야기가 나오는 이유도 여기에 있다.

넘쳐나는 정보, 의미를 찾아내라!

그렇다면 넘쳐나는 정보 중 무엇이 중요한지는 어떻게 알 수 있을까? 가장 확실한 방법은 정보를 찾는 사람이 충분한 식견을 갖춰서 옥석을 가리는 것이겠지만, 데이터가 어디에서나 사용되고 사람이 아닌 기계끼리도 데이터를 주고받는 상황에서는 그리 좋은 방법이 아니다. 현실적으로 가장 좋은 방법은 데이터를 명확하고 효율적으로 알아볼 수 있게 시각화하는 것이다.

데이터 시각화(data visualization)는 데이터 분석 결과를 일목요연하게 보여줌으로써 모든 데이터를 일일이 보지 않고도 이해할 수 있게 한다. 예컨대 기업에서 수백 건의 기사, 수천 개의 해시태그, 수만 건의 고객 리뷰를 읽지 않고도 시장 반응이나 소비자의 보편적인 생각을 비교적 정확하게 파악할 수 있는 이유는 복잡한 데이터를 시각적으로 요약해 보여주기 때문이다.

데이터 시각화의 목적은 '도표'라는 수단을 통해 정보를 명확하고 효과적으로 전달하는 데 있다. 가장 흔히 사용되는 방법이 '인포그래픽'과 '핵심 단어 시각화'다. 인포그래픽은 수많은 데이터를 한 장의 그림으로 요약하는 것을 말한다. 주로 그래프로 표현되는 통계자료에서 쉽게 볼 수 있는 형태로, 그림으로 표현되기 때문에 한눈에 데이터의 의미를 파악하기 용이하다.

인포그래픽이 그림을 이용해 전체적인 인상을 직관적으로 전달한다면, 핵심 단어 시각화는 활자 정보를 활용해 구체적인 정보를 한 번에 이

해할 수 있도록 제시한다. 문서와 웹 등에 기재된 문구와 단어를 분석해서 중요도에 따라 달리 배치함으로써 핵심 아이디어가 무엇인지 단번에 파악할 수 있도록 시각화한다. 특정 문서에서 많이 언급된 단어를 크게 표현해 한눈에 문서의 내용을 알 수 있게 한 것이 대표적이다.

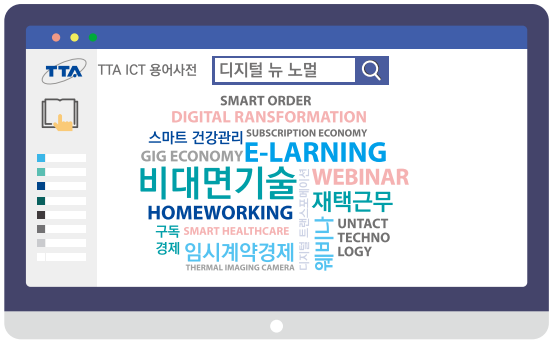
핵심 단어 시각화는 자료로부터 정보를 습득하는 시간을 줄여서 수많은 데이터가 만들어내는 트렌드를 단번에 알아볼 수 있게 한다. 자칫 혼란스러울 수 있는 데이터 덩어리에서 핵심 정보만 확인할 수 있으니 정보 확산을 촉진하기도 한다.

빅데이터의 지도, 핵심 단어 시각화

핵심 단어 시각화는 통상 방대한 분량의 데이터를 다루는 빅데이터를 분석할 때 데이터의 특징을 도출할 때 특히 유용하다. 일반적으로 문구와 단어의 중요도, 혹은 빈도에 따라 글자 색상이나 굵기 등을 변경해 특정 형태로 표시한다. 문서에서 사용 빈도가 높은 단어의 글씨 크기를 크게 표현함으로써 문서의 핵심 내용을 파악할 수 있다. 방대한 정보의 지도 역할을 하는 셈이다.

광범위하게 수집된 정보 중 주요 단어를 출현 빈도에 비례하는 크기로 시각화한 그래프를 '워드 클라우드(word cloud)'라고 하는데, 이는 핵심 단어 시각화와 동일한 개념이다. 웹에서는 메타 데이터 태그를 분석하기 때문에 '태그 클라우드(tag cloud)'로 불린다.

핵심 단어 시각화는 빅데이터의 트렌드를 한눈에 빠르게 파악할 수 있게 한다. 따라서 실시간으로 수많은 정보가 오가는 사이트에서 유용하게 사용된다. 수많은 사람들이 사진을 올리고 공유하는 '플리커(flicker)'에서 2004년 처음으로 핵심 단어 시각화 기능을 제공한 것도 자연스러운 일이다. 이후 테크노라티(technorati)



등 웹 사이트에서 사용되며 대중화됐다는 것이 정설이다. 현재는 빅데이터의 데이터 특징 분석, 보도자료의 키워드 분석, 소셜 네트워크 서비스(SNS, Social Networking Service) 인기 글 분석 등 다양한 분야에 널리 활용되고 있다.

물론 핵심 단어 시각화가 만능은 아니다. 문서의 주요 키워드 혹은 중요도를 한눈에 보고 이해할 수 있다는 것은 장점이지만, 단어 간 관계를 표현할 수 없고 빈도수 높은 정보가 특히 많이 노출되어 정보가 편향될 수 있다는 것은 단점이다. 제약 역시 적지 않다. 제대로 효과를 보려면 색상, 기호, 글씨 크기 등 시각적 요소를 보기 좋게 구성해야 하지만, 그렇다고 지나치게 기능적 측면을 강조하거나 아름답게 표현하는 데 집착하면 의미전달이 약해질 수 있다. 사용 가능한 문구도 한정적인 의미를 담은 명사와 형용사로 제한된다.

빅데이터 기술이 다양한 분야에 활용되면서 양질의 데이터를 수집하는 일이 중요해졌다. 그러나 데이터를 모아두기만 해서는 의미나 부가가치가 생기지 않는다. 데이터를 활용해서 의미를 찾아내야 가치 있는 정보를 발굴할 수 있다. 핵심 단어 시각화처럼 방대한 양의 데이터를 시각정보로 명료하게 제공하면 사용자가 빠르게 의미를 파악하고 사용처를 계획할 수 있다. 빅데이터에 날개를 달아주는 기술인 셈이다.