# 유비쿼터스 데이터베이스 시대를 준비하라!


국내 데이터베이스 시장은 정보 선진국인 미국이나 일본에 비해 전체 매출 규모나 데이터베이스 단위당 매출 규모가 매우 뒤떨어져 있다. 이는 이용 측면에서 국내 이용자들이 유료 정보에 대한 인식이나 비즈니스 활동에 있어서 정보의 중요성에 대한 인식이 아직 취약하고, 제공 측면에서는 고품질의 정보 콘텐츠가 적다는데 근본 원인이 있다. 기존에는 데이터베이스 사업자의 주요 수익 모델이 유료 회원 제도를 통한 회원 수입이나 정보 이용료를 통한 종량제 수입이 주류를 이뤘다. 하지만 최근에는 정보의 유통 매체가 웹으로 발전함에 따라 인터넷의 포털사이트를 통한 정보가 폭발적으로 증가하고 있다. 때문에 기존 정보제공 사업자들은 웹이라는 강력한 매체를 통해 자신의 정보에 새로운 비즈니스 모델을 부가해 수익을 창출하고, 콘텐츠간 융합 또는 기존의 오프라인 비즈니스 모델과 융합하는 등 다양한 형태의 서비스를 제공하고 있다.

글 / 이창한 한국데이터베이스진흥센터 기획조정실장

국내 데이터베이스 시장은 정보 선진국인 미국이나 일본에 비해 전체 매출 규모나 데이터베이스 단위당 매출 규모가 매우 뒤떨어져 있다. 이는 이용 측면에서 국내 이용자들이 유료 정보에 대한 인식이나 비즈니스 활동에 있어서 정보의 중요성에 대한 인식이 아직 취약하고, 제공 측면에서는 고품질의 정보 콘텐츠가 적다는데 근본 원인이 있다. 기존에는 데이터베이스 사업자의 주요 수익 모델이 유료 회원 제도를 통한 회원 수입이나 정보 이용료를 통한 종량제 수입이 주류를 이뤘다. 하지만 최근에는 정보의 유통 매체가 웹으로 발전함에 따라 인터넷의 포털사이트를 통한 정보가 폭발적으로 증가하고 있다. 때문에 기존 정보제공 사업자들은 웹이라는 강력한 매체를 통해 자신의 정보에 새로운 비즈니스 모델을 부가해 수익을 창출하고, 콘텐츠간 융합 또는 기존의 오프라인 비즈니스 모델과 융합하는 등 다양한 형태의 서비스를 제공하고 있다.

## 웹정보의 폭발적인 증가와 유비쿼터스 DB

미국 버클리대학의 2003년 정보량 계량 프로젝트 결과에 따르면 전세계적으로 종이, 필름, 광디스크, 자기 디스크 등의 매체를 통해 생성되는 정보의 양이 연간 5엑사바이트(Exabyte, 10의 16승)정도로 추정되며, 이중 0.34%가 웹을 통해 유통되고 있다.

하지만 이 가운데 정형화돼 이용자가 편리하게 이용할 수 있는 웹 정보의 비중은 더욱 낮다. 이러한 현상은 기본적으로 인터넷이 안고 있는 한계에 기인하는데, 즉 발생되는 정보를 웹으로 유통시키기 위해서는 인위적으로 체계화하는 과정이 필수적으로 따라야 하며, 이는 실세계 공간과 웹이라는 가상공간이 긴밀하게 결합

될 수 없기 때문이다. 기술의 발전에 따라 현재의 웹 환경은 무선 네트워크 환경과 결합하게 될 것이고, 궁극적으로 유비쿼터스 컴퓨팅 환경 아래 정보 콘텐츠가 생성되고 유통될 것이다.

즉 이용자에게는 보이지 않지만 무의식적으로 정보를 이용할 수 있게 하거나 혹은 데이터베이스가 스스로 이용자의 여건을 탐색해 여건에 맞는 정보를 제공하는 유비쿼터스 데이터베이스로 발전하게 될 것이다. 전문가들은 향후 다가올 유비쿼터스 컴퓨팅 환경에서 출현하게 될 유비쿼터스 데이터베이스는 제작자가 지금과 같이 인위적으로 정보를 정형화해 축적하지 않아도 유비쿼터스 컴퓨팅 요소들이 매년 수십 엑사바이트 급으로 발생하는 정보를 자동적으로 정형화해 축적하고, 이 축적된 정보가 이용자에게는 보이지 않지만 무의식적으로 정보를 이용할 수 있게 하거나 혹은 데이터베이스 스스로 이용자의 여건을 탐색해 여건에 맞는 정보를 제공하는 방향으로 발전할 것이라고 전망한다. 웹을 통해 정보를 제공하는 소위 웹 데이터베이스는 이러한 로드맵 상에서 유비쿼터스 데이터베이스로 지향하는 과정에서 중간 위치에 있다고 볼 수 있다.

## 유비쿼터스 DB로 가기 위한 해결 과제들

이러한 발전 과정상에서 해결돼야 할 과제로는 유용한 대량의 데이터 확보, 품질 향상, 데이터 포맷의 표준화, 검색 시스템의 개선 등을 들 수 있다. 이러한 과제는 개별적인 문제가 아니라 상호 연관돼 있다.

### 데이터 정형화를 위한 자동화 기술

현재 우리 주위에서 발생되는 대부분 정보는 비정형 형태로 생성된다. 실세계에서 말을 하거나 글을 쓰거나 어떤 행위를 할 때 발생되는 정보원의 형태는 비록 전자적 매체에 수록돼 유통된다고 할지라도 대부분이 비정형 형태에 의존한다.

미국 버클리대학의 2003년도 정보량 계량 프로젝트 결과에 따르면 2002년 전화, 라디오, TV 및 인터넷과 같은 전자 채널을 통한 정보 유통의 전체양은 18엑사바이트 정도로, 이중 98%가 전화선을 통해 P2P(Person To Person) 형태로 유통되고 있다고 한다. 하지만 이 통화 내용이 디지털 매체에 저장돼 있더라도 정형화된 데이터로 볼 수는 없다. 예를 들어 웹상에서 문서 정보를 검색할 경우 PDF 파일 형태를 자주 접하게 되는데 이는 디지털 형태의 정보이지만 정형화된 데이터는 아닌 것이다.

정보로서 그 가치를 극대화시키기 위해서는 문자 형태의 이미지 데이터를 문자 데이터로 변환시키는 기술이 있어야 하는데, 현재 대표적인 것이 OCR 기술이다. PDF 파일에 포함된 모든 문자 정보를 디지털 환경에서 자유자재로 활용 할 수 있는 정형 데이터로 전환시켜야 한다.

이러한 과제는 음성 정보나 영상 정보의 경우에도 동일하게 적용된다. 즉 음성 인식, 패턴 인식 등의 기초 인식 기술이 필요하게 된다. 대량의 정형 데이터 확보를 위해서는 비정형 정보를 정형화된 정보로 변환시키는 기술이 더욱 세련되고 고도화돼야 한다.

### 데이터 품질의 향상

발생 정보의 정형화 문제와 연장선상에서 고려해야 할 것이 데이터 품질 문제이다. 전세계적으로 가장 큰 규모의 웹 포털 사이트인 구글(Google)의 최대 과제는 웹상에서 대량으로 제공되는 정보를 수집해 고도의 품질을 유지하는 데이터베이스로 구축, 효율적인 검색을 가능하게 한다. 이를 위해 구글은 다양한 웹정보를 자동적으로 인식, 분류하고, 태깅 작업을 자동화하는 기술 개발에 많은 노력을 기울이고 있다.

구글의 사례에서 보듯이 웹상에서 유통되는 관련 정보를 데이터베이스로 구축하기 위해서는 특정 목적을 위한 범위 한정, 개별화, 정보 필터링 혹은 선택 등의 기능이 필요하다. 정보의 범위를 한정할 때 필요 정보를 누락시키지 않고 정보 범위를 설정하며, 정보 콘텐츠에 카테고리나 분류를 부여하는 것이 웹정보 콘텐츠의 품질 향상에 있어서는 매우 중요한 과제이다.

특히 정보량의 증가에 따라 범위의 한정, 선택, 분류 및 태깅 작업을 정확하게 수행할 수 있는 자동화 기술은 더욱 요구된다. 때문에 데이터 품질 향상을 위한 중요한 과제중의 하나가 데이터 포맷의 표준화이다. 정형화된 웹정보 콘텐츠가 효과적으로 공유 및 교환되기 위해서는 데이터 포맷이 표준화돼야 하는데, 이와 관련해 가장 관심 있게 보아야 할 것이 메타데이터 레지스트리(Metadata Registry)와 XML기반 웹 데이터베이스이다.

메타데이터 레지스트리는 메타데이터의 명세를 표준화하고, 중앙등록기관을 통해 등록, 인증하도록 해 데이터의 표준화를 포괄적으로 지원하는 시스템이다. 'ISO/IEC JTC1/WG2(데이터 관리 및 교환)'에서 'ISO/IEC 11179 - 메타데이터 레지스트리'가 국제표준으로 제정됐고, 미국 환경청의 EDR(Environmental Data Registry), 미국 교통부의 ITS(Intelligent Transportation System) 데이터 레지스트리 등 여러 공공 기관이 이 표준을 채택하고 있다. 메타데이터 레지스트리는 전자정부, 전자상거래, 사이버교육, 데이터웨어하우스, ERP 등 웹 기반에서 발생되는 데이터의 이질성 문제를 해결할 수 있는 중요한 개념으로 인정받고 있다.

한편, XML 기반 웹 데이터베이스는 차세대 웹 데이터베이스로서 현실적인 활용 방법론에 대한 연구가 진행중이다. 마이크로소프트사의 액세스(Access) 2002와 SQL 서버 2000, 오라클사의 8i와 9i 등 최근에 출시되고 있는 DBMS가 XML 도큐멘트를 지원하고 있다. XML은 웹상의 다양한 데이터의 교환 수단으로서 널리 활용될 것이며, 머지않아 대량으로 유통될 XML 기반의 도큐멘트를 효율적으로 관리할 수 있는 XML 기반의 웹 데이터베이스가 나타날 것이다. 현재는 XML 기반의 웹 데이터베이스와 관련한 기본적인 표준들이 제안되고 다양한 저장 모델, 검색 기법들과 함께 상업용 제품들이 제시돼 기술, 성능 및 실용 측면에서 평가를 받고 있다.

### 검색 시스템의 지능화

정보량의 증대와 정보 내용의 다양화에 따라 보다 효율적인 정보검색시스템도 요구된다. 종래의 정보검색 서비스와는 달리 웹정보 포털 검색 서비스인 구글 혹은 웹정보의 망라적(Exhaustive)인 보존을 위한 미국의 인터넷 아카이브(Internet Archive) 검색 시스템 웨이백 머신(Wayback Machine) 등에서는 대량의 퍼스널 컴퓨터를 이용해 대용량 정보처리를 하고 있는데, 이러한 신기술은 유비쿼터스 컴퓨팅 환경에 있어서 정보검색 시스템의 프로토타입을 시사하고 있다.

정보검색 시스템의 검색 효율을 향상시키기 위해서는 콘텐츠 작성시 더블린 코어와 같은 메타데이터의 채택, 콘텐츠 식별 번호의 부여에 의한 관련 정보와 링크 혹은 색인, 분류에 의한 콘텐츠의 체계화 등이 전제 조건이다. 이러한 전제하에 자연어 검색, 이미지 콘텐츠 검색, 검색 인터페이스 개선, 이종의 정보원 링크에 의한 조직화, 특정 목적을 위한 범위 한정 등의 검색 기능 개선, 검색 인터페이스의 커스터마이징, 콘텐츠의 자동 필터링 및 선택 기능 등 정보검색 시스템의 개량에 다양한 시도가 기대된다.

특히 최근 대학이나 연구소를 중심으로 차세대 웹으로서 검색 결과의 의미론적 해석을 컴퓨터가 자동적으로 수행할 수 있도록 하는 시멘틱 웹(Semantic Web)이 국내외에서 활발히 연구되고 있다. 향후 정보검색 시스템이 검색 결과의 의미론적 해석, 검색 질문식의 자동 전개, 인물 · 장소 · 사물 등의 특정화, 검색결과의 링크 기능 등을 어디까지 자동화할 수 있는가 등이 과제이다.

## 뉴 패러다임 위한 역량 집중해야

유비쿼터스 데이터베이스는 현재의 기술 발전 속도에 비추어 요원한 이야기가 아니다. 그러나 이러한 발전 과정에서 기술 중심의 과제뿐만 아니라 제도, 문화 등 다방면에 걸쳐 해결돼야 할 과제가 산재해 있다.

민간 정보 사업자들의 비즈니스 모델도 과거 C/S 통신 환경에서 웹 환경으로 변화했을 때 이상으로 다양하게 변화될 것이다. 유비쿼터스 컴퓨팅 환경에서 출현하게 될 네거티브한 요인들, 예를 들어 빅브라더(Big Brother)에 의한 정보 독점이나 프라이버시 침해, 지식 정보 재산권 관리 문제 등에 대한 대책이 강구돼야 할 것이다.

데이터베이스 산업은 자원의 효율적 활용, 기존 지식에 기초한 새로운 지식의 재창출 등에 필수적인 기반 산업으로서 물리적 자원이 부족한 우리나라의 여건에 비추어 국가 성장의 핵심 산업중의 하나로 성장시켜야 한다. 특히 기존의 패러다임을 변화시킬 유비쿼터스 데이터베이스에 대해 정부, 연구계, 산업계 등이 역량을 집중해야 할 시기이다.