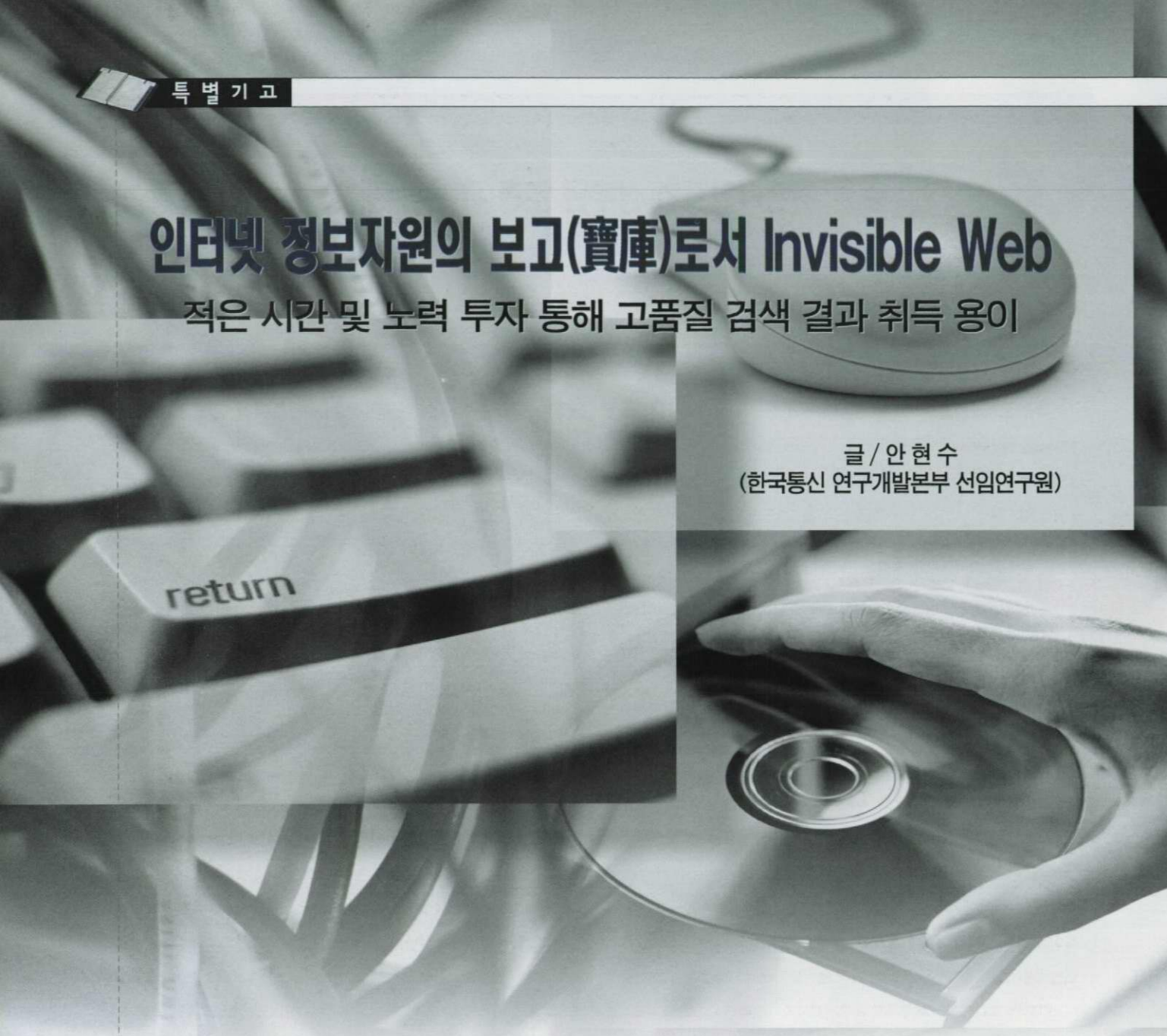
인터넷 정보자원의 보고(寶庫)로서 Invisible Web

적은 시간 및 노력 투자 통해 고품질 검색 결과 취득 용이

글 / 안 현 수
(한국통신 연구개발본부 선임연구원)



들어가는말

인터넷은 원래 미국 국방 프로젝트의 일환으로 1969년 ARPAnet으로 시작되었으나 1992년 NSFNet을 계기로 교육기관들이 연구 결과를 공유하는 목적으로 사용되었고 현재는 교육, 상용, 개인 용도 등 거의 전 분야에서 활용되고 있다. 이러한 인터넷은 1990년대 초반 Web의 등장과 상용화를 계기로 급속히 발전을 하고 있으며 인터넷상의 호스트 컴퓨터는 2002년까지 총 5억대까지 증가할 것으로 추산되고 있다.

또한 색인 가능한 웹 정보자원의 규모는 2001년말까지 40억 페이지가 될것이며 2003년에 가면 165억 페이지로 증가할 것으로 예상된다. Inktomi와 NEC의 연구 결과 2001년초에 이미 10억개의 웹 페이지가 인터넷상에 존재하는 것으로 평가되고 있다[1].

인터넷상의 정보자원이 이처럼 급격히 증가함에 따라 1990년대 중반까지 일반인들이 쉽게 이용할 수가 없었던 인터넷 검색엔진들이 비교적 짧은 기간 동안에 우리 일상 생활의 일부가 되었다. 즉, 이들 검색엔진이 없는 우리의 일상 생활은 상상하기가 힘들 정도가 되었다.

그러나, 현재 인터넷상에서 필요한 정보자원을 찾기 위해 이용되고 있는 대부분의 검색엔진과 관련하여 하나의 커다란 문제가 있는데 그것은 AltaVista, HotBot, Google 등과 같은 범용의 검색엔진으로는 검색이 불가능한 Invisible Web 웹 정보자원이 상당히 존재한다는 것이다. 그리고 더욱 심각한 것은 이

들 Invisible Web 정보자원이 인터넷상의 보통의 정보자원들(Visible Web) 보다 훨씬 빠른 속도로 증가한다는 점이다.

본 고에서는 Visible Web과 비교가 되는 Invisible Web의 정의와 특징에 대해 설명을 한다. 그리고 Invisible Web의 생성원리에 대해 언급을 하고 Invisible Web의 종류를 살펴본 후, 대표적인 Invisible Web 게이트웨이를 소개한다.

Invisible Web의 정의 및 특징

인터넷 검색과 관련하여 Visible Web과 Invisible Web의 용어들이 사용되고 있다. Visible Web이란 검색엔진들에 의해 선택이 되어 색인이된 웹사이트들로서 "공개적으로 색인이된 Web" 혹은

"Surface Web"이라고도 한다. 이에 대하여 Invisible Web이란 검색엔진들이 색인을 할 수 없거나 혹은 색인을 하지않는 정보자원을 말한다[2].

Invisible Web의 "invisible"한 부분은 해당 웹 사이트 자체가 아니라 이들 웹 사이트들이 연결되어 있는 데이터베이스의 컨텐트를 의미한다. 예를들어, 특정인의 전화번호를 찾기위해 Google 검색엔진을 검색하는 것과 특정인의 전화번호를 찾을 수 있는 원격의 데이터베이스를 검색할 수 있도록 해주는 특정 사이트를 찾기 위해 Google 검색엔진을 검색하는 것은 전혀 다르다. 즉, 특정인의 전화번호를 찾기 위해 Google을 검색하지는 않으며, 대신 Google 검색을 통해 한미르의 전화번호 검색 웹사이트(http://tel.hanmir.com/)를 찾게 된다.

전통적인 검색엔진으로는 찾을 수 없는 컨텐트를 의미하는 Invisible Web과 동일한 의미로 사용되는 용어로 Deep Web, Hidden Web, Shallow Web 등이 있으며 반대 개념으로는 전통적인 검색엔진으로 억세스가 가능한 컨텐트를 의미하는 Visible Web과 Surface Web 등이 있다.

이들 Invisible Web 정보자원은 다음과 같은 특징들을 가지고 있다[3]. 첫째, 저명하고 권위(authority)가 있는 정보원들에 의해 정보자원들이 유지 및 관리가 되기 때문에 일반적으로 Invisible Web 컨텐트의 품질은 우수하다. BrightPlanet의 연구결과에 따르면 Invisible Web의 품질이 우수한 컨텐트의 총량은 Visible Web의 그것보다 1,000에서 2,000배 더 규모가 크다.

둘째, Invisible Web은 포괄적(comprehensive)이기 때문에 특정 주제분야에서 내용의 깊이가 있는 컨텐트들을 갖게 된다.

셋째, 일반적으로 Invisible Web의 대상이 되는 정보자원의 범위가 한정되기 때문에 특정 분야에 집중된 데이터베이스들이 제공된다. 또한, 검색 대상 문헌들의 규모가 작기 때문에 검색시 높은 정확률과 재현률(precision/recall)을 구현할 수 있다.

넷째, 일반적으로 Invisible Web 정보자원이 독창적(unique)이기 때문에 웹상의 다른 곳에서는 구할 수 없는 컨텐트들이 Invisible Web에 존재한다.

다섯째, Invisible Web 정보자원이 최신성(currency)을 갖기 때문에 일반 검색엔진들에 비해 비교적 최신의 컨텐트를 이용할 수 있다. 그 이유는 검색하는 시점에 Invisible Web 정보자원의 상당 부분을 차지하는 동적 데이터베이스들이 데이터베이스내의 컨텐트를 대상으로 검색 결과를 즉시 생성하기 때문이다.

Invisible Web의 생성 원리 및 규모

인터넷상의 일부 웹 정보자원들이 검색엔진들의 색인에서 제외되는 이유는 무엇인가? 첫번째 이유는 데이터베이스내의 컨텐트를 직접 검색할 수 있도록 설계가 안되어있기 때문이다. 대신, 웹 개발자들은 검색과 동시에 만들어지는 맞춤 컨텐트를 제공하기 위하여 데이터베이스 기술들을 이용하고 있으며 My Excite(http://www.excite.com/)와 My Yahoo(http://my.yahoo.com/) 등이 이러한 유형의 서비스가 된다. 웹 사이트들이 더욱 복잡해지고 이용자들이 더욱 개인화된 요구를 해옴에 따라, 동적으로 생성된 컨텐트에 대한 이러한 경향은 더욱 가속화될 것이며 검색엔진들이 포괄적인 웹 색인을 생성하는 것이 더욱 어렵게 될것이다.

두번째로 특정 웹 페이지가 그래픽 파일, CGI 스크립트, MacroMedia Flash 파일, PDF 파일 등과 같이 검색엔진들이 색인할 수 없는 데이터 유형들로 구성된 경우에도 색인에서 제외된다. 비록 규모는 작지만 중요한 내용을 담고있는 컨텐트들이 웹상에서 Microsoft Word, Excel, PowerPoint 포맷 뿐만 아니라 RTF(Rich Text Format), PostScript나 PDF 포맷 등으로 존재한다. 그러나 주요 검색 엔진들 중, 오직 구글(Google)만이 이런 유형의 파일들을 색인 처리할 수 있며, 대부분의 다른 검색 엔진에서는 이들 파일들에 대한 검색이 불가능하다. 예를들어 현재 구글을 통해 2200만개 이상의 PDF 파일을 검색할

수 있다[4].

Invisible Web과 비교하여 Visible Web의 크기는 어떠한가? 가장 규모가 큰 웹 색인 데이터를 가지고 있는 구글 검색엔진을 이용하여 검색을 할 경우, 약 14억개의 웹 페이지를 검색할 수 있다. 그러나 구글은 실질적으로 모든 페이지들을 색인하지는 않는다.

주요 검색 엔진들로부터 나온 추정치에 의하면 전체 Visible Web의 크기는 약 25억에서 40억 페이지에 이르며 매일 약 700만 페이지씩 증가를 한다. 이에비해 BrightPlanet사가 수행한 연구결과 Invisible Web은 Visible Web 보다 400에서 550배 더 규모가 크며 약 5,430억개의 문헌으로 구성되어 있다[5].

Invisible Web의 종류

Invisible Web은 스팸(spam)과 포르노 등과 같은 가치가 없는 정보가 결코 아니며 반면에 대부분의 Invisible Web은 권위가 있는 정보원들에 의해 유지 및 관리되는 높은 품질의 정보자원들로 구성되어 있다.

일반 검색엔진에 의해서는 색인이 되지않는 웹 정보자원을 의미하는 Invisible Web을 크게 구분하면 다음과 같다. 첫째, 로그인을 통해서만 접근이 가능한 데이터베이스들. 상당 수의 웹 사이트가 이용자들로 하여금 해당 웹 사이트에 등록을 하도록 함으로써 비밀번호에 의한 접근 제한을 하고 있다. 이렇게 비밀번호로 접근이 제한되는 웹 사이트의 경우에는 검색엔진이 해당 사이트에 있는 컨텐트를 대상으로 색인을 할 수가 없게 된다. 둘째, 동적으로 생성되는 데이터. 검색엔진과 스파이더를 통해서는 동적인 데이터베이스에 들어갈 수 없으며 해당 정보에 억세스할 수 있는 유일한 방법은 데이터베이스 자체를 검색하는 것이다. 이를 위해 CGI 스크립트, 자바 스크립트, ASP 등이 있으며 이들은 URL내의 "?" 기호로 식별되는데, 검색엔진들이 이를 절단 표시로 취급하기 때문에 색인을 할 수 없게 된다. 셋째, 실시간 데이터(real-time data). 수명이 짧고 대규모 컴퓨터 저장 공간을 필요로하는 특성 때문에 주식 시세, 날씨 정보, 항공 여행 정보(Flight Tracker: http://www.trip.com/ft/home/0,2096,1-1,00.shtml) 등과 같은 실시간 데이터가 Invisible Web을 이룬다. 네째, PDF, MacroMedia Flash 파일, Office 파일, 스트리밍 미디어 파일 등. Google을 제외한 대부분의 검색엔진에서 PDF 문헌들을 색인에서 제외하고 있다. 또한 MacroMedia Flash 파일, Office 파일, 스트리밍 미디어 파일 등도 전통적인 검색엔진의 색인 대상에서 제외되고 있다. 다섯째, 목록들. 단행본, 음악, DVD, 비디오, 소프트웨어, 게임 등에 대한 검색이 가능한 데이터베이스를 가지고 있는 아마존(http://www.amazon.com)과 각 도서관의 온라인 열람목록(OPAC: Online Public Access Catalog) 등이 대표적인 예가 된다.

Invisible Web 게이트웨이

인터넷 정보자원에 대한 검색을 할 때, 어떤 경우에는 정보가 존재하지만 검색엔진을 통해 접근이 불가능한 경우도 있다. 웹 사이트 소유권자가 자신들의 정보가 공개되는 것을 원치 않을 수도 있다. 또는 웹 기술을 이용하여 검색엔진들이 특정 정보에 접근할 수 없도록 할 수도 있다.

즉, Invisible Web 정보자원이 존재하는 이유는 인터넷 검색엔진이나 웹 디렉토리의 기능이 부족해서가 아니라 전통적인 검색엔진을 통해서는 Invisible Web 정보자원을 아예 인식을 못하도록 되어 있기 때문이다. 따라서 이들 Invisible Web 정보자원에 효율적으로 접근하기 위해서는 적절한 도구를 갖춰야 하는데 바로 이들이 Invisible Web 정보자원에 대한 게이트웨이이다.

다음은 현재 인터넷상에서 활용이 가능한 대표적인 Invisible Web 게이트웨이이다.

■ Direct search

http://gwu.edu/~gprice/direct.htm

AltaVista, Google, HotBot 등과 같은 범용의 검색 도구를 이용하여 검색하거나 접근할 수 없는 데이터를 포함하고 있는 정보자원들의 검색 인터페이스에 대한 링크들의 모음. Direct search의 목표는 검색 창에 접근하기 위해 1페이지나 혹은 2페이지에 걸쳐 클릭을 하는 것 보다 웹 정보자원에 의해 제공되는 검색 창에 가능한 가까이 접근하는 것이다.

■ InvisibleWeb.Com

http://www.invisibleweb.com

InvisibleWeb 목록에는 전통적인 검색에서는 종종 간과되는 10,000개 이상의 데이터베이스와 검색 가능한 정보자원들이 포

함되어 있다. 각 정보자원은 전문 편집자들에 의해 분석되고 기술됨으로써 InvisibleWeb 목록의 모든 이용자들이 800개 이상의 주제에 관한 신뢰할만한 정보를 찾을 수 있도록 한다.

단순/고급 검색 기능 혹은 브라우즈 가능한 색인을 이용할 경우 이런 모든 정보자원에 쉽게 억세스할 수 있다. 다른 검색 엔진들과는 달리, 특정 웹 사이트내의 검색가능한 정보자원에 직접 억세스할 수 있으며 심지어 질의를 입력할 수 있는 검색 창을 제공하기도 한다.

■ Librarians' index to the Internet

http://www.lii.org

Librarians' index to the Internet은 주석(annotations)이 제공되는 검색 가능한 주제별 디렉토리로서 공공도서관 이용자들에 대한 유용성을 기준으로 전문 사서들이 선택하고 평가한 약 7,000개 이상의 인터넷 정보자원들로 구성되어 있다.

여기에서는 가장 우수한 컨텐트에 대한 링크만을 포함한다. 비록 순수한 Invisible Web 검색 안내 사이트는 아니지만 본 사이트에서는 각 정보자원을 가장 우수한 디렉토리, 데이터베이스, 특정 정보자원 등으로 분류한다.

물론, 데이터베이스 정보자원은 Invisible Web에 해당되며 고급 검색 기능을 이용하여 데이터베이스만을 대상으로 검색을 할 수 있다. 또한, 고급 검색 기능을 이용할 경우, 검색결과를 디렉토리의 특정 필드(저자명, 기술, 서명, URL 등)로 제한할 수 있다. Librarians' index to the Internet은 Invisible Web 데이터베이스를 찾기 위한 강력한 검색 도구다.

■ ProFusion

http://www.profusion.com

ProFusion은 InvisibleWeb.com을 운영하는 Intelliseek사가 제공하는 메타 검색엔진이다. 또한 ProFusion은 주요 검색엔진들에 대한 정력한 동시 검색 기능을 제공할 뿐만 아니라, TerraServer, Adobe PDF Search, Britannica.com, NY Times, 미국 특허 데이터베이스 등을 포함하여 1,000개 이상의 Invisible Web 정보자원들에 대한 검색 기능을 제공한다.

■ Alpha Search

http://www.profusion.com

Alpha Search의 가장 중요한 목표는 가장 우수한 인터넷 "게이트웨이" 사이트에 억세스할 수 있도록 하는 것이다. 이들 "게이트웨이" 사이트에서는 특정 분야, 주제, 혹은 아이디어와 관련된 모든 관련 사이트들을 한군데로 집중한다. 따라서 이를 이용할 경우, 하나의 게이트웨이 사이트에 접속만 하면 수백개의 웹 사이트에 접속을 할 수 있게 된다.

맺는말

Invisiblw Web 정보자원은 범용의 검색엔진들이 그들의 웹 페이지 컬렉션(색인)에 포함할 수 없는 혹은 포함할 의사가 전혀 없는 정보자원들을 의미한다. 그러나 실질적으로 Invisible Web은 우수한 품질의 공개된 정보자원들로 구성되어 있다. 또한 BrightPlanet 연구결과에 의하면 Invisible Web의 95%가 공개 정보자원이며 따라서 이들을 이용하기 위해 별도로 구독을 하거나 비용을 지불할 필요도 없다.

웹 검색자의 첫번째 도전은 Invisible Web이 존재한다는 사실을 이해하는 것이다. 만약 Yahoo와 Google 등과 같은 범용의 검색엔진만을 이용한 검색 경험을 가지고 있을 경우, 억세스가 가능한 웹 정보자원의 극히 일부에만 억세스를 해왔다는 것을 깨닫게 될 것이다.

따라서 인터넷 검색 시, Invisible Web 정보자원의 존재를 이해하고 이들 정보자원에 접근하는 방법을 알 경우에는 시간과 노력을 절약할 수 있을 뿐만 아니라 다른 방법으로는 구할 수 없는 고품질의 검색 결과까지도 얻을 수 있게된다.

참고문헌

[1] Shekhar, Shashank. The Deep Web: Surfacing Hidden Value. 2001. (http://www-courses.cs.uiuc.edu/~cs497kcc/Lectures/shekhar-0904.ppt)

[2] Pedley, Paul. The Invisible Web: Searching the hidden parts of the Internet. London: Aslib-IMI, 2001.

[3] Price, Gary and Sherman, Chris. The Invisible Web. 2001. (http://www.infotoday.com/ ili2001/presentations/sherrman-price.ppt)

[4] Sherman, Chris. Google Unveils More of the Invisible Web. SearchDay, No 128, October 31, 2001(http://www.searchenginewatch.com/searchday/01/sd1031-google-files.html)

[5] BrightPlanet. The Deep Web: Surfacing hidden value(White Paper). July 2000. (http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/deep webwhite paper.pdf)