Invisible Web 탐색도구의 성능 비교 및 분석

The Effectiveness of the Invisible Web Search Tools

노 정 순(Jung-Soon Ro)*

초 록

본 연구는 표준 웹탐색엔진에 색인되지 않는 Invisible Web에 대한 특성과 Invisible Web 탐색도구들을 파악하고, 이들 도구에서 Invisible Web 탐색의 성능을 비교 평가하기 위해 수행되었다. 표준 웹 탐색엔진인 Google과 Invisible Web 탐색엔진인 IncyWincy, Invisible Web 메타탐색엔진인 ProFusion과 Search.com에서 11개의 탐색질문이 탐색되었다. ProFusion과 Search.com, IncyWincy에서의 Invisible Web(메타)탐색 기능은 이 세 엔진에서 제공하는 웹 메타탐색기능과도 비교되었다. 탐색결과 Google이 Invisible Web 탐색에서 Invisible Web 탐색엔진보다 .15 -.35 높은 적합성순위정확률을 보였지만 통계적으로 유의한 차이는 아니었다(α=.055). Invisible Web 탐색엔진에서 웹 메타탐색은 Invisible Web(메타)탐색보다 통계적으로 유의한 수준에서 더 우수한 것으로 나타났다. 성능평가에 사용된 적합성순위정확률은 검색된 문헌의 질(적합성)과 적합문헌의 순위를 반영하는 정확률 척도로 사용될 수 있음을 보여주었다.

ABSTRACT

This study is to investigate the characteristics of the Invisible Web and many search services designed to serve as gateways to the Invisible Web and to evaluate searching the Invisible Web in the Services. The four services for searching the Invisible Web were selected to search the Invisible Web with 11 queries, that are Google as portals, ProFusion and Search.com as Invisible Web meta search engines, and IncyWincy as Invisible Web search engines. It was found that the effectiveness of Google's Invisible Web searching was better compared with the three Invisible Web search tools but the difference between the four systems was not significant(α=.055). The Invisible Web meta searching was better than the Web meta searching in the three search tools at the statistically significant level. The effectiveness measurement based on the ranks and relevance degree(quality) of relevant documents retrieved seemed appropriate to the ranked search results.

키워드: 탐색성능, Invisible Web, Deep Web, Google, Search.com, ProFusion, IncyWincy, searching evaluation, meta engine, effictiveness, first-n precision.

---

* 한남대학교 문헌정보학과 교수(jsr@mail.hannam.ac.kr)

■ 논문접수일자 : 2004년 8월 16일
■ 게재확정일자 : 2004년 9월 18일

1. 연구의 목적

인터넷의 발달로 인터넷을 통해 얻을 수 있는 자원 특히 웹(Web)자원의 폭발적인 증가는 학술정보 이용자를 원하는 정보가 있느냐 없느냐 보다는 어디에 있느냐 어떻게 찾느냐 하는 문제에 당면하게 하였다. 이를 위하여 웹자원을 수집·색인하여 탐색기능을 제공하는 다양한 탐색도구가 출현하였으나[1], 심사를 거치지 않는 본인출판(Self-publishing)과 언제 인터넷상에서 사라질지 모르는 불안정성(Unstability)[2] 때문에 웹 탐색도구들의 양질의 정보를 제공하려는 노력에는 한계가 있다.

특히 HTTP라는 프로토콜을 사용한 웹에서 탐색도구들이 웹문서를 색인하기 위해 사용하는 스파이더는 html화일을 읽고 그 안에 링크된 문헌들을 찾아가 색인하도록 프로그램된 프로그램이다. 이 때문에 비html화일이나, 링크된 사이트에서 패스워드를 요구하거나 탐색식을 통해서만 접근이 가능한 데이터베이스(DB) 시스템에 들어있는 정보, 인터넷에 머무르는 수명이 매우 짧은 주식, 날씨, 뉴스 등과 같은 실시간 정보 등은 스파이더에게 보이지 않아(invisible) 색인되지 않는다. 웹에 존재하나 스파이더에게 보이지 않아 색인되지 못하는 정보는 색인된 정보보다 훨씬 많고, 보다 양질의 정보가, 보다 빠르게 증가하며, 그 중 95%는 무료로 사용할 수 있다고 보고되어(Bergman 2001) 이용자에게는 중요한 정보원이 되고 있다.

이용자가 Invisible Web정보를 이용하기 위해서는 Invisible Web정보를 담고 있는 DB사이트를 찾아 직접 탐색을 수행하여야하나, 이용자는 적당한 주제별 DB사이트를 찾고, 각기 다른 탐색시스템을 사용하는 DB에서 탐색을 수행하는데 어려움을 겪고 있다. 이 때문에 주제별 DB엔진사이트를 안내하고 이용자를 대신하여 여러 DB엔진으로 탐색식을 보내어 탐색을 대행하는 Invisible Web 게이트웨이가 출현하였으며, 최근에는 인터넷 기술 향상으로 기존의 웹 탐색도구가 이러한 Invisible Web 정보까지 색인하여 탐색을 제공하기도 한다.

다양한 Invisible Web 탐색보조도구가 출현하고 소개되고 있지만 IW 탐색도구들의 탐색성능을 비교 분석한 연구는 거의 찾아볼 수 없다[3]. 본 연구는 Invisible Web에 대한 성과

---

1) Google directory의 Computer/Internet/Searching 아래에서 Directories로는 823건이, Search Engines으로는 326건이, Metasearch로는 51건이 색인되었고, Computer/Internet/On the Web/Web Portals에는 82건이 색인되었다. Yahoo directory에서는 Internet/WWW/Searching the Web/Search ing Engines & Directories아래에 453건이 색인되어있다(2004. 8. 3 현재).

2) 1998년부터 2002년까지 5년 동안 전년도에 존재했던 사이트가 다음 해에도 존재한 것은 51% -56%에 지나지 않았다. 2001년에 존재한 웹사이트의 49%가 2002년에는 더 이상 존재하지 않았다(OCLC, Web Characterization Project). <http://wcp.oclo.org/stats/misc.html>.

3) Google에서 "invisible web"으로 192,000건이 검색되고(2004. 8. 3), Dialog의 Library Literature & Information Science(file 438)와 Information Science & Technology Abstracts(file 202), ERIC(file 1)에서 70건이 검색되었지만(2003. 10. 17), Invisible Web에 대한 학술적 연구는 극소수였고, 특히 Invisible Web 탐색도구의 탐색성능을 평가한 연구는 찾지 못하였다. 검색된 문헌은 Invisible Web에 대한 특성과 Invisible Web자원을 소개하는 것이 대부분이었다. 국내학술잡지기사는 1건도 검색되지 않았다. 선행연구문헌에서 조사된 Invisible Web에 대한 특성은 2장에 요약되었다.

Invisible Web 정보의 탐색을 돕는 탐색게이트웨이를 파악하고, 포탈서비스를 제공하는 표준탐색엔진과 Invisible Web 탐색 게이트웨이에서 Invisible Web 탐색의 효과를 비교 평가하기 위한 것이다.

연구과제 :

1) 웹 탐색엔진과 Invisible Web (메타)탐색엔진에서 Invisible Web 탐색성능에는 차이가 있는가?
2) Invisible Web (메타)탐색엔진과 웹 메타탐색엔진에서 Invisible Web 탐색성능에는 차이가 있는가?
3) 적합성순위정확률은 웹페이지의 질과 출력순위를 반영하는 성능척도인가? 순위정확률과는 차이가 있는가?

본 연구에서 사용된 용어의 정의는 다음과 같다.

IW 메타탐색엔진 : 탐색식을 전문DB엔진으로 보내 검색결과를 통합하여 제공하는 시스템(예, ProFusion, Search.com).

IW 탐색엔진 : Invisible Web 정보까지 직접 색인하여 자체 DB에서 탐색을 제공하는 시스템(예, IncyWincy).

IW (메타)탐색엔진 : IW 메타탐색엔진과 IW 탐색엔진을 통칭.

웹 메타탐색엔진 : 탐색식을 표준 웹탐색엔진으로 보내 검색결과를 통합하여 제공하는 시스템(예, Metacrawler, ProFusion, Search.com, IncyWincy).

2. Invisible Web의 정의 및 특성

전통적으로 웹 탐색엔진은 "Submit your site"을 통한 웹페이지 작성자의 색인신청을 받아, 혹은 스파이더를 통해 html화일에 링크된 문서를 수집하여 색인한다. 이러한 표준탐색엔진이 일반적으로 색인하지 않는 정보의 바다 심연에 있는 웹정보는 Invisible Web으로 지칭되고 있다. 'Invisible Web'이란 용어는 1994년 Jill Ellsworth가 "전통적 탐색엔진에게는 보이지 않는(Invisible) 정보 컨텐츠"라는 말을 사용하면서 처음 사용된 것으로 보고되었다(Bergman 2001). 1998년 Intelliseek사의 'Invisibleweb.com'이라는 첫 게이트웨이가 출현하면서 Invisible Web이란 용어는 자리를 잡았다. 그러나 2000년 탐색엔진 CompletePlanet은 다른 탐색엔진에게는 보이지 않지만 그들에게는 보인다면서, 문제는 보이고 보이지 않고가 아니라 spidering기술이라며, Invisible Web 대신 "Deep Web"이라는 용어를 제안하였다. 표준 웹엔진의 크롤러가 정보의 바다를 항해하면서 잡아 올리는 바다 수면(surface)에 있는 정보에 비하여 바다 심연(deep)에 있는 거대한 정보를 언급한 것으로, 표준 크롤러가 끌어올리지 못하는 Deep Web 정보는 특수 전문DB로 데이터베이스화되어 DB탐색시스템(엔진)으로 검색되어 이용될 수 있으므로, CompletePlanet은 좀더 협의로 "DB에 들어있는 정보"로 정의하고 있다[4].

표준탐색엔진이 색인하는 Visible Web은 비교적 자유로운 구조의 html형식으로 작성되어

4) 한 때 표준탐색엔진에게 보이지 않았던 정보가 지금은 정보기술의 발달로 보이느냐 보이지 않느냐보다는 엔진의 색인정책상의 결정으로 색인여부가 결정될 수 있다는 점에서 Invisible이란 용어보다는 deep이란 용어가 더 타당

(unstructured) 링크를 통해 탐색되나(browsable), Invisible Web은 주로 관계형 DB의 구조로 저장되어(structured) 직접적인 탐색 질의를 수행하여 검색된다(searchable). 또한 Visible Web은 웹 서버에 저장된 정적인(statatic) 페이지임에 반하여, Invisible Web은 요구가 있을 때 동적으로 생성되어 제공되는 페이지이다(dynamic). Sharman과 Price(2004)는 Invisible Web이란 "기술적 제한이나 색인정책으로 인해 표준 탐색엔진이 색인하지 않는, 웹에 존재하는 텍스트 페이지, 화일 혹은 고품질 정보"라고 정의하고, Invisible Web를 4종류로[5] 분류하였다.

전통적인 표준탐색엔진에서 스파이더가 Invisible Web을 색인하지 못하는 이유는 4가지로 요약된다. 첫째 표준 웹엔진에서 스파이더는 html화일을 링크하여 문서를 읽어오도록 프로그램되었기 때문에 <html>태그로 시작되지 않는 pdf, ps, word, excel 등의 비html 텍스트화일이나, 사운드, 이미지, Flash 등의 화일을 읽어오지 못하였다.

둘째, 서버에 저장된 정적 페이지가 아니라 요구에 따라 동적으로 생성되는 페이지는 표준 탐색엔진이 색인하지 않는 Invisible Web의 일종이다. 동적 페이지란 URL에 "?"이 나타나는 페이지로, 스크립트 명령어가 사용됐음을 의미한다. URL에 들어있는 스크립트 언어는 탐색이나 출력 명령을 DB에서 실행시킨 후 그 결과를 웹 페이지로 생성해 낸다. 그러나 비윤리적인 웹마스터들이 같은 페이지를 반복해서 작성하도록 스크립트을 작성하여 스파이더를 끝이 없는 Loop(Spider traps)에 빠지도록 하기 때문에 탐색엔진은 ?가 있는 URL은 색인하지 않는다.

셋째 인터넷상에 존재하는 시간이 너무 짧아 스파이더가 미쳐 건져 올리기 전에 사라져 버리는 주식시세, 날씨, 뉴스, 항공기예약 등의 실시간 정보는 수명이 짧을 뿐만 아니라 대규모의 저장 공간을 필요로 하기 때문에 표준 웹 탐색엔진으로는 색인할 수 없었다.

넷째, 링크된 웹사이트가 패스워드나 로그인 혹은 탐색식을 요구할 때 스파이더는 스파이더가 모르는 뭔가를 입력하라는 시스템의 요구를 이해하지 못하기 때문에 더 이상 접근하지 못하고 탐색을 중단한다. 그러므로 로그인이나 탐색식을 입력해야 하는 DB 기반의 정보들은 색인이 불가능하였다.

그러나 탐색기술의 발달로 표준탐색엔진은 이론적으로 기술적으로 Invisible Web을 색인할 수 있고, 몇몇 표준탐색엔진은 실제로 색

한 것으로 보인다. 그러나 Google 탐색(2004. 8. 3일 현재)에서 "invisible web"으로는 0.39초 동안 192,000건이 검색되나, "deep web"으로는 39,800건이 검색되는 것으로 보아, Invisible Web이 더 통용되는 것으로 볼 수 있다. 따라서 본 연구에서도 Deep Web보다는 Invisible Web을 사용하였다.

5) 1. Opaque web: 색인될 수 있지만 색인하지 않은 화일로 구성된 웹 자원. 기술적으로 문제는 없지만 비용 측면에서 좋은 페이지가 아니라서 선별되지 못한 페이지, 아직 스파이더가 찾아오지 못한 새(new) 페이지, 다른 페이지에 링크되지 못하여 스파이더가 찾아오지 못한 Disconnected URL 등이다.

2. Private Web: Username이나 Password를 요구하여 스파이더가 들어오는 것을 막거나, "noindex" 라는 메타태그나 "robots.txt" 화일를 사용하여 탐색엔진이 색인하는 것을 막는 웹 자원.

3. Proprietary web: 이용자 등록을 요구하는 웹.

4. Truly Invisible Web: 기술적으로 스파이더가 색인하지 못하도록 프로그램된 웹. 메타데이타나 텍스트 데이터가 없는 페이지, 동적으로 생성된 페이지, 관계형 DB에 들어 있는 정보 등이다.

인하고 있다. AltaVista와 Google은 화일명이나 html 이미지태그에서 페이지작성자가 사용하는 alt 텍스트를 사용하여 이미지나 오디오, 비디오 화일과 같은 비텍스트화일을 색인한다. Google은 그래픽 이미지에서 OCR(광학문자인지장치)을 사용하여 색인어를 추출하며, Singingfish는 메타데이타를 사용하여 오디오 정보를 텍스트용어로 색인한다. AlltheWeb은 Flash화일에 있는 텍스트 부분을 색인하고, Google은 flash화일에 들어있는 링크(link)를 찾아 flash화일을 색인한다.

탐색엔진은 Invisible Web을 색인할 수 있음에도 비용측면에서 비실용적이거나 날씨, 주식시세와 같이 정보의 수명이 짧고 색인할 가치가 없다고 판단되는 정보는 색인하지 않는다. 또한 수백 페이지로 이루어진 pdf나 post-script(ps) 화일의 전문은 색인화일이 너무 커지기 때문에 Google은 전문 대신 텍스트를 120kb까지만, AlltheWeb은 110kb까지만 색인하고 중단한다.

주요 탐색엔진은 Spider traps을 유도하지 않는 믿을만한 동적 컨텐츠 생성 URL의 리스트를 명시하고 있는 "Paid Inclusion" 프로그램을 이용하여 동적으로 생성된 컨텐츠를 색인하는데 장애물을 줄이고 있다(Cartwright).

그러나 기술적으로 표준탐색엔진이 직면하고 있는 가장 큰 장애는 데이터 구조나 탐색기능이 각기 다른 수십만 개의 DB에 저장된 고품질 정보의 엑세스이다. OPAC DB에 들어있는 Invisible Web 정보를 모두 MARC에서 XML과 같은 형식으로 변환하여 Visible 정보로 제공하기도 하지만, 대부분의 표준 탐색엔진은 아직 ID나 패스워드, 옵션선택 등 이용자의 입력을 요구하는 양식을 포함하는 html 페이지에 대응하지 못하고 있다.

Invisible Web의 규모와 실체는 Bergman의 Deep Web 백서(2001)에서 처음으로 보고되었다. 2000년 당시 Invisible Web의 규모는 문서 수 550billion으로 Surface Web보다 400~550배 많고, Deep Web의 사이트 수는 200,000이상이며, 상위 60개 사이트의 크기는 총 748.5 terabytes로 전체 Surface Web의 40배이고, 양질의 컨텐츠는 Deep Web이 Surface Web보다 1,000~2,000배이며, Deep Web의 증가속도는 Surface Web의 증가속도보다 빠르며, Deep Web의 95%는 무료로 사용 가능하다고 보고하였다.

Brightplanet은 각종 주제별 웹기반 DB엔진은 삼십만이 넘는다고 설명하고 있다. (Brightplanet.com/deepcontent/index.asp)

3. Invisible Web 탐색보조도구

Invisible Web 정보는 이용자가 원하는 Invisible Web 컨텐츠를 담고 있는 DB엔진의 웹사이트에서 탐색식을 입력하여 직접 탐색을 수행하여 얻는 것이 기본적인 방법이다. 그러나 DB의 웹사이트를 알지 못하는 이용자를 위해서 주제별 DB엔진의 사이트를 안내한다든지, 각기 다른 DB엔진의 사용에 익숙치 못한 이용자를 대신하여 여러 DB엔진에 탐색식을 보내어 탐색결과를 통합하여 제공해 주는 탐색 보조도구들이 Invisible Web 탐색을 돕는다.

3. 1 표준탐색엔진의 포탈서비스

표준탐색엔진은 html문서로 작성된 DB엔진의 홈페이지를 색인하기 때문에 특수 DB엔진의 사이트를 엑세스하는 Invisible Web 탐색의 출발점이 될 수 있다. 그러므로 toxic chemicals 관련 DB 검색엔진을 찾기 위해서는 "toxic AND chemicals AND database"와 같이 "database"라는 탐색어를 추가하여 "toxic chemicals" 관련 DB를 탐색할 수 있다.

또한 <표 1>에서와 같이 Google, AltaVista, Alltheweb과 같은 표준탐색엔진은 비html 텍스트화일, 이미지, 동영상, 실시간정보를 선택적으로 색인하기 때문에 Invisible Web 정보를 직접 제공한다. 그러므로 "2003년도 미국의 연령별 실업률" 정보를 탐색하기 위해서는 "us AND unemployment rates by age" AND "filetype:pdf"와 같이 주제탐색어 외에 pdf화일로 탐색을 제한시킴으로 고품질의 미연방정부의 정보를 직접 얻을 수 있다.

〈표 1〉 웹 탐색엔진에서 색인하는 비html화일

	Google	Alltheweb	Altavista
Image	yes	yes	yes
Audio	no	mp3	mp3
Video	no	yes	yes
기타 탐색	목록 Groups	ftp	
기타 화일형식	pdf, ps, word, ppt, excel 등	pdf, flash	pdf
뉴스	yes	yes	yes
Online 상품	Froogle		
Directory	yes	no	yes
용어	lab		

3. 2 Invisible Web 게이트웨이

Invisible Web 탐색을 돕는 Invisible Web (IW) 게이트웨이는 크게 3가지로 구별될 수 있다. 첫째는 DB엔진들을 색인하여 디렉토리를 제공하는 IW 디렉토리시스템, 둘째는 디렉토리에서 선정된 주제범주의 전문엔진(DB)들에게로 탐색식을 보내어 탐색결과를 통합하여 보여주는 IW 메타탐색엔진, 셋째는 Invisible Web을 직접 색인하여 탐색을 제공하는 IW 탐색엔진이다. 디렉토리시스템으로는 Invisible-web.net, Beaucoup, Fossick, CompletePlanet 등이 있으며, 메타탐색엔진으로는 ProFusion과 Search.com이, IW 탐색엔진으로는 Incy-Wincy가 대표적이다.

3. 2. 1 DB엔진 디렉토리시스템

<표 2>는 대표적인 DB엔진 디렉토리시스템을 비교한 것이다. Beaucoup, Fossick, CompletePlanet은 주제별 DB의 디렉토리만 제공하는 시스템이지만, IW 메타탐색시스템인 ProFusion과 Search.com도 주제별 DB에 대한 디렉토리를 제공한다.

색인된 DB수로 보면 CompletePlanet이 70,000개 이상의 DB엔진을 색인하고 있고, ProFusion은 10,000여개, Fossick은 3,000여개, Beaucoup은 2,500여개, Search.com은 1,000여개의 DB엔진을 주제별로 분류하고 있다. CompletePlanet은 인문, 사회, 자연, 과학, 문학, 정치, 경제, 군사, 종교 전 분야에 걸쳐 학술적인 DB엔진들을 망라하고 있다. Search.com과 Beaucoup은 경제, 고용, 오락, 건강 등 일반적인 공통관심사를 다룬 DB

를 주로 색인하였으며, ProFusion은 일반관심사 이외에 농업, 생물학, 천문학분야, Fossick은 농업과 환경분야의 학술BD도 색인하고 있다.

ProFusion, Search.com, Beaucoup, Fossick에서 DB탐색은 디렉토리에서만 가능하나, CompletePlanet은 디렉토리와 함께 키워드탐색이 가능하다. CompletePlanet의 홈페이지에서 키워드탐색창은 DB탐색에 사용되지만, 다른 네 시스템의 홈페이지에서 키워드검색은 웹 탐색엔진으로 탐색식을 보내고 검색결과를 통합하는 웹 메타엔진 기능을 수행한다. ProFusion은 13개의 엔진(Altavista, About, Alltheweb, AOL, Adobe pdf, Lycos, Looksmart, MSN, Metacrawler, Netscape, Raging Search, Teoma, WiseNut)에게로, Search.com은 5개의 엔진(Google, MSN, Open Directory, Thunderstorm, WiseNut)에게로, Beaucoup은 6개의 엔진(Yahoo, About, Infoseek, Webcrawler, AlltheWeb, Lycos)에게로, Fossick은 7개의 엔진(Altavista, Excite, Infoseek, Hotbot, Lycos, Northenlight, Webcrawler)에게로 탐색식을 보낸다(2004. 8. 3 현재).

3. 2. 2 IW 메타탐색시스템

ProFusion과 Search.com은 DB디렉토리와 표준웹 메타탐색과 함께 Invisible Web 메타탐색까지 제공한다. Search.com은 홈페이지에서 Research by Topic을 클릭하여 얻은 Specialty Searches 디렉토리에서 분야별 DB엔진 리스트를 보여준다. 주제 분야를 선택한 화면에서 원하는 DB를 체크하고 탐색식을 입력하면 체크된 DB엔진으로 탐색식을 보내

〈표 2〉 IW 게이트웨이 비교

	ProFusion	Search	Beaucoup	Fossick	CompletePlanet	IncyWincy
Directory(범주)	21	14	15	11	42	16(ODP)
DB(엔진) 수	10,000?	1,000+	2,500+	3,000+	70,000+	46,000,000 페이지
메타엔진	Home> 웹 탐색엔진(13)	Home> the web(5)	Home>AlltheInternet(6)	Home>Meta search(7)	no	Home>meta search(5)
강점 DB	일반/학술: 인문, 예술, 경제, 고용, 정부, 법률, 오락, 건강, 생물, 농업, 천문학	일반: 경제, 고용, 음악, 오락, 게임, 건강, 다운로드	일반: 일반, 경제, 고용, 컴퓨터, S/W, 건강, 오락	일반/학술: 경제, 오락, 취미, 컴퓨터, S/W, 게임, 사회과학, 농업, 환경	학술: 인문, 사회, 자연과학, 문학, 군사, 정치, 종교, 역사	일반/학술: Form DB
DB 디렉토리검색	<----------디렉토리에서 범주 선택---------->					
IW 메타엔진	<--서브디렉토리에서--> DB선택		no	no	no	no
IW 자체DB에서 탐색	no	no	no	no	no	Yes
웹 메타탐색	<--------Home에서 키워드 탐색--------->				no	Home에서 키워드탐색

고 탐색결과는 엔진(source)별로 제공한다.

ProFusion은 홈페이지에 리스트된 디렉토리에서 주제 분야를 선택하여 얻은 화면에서 리스트된 모든 DB엔진을 다 선택하거나(all), 가장 빠른 5개 엔진으로만(Fastest 5), 혹은 체크된 엔진(You choose)만을 선택하여 탐색식을 보내고 탐색결과는 통합하여 적합성순으로 제공한다. Score별 탐색결과는 title별, URL별 정렬도 가능하다.

ProFusion과 Search.com은 홈페이지의 입력창에서는 표준 웹 메타탐색을, 분야별 서브디렉토리 화면의 입력창에서는 Invisible Web 메타탐색을 수행한다.

3. 2. 3 IW 탐색엔진

DB의 디렉토리탐색과 키워드탐색만을 제공하는 CompletePlanet, DB의 디렉토리 탐색과 웹 메타엔진 기능을 수행하는 Beaucoup과 Fossick, DB의 디렉토리탐색과 웹 메타탐색과 함께 Invisible Web의 메타탐색까지 제공하는 ProFusion과 Search.com과는 달리, IncyWincy는 IW 탐색엔진이다. IncyWincy는 Open Directory Project(ODP)에 색인된 4백만 웹페이지를 더 깊이 파고 들어가 4천 6백만 Invisible Web 페이지를 색인한 IW 탐색엔진이다. 4천 6백만 웹페이지는 ODP의 주제 디렉토리로 분류되었다. 홈페이지에서의 키워드탐색은 전체 색인에서 탐색하나, 선택된 서브디렉토리 화면에서의 키워드탐색은 전체 IncyWincy에 색인된 웹페이지 중 선택된 주제 분야로 한정하여 탐색한다.

한편 IncyWincy의 웹 메타탐색은 홈페이지에서 "Meta Search"를 클릭한 후 키워드탐색을 수행하면 Alltheweb과 Altavista, Google, Teoma와 IncyWincy 다섯 엔진으로 탐색식을 보내고 탐색결과를 엔진별(Source)로 정렬하여 제공한다. 엔진별 탐색결과는 Rank별, Title별 정렬이 가능하다.

4. Invisible Web 탐색 실험

4. 1 Invisible Web 탐색도구

본 연구에서는 연구과제1을 위하여 표준탐색엔진 Google과, IW 탐색엔진 IncyWincy, IW메타탐색엔진 ProFusion과 Search.com이 선택되었다. 국내DB를 안내하고 탐색을 대행하는 적절한 IW(메타)탐색엔진이 없기 때문에 외국의 IW(메타)탐색엔진을 선택하였다. IW 게이트웨이 중 Beaucoup과 Fossick, CompletePlanet은 Invisible Web DB엔진만 제공할 뿐 Invisible Web 자체를 탐색하지는 않기 때문에 제외되었다. ProFusion과 Search.com, IncyWincy는 Invisible Web (메타)탐색뿐만 아니라 웹 메타탐색기능도 수행하기 때문에 연구과제 2를 위해서도 적절한 Invisible Web 탐색도구로 생각되었다.

IW(메타)탐색엔진의 성능을 비교평가하기 위한 표준웹엔진으로는 Google이 사용되었다. Google은 표준 탐색엔진 중 가장 선호되고 성능이 우수한 탐색엔진으로 보고되고 있을 뿐만 아니라(Hawking de al. 2001, Eliopoulos & Gotlieb 2003, Vaughan 2004), <표 1>에서 본 바와 같이 다양한 Invisible Web을 직접 색인하고 있기 때문에 선택되었다. 미국

의 IW(메타)탐색엔진이 사용되었기 때문에 영어로 된 탐색문이 필요했고, 때문에 네이버나 엠파스같은 국내엔진은 영문문서의 색인에서 Google과 비교되는 것이 불합리하다고 생각되어 제외되었다.

Google은 총 3백 3십만 웹 페이지를 색인하고 있으며, 많이 링크된 문헌 순으로 적합성 순위를 부여하는 PageRank; 웹사이트로 직접 연결시켜 주는 "I'm feeling Good"; 적합성 피드백 탐색(Similar Pages 탐색); filetype과 site, date, 용어출현위치 등으로 탐색을 제한하는 제한탐색; 검색된 문헌이 없으면 자동으로 철자를 체크하는 기능 등을 가지고 있다.

<표 3>은 실험에서 사용된 4개 시스템의 탐색 특성을 비교한 것이다. 4 시스템 모두에서 대소문자를 구별하지 않으며, space는 AND연산자 역할을 하고, "+"와 "-" 전치기호, 구단위 인접연산자 " "를 사용한다. 불연산자 AND, OR, NOT은 Google을 제외한 세 시스템에서 사용가능하다.

Google에서는 복수/단수, 동의어, 철자 변형을 자동 통제한다. 불용어가 탐색어로 사용되면 탐색시 자동 제외된다. 그러나 불용어가 들어 있는 구(phrase)를 인접탐색할 때 불용어가 반드시 포함된 탐색을 원한다든지, 복수/단수 통제를 원하지 않을 때는 "+" 전치기호를 사용한다. NOT은 "-"를 사용하며, OR은 대문자 OR을 사용한다.

IncyWincy는 AND, OR, ANDNOT를 사용하나 괄호( )는 사용하지 못한다. 용어절단으로는 *기호를 사용하고, 필드탐색이나 제한탐색을 못하며, 불용어는 사용하지 않는다.

〈표 3〉 시스템의 탐색기능 비교

	Google	ProFusion	Search.com	IncyWincy
디렉토리 검색	ODP 디렉토리	자체 디렉토리 DB엔진 검색	자체 디렉토리 DB엔진 검색	ODP 디렉토리
홈페이지 검색창	Web탐색	메타탐색	메타탐색	메타탐색
서브디렉토리 검색창	Web탐색 주제 제한	IW 메타탐색	IW 메타탐색	IW 탐색 주제 제한
space=AND	Y	Y	Y	Y
+ - 사용	Y	Y	Y	Y
AND 사용	no	Y	Y	Y
OR 사용	OR(대문자만)	Y	Y	Y
NOT 사용	no	Y	Y	ANDNOT 사용
구탐색 " "	Y	Y	Y	Y
용어 우측절단	no			*
출력내용				
제목, 초록, URL	Y	Y	Y	Y
Source		Y	Y	Y
기타	Filetype size, date	Score		size IncyWincy 카테고리
정렬				
디폴트	PageRank	Score	Source	Source
기타		title URL		rank title
적합성피드백검색	Similar Pages	Similar Results	no	no

ProFusion과 Search.com은 탐색식을 다른 여러 DB엔진에게 보내기 때문에 특별한 주의가 필요하다. 어떤 엔진에서는 +, -, " ", 불연산자를 사용하나 대부분의 엔진은 지원하지 않기 때문에 여러 가지 방법으로 시도해 보라고 권하고 있다(Search.com의 help).

탐색결과는 제목과 초록, URL이 네 시스템 모두에서 제공되며, ProFusion은 Source와 적합성 점수를, Google은 화일의 유형과 크기, 날짜를, IncyWincy는 Source와 size, IncyWincy 카테고리를 제공한다.

검색결과는 Google에서 적합성 순위로 제공된다. Search.com은 메타탐색의 경우는 통합된 탐색결과를 제공하나, IW 메타탐색의 경우는 통합하지 않고 엔진별로 제공한다[6]. ProFusion은 탐색결과를 통합하여 Score별로 제공하며, Title과 URL로 정렬도 가능하다. IncyWincy는 탐색결과를 Source별로 제공하나, Rank 혹은 Title로 통합 정렬도 가능하다.

4. 2 탐색 질의와 탐색식

Invisible Web 종류와 선택된 4개의 탐색도구가 다루고 있는 정보의 유형과 주제를 고려하여 11개의 질의가 연구자에 의해 작성되었다(표 4 참조). 두 개의 질문은 Invisible Web을 담고 있는 DB 탐색엔진을 탐색하기 위한 것이고, 나머지 9개의 질문은 Invisible Web의 유형을 고려하여 DB에 들어있는 특정 Invisible Web(학술잡지기사원문, pdf정보, 이미지정보, 통계정보, 영화리뷰기사, 요리법, 실시간정보)을 탐색하기 위한 것이다.

탐색질의는 간략탐색 전략을 사용하여 파셋을 AND연산자로만 조합하였다. 그러나 사용된 파셋은 각 시스템에서 사용된 주제 범주를 고려하여 신축적으로 사용하였다. 예를 들어 탐색질문 1에서 ProFusion은 Science에서 탐색하였으므로 Physics와 Preprint, Database라는 3개의 파셋이 사용되었으나, IncyWincy에서는 Physics에서 탐색되었으므로 Preprint와 Database 두 개념만 사용되었다. Job, recipe, review, stock과 같은 파셋이 신축성있게 사용되었다.

각 탐색식은 4개 시스템에서 공통적으로 사용되는 스페이스를 AND 대신 사용하여 작성되었다. Google에서는 꼭 필요한 불용어가 생략되지 않도록 +"the us"("us"로 탐색되지 않도록), +"as we may think", +"the simplest little higgs"와 같이 "+"를 사용하였다. 그러나 DB엔진에서는 "+"기호 사용이 통일되어있지 않으므로 메타탐색을 수행한 다른 세 엔진에서는 "+"는 사용하지 않았다. AND로 조합되는 탐색용어의 순서는 탐색결과의 랭킹에 영향을 미칠 수 있으므로 네 시스템에서 통일되었다.

세 IW 탐색도구에서의 웹 메타탐색에는 동일 탐색식이 사용되었다. 모든 탐색개념 파셋이 space로 조합되었다.

탐색결과는 적합성 순위로 상위 10개로 제

6) Help 페이지에 의하면 Sort by Source, Relevance, Date 기능이 있다고 안내하고 있으나 탐색기간 중 이 기능은 제공되지 않았다.

〈표 4〉 탐색질문과 탐색주제 범주 및 탐색식

질문 1: 물리학 분야의 Preprint DB		
Google		physics preprint database
ProFusion	Science>	physics preprint database
Search.com		physics preprint database
IncyWincy	Physics>	preprint database
메타탐색		physics preprint database
질문 2: 구직 관련 DB		
Google		"job search" database
ProFusion	Career>Jobs>	
Search.com	Employment>Job Search>	
IncyWincy	Business>Employment>Job Search>	"job search" database
메타탐색		"job search" database
질문 3: "martin schmaltz"가 쓴 preprint "the simplest little higgs"의 본문		
Google		"martin schmaltz" +"the simplest little higgs"
ProFusion	Science>	"martin schmaltz" "the simplest little higgs"
Search.com		"martin schmaltz" "the simplest little higgs"
IncyWincy	Physics>	"martin schmaltz" "the simplest little higgs"
메타탐색		"martin schmaltz" "the simplest little higgs"
질문 4: 뉴욕에서 참고사서직 구인 정보		
Google		job reference librarian "new york"
ProFusion	Career>Jobs>	reference librarian "new york"
Search.com	Employment>Job Search>	reference librarian "new york"
IncyWincy	Business>Employment>Job Search>	reference librarian "new york"
메타탐색		job reference librarian "new york"
질문 5: Vannevar bush의 "as we may think" 전문		
Google		"vannevar bush" +"as we may think"
ProFusion	Science>	"vannevar bush" "as we may think"
Search.com		"vannevar bush" "as we may think"
IncyWincy	Physics>	"vannevar bush" "as we may think"
메타탐색		"vannevar bush" "as we may think"
질문 6: 2000년 인구 조사에서 재미교포 수		
Google		+"the us" population Korean 2000
ProFusion	Government>Federal Government>	"the us" population Korean 2000
Search.com	Reference>Encyclopedia>	"the us" population Korean 2000
IncyWincy	Reference>	"the us" population Korean 2000
메타탐색		"the us" population Korean 2000
질문 7: SARS의 분포 지도		
Google		sars distribution map
ProFusion	Health>	sars distribution map
Search.com	Health & Medicine> Gerneral Health>	sars distribution map
IncyWincy	Health>	sars distribution map
메타탐색		sars distribution map
질문 8: 김치 담그는 법		
Google		Kimch recipe
ProFusion	Living>Food & Beverage>	Kimch
Search.com	Reference>Recipes>	Kimch
IncyWincy	Cooking>	Kimch
메타탐색		Kimch recipe

질문 9: 영화 "talk to her" 리뷰기사		
Google		"hable con ella" review
ProFusion	Entertainment>Movie Reviews>	"hable con ella"
Search.com	Entertainment>Movies>	"hable con ella"
IncyWincy	Arts>Movies>Reviews>	"hable con ella"
메타탐색		"hable con ella" review
질문 10: 피카소의 자화상 그림		
Google	Images>	"pablo picasso" self-portrait
ProFusion	Arts & Humanities>Art>	"pablo picasso" self-portrait
Search.com	Downloads>Images>	"pablo picasso" self-portrait
IncyWincy	Arts>	"pablo picasso" self-portrait
메타탐색		"pablo picasso" self-portrait
질문 11: IBM 현재(2004. 8. 13) 주가		
Google		ibm stock
ProFusion	Business>Business Discussion>	ibm stock
Search.com	Business & Money>Stock Quotes>	ibm
IncyWincy	Business>Investing>Stocks & Bonds>	ibm
메타탐색		ibm stock

한 출력하였다. IW 메타탐색의 경우 탐색결과를 통합하여 적합성 순위로 정렬하였으나, Search.com의 IW 메타탐색에서처럼 통합되지 못하고 Source별로 결과가 제공된 경우엔 Source에 상관없이 출력된 순서대로 상위 10개를 최종 결과로 선택하였다.

Google은 질문 9의 이미지 검색을 제외하고 모두 Web에서 검색하였다. 그러나 일반적인 공통관심사로 DB가 제한된 Search.com에서는 과학 주제(질문 1, 3, 5) DB가 없기 때문에 웹 메타탐색으로 IW 메타탐색을 대신하였다.

IW 메타탐색에서 주제 범주 내 엔진의 선택은 일정시간 동안 선택한 엔진의 탐색결과가 일정량이 되는대로 탐색을 중단하기 때문에 IW 메타탐색의 결과에 영향을 준다. 특히 메타탐색의 결과가 source별로 10건씩 출력되기 때문에 첫번째 source의 상위 10건이 최종결과 상위 10건으로 선택되는 Search.com에서 DB엔진의 선택은 탐색결과에 큰 영향을 준다. ProFusion과 Search.com의 IW 메타탐색에서 주제 범주 내 엔진의 선택은 탐색질문에 적합한 엔진은 모두 선택하였다(5.1에서 자세히). 탐색옵션선택이 가능한 ProFusion에서 탐색옵션은 디폴트로 설정된 Show Results 1-10, Results per Source 10, Search Timeout 30초를 그대로 사용하였다.

4. 3 적합성 판정

검색된 사이트의 적합성은 사이트를 클릭하여 웹페이지를 보고 판단하였다. 적합성은 완전적합과 부분적합, 부적합으로 판정하였다. 인터넷에서 하나의 웹페이지는 19번의 클릭으로 다른 어떤 페이지로든 이동이 가능하다고 보고되었다(Albert, Jeong & Barabasi 1999). 이것은 부적합페이지에서 클릭 몇 번으로 적합페이지를 얻을 수 있음을 의미한다. 인쇄문헌과는 달리 웹페이지의 적합성은 적합정보를 얻는데 링크의 사용여부가 고려되어야 할 것으로 생각

되었다. 이를 위해 본 연구에서는 적합문헌을 완전적합과 부분적합으로 구분하였다. DB를 원하는 질문 1과 질문 2에서는 DB엔진의 홈페이지만 완전적합으로 간주하고, DB 홈페이지로 안내하는 색인페이지는 색인페이지에서 리스트된 DB를 클릭하면 다시 DB엔진의 홈페이지로 연결되기 때문에 부분적합으로 간주하였다. 질문과 직접 관련된 Invisible Web 페이지를 원하는 질문 3-11에서는 해당 Invisible Web 페이지만 완전적합으로 간주하고, 해당 Invisible Web 페이지를 담고 있는 DB엔진의 홈페이지나 색인페이지는 부분적합 처리하였다. DB에 들어 있는 Invisible Web의 탐색 성능을 연구하는 것이 본 연구의 목적이기 때문이다.

질문 1에서 preprint를 포함하여 잡지기사, 보고서, 책 등을 색인한 물리학 종합DB는 부분적합 처리하였다. 본문을 요구한 질문 3과 질문 5에서 본문으로 링크가 되어 있는 해당 논문의 요약문 사이트는 부분적합으로 간주하였지만, 다른 논문에서 참고문헌으로 링크된 경우는 부적합으로 간주하였다. 질문 6은 재미교포 수 1,076,872명을 제시한 페이지만 완전적합으로 간주하고, 제목은 Number of Korean-American이지만 인구수가 다른 페이지는 부분적합으로 간주하였다. 질문 7의 경우도 SARS의 분포지도를 직접 제공하면 완전적합, SARS에 관한 웹사이트 중 분포지도로 링크가 되어 있는 사이트는 부분적합으로 간주하였다. 검색된 사이트가 질문 관련 사이트는 아니지만 원하는 정보가 그 사이트의 홈페이지에 들어 있는 경우는 부분적합으로 간주하였다. 즉 질문 8에서 김치 담그는 법이 홈페이지에 나타나 있는 한국음악사이트와 질문 11에서 IBM 주가 snapshot이 프레임으로 포함되었지만 주가에 대한 일반인들의 코멘트나 질문, 응답을 내용으로 한 페이지는 부분적합으로 판정하였다.

동일 적합사이트가 반복되어 상위 10건에 들어 있을 경우, 반복된 사이트는 추가로 어떤 정보도 제공하지 않고 이용자에게 한번 더 링크하게 하는 수고만 끼치므로 적합사이트는 1개로 간주하고 나머지 반복된 사이트는 부적합으로 간주하였다. 변경된 사이트로 자동링크되지 않고 변경된 사이트만 안내한 경우는 Dead 사이트로 간주하였다.

연구자의 적합성판정에 대한 신뢰도는 제 3자의 적합성판정과의 일치도를 측정하여 증명할 수도 있겠지만, 본 연구에서는 각 질문에 대해 완전적합, 부분적합, 부적합의 판정기준을 분명히 세움으로 반복측정 대신 신뢰성 문제가 어느 정도 해결되기를 기대하였다. 즉 적합, 부분적합, 부적합 판정의 기준이 타당하다면, 또한 제 3자가 이 기준을 잘 따른다면, 잡지기사 원문은 적합, 원문에 대한 링크가 있는 요약문은 부분적합; 1,076,872란 숫자가 있으면 적합, 인구수를 나타내지만 그 숫자가 다르면 부분적합 등으로 판정하는데 판정자에 따라 다른 판정을 할 가능성은 매우 적을 것으로 생각되었다.

4. 4 성능분석 척도

IW 탐색도구의 탐색성능 평가척도로는 상위 10개 문헌의 적합문헌수, 순위정확률이 사용되었다.

물리학 분야의 Preprint DB엔진(질문 1)이나 뉴욕에서 참고사서 모집공고(질문 4)를 원

하는 질문에서는 검색된 적합사이트의 총 수만큼 다양한 정보가 제공되므로 검색된 적합사이트수가 중요할 것이다. 그러나 "The Simplest little higgs"의 본문을 찾는 질문 3과 같은 질문에서는 검색된 여러 적합사이트는 복권(copy)을 의미할 뿐 다른 어떤 정보도 제공하지 못하므로 검색된 적합사이트 수는 의미가 없을지도 모른다. 적합사이트수보다는 오히려 첫번째 적합사이트의 출력 순위가 더 중요할지도 모른다. 첫번째 적합사이트의 출력 순위는 적합정보를 얻기까지 클릭하여 접속한 사이트 수 즉 이용자의 탐색시간과 노력을 의미하기 때문이다. 이처럼 검색된 적합문헌수보다는 검색된 적합문헌의 랭킹순위를 보다 중요하게 반영하는 척도가 순위정확률이다.

11-포인트 평균정확률은 대표적인 순위정확률이지만 복잡한 계산을 필요로 하기 때문에 보다 간편한 순위정확률 척도를 사용한 연구들이 출현하였다(우유미,정영미 1998, Leighton & Srivastava 1999). 11-포인트 평균정확률을 포함하여 여러 연구에서 사용된 다양한 순위정확률 간의 상관관계를 조사한 연구에서 First-n 정확률을 비롯하여 여러 간편 척도는 11-포인트 정확률과 상관관계가 매우 높고, 특히 비교하는 탐색결과세트의 문헌 수가 동일하지 않을 때 First-n 정확률은 11-포인트 정확률보다 더욱 분별력이 좋은 성능측정 척도인 것으로 보고되었다(노정순 2000).

본 연구에서는 두 종류의 순위정확률이 성능측정 척도로 사용되었다. 첫 번째 순위정확률 First-n P(1)은 적합문헌과 부분적합을 적합으로 간주한 2진척도를 사용하였다. First-n P(1)은 검색된 문헌이 적합문헌일 경우 그 문헌이 top 10에서 1등이면 10점, 2등이면 9점, … 10등이면 1점으로 순위가중치를 부여한 후, 검색된 적합문헌의 순위가중치의 합을 분자값으로 하고, 10개 문헌이 모두 적합문헌일 경우 순위가중치 합계 55를 분모값으로 하여 나눈 값이다. 검색된 사이트가 10개 미만일 경우엔 10개 적합문헌의 순위가중치 합계(55)에서 부족분(10-검색된 사이트 수) 당 최소순위가중치 1점을 마이너스하여 분모값에 변화를 줌으로써 검색된 사이트수에 영향을 받는 정도를 반영시켰다.

$$First\text{-}nP(1) = \frac{\sum(\text{검색된적합사이트의순위가중치})}{55-(10-\text{검색된사이트수}n)}$$

그러므로 검색된 상위 10건 중 1등, 2등, 3등 문헌이 적합문헌일 경우 First-10 P(1)은 (10+9+8)/55 = 49.09%이나, 4건이 검색되고 그 중 1등, 2등, 3등 문헌이 적합문헌일 경우 First-4 P(1)은 (10+9+8)/55-(10-4)= 55.10%이다.

두번째 First-n P(2)는 순위정확률 First-n P(1)에 적합문헌은 2점, 부분적합은 1점, 부적합은 0점의 적합성가중치를 부여한 적합성순위정확률이다.

$$First\text{-}nP(2) = \frac{\sum(\text{검색된적합문헌순위가중치}\times\text{적합성가중치})}{55\times2-(10-\text{검색된사이트수}n)\times2}$$

그러므로 상위 10건 중 1등, 2등, 3등이 완전적합일 경우 First-10P(2)는 First-10P(1)과 동일하나, 1등, 2등, 3등이 부분적합일 경우 First-10P(2)는 First-10P(1)의 1/2이 된다. 적합성순위정확률은 양질의 적합정보를 검색하는 능력을 측정하는 척도로 사용되었다.

5. 데이터 분석 및 논의

4개 시스템에서 11개 질의의 탐색은 2004. 8. 2 - 8. 13에 수행되었다. 같은 질문은 4개 시스템에서 같은 날 수행되었다.

5. 1 검색된 적합사이트수

5. 1. 1 표준 웹 탐색과 Invisible Web 메타탐색 비교

<표 5>는 탐색별 검색된 사이트수와 적합사이트를 나타낸 것이다. 적합사이트 수는 완전적합과 부분적합을 구별하여 괄호 안에 표시되었다. 웹 표준탐색엔진과 IW (메타)탐색엔진을 비교하면, 11개 질문의 부분적합까지 포함한 적합사이트 총수는 Google이 78개로 가장 많고, Search.com 65개, ProFusion 61개, IncyWincy 37개 순이었다. Google이 IW 전용 탐색도구보다 더 많은 적합사이트를 검색하였다. 적합문헌 중 부분적합의 비율은 평균 20.79%이며, Google이 11.54%로 가장 작고, ProFusion이 31.15%로 가장 많았다.

질문 1과 3(물리학)과 질문 5(정보학)에서 Search.com은 해당 주제가 없으므로 IW 메타탐색 대신 웹 메타탐색을 수행하였고, ProFusion 역시 Science 범주에서 농업과 천문학, 생물학 이외의 분야를 위해 마련한 Include Web Search Engines을 선택하여 웹 메타탐색을 하였으므로, 이 세 질문에서 비교적 많은 적합문헌을 탐색하였다.

구직 DB엔진을 원하는 질문 2에서 ProFusion은 디렉토리에서 Career>Jobs 아래에 알파벳순으로 배열되어 있는 DB엔진 22개 중 상위 10개를, Search.com은 Employment>Job Search 아래에 있는 10개 엔진을 탐색결과세트의 상위 10건으로 간주하였기 때문에 100% 적합문헌을 얻을 수 있었다.

질문 3에서 IncyWincy는 한 건도 검색하지 못하고, 질문 5에서도 적합문헌 1건만 검색함으로, DB에 들어 있는 본문은 색인하지 않은 것

〈표 5〉 검색된 사이트수(적합사이트수)

질 문	웹 탐색	IW 메타탐색		IW 탐색	웹 메타탐식		
	Google	ProFusion	Search.com	IncyWincy	ProFusion	Search.com	IncyWincy
1	10(7/1)	10(6/2)	10(6/1)	10(4/2)	10(7/1)	10(6/1)	10(5/5)
2	10(9/0)	10(10/0)	10(10/0)	10(1/4)	10(9/0)	10(9/0)	10(9/0)
3	10(3/3)	9(0/6)	10(1/8)	0(0/0)	9(0/6)	10(1/8)	10(1/3)
4	10(7/0)	10(3/0)	0(0/0)	2(1/0)	10(6/0)	10(9/0)	10(5/0)
5	10(5/1)	10(3/1)	10(5/1)	10(1/0)	10(4/1)	10(6/0)	10(4/1)
6	10(2/2)	10(2/0)	10(1/2)	10(2/0)	10(1/0)	10(2/1)	10(4/1)
7	10(4/1)	10(5/2)	10(0/2)	10(1/0)	10(6/2)	10(7/0)	10(5/1)
8	10(8/1)	10(7/0)	10(10/0)	10(8/0)	10(8/0)	10(6/1)	10(7/1)
9	10(10/0)	2(2/0)	10(7/0)	10(6/1)	10(10/0)	10(8/0)	10(7/0)
10	10(9/0)	10(3/0)	10(10/0)	10(6/0)	10(8/0)	10(7/1)	10(8/0)
11	10(5/0)	10(1/8)	1(1/0)	0(0/0)	10(4/0)	10(5/0)	10(6/0)
계	110(69/9)	101(42/19)	91(51/14)	82(30/7)	109(63/10)	110(66/12)	110(61/12)
적합문헌 중 부분적합비율	11.54	31.15	21.54	18.92	13.70	15.38	16.44

※ 괄호 안은 <완전적합문헌수/부분적합문헌수)

으로 보였다.

질문 4에서 ProFusion에서는 Career>Jobs 관련 22개 DB 중 탐색질문과 관련 없는 4개의 DB(물리학과 건축 분야 Job DB와 미네소타와 뉴저지에 위치한 Job DB)를 제외한 나머지 엔진을 모두 선택하여 3개의 적합 페이지를 검색하였으나, Search.com은 Employment/Job Search 아래 2개의 DB(Tech Jobs과 샌프란시스코 지역의 Job DB)를 제외한 8개의 엔진으로 질문을 보냈으나 한 건도 검색되지 않았다. 그 이유는 DB의 URL주소가 다르거나(Fedworld Federal jobs), DB의 검색창이 직종 외에 지역을 콤보 박스에서 체크하게 설계되어 있어(Hotjobs.com과 Flip Dog.com) Search.com이 탐색식을 일괄적으로 DB엔진에게로 바르게 보내지 못하기 때문인 것으로 보인다. IncyWincy는 Job DB에 있는 정보는 색인/검색하지 못하는 것으로 보였다.

질문 6을 위해 ProFusion에서는 Federal Government 아래에 있는 DB 중 6개(Fedworld, MOCAT, FirstGov, SearchGov, Fedworld Information Searvice, GovExec)가 선택됐으며, Search.com에서는 Encyclopedia 아래 Encara, Fact Monster, Infoplease, Xrefe, Wikipedia가 선택되었다. 적합정보원이 Census DB나 Factfinder로 제한되어 있으므로 전체적으로 정확률은 낮지만 재현율은 높은 탐색결과를 얻었다.

질문 7을 위해 ProFusion은 Health 아래에서 Drugs와 Medical Terms, Alternative Medicine을 제외한 나머지 엔진을 선택하였으며, 7건의 적합문헌이 Hardin MD와 MSN에서 검색되었다. Search.com에서는 General Health에서 MeSH와 Healthfinder를 제외한 나머지 4개 DB를 선택하였으나 Hardin MD만 적합문헌 2건을 검색하였다. Hardin MD은 적합문헌 2건을 1등과 2등으로 정렬하였지만 탐색결과를 통합하지 못하는 Search.com은 3건(모두 부적합)을 검색한 Intellihealth 다음에 Hardin MD의 결과를 출력했기 때문에 적합문헌은 4등과 5등이었다. IncyWincy에서 검색된 10건 중 적합문헌은 1건이었다.

조리법(질문 8)과 영화리뷰기사(질문 9)는 모든 엔진에서 비교적 잘 탐색되었다. 단지 질문9에 대해 ProFusion에서는 Movie Reviews 아래 LA Times-Movie Review와 Film Critic, Internet Movie DB, Critics, Roger Ebert Movie Reviews by Title을 선택하였으나 Internet Movie DB와 Film Critic으로부터 각각 1개의 문헌만 검색되었다. Search.com은 Movie 아래 9개 DB를 모두 선택하여 7개의 적합문헌을 검색하였다. MRQE의 검색결과 10건이 상위 10건으로 출력되었기 때문이다.

질문 10에서 ProFusion은 Art 아래에서 선택 가능한 엔진의 제한 때문에 탐색결과가 좋지 못하였다. 선택 가능한 9개의 DB 중 Metropolitan Museum과 The National Gallery의 작가명, 작품명 DB 4개와 건축 관련 DB(ADAM)을 제외하고 남은 4개의 DB는 특정 미술관과 박물관의 소장품 DB와 박물관 백과사전이었으므로, 탐색결과를 특정 미술관/박물관의 소장품으로 제한하였다. Search.com에서는 Images 검색엔진 Picsearch와 Webshots이 100% 적합문헌을 검색하였다.

ProFusion에도 Downloads>Images라는 주제 범주가 있으나, 이 주제에서 IW 메타탐색은 Image 탐색엔진만 탐색하였다.

IBM 주가와 같은 실시간 정보(질문 11)는 IncyWincy는 전혀 색인하지 못하고 있었다. Search는 Stock Quotes 아래 유일하게 나열된 CNET Investor로 탐색식을 보내어 1건을 검색하였다.

5. 1. 2 웹 메타탐색

IW (메타)탐색엔진에서 웹 메타탐색은 ProFusion은 Home에서, Search.com도 Home에서(The Web이 디폴트), IncyWincy는 Home에서 Meta Search를 선택하여 수행하였다. <표 5>에서 보는 바와 같이 세 시스템 모두에서 IW (메타)탐색보다는 웹 메타탐색이 보다 많은 적합문헌을 검색하였다. Search.com은 78건, ProFusion과 IncyWincy는 동일하게 73건의 적합문헌을 검색하였다. Search.com 78건은 Google의 78건과 동일하였다. 적합문헌 중 부분적합문헌의 비율은 평균 15.17%로, IW 메타탐색의 평균 20.79보다 낮았다. 이것은 웹 메타탐색이 IW 메타탐색보다 더 적합성이 높은 문헌을 검색했음을 의미한다. 세 웹 메타탐색 중 ProFusion(13.70%)이 Search.com(15. 38%)과 IncyWincy(16.44%)보다 적합성이 높은 문헌을 검색하였다.

IncyWincy는 모든 질문에서 IncyWincy에서만 탐색한 것보다는 4개의 다른 엔진(Alltheweb, Altavista, Teoma, Google)에서 수행한 탐색결과와 IncyWincy의 결과를 통합한 메타탐색이 보다 많은 적합문헌을 검색하였다.

전체적으로 웹 메타탐색이 IW 메타탐색보다 적합문헌을 많이 검색하였으나, ProFusion은 질문 2, 질문 6, 질문 11에서, Search.com은 질문 2, 질문 8, 질문 10에서 IW 메타탐색이 더 많은 적합문헌을 검색하였다. 질문 2는 ProFusion과 Search.com에서 선택된 서브 디렉토리 아래 나열된 구직엔진을 그대로 검색된 사이트로 간주했기 때문이다. 질문 8은 Search.com에서는 Reference>Recipes라는 질문과 정확하게 일치하는 주제범주 항목이 있었기 때문으로 보인다. Search.com은 Google에서와 같이 image화일만을 개별 범주Downloads>Images>로 모았기 때문에 질문 10에 대해 IW 메타탐색이 보다 많은 적합문헌을 검색하였다. 질문 11에서 ProFusion의 IW 메타탐색이 많은 적합문헌을 검색한 이유는 Stock Discussion 범주에서 Yahoo Message Boards로부터 검색된 IBM주가에 대한 코멘트와 질의, 응답 페이지 9건이 주가snapshot을 프레임으로 포함하고 있어서 부분적합으로 판정됐기 때문이었다. ProFusion 메타탐색은 높은 정확률을 보였지만 실시간 주가 정보를 검색하기에 적당하지 못하였다.

5. 2 순위정확률

5. 2. 1 표준 웹탐색과 Invisible Web 메타탐색

<표 6>은 엔진별로 순위정확률 First-n P(1)을 비교한 것이고, <표 7>은 적합성순위정확률 First-n P(2)를 비교한 것이다. Google의 평균 순위정확률 .7355와 평균 적합성순위정확률 .7124가 각각 가장 우수하였다. IW 메타엔진 중 ProFusion(.6335)은 순위정확률에서

〈표 6〉 탐색별 순위정확률

질문	웹 탐색	IW (메타)탐색			웹 메타탐색		
	Google	ProFusion	Search.com	IncyWincy	ProFusion	Search.com	IncyWincy
1	.8000	.9273	.6364	.6182	.8727	.6364	1.0000
2	.8364	1.0000	1.0000	.4727	.8.27	.9091	.9455
3	.7455	.7455	.9636	.0000	.7455	.9636	.4364
4	.6545	.3818	.0000	.1915	.8000	.8545	.4727
5	.4909	.5091	.7818	.1818	.5091	.7818	.5091
6	.3273	.1091	.1636	.3273	.0545	.2182	.4545
7	.7091	.8364	.2364	.1636	.8182	.7091	.6727
8	.8909	.6909	1.0000	.9091	.8545	.8.27	.8182
9	1.0000	.4043	.6909	.7636	1.0000	.8336	.6909
10	.9455	.3818	1.0000	.6182	.9091	.8364	.8000
11	.6909	.9818	.2174	.0000	.4727	.6364	.5273
평균	.7355	.6335	.6082	.3860	.7190	.7502	.6661

〈표 7〉 탐색별 적합성순위정확률

질문	웹 탐색	IW (메타)탐색			웹 메타탐색		
	Google	ProFusion	Search.com	IncyWincy	ProFusion	Search.com	IncyWincy
1	.7545	.8636	.6182	.5273	.8991	.6182	.6909
2	.8364	1.0000	1.0000	.3000	.8727	.9091	.9405
3	.5909	.3727	.5364	.0000	.3727	.5364	.2909
4	.6545	..3818	.0000	.1915	.8000	.8545	.4727
5	.4818	.4636	.7273	.1818	.4545	.7273	.4182
6	.3545	.1091	.2636	.3273	.0545	.2545	.5455
7	.6455	.7364	.1182	.1636	.7545	.7091	.6000
8	.8818	..6909	1.0000	.9091	.8545	.8273	.7364
9	1.0000	.4043	.6909	.7545	1.0000	.8336	.6909
10	.9455	.3818	1.0000	.6182	.9091	.7818	.8000
11	.6909	.5818	.2174	.0000	.4727	.6364	.5273
평균	.7124	.5442	.5611	.3612	.6686	.6989	.6108

Search.com(.6082)과 IncyWincy(.3860)보다 우수하였다. 그러나 적합성순위정확률은 Search.com(.5611)이 ProFusion(.5442)과 IncyWincy(.3612)보다 우수하였다. 이것은 Search.com보다는 ProFusion에서 적합문헌 중 부분적합문헌의 비율이 높음을 의미한다. IncyWincy는 두 순위정확률 모두에서 다른 세 시스템에 비해 .20 이상 낮은 정확률을 보임으로써 IW 탐색엔진으로 좋지 않은 것으로 보였다. 특히 Google과 비교할 때 IncyWincy는 Google의 절반 수준의 성능을 보였다. 그러나 SPSS 분석결과 4 시스템간의 순위정확률은 유의한 차이를 보이지 못하였다(First-n P(1)에서는 α= .067, First-n P(2)에서는 α= .055)[7].

5. 2. 2 웹 메타탐색

<표 6>에서와 같이 엔진별 웹 메타탐색의 순위정확률 First-n P(1)은 Search.com이 .7502로 가장 많고, ProFusion이 .7190, IncyWincy가 .6661로 가장 낮았지만, 그 차이는 미비하여 매우 비슷하였다. <표 7> 적합성순위정확률 First-n P(2)에서도 Search. com이 .6989, ProFusion이 .6686, IncyWincy가 .6108로, 순서와 차이는 순위정확률에서와 비슷하였다. SPSS의 ANOVA 분석결과 세 메타탐색 간의 차이는 보이지 않았다(순위정확률에서는 α= .689, 적합성순위정확률에서는 α= .652).

<표 8>은 세 IW (메타)탐색 엔진에서 IW (메타)탐색과 웹 메타탐색의 평균간의 차이를 순위정확률과 적합성순위정확률로 표시한 것이다. 두 정확률 모두에서 웹 메타탐색이 IW (메타)탐색보다 평균 .17 더 높은 정확률을 보였다. IncyWincy에서 차이(.2801와 .2496)가 가장 많이 났고, Search.com이 두번째였다. ProFusion에 비해 IncyWincy와 Search. com에서 메타탐색이 높은 성능차이를 보인 것은 IncyWincy와 Search.com은 Google로 탐색식을 보내기 때문인 것으로 보인다. 메타탐색과 IW (메타)탐색 간의 차이는 T검증 결과, 유의수준 .001(순위정확률)과 .015 (적합성순위정확률)에서 유의하였다.

5. 3 두 정확률 척도 간의 차이

<표 9>에서와 같이 7개의 모든 탐색에서 적합성순위정확률이 순위정확률보다 조금 낮

〈표 8〉 IW 탐색과 웹 메타탐색의 정확률 비교

질문	순위정확률			적합성순위정확률		
	IW 메타탐색	웹 메타탐색	차이	IW 메타탐색	웹 메타탐색	차이
ProFusion	.6335	.7190	.0855	.5442	.6686	.1244
Search.com	.6082	.7502	.1480	.5611	.6989	.1378
IncyWincy	.3860	.6661	.2801	.3612	.6108	.2496
평균	.5426	.7118	.1692	.4888	.6594	.1706

〈표 9〉 순위정확률과 적합성순위정확률의 비교

	웹 탐색	IW (메타)탐색			웹 메타탐색			평균
	Google	ProFusion	Search.com	IncyWincy	ProFusion	Search.com	IncyWincy	
순위정확률 (순위)	.7355 (2)	.6335 (5)	.6082 (6)	.3860 (7)	.7190 (3)	.7502 (1)	.6661 (4)	.6426
적합성순위정확률 (순위)	.7124 (1)	.5442 (6)	.5611 (5)	.3612 (7)	.6686 (3)	.6989 (2)	.6108 (4)	.5939
차 이	.0231	.0893	.0471	.0248	.0504	.0513	.0553	.0487

7) α= .055는 .05보다 크기 때문에 차이가 없다고 단정짓기 보다는 .055가 얼마나 .05와 가까운지 생각해 볼 필요가 있다. 사회과학분야에서는 유의수준을 0.1로 사용하기도 한다.

아졌다(평균.0487). 이는 검색된 적합문헌 중 부분적합이 차지하는 비율이 높지 않음을 의미한다. ProFusion의 IW 메타탐색이 가장 높은 차이가 나는데, 이것은 부분적합문헌의 비율이 가장 높기 때문이다. Google이 적합문헌 중 부분적합문헌의 비율이 가장 낮음을 알 수 있다.

두 순위정확률 간의 차이는 크기 않지만 엔진 간의 탐색성능 순위에는 차이가 있었다. Search.com의 웹 메타탐색은 Google과 마찬가지로 78개의 적합문헌을 검색하였다. 그러나 순위정확률에서는 Search.com 메타탐색이, 적합성순위정확률에서는 Google이 가장 좋은 성능을 보였다. 이것은 Search.com 메타탐색보다 Google 탐색에서 완전적합문헌의 비율이 높기 때문이었다.

두 정확률 사이의 Pearson의 상관계수는 .927로 두 정확률 척도는 매우 유사함을 보여주고 있다. 적합성순위정확률이 검색된 문헌의 질(적합성 정도)까지 반영하는 척도로 사용될 수 있음을 보여주고 있다.

6. 결론 및 제언

Invisible Web 탐색을 연구한 본 연구에서 IW 메타탐색엔진 ProFusion과 Search.com, IW 탐색엔진 IncyWincy는 Invisible Web 탐색에서 표준웹엔진 Google보다 우수하지 못하였다(α= .055). 통계적으로 유의하지는 않지만 오히려 Google이 보다 높은 순위정확률을 보임으로써, 여러 선행연구에서 가장 우수한 웹탐색엔진으로 보고된 Google이 Invisible Web 정보탐색에서도 IW (메타)탐색엔진보다 우수한 것으로 보였다. 특히 IncyWincy보다는 2배나 높은 정확률을 보였다. 특정 주제에 관계없이 전체적으로 Google은 높은 성능을 보였으며, 주가와 같은 실시간정보, 이미지정보, 구직정보, 학술정보 탐색에서는 탁월한 성능을 보였다.

IW (메타)탐색엔진에서 웹 메타탐색은 IW 메타탐색보다 통계적으로 유의한 수준에서 우수하였다. IW (메타)탐색에 비해 웹 메타탐색에서 Search.com과 IncyWincy가 ProFusion보다 상대적으로 높은 성능 차이를 보인 것은 ProFusion은 Google로 탐색식을 보내지 않지만 Search.com과 IncyWincy는 Google로 탐색식을 보내기 때문인 것으로 보인다.

IW (메타)탐색엔진 중에서는 ProFusion과 Search.com이 IncyWincy보다 높은 정확률을 보이고 있다. IncyWincy는 Invisible Web 탐색이 좋지 못하며 특히 실시간 정보나 DB화된 학술정보 등은 거의 색인하고 있지 않는 것으로 보였다. Search.com은 구직정보에서, ProFusion은 영화리뷰 분야에서 DB를 보충할 필요가 있는 것 같다.

본 연구결과 제안된 Invisible Web 탐색전략은 다음과 같다.

첫째, 기본적으로 Google을 출발점으로 탐색을 시작한다.

둘째, Google로 충분한 정보를 얻지 못하고 추가로 정보가 필요할 경우, IW 메타엔진의 디렉토리에 주제 영역이 분명하면 IW 메타탐색한다.

셋째, IW 메타엔진의 디렉토리에 적절한

주제 영역이 없을 경우엔 웹 메타탐색을 수행한다.

본 연구결과는 영어로 된 외국정보를 검색하는데 제한되어 해석 응용되어야 하지만, 국내 BD에 대한 IW메타엔진이 없는 국내환경에서는, Google의 Invisible Web정보를 색인하는 정책과 기술로 볼 때, Google Korea와 같은 웹 탐색엔진을 출발점으로 Invisible Web 탐색을 시작하는 것이 한 방법일 것이다. Google Korea 뿐 만 아니라 Yahoo Korea, 네이버, 엠파스와 같은 국내 표준웹엔진의 Invisible Web 서비스에 대한 연구가 수행될 필요가 있겠다.

본 연구에서는 적합성순위정확률이 검색된 문헌의 순위와 문헌의 질(적합성)을 반영하는 척도로 사용될 수 있음을 보여주었지만, 랭킹 시스템의 성능척도로는 보다 종합적인 연구가 필요할 것 같다. 물리학 preprint 엔진이나 구직 엔진, 구직 정보, SARS 분포지도, 김치 담그기, 영화 리뷰, 피카소 그림을 요구한 질문에서는 여러 개의 적합문헌이 각기 다른 정보를 제공하기 때문에 적합문헌수는 의미가 있을 것이다. 그러나 학술논문의 본문, 재미교포수, 주가를 원하는 질문에서는 1등으로 출현하는 적합문헌 1건이면 이용자 측면에서는 충분할 것이다. 1건이 검색되고 그것이 적합인 경우 본 연구에서 사용된 순위정확률 34.78%(=10/46)보다는 표준정확률 100%가 더 타당할 수도 있다. 검색된 문헌의 적합성 척도와 검색된 문헌수, 적합문헌의 순위 등이 성능척도에 어떤 영향을 주는지와 이를 잘 반영할 수 있는 순위정확률에 대한 연구가 필요하겠다.

무엇보다도 본 연구에서는 시스템이나 이용자의 Invisible Web에 대한 이해가 부족한 국내 현실 때문에, 또한 인쇄문헌의 적합성 판정 기준과는 다른 웹사이트의 적합성 판정 기준을 연구하기 위해 연구자가 직접 적합성을 판정하는 것이 좋을 것으로 판단되어, 외국시스템을 대상으로, 연구자에 의한 질문작성과 탐색으로 설계된 실험실 실험이 수행되었다. 본 연구를 시작으로 웹사이트의 적합성판정에 영향을 미치는 요소들에 대한 다각적인 연구(질적연구나 이에 대한 검증연구 포함)와, 보다 다양한 탐색질문과 탐색시스템을 사용하여 일반 이용자들의 Invisible Web탐색에 대한 종합적인 현장실험연구가 수행되기를 바란다.

시스템 운영 측면에서 국내 DB를 안내하고 탐색을 대행하는 IW 메타탐색엔진의 출현과, 이런 IW 메타탐색엔진이 없는 상황에서 국내 표준탐색엔진이 보다 적극적으로 Invisible Web정보를 색인해 주기를 기대해 본다.

참 고 문 헌

노정순. 2000. 순위화시스템의 효과측정척도에 관한 연구. 『정보관리학회지』, 17(4): 67-81.

우유미, 정영미. 1998. 웹 검색엔진의 피드백 기능평가. 『제5회 한국정보관리학회 학술대회 논문집』, 69-72.

Albert, Reka, Hawoong Jeong, and Albert-Laszlo Barabasi. 1999. "Diameter of the World-Wide Web." *Nature*, 401(Sept): 130-131.

Bar-Ilan, Judit. 2002. "Methods for measuring search engine performance over time." *Journal of the American Society for Information Science and Technology*, 53(4): 308-319.

Bergman, Micheal K. 2001. "The Deep web: surfacing hidden value." *Journal of Electronic Publishing in University of Michgan*, July, 2001. [online]. [cited 2004.8.3]. <http://www.press.umich.edu/jep/07-01/bergman.html>. <http://www.brightplanet.com/technology/deepweb.asp>.

Brightplanet.com/deepcontent/index.asp. [cited 2004.8.3].

Cartwright, Jason. "Search engine indexing and dynamic sites." VIA net.Works. [online]. [cited 2004.8.20]. <http://www.webdev.vianetworks.co.uk/info/article-dynamic-seo.asp>.

Eliopoulos, D. and C. Gotlieb. 2003. "Evaluating web search results rankings." *Online*, 27(2): 42-46.

Gauch, S. and G. Wang. 1996. "Information fusion with Profusion." Webnet 96 Conference, San Francisco, CA, OCtover 15-19, 1996. [online]. [cited 2004.7.13]. <http://www.ittc.ku.edu/publications/documents/Gauch1996_WebNet96.pdf>.

Hawking, D., N. Draswell, and K. Griffiths. 2001. "Measuring search engine quality." *Information Retrieval*, 4(1): 33-59.

Leighton, H. V. and J. Srivastava. 1999. "First 20 precision among world wide web search services." *Journal of the American Society for Information Science and Technology*, 50(10): 870-881.

OCLC, Web Characterization Project 의 Statistics. [online]. [cited 2004.8.3]. <http://wcp.oclc.org/stats.html>.

Oppenheim, C., A. Morris, C. Mcknight & S. Lowley. 2000. "The Evaluation of WWW search engines." *Journal of Documentation*, 56(2): 190-211.

Search Engine Showdown. "Search engine statistics: Dead links report." [online]. [cited 2004.7.12]. <http://searchengineshowdown.com/stats/dead.shtml>.

Sherman, Chris and Gary Price. 2001. *The Invisible Web: uncovering sources search engines can't see.* CyberAge Book, 2001.

Sherman, Chris and Gary Price. 2003. "The Invisible Web: uncovering sources search engines can't see." *Library Trends*, 52(2): 282-298.

Sullivan, D. 2002. "Neilsen//NetRatings: searching engine ratings." [online]. [cited 2004.8.2]. <http://www.searchenginewatch.com/reports/netratings.html>.

Thelwall, M. 2002. "In Praise of Google: finding law joural Web sites." *Online Information Review*, 26(4): 271-272.

Vaughan, Liwen. 2004. "New measurements for search engine evaluation proposed and tested." *Information Processing and Management*, 40(2): 677-691.