

# 인터넷 데이터베이스와 온라인 데이터베이스의 특징 및 검색효율 비교

안태경/ 대외경제정책연구원 정보자료팀장  
김현희/ 명지대학교 문헌정보학과 정보통신교육연구센터

한국정보검색위원회에서는 위원간의 의욕 고취와 새로운 검색 및 데이터베이스 관련기술 보급을 위해 매월 연구발표회를 개최하고 있다. 본 코너에서는 주제논문을 게재함으로써 정보 검색과 관련된 정보를 제공하기 위해 마련된 것이다. <편집자>

## 1. 서론

### 1.1 연구의 목적

인터넷은 보급 후 겨우 수년에 지나지 않았는데 상용데이터베이스에 없는 정보와 상용데이터베이스와 거의 같은 정보를 무료로 입수할 수 있는 경우도 있으며 새로운 정보원으로서 중요성을 더해가고 있다. 온라인 정보원의 선택이 넓어짐에 따라 전문 검색사들은 어떤 때 무슨 온라인 정보원을 이용해야 하는지에 관한 안내가 필요하게 되었다. 무엇이 가장 좋은 것인가? 어떤 온라인 정보원이 가장 최신 정보를 보유하고 있는가? 온라인 정보원 모두가 훌륭한 정보원으로 이용되는가? 각각의 비용은 얼마인가?

좀 더 효율적인 데이터베이스 탐색을 위해서는 각 정보원의 특성을 이해하는 것이 무엇보다도 중요하다. 따라서 본 연구에서는 먼저 인터넷 데이터베이스와 상용 데이터베이스의 특성을 기존 연구 결과들에서 살펴본 후 두 탐색의 검색 효율성을 측정하기 위해서 도서관 이용자에게 의뢰 받은 정보 질문들을 이용하여 실험을 한 후 두 시스템의 탐색 결과를 분석하여 검색소요시간, 최종검색 건수, 적합성 등을 비교해 본다.

본 연구 결과는 초고속 인터넷상에서 운영될 데이터베이스 선택 톨의 지식베이스의 기초 자료로 이용될 계획이다.

### 1.2 연구의 방법 및 범위

본 연구는 주제분야를 경제, 경영분야로 제한한 후, 인터넷 데이터베이스와 상용 온라인 데이터베이스의 탐색 특징과 효율성을 문헌조사 및 검색실험을 통해 고찰해 본다. 자세한 내용은 다음과 같다.

첫째, 문헌조사를 통하여 인터넷과 상용 온라인 검색시스템의 특징 및 차이점을 살펴본다.

둘째, 두 검색시스템의 검색효율을 평가하기 위하여 검색 실험을 한 후 그 결과를 분석한다. 실험 내용을 살펴보면 먼저 경제학 관련 연구기관에서 연구자들이 일반적으로 요청하는 질문 중 6가지 실제 질문을 추출하였다. 그 다음 4명의 정보검색사에게 의뢰하여 두 검색시스템을 이용하여 탐색을 하게 한 후 검색소요시간, 검색도구, 검색식, 검색경로, 최종검색건수, 적합성을 비교 분석하였다. 마지막으로 질문자에게 최종 적합성을 평가받아 검색효율에 반영하였으며 그 결과를 종합하여 검색효율을 측정하였다.

## 2. 관련 연구

### 2.1 인터넷과 상용 데이터베이스 특징 비교

#### 1) 교전 연구

검색엔진에 의한 인터넷 탐색과 상용데이터베이스 탐색은 다음의 네가지 점에서 차이가 있다.

- 수록정보
- 색인파일
- 명령어체계
- 결과의 표시방법

#### (1) 수록정보

상용데이터베이스는 어떠한 데이터베이스라 하더라도 수록하고 있는 정보의 범위는 한정되어 있으며 균등한 성질의 정보가 시계열로 축적되어 있는 특징이 있다.

이에 반해 인터넷 정보는 내용, 질, 형식 등이 다양하고, 검색엔진은 이러한 이질적인 정보를 그대로 수용하고 있고 검색엔진별로 수록 정보의 범위에 차이가 있다. 따라서, 특정 검색엔진을 사용하여 탐색한 경우 만약 원하는 정보가 검색되지 않았다면 그 이유가 탐색방법에 있는지 검색엔진에 수록되어 있지 않기 때문인지, 아니면 그와 같은 정보가 인터넷에 존재하지 않기 때문인지 확실하게 판단할 수가 없다. 또한 인터넷에서는 특정 정보를 망라적으로 수집하고 싶은 경우라 하더라도 정보원을 한정하지 못하기 때문에 잡음이 많은 탐색의 망라성으로 인해 극단적으로 압축시킨 탐색을 하지 않을 수밖에 없다.

검색효율에는 인터넷의 정보 모두를 대상으로 하는 것이 아니고 특정의 분야에 한정된 범위를 탐색하는 편이 훨씬 효율적이다. 디렉토리 서비스와 디렉토리형 검색엔진의 분류는 어떤 의미에서 그러한 요구에 대한 해결의 하나의 시도라고 생각할 수 있다.<sup>1)</sup>

또한 최근에는 GOVBOT(미국의 정부계열 사이트들 망라한 검색엔진) 등 특정분야에 특화된 전문검색엔진이 증가하고 있는데 이와 같은 전문 검색엔진은 정보원을 제한하는 또 다른 방법이라 할 수 있다.

## (2) 색인 파일

상용 데이터베이스는 각 레코드, 레코드로부터 작성된 탐색용 색인파일, 결과 표시파일, 전부가 호스트내에 존재하여 레코드, 색인파일, 표시파일의 대응관계가 명확하다.

검색엔진에서 색인파일과 표시파일은 html 문서에서 태그를 나타낸 텍스트에서 만들어지는데 텍스트중에는 html의 코멘트문과 META태그 중의 문자도 포함된다. 색인파일과 표시파일은 그 페이지의 URL과 대응된다. 검색엔진에는 상용데이터베이스의 레코드에 해당하는 부분은 엔진내에 갖고 있지 않으며 URL에서 링크를 따라 찾게 된다.

색인파일은 탐색의 정확도를 결정하는 중요한 부분인데 로보틱 엔진에는 페이지 1문으로부터 작성되는 경우 및 페이지의 맨 앞에서 문자수를 한정해서 작성되는 경우 등이 있다.

상용데이터베이스의 색인은 키워드의 절단 방식이라는 점에서 데이터베이스마다 차이가 있는데 원칙적으로 명사의 단어 단위에서 절단한다. 색인은 레코드 가운데의 단어 마다에 그 단어의 출현빈도수, 그 단어가 레코드 중 몇번째에 있는가의 위치정보 등을 기록하여 인버티드파일(치환 파일)로 작성된다. 인버티드파일은 타이틀, 저자명 등 필드로 구성되어 있다. 상용데이터베이스에서 탐색은 입력된 키워드와 인버티드필드 중의 단어와 대조되는 형태로 행하여 진다.

한편 검색엔진도 똑같이 색인파일이 만들어진다. 그러나 색인 방법에서는 반드시 명사의 단어 단위에서 만들어진다고만 할 수는 없다. 특히, 검색엔진에서는 문장을 품사단위에서 분할하여 형태소해석의 기술과 문장을 문자단위에서 분할하여 1문자, 2문자단위의 색인파일을 작성하는 등 자연언어처리의 연구성과를 채택한 것이 많으며 기존 상용데이터베이스에서 사용하지 않는 색인

파일을 만들고 있는 경우가 있다. 그 결과, 예로는 [아름답다]라는 형용사와 [아름답게된다]라는 동사에서의 탐색과 [최고의 문명]은 [최고][의][문명]과 같이 색인되고 동시에 탐색하면 입력문자 [최고의 문명]도 입력문자처리 프로그램에서 [최고][의][문명]이라고 한번 분해된 다음 색인파일과의 대조를 하게된다.

형용사와 부사, 동사 등에서 탐색이 보다 정확한 검색결과를 갖고 올 것인지는 좀 더 검증이 되어야 하겠지만 적어도 기존의 상용데이터베이스하고는 전혀 다른 관점에서 탐색이 가능하게 되어 있다는 것을 이해하는 것이 필요하다.

## (3) 명령어체계

인터넷 검색엔진의 명령어체계는 대체로 상용데이터베이스와 같다. 대부분의 검색엔진에는 AND, OR, NOT 등의 불리언 연산과 인접연산이 가능하다. 검색엔진에서는 상용데이터베이스에서 일반적으로 있을 수 있는 표제, 저자의 필드를 지정하는 등의 제한탐색이 없다는 점이다. 상용데이터베이스에서는 필드를 지정하는 것으로 검색결과를 좁혀가거나 세밀한 탐색이 가능하다. 이와 유사한 기능으로 검색엔진 중에는 html의 표제 태그와 URL 사이트 등에서 제한 기능이 있는데 세밀하게 사용되는 기능이라고 할 수 없다.

## (4) 검색결과와 표시

검색엔진에는 수록정보의 내용과 형식이 다양하다는 것, 자연어(free word)로서만 탐색을 할 수 밖에 없다는 점에서, 적합성을 판단하는데 있어서도, 검색된 결과의 표시에도 각자 스스로 노력으로 얻어지는 경우가 많다.

그 하나로 적합도를 평정화하는 기능이 있다. 이러한 기능은 상용데이터베이스인 DIALOG에서 일부 사용되어지고 있는 예외는 볼 수 있지만 모든 종류의 정보를 수록하고 있는 검색엔진에는 검색결과를 판단하는 시간을 줄여나가는 기능으로서 많은 엔진에서 채용되고 있다.

평점을 부여 하는 것은 키워드의 출현빈도와 문서 전체에 대한 키워드의 수, 키워드가 문서 중에 어디에 있는가 등에 가중치를 주어 그 데이터

를 통계적으로 처리하여 산출하고 있다고 알려져 있다.

또한 표시내용에는 검색문서의 크기도 기재되어 있을 수 있다. 평점과 문서 크기의 표시는 검색결과 중 적합도가 높은 것 만을 본다던지 크기를 참조하여 필요한 자료를 선택하는데 이용할 수 있다. 이러한 기능은 상용데이터베이스에도 유용한 것만 아니라 정보내용도 형식이 각각이고 자연어 탐색만 할 수 있는 인터넷에서는 적극적으로 활용하여야할 서비스이다.

## 2) 펠드만(Feldman) 연구

펠드만(Feldman)은 인터넷 검색시스템과 전통적인 상용온라인 검색시스템을 비교하였다. 전문 검색사는 두 개의 온라인 서비스(DIALOG와 Dow Jones Interactive)와 적어도 하나 이상의 웹 검색엔진을 사용하여 최종이용자에게 실제 문제에 대한 대답을 하였다. 전문 검색사는 최종이용자에게 온라인서비스를 이용하여 찾은 대답과 웹 검색엔진으로 찾은 대답에 대한 결과의 적합성을 순위로 나타낼 것을 요청하였다. 결과는 최종이용자에게 적합한 자료, 정보를 찾는 데 소요된 시간 및 각 시스템에 강한 주제 분야가 조사되었다. 검색 결과의 자세한 내용은 다음과 같다.

### (1) 최종이용자에게 적합한 자료

두 개의 기존 탐색서비스(Dow Jones Interactive와 DIALOG)에서 탐색된 자료가 웹 검색엔진을 이용하여 탐색된 자료보다 더 높은 적합률을 나타냈다. 점수는 웹상의 문헌 1146에 대해 기존 온라인 검색시스템에 할당된 점수는 1400이었다. 물론 이 비율은 탐색마다 다양하였다. 그러나 일반적으로 DIALOG/Dow Jones Interactive 상에서의 탐색이 3과 5사이에 평가된 문헌에서 이루어지는 경향이 있음을 확인하였다. 웹 문헌들은 1에서 5사이에 발견된다. 우리의 결론은 높은 질의 정보면에서 볼 때 전통적인 온라인 데이터베이스에 더 의존할 수 있는 것처럼 보인다. 한가지 예외는 회사 관련 또는 널리 알려진 현재 사건에 관한 정보는 웹에서 더 잘 갱

1) World Wide Web의 인터넷의 정보탐색도구는 크게 나누어 (1)미러사이트 (2) 디렉토리서비스 (3)검색엔진의 3가지가 있다. 미러사이트는 특정분야의 홈페이지 URL을 수집 정리한 것으로 링크를 기반으로 하고 있다. 디렉토리서비스는 미러사이트의 규모를 넓혀 사이트 URL을 분류하여 조직화한 것이라고 할 수 있다. 검색엔진은 디렉토리형 검색엔진과 로보틱 검색엔진으로 구분된다. 디렉토리형 검색엔진과 디렉토리서비스의 차이는 명확하지는 않지만 구분하자면 일반적으로는 특정분야의 정보만을 수록하여 탐색은 분류된 링크에 따라 하는 것을 디렉토리 서비스, 수록범위를 한정하지 않고 다량의 정보를 수집하여 분류에 기반한 탐색을 기본으로 하면서 키워드검색도 가능한 것을 디렉토리형 검색엔진이라 부르고 있다.

신되어지는 것 같다.

## (2) 요구되는 탐색시간

일반적으로 웹상에서 탐색한 정보는 기존의 전문 온라인서비스에서 같은 정보를 탐색하는데 걸린 시간보다 두 배는 더 걸린다. 웹상에서의 전체 탐색시간은 20.5시간이다.

DIALOG/Dow Jones Interactive에서는 단지 9.9시간이다. 웹에서는 최종이용자에게 프린팅하고 전송하기 위해 문헌을 재포맷하는데 걸린 시간은 포함되지 않았기 때문에 웹상에서의 전체 탐색시간은 더 늘어날 것이다.

## (3) 각 시스템에 강한 주제 분야

### ① 기존 온라인 서비스 이용

- 예측, 안정된 문헌, 연속간행물
- 역사, 역사적 사실
- 기업의 역사와 배경
- 시장보고서, 산업보고서
- 권위있는 최신 약학연구
- 리서치, 학술논문, 여러분야의 학문연구
- 재무산업 범위
- 높은 질적 정보
- 시간이 요인일 때 빠른 탐색

### ② 웹 이용

- 그림과 삽화
- 생산품 정보 (기업으로부터 직접)
- 소규모 또는 사기업에 관한 정보 (기업으로부터 직접)
- 최신 의학 통계
- 최근 저자에 의해 알려진 학술논문의 전문 (全文)
- 특정 사이트에 존재하는 정보
- 공학, 컴퓨터 특히 기술보고서와 생산품 정보

### ③ 기존 온라인 소스와 웹의 이용

- 표준
- 일반적인 주제에 대한 논문
- 인기있는 주제
- 조직
- 디렉토리 정보
- 리뷰/평가/방법 정보
- 정부 조례와 타 정부기관의 정보
- 경쟁지식

## 3) 추(Chu) 연구

추는 온라인 검색 서비스와 인터넷 검색 서비스의 차이점을 다음 세가지 측면에서 서술하였다.

### (1) 데이터베이스 생성과 구조

온라인 데이터베이스의 생성은 주로 문헌집합부터 디스크립터까지를 망라하여 인간의 작업에 의해 이루어지고 레코드와 필드로 구성된 계층적 구조를 가진다. 반면 인터넷 검색서비스는 자동화된 방법으로 구축되며 하이퍼텍스트나 하이퍼미디어 등의 네트워크 구조로 이루어진다.

### (2) 탐색 기능

기본탐색능력에 있어 불리인 탐색과 절단 탐색등의 기능은 양자 모두 제공하고 있으나 온라인 데이터베이스가 좀더 폭넓은 탐색방법을 제공하고 있다.

필드 탐색의 경우 온라인 데이터베이스는 비교적 정교한 필드 탐색(제한 탐색)을 허용하고 있는 반면 인터넷 데이터베이스에서는 극히 제한된 유사 필드 탐색(예, 타이틀 및 URL)만을 허용하고 있다.

집합의 처리 및 탐색 처리는 온라인 탐색에 있어서는 필수불가결한 탐색수단이지만 인터넷 상에서는 이러한 탐색을 위한 집합을 생성하고 저장하지 않으며 이를 위한 처리 메커니즘을 제공하는 것은 극히 드물다.

통제어휘 탐색의 경우 온라인 데이터베이스 서비스에서는 오래전부터 제공해온 기능이지만 인터넷 상에서는 몇몇 탐색 서비스에서 부분적이고 특수화된 분야에서 이를 제공하고 있다. 한 예를 들면 BUBL Link은 데이터베이스를 조직하기 위해서 DDC 분류법과 LCSH(Library of Congress Subject Headings)을 이용하고 있다.

가중치 탐색은 핵심 디스크립터에 의해서 DIALOG상에서 이루어지고 있고, 인터넷상에서는 초창기에는 이 특징을 지원하지 않았지만 현재는 인터넷 이용자들이 이를 이용할 뿐만 아니라 일관된 인터넷 탐색 서비스를 위한 형태로 이를 제공하고 있다.

Mapblast와 같은 인터넷 탐색 서비스는 Fuzzy 탐색을 이용하여 오자로 기입한 질문에 대해서도 정확한 탐색 결과를 얻을 수 있도록 하고 있다. 다중 데이터베이스 탐색을 위해 DIALOG의 경우 60개의 데이터베이스를 이용

자가 탐색할 수 있고 Metafind나 Inference Find와 같은 메타검색엔진은 이러한 탐색을 위한 서비스를 제공하고 있다.

검색 수행은 정확성과 관련성에 의해 평가될 수 있다. 낮은 정확성으로 인터넷 서비스는 악명이 높는데 이는 단일한 질문에 대한 수천건의 검색결과가 검색되기 때문이다.

출력옵션에 있어 온라인 데이터베이스 서비스는 미리 지정된 형태나 사용자 지정 형태로 출력할 수 있지만 인터넷 서비스 상에서는 하이퍼링크와 같은 기본적인 정보만을 담고있는 간단한 포맷이나 요약이나 문헌의 형태 등을 포함한 좀더 세부적인 두 가지 형태만을 전형적인 출력형태로 제공하고 있다.

### (3) 이용자의 노력

이용자의 측면에서 두 서비스를 비교할 때 비용과 이용방법에서 차이를 들 수 있다. 비용에 있어 인터넷 서비스는 일단 접속만 하면 부가비용이 없고 이용자에게 친숙한 Hyper-graphical user interface를 이용하는 반면 온라인 서비스는 이용요금에 별도로 부가되며 별도의 교육이 없는 이용과 검색에 있어 어려움이 있다. 적합성 피드백을 위한 수단으로 온라인 검색에 있어서는 이전 탐색에서 추출한 새로운 용어로 탐색 전략을 수정하는 방법을 이용하고 있지만 인터넷 상에서는 Excite의 'More like this' 기능처럼 별도의 이용자의 노력없이 검색 엔진내에서 자체적으로 제공하고 있다.

번역의 경우 온라인 데이터베이스 서비스에서는 이루어지고 있지 않지만 인터넷 서비스에서는 영어를 기본으로하여 몇몇 언어에 대한 번역을 제공하여 이용자의 편의를 돕고 있다. 중복 방지 기능은 온라인상에서는 개개 및 다중 데이터베이스 탐색에서 지원을 하고 있지만 인터넷상에서는 몇몇 선택된 메타툴(예, Inference Find)에서만 이를 지원하고 있다. 문헌 중복은 인터넷 서비스 상에서 심각한 문제를 불러일으키고 있다. 즉 인터넷 이용자들은 탐색 결과에서 비적합 문헌외에 중복된 자료를 걸러내기 위해서 많은 시간을 낭비하고 있다.

## 2.2 검색 효율 측정 방법

### 1) 정보검색시스템

현재까지 수많은 연구에서 크랜필드형

(Cranfield Model) 탐색실험에는 많은 문제가 수없이 지적되어 왔다. 재현율과 정도율이 초점을 맞추어 그 문제점을 아래에 나타내었다.

(1) 재현율산출에는 모든 적합문헌의 정보가 반영되어야 할 필요가 있다. 이에 대해 현실적으로 정보탐색에서는 재현율의 산출이 매우 곤란하다.

(2) 재현율과 정도율은 탐색되었는지, 어떤지를 문제로 하기 때문에 순위를 부여하여 출력을 적절하게 평가하는 것이 가능하지 않다.

(3) 1950년대의 배치방식의 실험환경에 있어서 척도이며, 인터랙티브 정보검색시스템에의 응용가능성에 대해서는 의문이 남는다.

(4) 평가에는 양쪽의 필요에 있어서, 두 개의 척도는 반비례 관계에 있어서 시스템의 비교를 곤란하게 하고 있다.

## 2) 인터넷 검색엔진

기존의 검색엔진에 대한 연구에서는 많은 경우 탐색 성능에 대한 평가라기 보다는 인터페이스의 사용에 편리한 탐색기능, 탐색속도, 수록수, 갱신빈도 등의 관점에서 평가를 수행하여 왔다. 그러나 검색엔진이 정보검색시스템이고, 정보검색시스템이 이용자가 필요한 정보를 취합선택하는 것을 목적으로 한다면, 어떠한 형태로도 그 성능을 측정하는 것이 먼저이며 평가에는 당연히 탐색성능이 포함되어야 한다고 생각된다.

기존 크래필드형 탐색실험에서 탐색성능은 재현율과 정도율에서 측정되어왔는데 검색엔진의 경우에는 정도율은 산출되는데 재현율은 산출되지 않는다. 인터넷상의 자원은 팽창하고있고 재현율을 산출하기 위하여 모든 정보를 찾아본다는 것은 불가능하기 때문이다. 실제로 검색엔진을 평가하는 연구에서는 재현율은 방치되어 있거나 생략되어있다. 그 결과로서 검색엔진에 관한 평가연구에는 정도율만이 사용되어지고 있는데 원래 재현율과 정도율은 양쪽의 관계로부터 평가를 수행하는 것이며 정도율만으로 평가로는 충분하다고 할 수 없다.

또한 검색엔진의 대부분은 순위부여 출력을 하고 있는데 앞에서 언급되어진 재현율, 정도율

은 순위부여 출력에 대응 할 수 없는 척도이고 어느 시스템이 고순위에 적합한 웹페이지를 많이 출력한다 하였다 하더라도 그것이 적절한 평가로 반영될 수 없는 문제점이 있다.

다시 말하면 기존의 검색엔진의 평가에 관한 연구에서는 탐색성능 이외의 척도에서 평가가 수행되고있는 경우가 대부분이며, 더욱이 탐색성능에 관한 평가가 수행되는 경우에도 정도율만을 사용하고 있기 때문에 충분한 평가가 되고 있다고 말할 수 없다.

## 3. 검색성능실험

### 3.1 실험의 대상 탐색도구

본 연구에서 실험대상도구인 인터넷 검색시스템은 검색자가 자신이 주로 사용하는 탐색도구로 특정하게 제한을 두지 않았다. 이는 검색자들이 주로 사용하고 있는 탐색도구가 어떠한 것인지, 질문에 어떠한 효율을 나타내는지 조사하기 위해서이다.

상용온라인 데이터베이스로는 ProQuest Direct(PQD), EBSCOhost, OCLC First Search 3개의 데이터베이스로 한정하였는데 이는 이 세 데이터베이스가 경제, 경영학 분야와 학술전반을 포괄하는 데이터베이스이기 때문이다.

### 3.2 실험질문

실험질문은 실제 경제학 관련 연구기관 연구자들이 요구한 질문을 대상으로 선정하였으며 실험에 사용된 질문은 다음과 같다.

Q1) 한국의 외환위기 직전(97년 11월)부터 2개월간 달러대 원화 환율의 일자별 데이터를 구하고자 합니다.

Q2) OECD 경제전망 자료 중 1999년 한국을 포함한 아시아 경제전망 지표를 구하고자 합니다.

Q3) 김대중 대통령이 Foreign Affairs에 기고

한(95년경) 자료의 원문을 입수하고자 합니다.

Q4) 세계의 think-tank 목록을 입수하고자 합니다.(국가별 기관리스트와 연구기능 등)

Q5) 최근 아시아 금융동향과 관련 JP Morgan 발행 Asian Financial Markets 최신호(3/4, 4/4분기)자료를 입수하고자 합니다.

Q6) 1998년 이후 발표된 해외직접투자관련 외국학술논문의 서지사항을 조사하고자 합니다.

### 3.3 실험방법 및 절차

각 실험질문을 4명의 정보검색사에게 의뢰하여 6개의 질문을 인터넷 데이터베이스와 상용 온라인 데이터베이스에서 검색하도록 하였다. 여기에서 질문에 대한 적합성은 일차적으로 객관성을 유지하기 위하여 각자가 판단하도록 하였으며 최종평가는 질문을 요청한 이용자에게 최종평가를 하도록 하였다. 성능평가에 사용된 기준은 탐색 시간, 적합성<sup>2)</sup> 등을 사용하였다. 이 중 적합성1은 검색사가 판정한 적합성이고 적합성2는 질문자가 최종 판정한 적합성이다.

## 4. 실험결과 및 분석

### 4.1 검색결과

#### 1) 질문 1

질문1은 fact 데이터를 구하는 질문으로 인터넷 검색에서는 검색자 중 2명이 로봇형 검색엔진을 사용하였고 S1검색자의 경우에는 심마니의 디렉토리를 이용하여 질문을 탐색하였다. S2 검색자의 경우에는 이미 북마크를 해두고 있었다. 상용 온라인 검색에서는 적합성1의 경우 S4 검색자가 검색한 문헌의 전문을 결과로 제시하였지만 시계열 데이터가 일자별 로 정리되어 있지 않아 적합성2에서 비교적 적합(70)으로 판정되었다.

2) 적합성 : 검색결과에 대한 적합성으로 다음과 같은 구간척도를 사용하였다.

0 : 미해결	1-19 : 매우비적합	20-39 : 비적합	40-59 : 보통
60-79 : 비교적 적합	80-99 : 적합	100 : 매우적합	

〈표 1〉 질문1의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	15	14	15	12	5	4	4	4
탐색도구	심마니	북마크	Altav	Altav	PQD	EBhost	PQD	PQD
최종검색건수	1	1	1	1	59	25	3	37
적합성1	100	100	100	100	30	50	10	100
적합성2	100	100	100	100	10	50	10	70

〈표 2〉 질문2의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	14	30	12	12	5	4	4	4
탐색도구	북마크	디렉토리	북마크	Altav	EBhost	PQD	PQD	PQD
최종검색건수	1	1	1	1	10	9	5	20
적합성1	100	100	100	100	50	100	10	35
적합성2	100	100	100	100	50	80	80	0

〈표 3〉 질문3의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	20	30	10	12	10	4	5	5
탐색도구	Alta	Alta	Alta	Altav	EBhost	PQD	PQD	PQD
최종검색건수	1	1	1	1	55	1	1	1
적합성1	100	50	100	100	0	100	100	100
적합성2	50	50	50	50	0	100	100	100

〈표 4〉 질문4의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	30	15	12	14	4	4	5	3
탐색도구	Alta	Alta	Alta	Altav	PQD	PQD	PQD	EBhost
최종검색건수	1	1	1	1	50	9	27	1
적합성1	100	100	100	100	10	0	0	50
적합성2	100	100	100	100	0	0	0	0

〈표 5〉 질문5의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	15	15	12	14	4	5	5	3
탐색도구	Alta	Alta	Alta	Altav	EBhost	PQD	PQD	EBhost
최종검색건수	2	2	2	2	3	14	27	3
적합성1	100	100	100	100	0	0	0	0
적합성2	100	100	100	100	0	0	0	0

〈표 6〉 질문6의 검색결과

데이터베이스	인 터 넷				상용 온라인			
	S1	S2	S3	S4	S1	S2	S3	S4
검색자	S1	S2	S3	S4	S1	S2	S3	S4
검색시간(분)	12	10	12	12	4	5	6	3
탐색도구	Altav	Altav	Altav	North L	EBhost	EBhost	EBhost	PQD
최종검색건수	30	30	30	30	133	177	57	50
적합성1	100	100	100	0	50	50	70	100
적합성2	100	100	100	0	70	70	60	100

## 2) 질문2

질문2는 연구기관자료로서 전문(full-text)를 압수하고자 하는 질문으로 인터넷에서는 완전히

전문을 입수할 수 있었다. 또한 탐색도구는 검색자들이 북마크와 자주 사용하는 디렉토리를 사용하였는데 이는 관련 분야에 대한 자료를 수시로 조사하기 때문이었다. 상용온라인 검색에서는

S2, S3 검색자가 적합성1을 매우만족(100)으로 판정하였지만 최종이용자는 문헌의 최신성 결여로 적합(80)으로 판정하였다.

## 3) 질문 3

질문3은 학술지의 기사전문(full-text)을 입수하고자 하는 질문으로서 인터넷 검색에서는 검색자 모두가 "Foreign Affairs"의 홈페이지에서 2차자료를 검색한 뒤 2차자료 또는 프린트된 원문을 제공하는 것으로 적합성1을 판정하였다. 상용 데이터베이스 검색에서는 탐색도구를 PQD를 사용한 경우 모두 매우 적합한 것으로 판정되었다.

## 4) 질문 4

질문4는 디렉토리 유형의 정보원을 조사하는 질문으로 인터넷의 검색에서는 매우 적합한 것으로 판명되었다. 상용 온라인 검색에서는 서지데이터만을 얻을 수 있었다.

## 5) 질문 5

질문 5는 공식적으로 판매, 배포되지 않는 기업의 동향분석 자료의 입수를 요청하는 질문으로 인터넷 검색엔진을 사용하여 검색한 미러사이트에서 매우 적합한 결과값을 얻을 수 있었다.

## 6) 질문 6

질문6은 외국학술지의 서지데이터를 입수하고자 하는 질문으로 인터넷 검색에서는 검색이 제대로 이루어지지 않았다. 이는 인터넷의 특성상 별도로 학술지를 구분하고 있지 않기 때문이다. 상용 온라인의 경우 비교적 높은 적합도를 나타내었고 이외에도 일부 전문(full-text)까지도 검색 할 수 있었다.

## 4.2 검색효율 측정

검색효율의 측정은 전통적인 방법으로는 검색 결과의 정도율과 재현율로서 효율을 측정하는 것이 바람직하다. 그러나 현재 인터넷의 경우 재현율의 측정이 불가능하고 동시에 상용온라인 데이터베이스에 적용하는 것도 적절하지 않다고 판단되어 본 연구에서는 두 검색시스템을 검색시간, 적합도를 기준으로 측정하고자 한다.

## 1) 검색시간

현재 인터넷은 정보검색에 많은 시간을 소요하고 있다. 실제로 상용 온라인 데이터베이스와 두 시스템 간에 어느 정도 검색시간이 소요되는지 <그림 1>에 나타내었다. 6개의 질문 모두를 검색하는데 소요된 시간은 인터넷은 평균 14.67분, 상용온라인은 평균 4.75분으로 인터넷은 상용온라인 보다 약 3배의 검색시간을 소요하고 있다. 정보검색속도의 문제는 검색성능과도 관련이 있지만 비용의 문제가 제기 되고 있다.

## 2) 적합성 측정

적합성 측정은 두 가지로 <그림 2>의 정보검색사가 판정한 적합성과 <그림 3>의 정보이용자가 판정한 적합성으로 측정을 하였다. <그림 2>.<그림 3>에서 나타난 바와 같이 두 적합성 모두 유사한 관계를 나타내고 있다. 질문3(Q3)의 경우 상용데이터베이스에서 차이를 나타내고 있는 것은 정보검색사(S1)가 부적합한 데이터베이스를 선정하였기 때문이다.

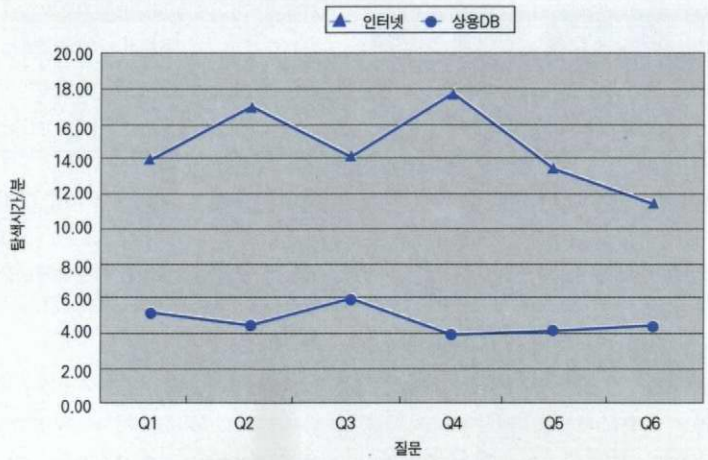
또한 데이터베이스 정보의 유형별 검색측정을 살펴보면 질문1(Q1)은 수치데이터를 요구하는 질문으로 인터넷의 경우 적합성은 모두 매우적합(100)으로 나타났고, 상용온라인의 경우 보통(59)이하로 나타나 인터넷 데이터베이스에 강한 것으로 나타났다.

질문2(Q2)는 연구기관자료로서 전문(full-text)을 입수하고자 하는 질문으로 인터넷에서는 적합성이 매우적합(100)으로 나타났으며, 상용온라인의 경우 검색자 적합성은 비교적 적합(75), 질문자 적합성은 보통(52)으로 그 이유는 최신성의 결여로 들 수 있다.

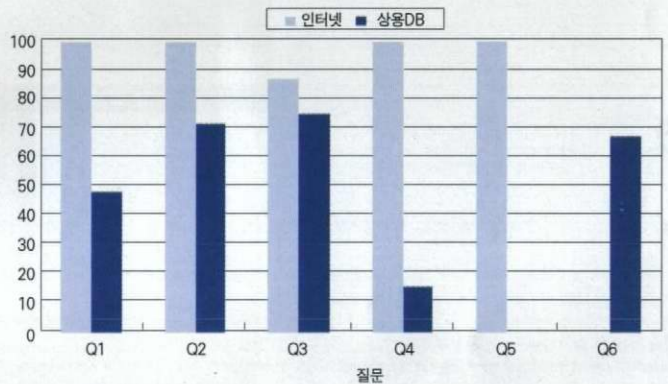
질문3(Q3)은 학술지의 기사전문(full-text)을 입수하고자 하는 질문으로 인터넷에서는 정보검색사는 적합(89), 질문자는 보통(50)으로 나타났으며 상용온라인에서는 비교적 적합(74)한 것으로 나타났다.

질문4(Q4)는 디렉토리유형의 정보원을 조사하는 질문으로 인터넷의 경우 모두 매우적합(100)을 나타내고 있었으며, 상용온라인에서는 매우 비적합(15)한 것으로 나타났다.

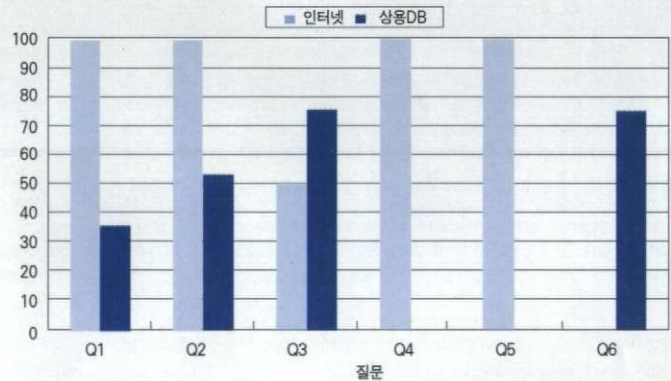
질문5(Q5)는 기업의 동향분석자료 요구 질문으로 인터넷에서는 모두 매우적합(100)을 나타냈고, 상용온라인에서는 적합성을 나타내지 않았다.



<그림 1> 인터넷과 상용온라인 데이터베이스 검색시간



<그림 2> 정보검색사 적합성 판정



<그림 3> 질문자 적합성 판정

질문6(Q6)은 외국학술지의 서지데이터를 입수하고자 하는 질문으로 인터넷에서는 검색이 이루어지지 않았고 상용온라인에서는 비교적 적합(68) 이상으로 서지데이터 검색에서는 상용온라인 데이터베이스가 강한 것으로 적합성이 나타났다.

이상과 같이 데이터베이스의 유형별 정보검색

의 적합성을 측정하여 보면 인터넷 데이터베이스에는 수치데이터, 연구기관자료, 디렉토리형 정보원, 기업의 동향분석자료 등에 강한 것으로 나타났다. 상용온라인 데이터베이스에는 학술지의 전문(full-text), 학술지의 서지데이터에 우수한 것으로 적합성이 평가되었다.

## 5. 결론 및 제언

인터넷 상에서 이용할 수 있는 정보가 증가함에 따라 보급후 겨우 수년에 지나지 않았는데 상용데이터베이스에 없는 정보와 상용데이터베이스와 거의 같은 정보를 무료로 입수할 수 있는 경우도 있으며 새로운 정보원으로서 중요성을 더해가고 있다.


본 연구에서는 좀 더 효율적인 데이터베이스 탐색을 위해서 각 정보원의 특성을 이해하고, 인터넷 데이터베이스와 상용 데이터베이스의 특성을 이해하기 위하여 기존 연구 결과들에서 살펴 보았다. 또한 두 탐색의 검색 효율성을 측정하기

위해서 이용자에게 의뢰받은 질문들을 이용하여 실험을 한 후 두 시스템의 탐색 결과를 분석하여 검색소요시간, 적합성을 비교해 보았다.

본 연구의 실험결과, 검색소요시간은 4명의 정보검색자가 6개의 질문 모두를 검색하는데 소요된 시간은 인터넷 평균 14.67분, 상용온라인 평균 4.75분으로 인터넷은 상용온라인 보다 약 3배의 검색시간이 소요되었다. 또한 정보검색속도의 문제는 검색성능과도 관련이 있지만 비용의 문제가 제기 되고 있다.

적합성 측정에서는 검색자와 질문자 적합성을 측정하였는데 모두 유사한 평가를 하고 있었다. 데이터베이스의 유형별 정보검색의 적합성 측정은 인터넷 데이터베이스에는 수치데이터, 연구기

관자료, 디렉토리형 정보원, 기업의 동향분석자료 등에 강한 것으로 나타났다. 상용온라인 데이터베이스는 학술지의 전문(full-text), 학술지의 서지데이터에서 우수한 것으로 적합성이 평가되었다.

현재 인터넷 데이터베이스와 상용온라인 데이터베이스의 검색효율 측정에는 많은 문제점이 제기 되고 있으며 연구가 수행되고 있다. 좀더 세계적인 정보검색 효율을 평가하기 위해서는 먼저 인터넷 검색엔진에 대한 검색효율 실험 모형이 개발되어야 하며 아울러 인터넷의 검색비용 측정에 대한 기준이 마련되어야 한다. 

## 참고 문헌

- 橋田昌 . 商用DB 検索システムと比較したの検索エンジン. 専門図書館 No.171, 1998-III pp.61-67
- 安形 輝 等著 . 인터넷의 서치엔진의 평가尺度, 1998 情報學シンポジウム講演 論文集, 東京: 情報處理學會, 1998. pp.137-144
- . 인터넷의 서치엔진의 평가尺度: ESL(Expected Search Length)을使った檢索實驗. Library and Information Science, No. 37 1997. pp. 1-11
- 이명희. 네트워크 데이터베이스에서의 주제별 디렉토리화 키워드 검색엔진의 탐색효율에 관한 탐색적 연구. 한국문헌정보학회지 31(2) 1997. pp.176-197
- 이은주. WWW탐색도구의 탐색성능에 관한 실험적 연구. 연세대학교 석사학위논문, 1997.
- 日本データベース協會. 最新オンライン情報源活用法: 인터넷에서부터データベースまで. 東京: 日外アソシエーツ, 1998.
- Anagnostelis, Betsy. Evaluation criteria for different versions of the same database: a comparison of Medline services available via the World Wide Web. In Online information 97 21st International Online Information Meeting Proceedings London, 9-11 December 1997, Learned Information Europe Ltd., pp.165-173
- Arnold, Stephen E. and Arnold Erik S. Search engines: lost and found in cyberspace. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 37-53
- Chu, Heting. Internet search services vs. online database service. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 69-75
- Feldman, Susan. The internet search-off: results and ruminations. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 125-130
- . The internet search-off, Searcher, Vol.6 No.6 June 1998. pp.28-35
- Harter, Stephen P. and Hert Carol A. Evaluation of information retrieval systems: approaches, issues, and methods. Annual Review of Information Science and Technology, Vol. 32, 1997. pp. 3-98
- Jasco, Peter. Mapping algorithms to translate natural language questions into search queries for web database. In Proceedings of the 18th National Online Meeting New York, May 13-15, 1997. Information Today Inc., pp. 189-199
- Jacobs, Paul S. Browsing vs. surfing: the next generation of search and retrieval. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 169-175
- Liddy, Elizabeth D. Beyond retrieval. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 229-233
- Platt, Nina. Search engines for intranets. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 265-273
- Prestamo, Anne Tubbs. Development of web-based tutorials for online database. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 275-287
- Rieb, Soo Young. User-computer-librarian interaction in end-user online searching. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 229-233
- Spink, Amanda and Bateman Judy. User's searching behavior on the exite web search engine. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 375-386
- Su, Louise T. Developing a comprehensive and systematic model of user evaluation of web-based search engines. In Proceedings of the 18th National Online Meeting New York, May 13-15, 1997. Information Today Inc., pp. 335-344
- Tragert, Joseph. CD-ROM vs. online vs. internet: continuous change. In Proceedings of the 18th National Online Meeting New York, May 13-15, 1997. Information Today Inc., pp. 345-356
- Weiner, Michael L. and Rusch, Peter F. New searching technologies and interfaces. In Proceedings of the 18th National Online Meeting New York, May 13-15, 1997. Information Today Inc., pp. 369-373
- Williams, Martha E. Highlights of the online database industry and the internet 1998. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 1-6
- Xie, Hong(Iris). Access to online database: an exploration of user's experiences with web interfaces. In Proceedings of the 19th National Online Meeting New York, May 12-14, 1998. Information Today Inc., pp. 481-493