인터넷상의 메타탐색엔진의 검색효율성 비교연구

김 성 희 *

〈목 차〉

Ⅰ. 서 론

최근들어 컴퓨터 성능의 향상과 통신기술의 발달로 인터넷 이용자가 급속히 증가하고 있다. 인터넷이란 세계적인 정보자원(information sources)들의 group, 즉 전세계적으로 퍼져있는 네트워크의 집합이다. 인터넷을 사용하는 이유는 첫째, 사람들끼리 서로 대화하도록 하기위해서이다. 인터넷은 지구상의 수백만의 사람들이 서로 전자우편을 보내고 받는다. 둘째, 자원들을 서로 공유하기 위해서이다. 인터넷은 세계 곳곳에 흩어져 있는 자원들을 네트워크를 통해 서로 공유할 수 있게 한다. 최근들어서는, 인터넷에 존재하는 일반 text 형태의 문서, 그림, 음성, 오디오등의 각종 데이터를 URL(Uniform Resource Locater: 인터넷에 있는 임의의 정보의 주소를 지정하는 방식) 를 이용하여 하나의 문서 형태로 통합적으로 제공한다. 대표적인 서비스로는 WWW(World Wide Web)로 Web이라 부르기도 하고, W3로 표기하기도 한다. WWW은 hypertext에 기반한 인터넷 서비스로 hypertext는 다른 데이터와의 연결관계를 가지고 있는 데이터라 할 수 있으며 hypertext문서는 데이터와 다른 문서와 연결 내용이 들어 있다. 이와같이 WWW는 hypertext 문서의 집합으로 인터넷 상에 모든정보(거기에다 컴퓨터에 추가하고 싶은 특정한 정보)를 규합하려는 하나의 시도로써, 분산 멀티미디어 하이퍼텍스트 시스템(distributed multimedia hypertext system)이라고도 한다. 여기서, 분산은 WWW상의 정보는 어느 한곳에 집중되어 있는 것이 아니라 전세계에 흩어져 존재한다는 의미이며, 멀티미디어라함은 WWW상에서 표현되는 정보는 일반 문자 뿐만 아니라 그래픽, 음성, 동화상등 사람이 표현할 수 있는 다양한 표현 방법을 사용하는 것을 말하며, 하이퍼텍스트는 문서들사이에 서로 연결되어 있음을 의미한다. WWW의 사용이유는 하이퍼텍스트 기사읽기와 인터넷

* 중앙대학교

자원에 접속하는 것이라고 할 수 있다. 첫째, 하이퍼 텍스트가 보통 일반 문서와 다른 점은 문서들이 연결되어 있다는 것이다. 문제는 하이퍼텍스트 안에 연결 관계를 설정해 놓은 것은 시간을 많이 소모한다는 점과 하이퍼텍스트 연결이 얼마나 적절한가 하는 점이다. 하이퍼텍스트 문서는 하이퍼텍스트 이용자의 생각과 그 문서에 연결시키는 사람과 얼마나 밀접한가에 따라 하이퍼텍스트가 유용할 수 도 있고 그렇지 않을 수도 있다는 것이다. 둘째, WWW를 이용하는 또 다른 이유는 WWW상에서 다른 다양한 인터넷 서비스에 접근할 수 있다는 것이다. 예를들면, WWW를 통해, telnet, gopher, ftp, E-mail과 같은 서비스에 접속을 할 수 있다는 것이다. 이와같이 WWW는 인터넷에서 제공하고 있는 여러 가지 서비스를 하나로 통합하여 사용할 수 있는 장점을 제공하므로 인터넷에서의 이용은 기하 급수적으로 증가하고 있다.

한편, 인터넷사용이 급증과 더불어, 인터넷 정보 또한 급증하면서 이용자들이 원하는 정보를 신속하고, 정확하게 제공할 수 있도록 도와주는 탐색엔진들이 제공되고 있다.

앞에서 언급했듯이 인터넷은 전세계적으로 퍼져있는 네트워크의 집합으로 인터넷상의 정보는 한 곳에 집중해서 존재하는 것이 아니라 전세계적으로 퍼져있다. 따라서, 필요한 정보가 인터넷상에서 어디에 있는지(know-where), 또 어떻게 검색할 수 있는지(how to get)를 모를경우 인터넷의 정보를 유용하게 활용할 수 없게 된다. 따라서, 전세계적으로 퍼져있는 인터넷정보를 신속하고 정확하게 검색할 수 있도록 도와주는 검색도구가 그동안 많이 발전해 왔다. WWW가 나오기 전에는 주로 텍스트환경의 문자형 검색도구가 개발되었다. 예로는 아키(archie), 고퍼(gopher), 베로니카(veronica), 후이즈(whois)등을 들 수 있다. 그 후 WWW가 등장하면서 Web을 통한 검색도구가 많이 개발되었는데 이를 탐색엔진(search engine)이라고 한다. 대표적인 WWW 탐색엔진으로는 야후(Yahoo), 알타비스타(Altavista), 라이코스(Lycos), 익사이트(excite), 인포시크(InfoSeek), 심마니, 까치네 등을 들 수 있다. 그러나 이런 많은 탐색엔진들은 각자, 데이터베이스 구축방법, 탐색기법, 검색대상, 출력내용등이 다르기 때문에, 검색결과가 다양한 것으로 나타났다.(Courtois et al., 1995). 또한 이들 탐색엔진들은 서로 다른 검색결과의 제공 이외에도 상이한 사용자 인터페이스로 인하여 사용자가 원하는 모든 정보의 검색은 어렵게 되었다. 이런 이유로 사용자가 쉽게 사용할 수 있는 단일화된 인터페이스를 제공하며 여러 탐색엔진의 검색 결과를 종합해줄 수 있는 메타 탐색엔진이 등장하게 되었다. 대표적인 메타 탐색엔진으로는 MetaCrawler, Savvy Search, All in One, Internet Sleuth등이 있으며, 국내에서는 미스 다찾니가 있다. 본 논문에서는 인터넷 탐색엔진의 정의 및 종류, 메타탐색엔진에 대해 살펴보고 메타탐색엔진중에서 Savvy Search와 MetaCrawler에 대한 특성 및 검색 효율성에 대해 살펴보고자 한다.

Ⅱ. 탐색엔진의 정의 및 종류

인터넷 상에 정보의 양과 종류, 정보제공자의 수가 급격히 증가하면서 인터넷 사용자들은 정보검색도구로서 탐색엔진(search engine)을 많이 이용하고 있다. 특히, 최근들어 하이퍼텍스트형태의 정보 조직 및 브라우징(browsing)이 가능한 WWW(World Wide Web)를 이용하여 정보를 제공하는 사이트가 기하급수적으로 늘어나면서 WWW를 이용해서 인터네트 상에 널리 퍼져있는 정보를 사용자에게 찾을 수 있도록 도와주는 도구인 탐색엔진들이 개발되고 있는 추세이다. 현재 전 세계적으로 많은 탐색엔진들이 동작하고 있고, 각종 정보에 대한 데이터베이스를 구축하고 있고 계속적으로 갱신함으로써 사용자에게 최신의 정보를 제공하고 있다. 탐색엔진은 로봇 에이전트(Robot agents)라는 웹을 순회(traverse)하는 프로그램에 의해 정보를 수집한다. 로봇 에이전트는 지정한 URL을 순회하며 각 홈페이지들의 정보를 수집하고 수집된 정보에 대해서 탐색엔진은 사용자가 편리한 방법으로 검색할 수 있는 인택스를 제공한다. 이상에서와 같이 탐색엔진이란 특정한 정보를 체계적으로 분류해놓고, 정보를 신속하고 정확하게 찾을 수 있도록 도와주는 도구이다. 현재 인터넷에는 약 300여개의 탐색엔진이 서비스되고 있다. 이들 탐색엔진들은 각자 특성이 있고, 사용법도 모두 다르기 때문에 이들 탐색엔진을 사용하기 위해서는 각 엔진들의 특성과 사용법을 알아야 한다.

한편, 탐색엔진을 분류하는 공통된 기준은 없다. 일반적으로 탐색엔진은 정보의 보유측면, 동작형태, 정보구축범위등에 따라 다양하게 구분된다.

1. 정보보유측면에 따른 분류

로봇 에이전트를 통해 자료를 수집하거나 사용자가 정보를 등록하게 함으로써 자기 자신의 데이터베이스를 구축하고 있는경우와 다른 탐색엔진에서 보유하고 있는 데이터베이스를 이용하여 사용자에게 서비스를 하는 형태로 구분할 수 있다.

로봇 에이전트를 이용해 자료를 수집하는 탐색엔진의 경우는 자료를 수집하는 로봇 에이전트, 수집된 자료를 저장하는 데이터베이스, 그리고 사용자가 질의(query)를 했을 때 자료를 검색해주는 IRS(Information Retreival System)로 구성된다. 다음에 설명하게 될 주제별 탐색엔진과 키워드형 검색엔진이 이에 속한다. 한편, 다른 탐색엔진의 정보(데이타베이스)를 이용하는 탐색엔진은 전체적으로 질의를 각 탐색엔진에 보내는 부분과 검색된 결과에 대해서 통합해서 사용자에게 보여주는 부분으로 구성된다. 메타검색엔진이 여기에 포함된다.

2. 동작방식에 따른 분류

탐색엔진은 어떤 동작형태를 갖고 있느냐에 따라 주제별 탐색엔진, 키워드형 엔진으로 구분된다.

주제별 탐색엔진(subject-oriented searching engine)은 인터넷에 있는 정보를 사회, 문화, 예술, 스포츠, 정치등 큰 주제에 따라 분류해 놓은 목록을 제공하는 탐색엔진이다. 즉, 특정주제별로 각 페이지들을 분류하여 정리해 놓은 탐색엔진으로 예술, 정치, 경제, 스포츠등 각 분야별로 분류되어 있는 항목을 마우스로 클릭하여 그 분야의 세부 항목으로 들어가서 원하는 정보를 찾는 방식이라 볼 수 있다. 주제별 탐색엔진은 정보를 검색하기 위해 특별한 주제어나 중심어 등을 발견할 수 없을 때나 알기 힘들 때 사용하면 효율적이다. 대표적인 주제별탐색엔진으로는 야후(Yahoo), 라이코스(Lycos), WWW Virtual Library등이 있다. 주제별 탐색엔진의 장점으로는 검색하고자 하는 내용에 대해 특정한 주제어, 키워드등을 표현하기 힘들더라도 대분류 정도만 알수 있으면 정보를 검색할 수 있다는 것이다. 단점으로는 원하는 정보를 얻기까지 '대분류 -> 중분류 -> 소분류 -> 정보' 와 같이 여러 단계를 거쳐야 하므로 중간에 잘못 선택하면 원하는 정보를 찾기 어렵게 된다. 이와 같이 주제별 탐색엔진에는 메뉴형식으로 계속해서 선택하다보면 원하는 정보를 검색하지 못하는 경우가 있다. 이런 단점을 보완하기 위해 대부분의 주제별 탐색엔진은 키워드를 통해 정보를 검색할 수 있는 기능을 제공하고 있다.

키워드형 탐색엔진(Keyword Search Engine, Word oriented searching engine)은 특정한 키워드를 입력하여 그에 해당하는 정보를 검색하는 엔진으로 웹 페이지의 타이틀 및 본문에 있는 문구를 하나의 데이터베이스로 구축해놓고 특정 주제어 또는 검색어를 입력함으로써 원하는 정보를 검색하는 엔진이다. 키워드 탐색엔진은 더욱 정확하게 정보를 찾아낼 수 있도록 불리언 연산(AND, OR, NOT), 따옴표등 여러 가지 검색 옵션을 지정해줄 수 있으며, 검색결과에 대한 신뢰도 점수나 가중치를 나타내준다. 신뢰도 점수나 가중치란 해당 검색결과가 얼마나 정확한지 알려주는 점수이다. 이들 점수가 높을 수록 정확한 검색결과를 의미한다. 키워드 탐색엔진의 장점으로는 몇개의 키워드(검색어)를 통하여 원하는 정보를 신속하게 검색할 수 있다. 단점으로는 색인(해당 탐색엔진이 웹에 있는 정보를 분류해 놓은 목록)이 정확하지 않거나 검색하고자 하는 정보에 대한 키워드가 부적합할 때는 원하는 정보를 검색할 수 없다. 따라서, 해당 탐색엔진이 얼마나 많은 데이터베이스를 얼마나 최신정보로 구축하고 있는가가 중요하다. 또한, 키워드 탐색엔진은 키워드를 정확하게 알고 있어야 원하는 정보를 검색할 수 있는 단점이 있다. 이를 보완하기 위해 키워드 탐색엔진은 주제별 검색을 함께 제공하고 있다. 예를들어, 우리나라 탐색엔진인 심마니의 경우 그동안 키워드를 통한 검색방식만을 제공해 오다가 1996년 6월부터 과학, 교육, 역사, 종교, 컴퓨터등 16가지 분야로 나누어진 주제별 분류를 지원하고 있다. 대표적인 일반형 키워드 탐색엔진으로는 알타비스타(Altavista), 라이코스(Lycos), 웹크롤러(Web crawler)등이 있다.

메타 탐색엔진은 여러개의 탐색엔진을 이용해서 검색하는 도구이다. 메타 탐색엔진은 크게, 단순히, 여러개의 탐색엔진을 한곳에 모아 검색하게 만든 순차적 탐색엔진(Sequential search engine)과 로봇 에이전트를 이용, 다른 탐색엔진을 참조하여 정보

를 직접 찾아주고, 그 결과를 다시 정리해서 보여주는 동시 탐색엔진(Simulteneous search engine)으로 구분된다. 순차적 탐색엔진은 자기자신은 데이터베이스를 구축해 놓지 않고 여러 가지 엔진의 검색어 입력상자만을 따로 뽑아서 제공하는 것이다. 예컨데, 야후, 라이코스, 알타비스타등의 검색어 입력부분만을 따로 분리해서 모아놓은 것이라 볼 수 있다. 따라서 각각의 탐색엔진에 일일이 접속하지 않고도 한 화면에서 이용할 수 있다. 특징으로는 야후나 알타비스타, 심마니등 주제별 탐색엔진이나 키워드형 탐색엔진이 로봇을 이용하여 구축한 자체데이타 베이스를 갖고 있는 반면, 순차적 탐색엔진은 인터넷에서 있는 정보를 직접 수집하지 않으므로 색인화된 자체 데이터베이스를 갖고 있지 않다는 것이다. 장점으로는 각각의 탐색엔진을 옮겨 다니면서 검색할 필요없이 한 화면 안에서 각각의 탐색엔진을 이용할 수 있다. 또한 웹에 있는 HTML 문서만을 대상으로 검색하는 것이 아니라 공개소프트웨어나 뉴스그룹 또는 학술문서까지도 검색해주는 광범위한 검색영역을 갖고 있다. 단점으로는 자신 고유의 데이터베이스를 갖고 있지 않기 때문에 각각의 탐색엔진에서 사용할 수 있는 여러 가지 검색옵션을 모두 지원해주지 못하므로 정확한 검색을 신속하게 검색하는데는 한계가 있다다. 일반적으로, 순차형 탐색엔진은 야후와 같은 주제별 분류는 제공하지 않는다. All-in-one, CUSI, 서치콤등이 이에 속한다.

동시적 탐색엔진(Simultaenious Search Engine)은 정보 데이터베이스를 자체적으로 갖고 있지 않다는 점에서는 순차적 탐색엔진과 동일하나 순차형 탐색엔진이 단순히 여러개의 탐색엔진을 정리/분류하여 한 곳에 모아놓은 것인 반면 동시적 탐색엔진은 로봇 에이전트를 이용, 다른 탐색엔진을 참조하여 정보를 직접 찾아주고, 그 결과까지 보여준다. 또한 순차적 탐색엔진은 각각의 탐색엔진마다 하나씩의 검색어 입력상자가 제공되지만 동시적 탐색엔진은 검색어 입력상자가 하나만 있다. 이런 이유에서 동시적 탐색엔진을 통합탐색엔진 또는 멀티쓰레드(multi thread)형 탐색엔진이라고도 한다. 장점으로는 한번의 키워드 입력만으로 다양한 탐색엔진을 참조하여 검색을 진행하므로 간편한 정보검색과 다양한 탐색엔진에서의 정보를 검색할 수 있다. 단점으로는 여러개의 탐색엔진을 참조하게 되므로 검색속도가 느리며, 여러개의 탐색엔진에서 검색된 결과가 화면에 출력되므로 원하는 정보를 찾기 어려울 경우도 있다. 동시적 탐색엔진으로는 미스다찾니, All 4 One, IBM infoMarket, Savvy Search, MetaCrawler등이 있다. 한국의 대표적인 동시적탐색엔진으로는 미스다찾니가 있는데 이 경우 검색어 입력창에 키워드를 입력하면 심마니, 알타비스타, 코시크, 정보탐정, 뉴스탐색엔진등에 정보검색을 의뢰한 다음 각 엔진별로 검색결과를 화면에 보여준다.

3. 정보의 구축범위에 따른 분류

웹문서의 내용을 전부 색인화하는냐 또는 일부만을 대상으로 색인화해서 그 정보를 제공하느냐에 따라 전문탐색엔진과 초록탐색엔진으로 구분된다.

전문(full text)탐색엔진은 html문서의 내용 모두를 대상으로 정보를 색인화하고 데이터베이스를 구축하여 정보를 제공하는 탐색엔진으로 이는 웹페이지의 모든 단어를

색인화하기 때문에 서로 가까운 거리나 몇 단어 사이에 있는 정보를 찾게 해주는 기능, 특정 키워드의 바로 뒤에 또 다른 단어가 오는것만을 검색해주는 기능등을 제공한다. Full text 탐색엔진의 예로는 알타비스타, 인포시크(Info Seek), 오픈텍스트(Open text)등이 있다.

초록탐색엔진은 웹문서의 내용중 초록 또는 요약문만을 대상으로 정보를 색인화하고 데이터베이스를 구축해서 정보를 제공하는 탐색엔진이다. 초록검색은 쉽고 빠르게 키워드와 일치하는 정보를 검색해주며, 이용자가 원하는 내용에 접근하는 정보를 검색해 준다는 장점이 있지만 웹페이지내의 모든 단어를 색인화하는 전문검색 엔진에 비해 검색 가능한 키워드의 갯수가 적고, 자기에게 필요한 내용이 들어있는데도 자칫 키워드로 선정되지 못한 단어를 사용하면 원하는 정보를 검색할 수 없다는 단점이 있다. 초록형 탐색엔진의 예로는 야후, W3가상도서관등이 있다.

4. 기타

이상에서 설명한 탐색엔진 이외에 탐색엔진은 어디에 있는 정보를 대상으로 하고 있는지에 따라 웹문서를 대상으로한 탐색엔진, 유즈넷 뉴스에 있는 정보를 대상으로 하는 탐색엔진, Anonymous FTP에 있는 정보를 찾아주는 FTP 탐색엔진, 인터넷 이용자에 대한 신상 및 특정한 사람을 찾아주는 인명 탐색엔진등으로 분류될 수 있다. 또한 인터넷에 있는 여러 가지 문서중에서 각종연구 및 학술단체에서 발표한 기술 보고서를 검색하는 엔진등이 있으며, 각종 그림파일이나 그래픽자료, 영화에 관련된 자료, 상용프로그램를 전문으로 검색하는 엔진등이 있다.

Ⅲ. 메타탐색엔진: Savvy Search와 MetaCrawler

한번의 검색으로 여러 탐색엔진을 동시에 검색, 대부분 멀타쓰레드(MultiThread)기법을 사용함으로써, 한번의 검색시간으로 여러곳을 검색할 수 있다. 메타탐색엔진은 다량의 정보를 찾을 수 있으나 처리속도가 다소 느리다. 대표적인 메타탐색엔진으로는 MetaCrawler, Savvy Search, 미스다찾니등이 있다. 여기서는 Savvy Search와 MetaCrawler에 대해서만 살펴보기로 하겠다.

1. Savvy Search

새비서치는 미국 콜로라도주립대에서 개발한 탐색엔진으로 여러개의 엔진을 참조해 동시검색을 하는 엔진이며, 검색어를 입력할 수 있는 상자는 하나이다. 사용자의 검색환경을 다양하게 확장시킬 수도 있고 각 탐색엔진에서의 출력결과를 비교/분석할 수 있다. 다음은 Savvy Search의 특성 및 탐색기능과 출력옵션에 관한 내용이다.

(1) 새비서치는 야후와같은 분야별 목록은 제공하지 않으며, 키워드 입력을 통해 정

보를 탐색하는 기능만을 제공한다.

(2) 새비서치는 8개분야 11개의 탐색엔진을 참조할 수 있다. 특히, WWW에 있는 문서만을 대상으로하는 것이 아니라 인터넷에서 사람과 관련된 WhoWhere, 소프트웨어를 전문으로 찾아주는 FTPSearch95, 각종 쉐어웨어를 찾아주는 shareware.com등에도 정보검색을 의뢰할 수 있기 때문에 한번의 명령으로 다양한 분야를 검색할 수 있다.

(3) 새비서치의 화면은 기본적으로 영어로 나타나 있지만, 프랑스어, 독일어, 일본어 등 20개 언어로 제작된 홈페이지를 제공하기 때문에 이용자의 취향에 맞게 언어를 변경해줄 수 있다. 물론, 새비서치가 20개국의 언어를 지원한다는 것은 홈페이지의 설계가 20개국의 언어로 되어 있다는 뜻이며, 20개국의 언어로 작성된 정보를 검색할 수 있다는 것은 아니다.

(4) 새비서치의 홈페이지에서 아무런 검색옵션을 사용하지 않은 채 키워드를 입력하고 검색을 시작하면, AND 조건으로 WWW에 있는 정보를 대상으로 하여, excite, Yahoo, Altavista등 11개의 탐색엔진에 동시에 검색을 의뢰한다. 즉, 유스넷이나 소프트웨어등은 대상으로 하지 않는다.

(5) 홈페이이지에 있는 Sources and Types...를 클릭하면 WWW Resources, People 등 10개의 분야를 검색대상으로 설정해 줄 수 있는 옵션이 나타난다. 따라서 필요하다면 10개 모두를 선택할 수 있다.

(6) 새비서치는 입력된 검색어들을 AND 조건으로 찾을 것인지, OR조건으로 찾을 것인지, 아니면 여러 개의 단어들을 하나의 구로 간주하여 검색할 것이지를 지정할 수 있다. 그러나 검색어 입력상자에 불리언 연산자를 직접 입력하는 것이 아니라, 검색조건 설정 메뉴에서 all query terms를 선택하면 AND조건으로, any query terms를 선택하면 OR조건으로, all query terms as a phrase를 선택해주면 구(Phrase)로 지정된다. 만약 여러개의 키워드를 띄어쓰기한 채 입력한 후 아무런 검색조건을 지정해주지 않으면 AND조건으로 검색되며, 검색어입력상자에 the, of, it, for등 stop lists가 입력되면 검색과정에서 무시된다. 새비서치는 대소문자 구별이 없으며, 특정한 단어가 들어있는 정보를 제외시키는 NOT기능은 지원되지 않는다.

(7) 쪽당 결과갯수는 각 탐색엔진별로 한페이지당 10개, 20개, 30개, 40개, 50개 등으로 검색결과가 출력되는 옵션이 있으며, 기본값은 10개이다. 결과출력에 대한 설명은 간단히, 보통, 자세히 중 하나를 지정할 수 있다.

(8) 새비서치는 기본적으로 18개의 탐색엔진을 참조해 동시에 검색을 진행하기 때문에 검색결과의 출력도 각 탐색엔진별로 나타난다. 모든 탐색엔진에서 출력결과를 한꺼번에 보여주려면 시간이 많이 소요되기 때문에 3개씩 짝을 이루어 하나의 화면에 보여주게 된다. 즉, 새비서치는 모두 11개의 탐색엔진이 3개씩 짝을 이루어 4개의 그룹으로 나누어져 있다. 각 그 룹을 클릭할 때마나 짝을 이루는 3개 탐색엔진에서의 출력결과가 표시된다. 검색에서 통합결과(Integrated

results)로 지정하면, 짝을 이루고 있는 3개의 엔진에서 검색된 결과 중 서로 중복된 내용을 제거하게 된다. 따라서 짝을 이루고 있는 3개의 탐색엔진 내에서는 중복된 결과가 없지만, 다른 짝에서의 검색결과는 중복된 내용이 있을 수도 있다. 여기서 검색된 결과는 각 탐색엔진별로 나타나는 것이 아니라 새비서치가 자체 선정한 높은 점수 순으로 나타난다. 새비서치는 이와같이 여러개의 탐색엔진에서 검색한 결과가 동시에 표시된다는 장점이 있다. 그러나 한편으로 중복되는 검색결과도 있을 수 있다.

2. MetaCrawler

메타크롤러는 워싱턴 대학에서 Erik Selberg와 Oren Etzioni가 개발한 메타탐색엔진으로 95년 11월부터 서비스 되고 있다. Savvy Search가 11개의 탐색엔진을 참조하여 동시검색을 하듯이 MetaCrawler도 Open Text, Lycos, Webcrawler, Infoseek, Excite 등 6개의 탐색엔진을 이용한 동시검색을 지원한다. MetaCrawler의 특성 및 다양한 검색옵션과 출력옵션에 대한 설명을 살펴보면 다음과 같다.

(1) 키워드 입력을 통해 원하는 정보를 찾는 기능만 제공한다.
(2) 공개된 탐색엔진(상용탐색엔진이 아닌 무료탐색엔진)을 활용해 검색을 진행하며, 검색된 결과를 체계적으로 분류하여 화면에 보여준다.
(3) Savvy Search는 11개의 많은 수의 탐색엔진을 참조할 수 있지만, MetaCrawler와 같은 다양한 검색옵션을 제공하지 못한다.
(4) WWW 분야 6개의 탐색엔진을 참조한다. Savvy Search는 WWW 정보라도 Software, Reference, People, Entertainment등 8개 분야로 검색대상을 나누어 놓았지만 MetaCrawler는 이런 분류를 하지 않고 있다.
(5) 다양한 검색옵션을 제공한다. MetaCrawler는 검색어 입력상자에 입력한 키워드를 AND(all of these words), OR(Any of words), 구(as a phrase)로 설정할 것인지를 지정할 수 있다. 또한키워드 입력시 -, ( ), +, -등도 사용할 수 있다. 어떤 지역(아메리카, 유럽, 아시아등), 어떤기관(회사, 교육기관, 네트워크 관리기관, 군사기관)에 있는 정보를 대상으로 할 것인지도 지정할 수 있는데 구체적으로 살펴보면 다음과 같은 검색대상을 지정할 수 있다.
  ① 지역선택(search region): 절전세계, 자신이 속한 대륙(your continent), 나라(your country), 도메인(domain)등의 옵션과 북아메리카, 남아메리카, 아시아, 유럽, 오세아니아, 남극등 7개의 대륙중 하나를 지정할 수 있다. 기본값은 모든 전세계지역을 대상으로 한다.
  ② 서치사이트(search site)도메인 이름에 따른 기관분류 옵션을 제공한다. 기본값은 모든곳을 대상으로 한다.
(6) 검색시간은 몇 분간의 시간을 통해 검색할 것인지와 어느 정도의 정확성을 유지하면서 검색할 것인지를 지정해준다. 검색시간은 1, 3, 5, 7, 10분등 다섯 가지

가 제공되며 검색강도는 loose, medium, strong등 3가지가 제공된다.

(7) 검색결과를 북마크로 지정해 둘 수 있다. MetaCrawler는 자체적인 출력화면을 통해 여러 가지 탐색엔진으로부터 검색된 결과를 신뢰도에 의해 체계적으로 순위를 매긴 다음 화면에 출력해 주며, 이런 출력된 결과는 북마크로 지정해 줄 수 있다.

(8) MetaCrawler는 검색결과를 한 화면에 몇 개씩 보이게 할 것인지, 검색된 결과에 어느정도 분량의 설명을 붙일 것인지 지정해주는 옵션을 제공한다. MetaCrawler의 기본값에 따라 각 탐색엔진당 10개 정도의 결과가 출력된다. 따라서 MetaCrawler가 검색을 의뢰하는 엔진은 모두 6개이므로 이론상 60개의 결과가 나타난다.

(9) MetaCrawler는 여러 가지 탐색엔진으로부터 검색된 결과를 신뢰도에 따라 재구성한 다음 자체적인 검색결과출력화면을 통해 보여주게 된다. Savvy Search는 각 탐색엔진별로 검색결과가 따로따로 출력할 수 있지만 MetaCrawler는 탐색엔진의 종류를 가지리 않고 신뢰도 순으로 출력된다.

(10) 검색결과에 출력되는 내용: 검색된 문서제목, 어떤 탐색엔진에서 검색되었는지 알려주는 문구, 검색된 문서의 앞부분 50여 단어나 해당문서에 대한 설명, 문서의 신뢰도점수, URL, 해당 문서를 찾아낸 탐색엔진의 이름등이다.

Ⅳ. 연구설계

본 연구는 WWW 메타탐색엔진인 Savvy Search와 MetaCrawler가 주어진 탐색질문에 얼마나 적합한 문헌을 검색해내는가를 조사하기위한 것이다. 본 연구에서 사용된 탐색질문은 기존의 인터넷 탐색엔진 효율성 실험에서 사용되었던 질문들중에서 3개를 선정하였다.

적합성 판단은 탐색결과순위에 따라 출력된 간단한 내용만으로는 어려운 경우가 있으므로 출력된 문서에 연결된 링크를 따라 해당사이트로 가서 문서 전체를 살펴봄으로써 적합 여부를 판단하였다. 또한, 링크로 연결된 정보가 이용이 불가능한 경우(이를 deadlinks라 한다)에는 해당 탐색엔진의 최신성 유지라든지, 정확한 결과를 보여주는 성능에서 부정적이므로, 그 문서는 부적합한 문서로 처리하였다.

본연구는 현재 대표적인 메타탐색엔진인 얼마나 적합한 정보를 검색할 수 있는지를 측정하기 위한 것이므로, 본 연구에서 독립변인은 메타탐색엔진으로, Savvy Search와 MetaCrawler이고, 종속변인으로는 검색된 적합한 문헌수, 정도율, 재현율, 중복탐색의 정도와 Deadlinks 정도였다. Savvy Search와 MetaCrawler는 둘다 메타탐색엔진이라는 점에서는 동일하나, 참조하는 탐색엔진의 수, 검색대상과 출력내용이 다르다. 본 연구에서, Savvy search의 경우, 탐색문은 모두 소문자로 입력하였고, 출력내용은 보

통(normal)로 지정하였다. Savvy search는 11개의 탐색엔진에 3개씩 짝을 이루어 하나의 화면에 보여지게 되는데, 이때, 검색결과는 통합해서 나타나게 하였다. 따라서, 짝을 이루고 있는 3개의 탐색엔진에서 검색된 결과중 서로 중복된 내용을 제거하게 하였다. 이때, 검색결과는 자체적으로 선정한 높은 점수 순으로 나타나고 있다. MetaCrawler의 경우에도 대소문자 구분이 없으므로, 모두 소문자로 입력하였고, 출력은 보통(normal)으로 하였으며, 검색결과는 자체내에서 신뢰도 점수를 부여해서 가장 적합한 문헌순서로 출력하였다. 또한, MetaCrawler가 정보를 검색할 때 참조하는 탐색엔진의 수는 6개이다. 검색된 적합한 문헌의 수를 측정하기위해 사용된 문헌은 검색된 총 문헌중 상위 10개의 문서로 제한하였다. 정도율은 메타탐색엔진이 하나의 탐색질문에 대해 검색된 문서중에서 적합한 문서의 비율이다. 재현율은 실제의 완전한 재현율을 측정하기위해서는 특정 탐색질문에 대한 특정 시스템안에 있는 모든문서를 검토해서 실제의 적합한 문서를 파악해야 된다. 따라서, 이것은 불가능하므로, 여기서는 상대 재현을 사용하였다. 상대 재현율은 두 탐색엔진을 이용하여 검색된 적합한 문서에 대해 각 탐색엔진에서 출력한 적합문서 양의 비율이다. 중복탐색은 동일한 문서가 두 번 또는 그이상 중복해서 나타나는 문서를 의미하며 마지막으로, deadlinks 정도는 각 검색된 문서를 연결했을 때 연결이 안되고 error message가 나타나는 것을 의미한다.

본 연구의 제한점으로는 첫째, 검색효율성을 측정하기 위해 사용된 탐색질문의 3개만을 사용했으므로, 이 연구결과를 일반화 시키는데는 더 다양하고 많은 탐색질문을 이용해서 검색효율성을 측정해야 할 것이다. 둘째, 본 연구결과는 기술통계(descriptive statistics)에 국한해서 해석을 하고 있으므로 타당성 있는 연구결과를 도출하기 위해서는 과학적이고 정밀한 가설을 토대로 검증해야 할 것이다.

V. 탐색질문의 구성

탐색질문은 기존에 탐색탐색엔진의 검색효율성 비교실험에서 사용되었던 질문들 중에서 선정하였다. 탐색질문의 수는 3개이다. 본 연구에서는 웹문서를 대상으로하여 메타탐색엔진의 효율을 평가하므로, 탐색대상을 모두 웹으로 한정하여 실험을 수행하였다. 본 연구에서 사용되는 메타탐색엔진인 Savvy Search와 MetaCrawler는 검색옵션이 일반탐색엔진에 비해 다양하지 못하고, 단지 Boolean 연산자를 이용할 수 있는정도이다. 따라서, 각 탐색엔진의 탐색식구성은 동일하다. 검색된 문서의 내용은 보통(normal)로 지정하였으며, 자동으로 우선순위를 정해서 출력되었다. 다음은 본 연구에서 사용된 탐색질문과 탐색식이다.

(1) 탐색질문: A Recipe for Coconut Cream Pie

탐색식: recipe coconut cream pie

(2) 탐색질문: Effect of Divorce on Childere

탐색식: Effect Divorce Childeren

(3) 탐색질문: Mason(Freemasonry)

탐색식: Masor Freemasonry

이상의 탐색식 작성에서 탐색질문 1번과 2번은 검색옵션에서 all query terms를 선택하였고, 3번은 any query term를 검색옵션으로 지정하였다. 즉, all query terms는 각단어를 AND로 연결해서 검색하라는 의미이고, any query term은 각 키워드를 OR로 검색하라는 의미이다.

첫 번째 탐색질문인, "recipe for cocounut cream pie"는 코코넛 크림파이를 만들기 위한 재료를 검색하는 것이다. 따라서, 이 질문에서 적합한 문서는 실제로 코코넛 크림파이를 만들기 위한 재료가 포함되어 있으면, 적합한 문헌으로 판정하였고, 기타 코코넛 크림파이를 파는 제과점이나 레스토랑, 또는 코코넛 과 크림이 재료로 들어가 있는 다른 종류의 음식은 적합한 문서에서 제외시켰다.

두 번째 탐색질문인 "Effect of divorce on childern"에 대해서는 이혼이 어린이에게 미치는 영향에 관련된 문헌일 경우에는 적합한 문헌으로 간주하였고, 기타 검색리스트에는 effect of divorce on children이 포함되어 있으나 실제내용이 다를 경우 부적합한 문헌으로 판정하였다.

세 번째 탐색질문에 대해서는 Mason또는 Freemasonry에 관련된 문서는 지역이나, 언어, 나라, 특정단체에 상관없이 모두 적합한 문서로 판정하였다.

Ⅵ. 데이터분석

본 논문은 메타검색엔진인 Savvy search와 MetaCrawler의 검색효율성을 측정 하였다. 검색결과는 상위 10개의 문헌으로 제한하였으며, 측정기준은 검색된 적합문헌수, 정도률과 상대재현율 및 중복탐색수와 deadlinks 정도였다. 이때 검색된 적합문서 총수에서 중복되는 내용(URL까지 동일한 문서)은 제외하였다.

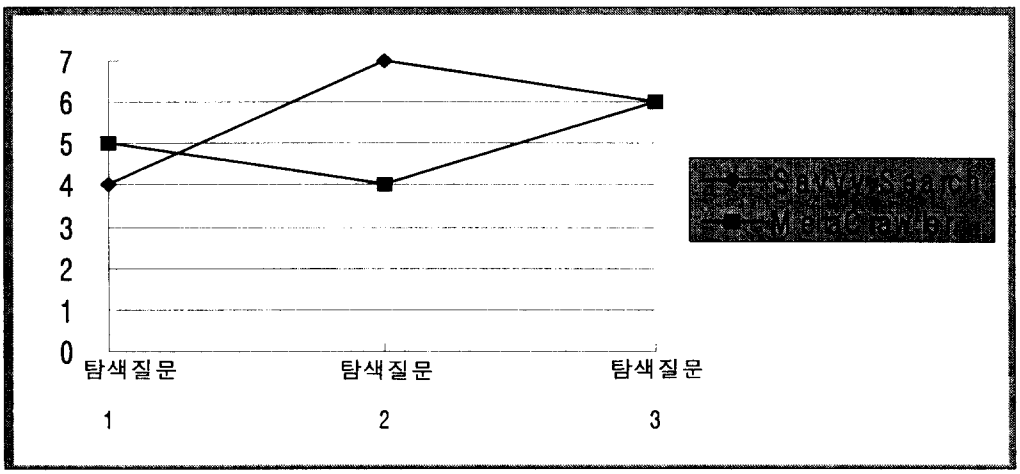
1. 검색된 적합한 문서수

Savvy Search와 WebCrawler가 각각 검색한 상위 10개까지의 문서를 판단해 보았을 때, 각 질문당 검색된 적합문헌의 평균은 Savvy Search의 경우, 5.7건이고, WebCrawler의 경우는 5건이었다. 구체적으로, 살펴보면, Savvy Search의 경우, 탐색질문 1에 검색된 적합한 문헌은 4개였고, 질문2에 대해서는 7건, 질문3에서는 6건이 적합한 문헌으로 나타났다. WebCrawler의 경우, 탐색질문 1에 대해서 검색된 적합한

문헌은 5건, 질문 2에서는 4건, 질문 3에 대해서는 6건으로 나타났다. 이 결과는 비교적 높은 수의 문서가 적합한 문헌으로 검색되었다고 볼 수 있다. 검색된 적합한 문서의 결과를 표와 그림으로 나타내면, <표 1>과 <그림 1>과 같다.

<표 1> Savvy Search와 MetaCrawler에 대한 적합문서수

	탐색질문 1	탐색질문 2	탐색질문 3
Savvy Search	4	7	8
MetaCrawler	5	4	6



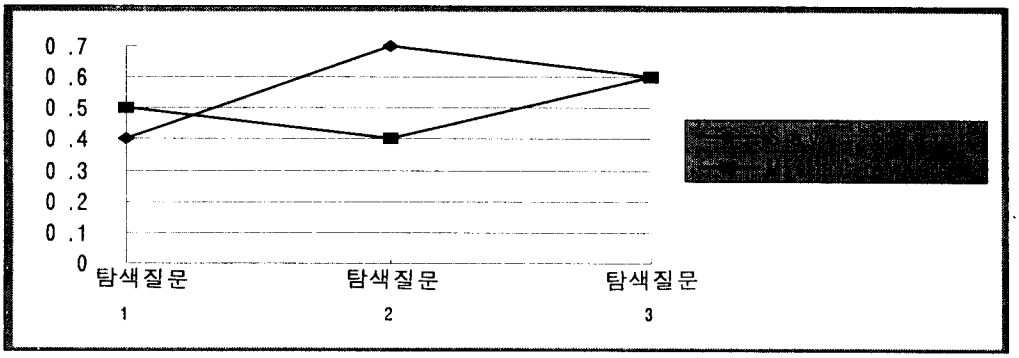
<그림 1> Savvy Search와 MetaCrawler에 대한 적합문서수

2. 정도률

<표 2> 및 <그림 2>에서 나타난 바와 같이, Savvy Search의 평균정도률은 0.6이고, WebCrawler의 경우는 0.5이다. Savvy Search의 경우, 탐색질문 1에 경우, 정도률은 0.4였고, 질문2에 대해서는 0.7, 질문 3에 대해서는 0.6건이 었다. WebCrawler의 경우, 탐색질문 1, 2, 3에 대해서 각 정도률은 0.5, 0.4. 0.6이었다. 이상에서와 같이, 정도률도 평균적으로, Savvy Search가 높은 것으로 나타났다. 이런결과는, 검색된 적합한 문서의 수에 따른 것이다.

<표 2> Savvy Search와 MetaCrawler에 대한 정도률

	탐색질문 1	탐색질문 2	탐색질문 3
Savvy Search	0.4	0.7	0.6
MetaCrawler	0.5	0.4	0.6



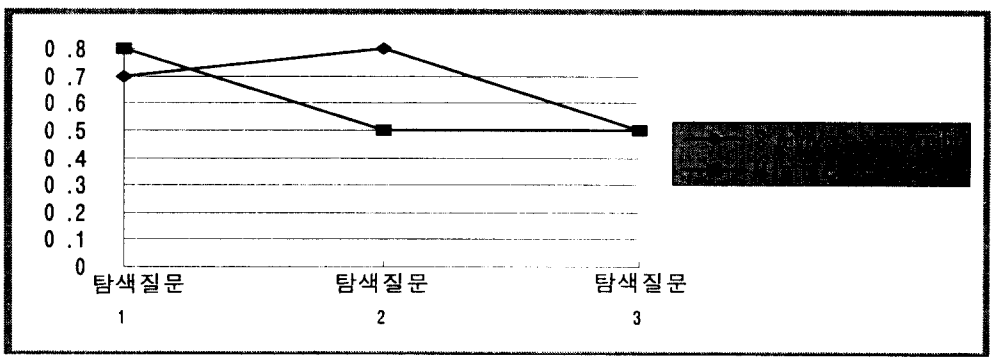
< 그림 2> Savvy Search와 MetaCrawler에 대한 정도률

3. 재현률

Savvy Serch의 평균 재현률은, 0.7이었고, WecCrawler의 경우, 0.6이었다. Savvy search의 경우, 탐색질문 1은, 0.7, 탐색질문 2에 대해서는 0.8, 탐색질문 3에 대해서는 0.5였다. WebCrawler의 경우, 탐색질문 1, 2, 3에 대한 재현률은 각각, 0.8, 0.5, 0.5였다. 재현률 역시, Savvy Search가 WebCrawler보다 높은 것으로 나타났다. <표 3>과 <그림 3>은 재현률에 대한 결과이다.

<표 3> Savvy Search와 MetaCrawler에 대한 재현률

	탐색질문 1	탐색질문 2	탐색질문 3
Savvy Search	0.7	0.8	0.5
MetaCrawler	0.8	0.5	0.5



<그림 3> Savvy Search와 MetaCrawler에 대한 재현률

4. 중복탐색과 Deadlinks

Savvy Search의 경우, 탐색질문 1의 경우, deadlinks는 없었으나, 중복되는 문서가

한 건 있었다. 즉, 동일한 내용과 URL임에도 불구하고, 두 번 중복되어서 검색된 문서가 1건 있었다. 질문 2의 경우, 중복문서는 없었으나, deadlinks가 1개 있었으며, 질문 3의 경우에는 중복문서는 없었으나, deadlinks는 2개 있었다. 따라서, Savvy Search의 경우, 총 중복문서는 2건이었고, deadlinks수는 3개였다. 한편, MetaCrawler의 경우, 탐색질문 1의 경우는, 중복문서는 1개 였고 dadlinks는 없었다. 질문2의 경우, deadlinks는 1개 있었으나, 중복문서는 없는 것으로 나타났다. 질문3의 경우는 중복문서도 deadlinks도 없는 것으로 나타났다. 따라서, MetaCrawler의 경우, 총 중복문서는 1개이고, Deadlinks도 1개 인 것으로 나타나, Savvy Search보다 최신성 유지라든지, 정확한 결과를 보여주는 성능면에서는 높은 것으로 나타났다. <표 4>는 Savvy Search와 MetaCrawler에 대한 중복탐색 및 deadlinks의 정도를 요약한 것이다.

<표 4> Savvy Search와 MetaCrawler에 대한 총 deadlinks와 중복탐색건수

	Deadlinks 수	중복탐색건
Savvy Search	3	2
MetaCrawler	1	1

Ⅶ. 결과해석

이상에서 Savvy Search와 MetaCrawler의 검색효율성을 검색된 적합한문서수, 정도율, 재현률, 중복탐색 및 deadlinks정도를 기준으로 살펴보았다. 그 결과 전체적인 정도률과 재현률은 비교적 높은 것으로 나타났다. 이미 기존연구를 통해 알 수 있듯이, 기존의 연구결과에 의하면 일반탐색엔진의 검색효율성는 정도율과 재현률을 기준으로 했을 때 매 우 저조한 것으로 나타났다. 예를들어, 김성은, 정영미가 조사한 바에 따르면, 2개의 탐색질문에 대해 9개의 탐색도구간의 검색효율성을 측정할 결과, 정도률이 0.5이상된 탐색엔진은 Hotbot과 Open Text뿐이것으로 나타났으며, 재현률은 9개 탐색엔진모두 0.5를 넘지 못했다. 이런 결과와 비교했을 때, 본 논문에서 측정한 메타탐색엔진은 정도률과 재현율이 모두 0.5이상이므로 비교적 높다고 할 수 있으며, 특히, 재현률은 비록 상대재현율이긴 하지만 매우 높은 것으로 나타났다. 이와 같이, 메타탐색엔진이 높은 검색효율성 보이는 몇가지 요인을 분석해 보면 다음과 같다.

첫째, 본연구의 연구에서, 재현률과 정도률이 비교적 높은 이유는 메타탐색엔진으 특성때문인 것으로 보인다. 메타탐색엔진은 일단 질문을 입력하면, 여러곳의 탐색엔진에 의뢰해서 검색결과를 통합하기 때문인 것으로 보인다. 이런결과는 기존여구의 결과와 비교해 보면 알 수 있다. 즉, 기존의 개개의 탐색엔진 효율성 연구를 살펴보면, 각기 엔진마다 다른 결과를 보여줄 뿐 아니라, 중복도가 낮아서 하나의 탐색엔진만을 이용했을 경우, 그 효율성이 낮은 것으로 나타났다. 즉, 각 탐색엔진들은 각각 중점을

두고 있는 서비스가 다르고 색인 데이터베이스의 구성방법등에 차이가 있기 때문이다 (Leonard, 1996). 따라서, 이들 각기다른 특성을 갖고 있는 탐색엔진을 동시에 이용했기 때문에 검색효과가 높은 것으로 보인다.

둘째, Savvy Search와 MetaCrawler의 검색효율성을 비교했을 때 Savvy Search가 높은재현률과 정도률을 나타내었다. 이는, Savvy Search는 키워드를 입력했을 때 11개의 탐색엔진을 참조하여 검색하는 반면 MetaCrawler는 6개의 탐색엔진만을 사용하기때문인 것으로 보인다. 이런결과는 미래의 메타검색엔진은 더욱다양하고 많은 탐색엔진을 참조하는 엔진이 개발되어야 한다는 것을 의미한다.

셋째, 중복탐색과 deadlinks 정도는 MetaCrawler가 Savvy Search보다 적은 것으로 나 Savvy Search보다 최신성 유지라든지, 정확한 결과를 보여주는 성능면에서는 높은 것으로 나타났다. 탐색엔진의 중복탐색은 2가지 의미를 갖는다. 첫째, 동일한 도메인이나 사이트에 질문에 관련된 여러개의 문서가 가기 다른 파일명으로 저장되어있는 경우 각 파일을 검색하기 위해 동일한 도메인이 중복하여 탐색되는 것을 말한다. 둘째, 동일한 사이트나 문서가 오류에 의해 중복되어 검색되는 경우이다. 동일한 도메인이나 사이트의 중복탐색은 로봇프로그램이 적적한 히스토리 파일을 관리하지 않는 경우에 발생하는 것으로 지적되었다. (McMurdo, 1995). 일반적으로 이용자가 하나의 도메인 혹은 상위 경로명에 접근하게 되면 제공되는 링크를 따라 여러 파일에의 접근이 가능하기 때문에 위와 같은 탐색방법은 효과적이지 못하다. 그리고 어떤 경우에는 같은 사이트나 문서임에도 불구하고 URL 표기의 차이 등으로 인해 두 번 이상 검색되거나 또는 같은 URL인데도 두 번 이상 검색되는 경우가 발생하게 된다. 예를 들면 하나의 IP주소에 대해 여러개의 DNSS alias가 할당 되어 있는 경우 로봇 프로그램이 이를 인식하지 못하면 이러한 중복결과가 나타나게 되는 것이다.(McMurdo, 1995). 웹탐색은 대부분의 경우 그 검색결과가 매우 많은 것이 특징이므로 이러한 중복탐색은 탐색기능의 효율을 저하시키는 요인 중의 하나라고 할 수 있다. 이런 결과는 비록 메타검색엔진이 중복탐색결과를 배제하는 옵션을 지정하였더라도 중복탐색수가 나타난다는 것을 의미한다고 볼 수 있다. 한편, Deadlinks가 나타나는것에 대해서는 앞에서도 언급하였지만 각 문서의 URL은 자주 바뀌는 경우가 많으므로 갱신주기를 자주하여야 할 것으로 보인다.

Ⅷ. 결 론

본 연구는 메타탐색엔진인 Savvy Search와 MetaCrawler의 검색효율성을 측정하기 위해, 기존의 인터넷 탐색엔진 검색효율성에 이용되었던 탐색질문중 3개를 선정하여, 검색된 적합문서, 정도률, 재현률, 탐색중복과 deadlinks를 기준으로 살펴 보았다. 그 결과를 요약하면 다음과 같다.

첫째, 각 탐색엔진이 검색해낸 문서중에 상위 10개를 기준으로 적합성 판단을 했을 경우, 각 질문당 검색된 적합문헌의 평균은 Savvy Search의 경우, 5.7건이고, MetaCrawler의 경우는 5건이었다.

둘째, 평균정도률과 재현률은 Savvy Search가 MetaCrawler보다 높은 것으로 나타났다. Savvy Search의 평균정도률과 재현률은 0.6, 0.7이었으며, MetaCrawler의 경우 평균정도률과 재현률은, 0.5와 0.6이었다. 이런 결과는 Savvy Search의 특성때문인 것으로 보인다. 즉, Savvy Search가 MetaCrawler에 비해 많은 탐색엔진을 참조해서 자체적으로 각 탐색질문에 가장 적합한 문서가 검색된 것으로 판단되는 엔진을 선정해서 검색하기 때문인 것으로 보인다.

셋째, Savvy Search의 경우, 총 중복문서는 2건이었고, deadlinks수는 3개였다. MetaCrawler의 경우, 총 중복문서는 1개, Deadlinks도 1개 인 것으로 나타나, Savvy Search보다 최신성 유지라든지, 정확한 결과를 보여주는 성능면에서는 높은 것으로 나타났다.

넷째, 메타탐색엔진을 일반 키워드나 주제별 탐색엔진과 검색효율성을 기준으로 비교했을 경우, 정도률과 재현률이 비교적 메타탐색엔진이 높은 것으로 나타났다. 이는, 각각의 탐색엔진이 색인구축방법, 검색옵션등이 다르기 때문에 각각 동일한 질문에 대해서도 상이한 검색결과가 나타나는 것으로 보인다. 따라서, 다양한 질문의 유형에 따라 다양한 탐색도구를 선정하여 실행하는 메타탐색엔진이 앞으로는 상당히 활발해 개발되고 활용될 것으로 기대된다.

참고문헌

정영미, 김성은 (1997) WWW 탐색도구의 색인 및 탐색기능평가에 관한연구, 한국문헌정보학회지, 제 31권 제 1호: 153-184

Courtois M. P. (1996) "Cool Tools for Web Searching: An Update.", Online 20(3): 29-31

Courtois M. P., et al., (1995) "Cool Tools for Searching the Web: A Performance Evaluation.", Online 19(6): 14-32

Daniel Dreilinger. Savvy Search Home Page. http://www.cs.colostate.edu/dreiling/smartform.html

Koster, M. 1995 Why Simultaneous Search Engines are not so great http://www.nexor.com/public/cusi/simultaneous

Leonard, A. J. (1996) "Where to Find Anything on the Net". http://www.cnet.com/contents/reviews/

McMurdo, G. (1995) "How the Internet was Indexed" Journal of Information Science 21(6):479-489

Robertson, M. (1996) A Review of the Internet Sleuth, http://www.fis.utoronto.ca/courses/LIS/2108/1996/

Selberg, E. and Oren Etzioni (1995) Multi-Service Search and Comparison Using the MetaCrawler http://www.cs.washington.edu/home/

Zone, P. et al., (1996) "Advanced Web Searching: Tricks of the Trade", Online 20(3):14-28