데이터마이닝기법을 이용한 검색엔진의 검색효율성 측정에 관한 연구

A Study on the Retrieval Effectiveness in the Search Engines Using Data Mining Techiniques

김성희(Seong-Hee Kim)* · 이수연(Su-Yeon Lee)**

〈목 차〉

초 록

본 연구에서는 데이터마이닝 기법을 이용한 검색엔진인, Northernlight와 Google과 일반메타탐색엔진인 Metacrawler를 정도률, 재현률을 기준으로 검색효율성을 측정하였다. 분석결과 데이터마이닝 기법을 이용한 검색엔진이 일반검색엔진에 비해 검색효율성이 높은 것으로 나타났다. 따라서, 데이터마이닝 기법을 이용한 검색엔진이 앞으로 검색효율성을 높이는데 기여를 할 수 있을 것으로 기대된다.

주제어 : 검색엔진, 검색효율성, 데이터마이닝기법

Abstract

This study is intended to compare the effectiveness of the Northernlight and Google, which are based on Datamining technique with a Metacrawler, one of metasearch engines. As a result, searches responding to queries in the Northernlight and Google produced a higher precision and recall as compared with searches responding to queries in the metacrawler. The results show that the Datamining techniques can help improve information retrieval effectinveness.

Key Words : data mining techniques, retrieval effectiveness, search engine

* 동덕여자대학교 정보학부 조교수. shkim@www.dongduk.ac.kr.
** 동덕여자대학교 대학원 문헌정보전공

Ⅰ. 서 론

정보기술의 빠른 발전은 업무의 자동화를 촉진시켜 엄청난 양의 데이터를 전자적으로, 수집, 축적하는 것을 가능하게 하였다. 그러나, 이러한 기하급수적인 데이터의 증가로 인해 우리가 원하는 정보를 신속하고 정확하게 검색하기 어려운 실정이다. 또한 수집된 데이터를 어떻게 처리할 것 인가하는 문제와 수집된 방대한 정보집합체들의 함축된 의미 파악문제 즉, 잘 고안되어 축적된 데이터가 전통적인 분석기술에 의해 감지되지 않는다는 문제가 제기되었다. 이러한 가운데 데이터의 의미를 효과적으로 이해하고 데이터의 특성을 체계적으로 구조화하며, 구조화된 데이터를 통해 유용하게 사용되는 데이터마이닝 또는 지식발견기술(knowledge discovery)이 많이 적용되어 왔다.

이러한 데이터마이닝기법은 최근들어 hypertext형태의 정보조직 및 브라우징(browsing)이 가능한 WWW(World Wide Web)를 이용하여 정보를 제공하는 사이트가 기하급수적으로 늘어나면서 WWW를 이용해서 인터넷상에 널리 퍼져있는 정보를 사용자에게 찾을 수 있도록 도와주는 도구인 검색엔진에 적용되고 있다. 예를들면, Peggy Zorn (1999)은 데이터마이닝의 개념과 이를 토대로 구현된 검색엔진인 Northernlight을 소개하고 있다. 그러나, 아직까지 본 연구에서 수행하고자하는 데이터 마이닝기법을 이용한 검색엔진의 효율성을 측정한 연구는 수행되고 있지 않았다. 따라서, 본 연구에서는 정보를 구축할 수 있는 새로운 데이터베이스 기술개념인 데이터마이닝의 개념과 그 구축방법, 이를 기반으로 구축된 인터넷 검색엔진인 Northernlight과 Google의 검색효율성을 기존의 메타검색엔진인 Metacrawler와의 비교를 통해 그 효율성을 살펴보고자 한다.

Ⅱ. 이론적 배경 및 선행연구

1. 데이터마이닝 개념 및 데이터마이닝 기법

데이터마이닝(datamining)이란 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 의미하며, "mine" 이란 언어의 사전적 의미에서 알수 있듯이, 거대한

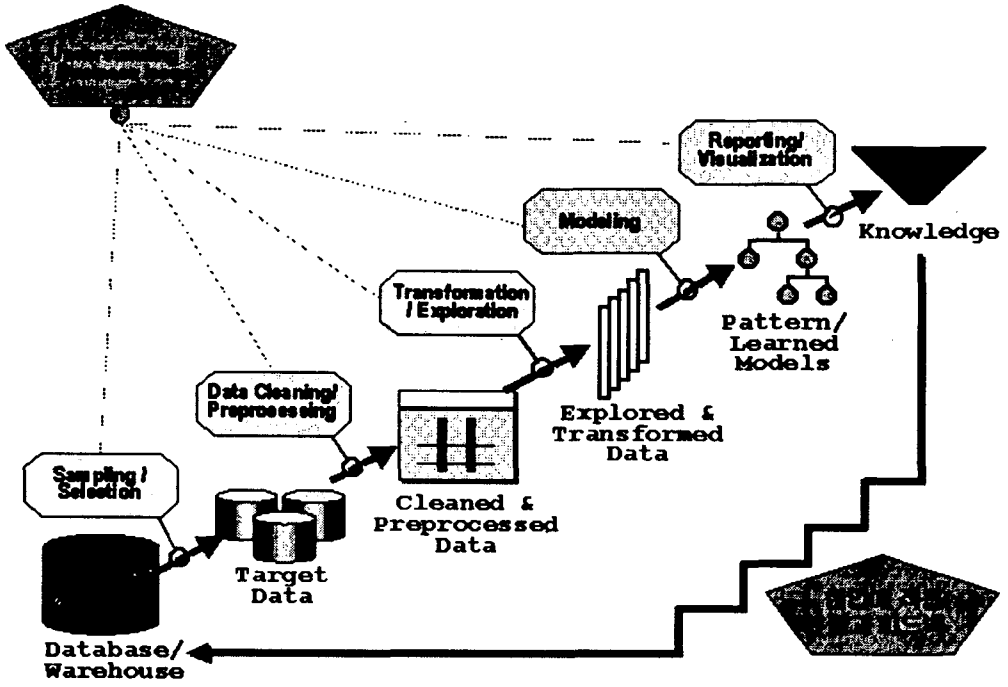
더미 속에서 가치 있는 무언가를 캐낸다라는 것을 의미한다. 즉, 데이타마이닝이라는 것은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정, 데이터간의 숨겨진 관계, 혹은 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 뽑아내기에는 너무나 복잡한 관계를 찾아내고, 이 관계를 바탕으로 앞날을 예측하는 기술이며, 대용량의 데이터로부터 이들 데이터내에 존재하는 관계, 패턴, 규칙등을 탐색하고 찾아내어 모형화함으로서 유용한 지식을 추출하는 일련의 과정들이라고 정의할 수 있다. 이는 하나의 분석기법을 의미하는 것이 아니라 여러기법과 방법들의 적절한 조합으로 이루어진 일련의 과정(process)이라 말할 수 있다.

데이터마이닝에 적용되는 기법으로는 시계열분석등 각종 통계기법과 데이터베이스 기술 뿐만 아니라 산업공학, 신경망, 인공지능, 전문가시스템, 퍼지논리, 패턴인식, 기계적 학습(machine learning), 불확실성 추론(reasoning with incertainty), 정보검색에 이르기까지 각종 정보기술과 기법들을 사용하게 된다. 또한 경영전략, 마케팅 기법등의 최신 경영기법들의 이용되기도 한다. 이러한 데이타마이닝을 통하여 거대한 데이터베이스에 숨어있는 전략적인 정보를 발견할 수 있게 된다.

데이타마이닝과 관련된 용어로는 데이터웨어하우징, 의사결정 지원시스템, OLAP,지식관리 등이 있으며 데이터마이닝을 적용하기위해서는 "대용량의 데이터"를 구축해놓은 데이터창고인 데이터웨어하우스(DatawareHouse)구축이 필요하다.

데이터마이닝은 text mining과 web mining으로 구분하고 있는데, 먼저 text mining은 온라인 정보원 모두를 대상으로 정보속에 내재되어 있는 메세지와 주요관계를 파악하여 이용자의 이해를 도우며 정보와 그 속성이 일정한 패턴과 기법으로 배열할 수 있는 기술 테크닉으로, 이용자에게 개념이해에 대한 편의를 도모하고자 하는데 그 중점을 두는 것이라 할 수 있다. Web mining은 인터넷 사용자의 사이트 방문시 남게되는 로그화일등의 로그데이터를 양적(quantative)으로 분석하여 이를 기반으로 한 모델과 이론을 개발하는데 중점을 두어 이용자의 행동패턴에서 일정한 패턴을 추출하여 이를 모델화하여 가시적으로 제시하는 기법이라 하겠다.

데이타마이닝의 구축수행과정은 <그림 1>에서 보듯이 표본선택, 데이터정제, 데이터변형, 모델링, 가시화 및 평가의 과정을 거쳐 수행된다.



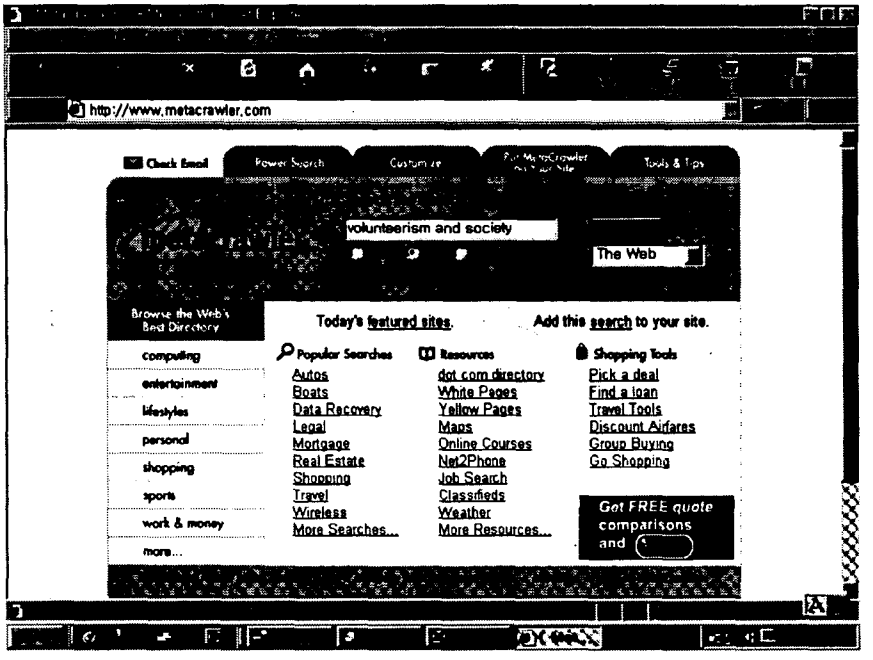
<그림 1> 데이터마이닝 구축수행과정

2. 연구대상 검색엔진의 특성

2.1. 메타검색엔진: MetaCrawler(www.metacrawler.com)

MetaCrawler는 워싱턴대학에서 Erik Selberg 와 Oren Etzioni가 개발한 메타검색엔진으로 95년 11월부터 서비스되고 있다<그림 2>. 한번의 검색으로 여러 검색엔진을 동시에 검색, 대부분 멀티쓰레드(Multithread)기법을 사용함으로써, 한번의 검색시간으로 여러곳을 검색할 수 있다. 메타검색엔진은 다량의 정보를 찾을 수 있으나 처리속도가 다소 느리다.

MetaCrawler는 자체데이타베이스를 보유하고 있지 않으며 다만 오픈텍스트, 라이코스, 웹글롤러, 인포시크, 익사이트, 알타비스타, 야후등 14개에 이르는 검색엔진을 통해 검색을 명령하고 결과를 수집하므로 제공되는 검색형식이 매우 단순하다. 광범위한 검색은 가능하지만, 각 검색엔진의 특징에 맞는 세부적인 연산자와 검색방법 사용하기 어렵다.

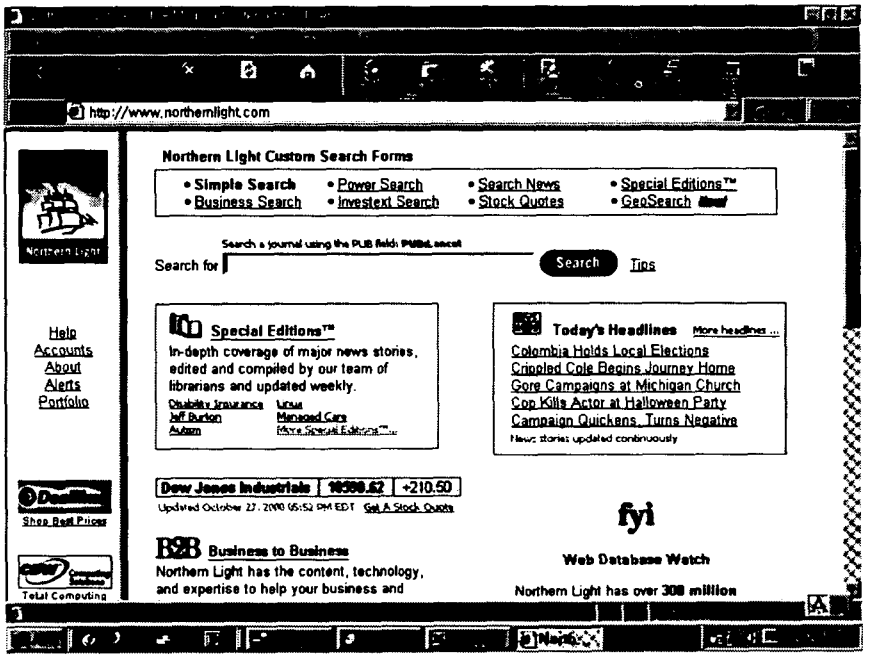


<그림 2> MetaCrawler의 메인화면

2.2. 데이타마이닝기법을 이용한 검색엔진

(1) NorthernLight Search(http://www.northernlight.com)

<그림 3>에서 보여지는 Northernlight은 1995년 개발된 검색엔진으로서 1400억이상의 웹 페이지와 6,900여개의 원문데이타베이스를 제공하고 있다. 걸리버라는 로봇프로그램을 사용하여 웹사이트, 수 많은 저널과 전문잡지로부터 정보를 제공받고 있다. 다른 검색엔진과는 달리 Northernlight는 전문사서에 의해 미리 주제분류된 폴더(folder)에 각 정보를 배열하는 데이타마이닝 기술을 사용한 검색엔진이라 하겠다. 이 폴더는 주제, 문서형식, 소스, 언어별로 분류되며. 대략 20,000개 이상의 광범위한 계층적 관계어와 200,000-300,000에 이르는 첨가어로 수록된다. 이러한 색인은 사람에 의해 수작업으로 생성되지만, 데이타베이스는 컴퓨터에 의해 자동적으로 생성하게 된다. Northernlight의 폴더는 각각의 검색결과를 미리 지정된 검



<그림 3> Northernlight의 main화면

검색결과의 색인을 고려한 알고리즘에 근거하여 구성된다. 이러한 데이타마이닝 기술을 도입한 폴더의 생성과 로봇기술을 이용한 다른 일반검색엔진의 결합은 별도의 통제어를 사용하지 않고 이용자에게 정확한 검색결과를 제공한다.

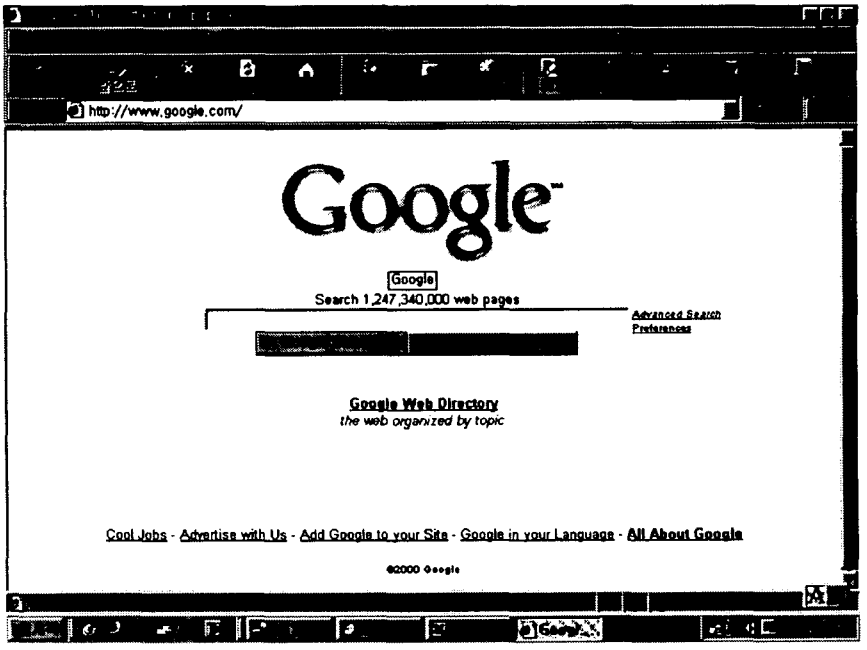
웹을 대상으로 검색을 실시하는 이 검색엔진의 특징은 Special Collection이라는 6,900여개의 별도 간행물로부터 전문을 제공하고 있다는 것이다. Special Collection에는 Business Magazine, Trade Journal, Newswire, Academic Journal 등이 포함되어 있다. 또한 33개 통신사의 최근 2주간 뉴스를 무료로 이용할 수 있다. 별도의 가입이 필요하며, 가입은 회사, 도서관 등이 포함되어 있는 기관과 개인 등 2가지로 분류된다. 요금은 'pay per document'로 $1~$4까지 지불한다.

(2) Google (http://www.google.com)

Google은 1999년 9월 21일 베타 단계를 끝내고 라이브 버전으로 시작한 새로운 검색 엔진이다. 미국 Stanford 대학의 두 연구원, Larry Page와 Sergey Brin이 1998년에 시작한 검색엔진으로 Yahoo!처럼 특이한 이름을 가지고 있는 이 검색엔진은 Internet, Web, Cars 등 세세한 내용이 아닌 일반 정보를 찾을 때 가장 적합한 검색 엔진으로 평가를 받아왔다. 독특한

"PageRank" 기술을 이용하여 사이트 순위를 매기는 것으로 유명하다.

Google에서는 재미있는 개성을 많이 찾아볼 수 있는데, <그림 4>의 Google의 main화면에서 보여지는 "I'm feeling lucky" 버튼은 다른 검색 엔진에서 찾아볼 수 없는 것으로 검색결과를 직접 보여주지 않고 그중 첫번째 웹 페이지로 바로 이동시켜주는 기능이다. 그리고 "cached link"를 제공하여 이미 사라진 사이트이거나 서버/네트워크 일시장애, 중단으로 접속이 잘 안되는 것을 어느 정도 막아주기도 한다. Google을 이용하면 "404 Not Found"에러를 그만큼 적게 볼 수 있다.



<그림 4> Google의 main화면

이상에서 데이터마이닝 기법을 이용한 검색엔진과 메타검색엔진에 대해 살펴보았다. 데이타마이닝 기법은 방대한 양의 데이타 속에서 쉽게 드러나지않는 유용한 정보를 찾아내는 과정, 데이타간의 숨겨진 관계, 혹은 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 파악하기에는 너무나 복잡한 관계를 찾아내고 이 관계를 바탕으로 앞날을 예측하는 기술이다. 따라서, 데이터마이닝 기법을 이용한 검색엔진은 적합한 문서를 검색하기위해 다양한 인공지능기법을 활용함으로써 인간의 판단에 기초를 두고 검색을 하려는 특징이 있다. 즉, 데이터마이닝 기법을 이용한 검색엔진 적합한 문서를 검색하기위해 다양한 인공지능기법을 활용함으로써 인간의 판단에 기초를 두고 검색을 함으로써 대용량의 데이타로부터 이들 데이타 내에 존재하는 관계, 패턴, 규칙등을 탐색하고 찾아내어 모형화함으로서 유용한 지식을

검색하려는 특성이 있다고 볼 수 있다.

3. 선행연구

아직까지 데이터마이닝기법을 이용한 검색엔진의 효율성에 관한 연구는 거의 없고 일반 검색엔진의 효율성에 관한 연구만 국내외에서 다양하게 이루어졌다. 국내의 경우에는 정영미, 김성은(1997)이 Zorn(1996)이 사용하였던 검색질문을 사용하여, 일반키워드 검색엔진과 카테고리 검색엔진 9개를 대상으로 그 효율성을 측정하였다. 탐색실험 결과 탐색질문의 유형에 관계없이 Altavista, Hotbot, OpenText가 비교적 좋은 검색효율을 보였으나, 대부분의 검색엔진이 질문의 성격과 작성된 탐색문에 따라 탐색결과에 있어 많은 차이를 보인다는 것을 발견하였다. 이 후 이명희(1997)는 Yahoo와 Altavista를 대상으로 그 검색효율을 측정하였으며, 그 결과 AltaVista는 특정적이고 전문적인 용어의 탐색에 적합한 반면, Yahoo는 일반적이며 추상적인 용어의 탐색에 적합하다는 결론을 도출하고 있다. 김성회(1997)는 메타검색엔진 Savvy Search 와 Metacrawler를 대상으로 검색효율을 측정한 연구를 수행하였다. 연구의 결과로 다양한 검색엔진의 하이브리드(hybrid)한 방법을 숙지함으로써 인터넷검색에 있어 검색의 효율성을 높일수 있음을 제시하고 있다. 이상의 각 연구에서 나타난 검색엔진의 효율성은 엔진의 성능에 따라 각기 다른 결과를 도출하고 있음을 알수 있었다.

국외의 경우에는 Chu와 Rosenthal(1996)은 로봇에 의해 운영되는 검색엔진인 Altavista, Lycos, Excite를 대상으로해서 그들의 탐색능력(불리언 논리, 용어절단, 제한탐색, 단어와 구탐색)과 검색효율(정확률과 응답시간)을 측정하였다. 대학도서관의 참고업무중의 실제 발생된 10개의 질문을 대상으로 수행된 이 연구에서 적합성의 판정은 두 연구자 자신에 의해 이루어졌다. 그들은 Altavista와 Excite는 각 질문당 10개 이상의 문헌을 검색해 내었으나 Lycos는 수록범위가 가장 넓은데도 불구하고 어떤 질문에 대해서는 하나의 문헌도 검색해내지 못하는 것을 발견하였으며, 검색성능이나 검색효율에 있어서 Altavista가 가장 성능이 뛰어남을 밝히고 높은 정도율을 요구하는 이용자에게는 Altavista가 가장 권고할 만하다고 주장하였다. Kimmel(1996)은 7개의 검색엔진을 대상으로 단일단어로 구성된 문장을 가지고 검색엔진의 성능을 테스트하였다. Kimmel (1996)는 WWW Worm , Lycos, Open Text등 7개의 검색엔진을 대상으로 단일 단어로 구성된 문장에 대한 이들의 성능을 테스트하였다. 테스트 결과 키워드 검색엔진에 있어서 Lycos가 가장 강력한 검색엔진이라고 주장하였다. Zorn(1996)은 AltaVista, Lycos, InfoSeek, OpenText등을 대상으로 복합탐색(advanced search)의 특성을 살펴보았다. 검색엔진의 복합탐색기능을 사용하여 3개의 탐색질문이 각각 탐색되었는데, 검색엔진의 색인과 평가과정은 시스템마다 다르기 때문에 어떤 단일 검색엔진도 최고의 효율

성을 가질 수 없다고 하였다.

이상에서 살펴본바와같이 기존의 연구들은 일반적인 검색엔진의 성능을 주로 다루었으며 데이터마이닝기법을 이용한 검색엔진의 효율성은 측정되지 않았다.

Ⅲ. 연구설계

본 연구는 기존의 메타검색엔진과 새로이 소개된 데이타마이닝기법을 이용한 검색엔진과의 검색효율성을 비교, 분석한 것이다. 본 연구에서 사용된 검색질문은 기존의 인터넷 검색엔진 효율성 실험에서 사용되었던 질문들중에서 일부를 선정하고 일부는 현장사서 및 검색사에게 직접 검색질문을 의뢰하여 실험을 실시하였다. 적합성 판단은 검색결과순위에 따라 출력된 타이틀과, 간략사항만으로 어려운 경우가 있으므로 출력된 문서에 연결된 링크를 따라 해당 사이트로 가서 웹페이지 전체를 살펴봄으로써 적합여부를 판단하였다. 또한, 링크로 연결된 정보가 이용이 불가능한 경우(이를 deadlinks라 한다.)에 그 문서는 부적합한 문서로 처리하였다.

본 연구에서 사용된 검색엔진유형은 Metacrawler, Northernlight, Goolgle이다. 검색효율성 측정기준은 1)검색된 문헌의 수, 2)검색된 적합문헌의 수, 3)정도율, 4)재현율, 5)deadlinks정도이다. 여기서, 검색된 적합문헌의 수는 검색효율의 측정에 있어서 기본이 되는 가장 중요한 요소이다. 본 연구에서 검색된 적합한 문헌의 수를 측정하기 위해 사용된 문헌은 검색된 총 문헌중 상위 10개의 문서로 제한하였다. 재현률은 실제 어느 검색엔진의 재현율을 측정하기 위해서는 특정의 검색질문에 대한 전체시스템 내에서의 적합한 문헌의 전체수를 파악해야 하는데, 이는 현실적으로 불가능하다. 따라서, 본 연구에서는 재현율 대신에 상대재현율을 사용하였다. 상대재현율은 두 검색엔진을 이용하여 출력된 적합문헌양에 대해 각 검색엔진이 출력한 적합문헌양의 비율로 계산한다. Deadlinks정도는 각 검색된 문헌을 연결했을 때 연결이 안되고 error message(not find 404와 403 Forbidden, 505 Not gateway등)가 나타나는 것을 의미한다.

본 연구의 제한점으로는 첫째, 연구대상을 여러 검색엔진중 메타검색엔진으로 한정한 이유는 기존에 수행된 연구에서 메타검색엔진의 검색효율성이 타 검색엔진에 비해 높게 나타났음에 근거하였고, 이러한 메타검색엔진중 최초 그 검색효율성이 높은 Savvy Search와 Metacrawler를 선정하였으나, 연구수행도중 2000년 7월이후부터 기존에 서비스되던 Savvy

Search의 URL[1]이 변경되어 불가피하게 Metacrawler만을 연구대상으로 선정하였다. 또한, 데이타마이닝 기법을 이용한 검색엔진의 선정시 Northernlight과 Google에 한정한 이유는 기존의 Peggy Zorn (1999)과 다른 매체에서 이미 데이타마이닝 기법을 이용한 검색엔진으로 제시한 것에 근거하여 선정하였다.

1. 검색질문 구성

본 연구에서 사용한 검색질문은 기존의 이명희(1997)와 김성희(1997)의 검색효율 측정시 사용했던 검색질문과, SK투자개발사업그룹의 직원 33명을 대상으로 가장 요청빈도가 높았던 검색질문을 토대로 작성되었다. 이중 본 연구에서는 다음의 10개의 질문을 채택하여 검색엔진의 검색효율을 측정하였다.

질문1. 사회내에서의 자원봉사자의 실태에 관한 자료를 검색하시오.
질문2. 세계 통신업계에 관한 신기술 관련정보, 각국 통신정책, 대표적 통신 사업체 등 통신분야에 관련된 정보를 검색하시오.
질문3. 미국의 2000년도 동절기 기상전망에 관한 자료를 검색하시오.
질문4. 영국런던대학 산하교육연구소에 관한 연구정보를 검색하시오.
질문5. 코코넛 크림파이를 만들기 위한 재료를 검색하시오.
질문6. 제약산업의 규모를 알수 있는 향후 전망자료를 검색하시오.
질문7. 자동차 리콜실적에 관한 자료를 검색하시오.
질문8. 이동통신을 이용한 광고현황에 관한 자료를 검색하시오.
질문9. 캐릭터를 활용한 스포츠마케팅 사례에 관한 자료를 검색하시오.
질문10. 미국의 유아전문교육기관에 관한 자료를 검색하시오.

2. 검색과정 및 검색식 작성

위의 검색엔진에 관한 일반적인 내용에서 살펴본 바와 같이 Metacrawler, Google, Northernlight은 기본검색(basic search), 확장검색(advanced search)의 제공, 불리언 연산자

1) 기존에 서비스되던 http://www.cs.colostate.edu/dreiling/smartform.html에서 현재는 Cnet社에 통합되어 http://www.cnet.com에서 서비스되고 있음.

의 적용, 우선순위를 적용한 검색출력물의 배열등의 공통점이 있었다. 반면에 색인기법이나 확장검색의 검색형식면에서 상이한 검색형식을 제공하고 있었다. Metacrawler의 경우 확장검색에서 제공해주는 검색형식(search tip)이 검색엔진과 페이지, 검색속도로 제한을 하는 반면에 Google의 경우는 단어(words)와 구(phrase)등의 검색형식(search tip)을 제공하고 있었으며 Northernlight의 경우에는 일반적인 키워드 검색엔진에서 제공하는 제목(title), 출판사(publication), URL등의 검색형식(search tip)이 제공되는등 상이한 확장검색형식(seardh tip)을 제공하고 있었다.

상이한 검색엔진의 검색형식(search tip)에 대해 일관성 있는 실험을 수행하고자 상이의 3개 검색엔진의 검색형식을 기본검색(basic search)의 조건으로 통일하고 불리언 연산자의 적용을 검색질문에서 포함시켜 적용하였다. SK직원을 대상으로 수집된 검색질문의 경우에는 신사업의 개발과 관련된 검색을 의뢰한 질문이 많은 관계로 적합문헌을 선별하는데 있어 기검색질문에 비해 생소하고 전문적인 용어를 인지하고 있어야 적합문헌을 선별한는데 전문검색사의 도움이 필요했다. 이에 현재 SK그룹의 전문정보검색사와 숙명여자대학교 참고정보실에 근무중인 사서를 대상으로 본 검색질문에 대한 검색과정을 검증하였다.

구체적인 검색식 작성은 다음과 같다.

질문1. 사회내에서의 자원봉사자의 실태에 관한 자료를 검색하시오.
검색식 : volunteerism and society

질문2. 세계 통신업계에 관한 신기술 관련정보, 각국 통신정책, 대표적 통신사업체 등 통신분야에 관련된 정보를 검색하시오.
검색식 : telecommunication and (resource* or policy)

질문3. 미국의 2000년도 동절기 기상전망에 관한 자료를 검색하시오.
검색식 ; weather and forecast and annual and US 2000

질문4. 영국런던대학 산하교육연구소에 관한 연구정보를 검색하시오.
검색식 ; "London University" and "Institute of Education"

질문5. 코코넛 크림파이를 만들기 위한 재료를 검색하시오..
검색식 ; recipe coconut cream pie

질문6. 제약산업의 규모를 알수 있는 향후 전망자료를 검색하시오.
검색식 ; Industry pharmacy future trend*

질문7. 자동차 리콜실적에 관한 자료를 검색하시오.
검색식 ; car recall* -child

질문8. 이동통신을 이용한 광고현황에 관한 자료를 검색하시오.
검색식 ; mobile advertising(advertisment)

질문9. 캐릭터를 활용한 스포츠마케팅 사례에 관한 자료를 검색하시오.
검색식 ; character "sports-marketing" case*

질문10. 미국의 유아전문교육기관에 관한 자료를 검색하시오.
검색식 ; childhood education USA +organization

우선, 앞서 제시한대로 기존 연구와 수집된 검색질문을 토대로 2명의 검색사에게 의뢰, 검색을 수행하였다. 각각의 상이한 검색출력건수는 당연한 결과이므로 먼저 예비 검색수행을 통해 검색질문 및 검색과정에 대해 숙지를 한 후에 질문식에 대한 토의를 거친뒤 실제 검색을 수행하게 되었다.

Ⅳ. 데이터분석

1. 검색된 문헌의 양

3개의 검색엔진에 의해 검색된 문헌의 양은 <표 1>에서 보는 바와 같이 질문마다 다른 양상을 보여주고 있으나 대체로 메타검색엔진인 Metacrawler에 의해 검색된 문헌의 양이 기타 2개의 검색엔진인 Google과 Northernlight에 의해 검색된 문헌의 양에 비해 월등히 적게 나타나고 있었다. 이는 메타검색엔진의 특성상 자체 데이터베이스를 보유하지 않고 다른 검색엔진에 검색을 의뢰하여 짧은 시간에 이를 토대로 검색결과를 제공해주며 각각의 검색엔진에서 제공되는 동일한 검색결과를 하나의 데이터로 간주하여 제공해주므로 자체 데이터베이스를 구축하여 제공하는 Google과 Northernlight에 비해 검색된 문헌의 양이 상대적으로 적게 나타났다.

<표 1> 검색된 문헌의 양

	질문1	질문2	질문3	질문4	질문5	질문6	질문7	질문8	질문9	질문10
Metacrawler	54	27	30	17	66	23	35	55	33	34
Google	33,900	59,800	48,100	1,270	11,100	5,850	246,000	405,000	296	28,400
Northernlight	25,716	1,258,262	45,949	2,883	11.434	22,590	356,369	307,960	503	14,264,723

2. 검색된 적합문헌의 양

각 검색엔진이 검색해낸 문헌의 양 중에서 출력순위 상위 10위까지를 판단해 보았을 때 <표 2>에서 보듯이 각 질문당 검색된 적합문헌의 평균은 Metacrawler은 3.0건이고 Google의 경우는 5.3건이며 Northernlight의 경우에는 5.2건으로 나타났다. 각각의 질문에서 출력된 적합문헌에 대해 살펴보면 질문6을 제외하고는 대체로 Google과Northernlight은 Metacrawler에 비해 평균 2배정도의 높은 적합문헌을 검색해내고 있었다. 이는 Google과 Northernlight가 출력해낸 많은 양의 문헌중에서도 각각에서 독특하게 개발 제공하고 있는 적합성 순위방식을 채택하여 적합문헌을 출력해내고 있음을 보여주는 것이다. 즉 Google의 경우에는 관련성 피드백시스템을 이용한 순위부여방식과 Northernlight의 경우에는 디스크립터를 색인대상으로 하여 부여된 적합성 퍼센테이지가 높은 문헌을 출력해 내는 방식을 채택하고 있다. 각각의 검색엔진에서 채택하고 있은 높은 적합성순위방식을 이용하여 검색결과 출력시 상위에 리스트화해줌으로써 검색수행자에게 보다 효과적이고 정확한 검색문헌을 제공하고 있다.

<표 2> 검색된 적합문헌의 양

	질문1	질문2	질문3	질문4	질문5	질문6	질문7	질문8	질문9	질문10	평균
Metacrawler	6	4	4	-	6	4	2	2	-	2	3.0
Google	7	8	7	2	8	4	5	5	1	6	5.3
Northernlight	7	9	5	3	8	5	4	4	3	4	5.2

3. 재현률

아래의 <표 3>은 이 연구에서 사용한 검색엔진인 Metacrawer와 Google, Northernlight의 측정된 재현율을 요약한 것이다. Metacrawler의 평균 재현률은 0.19이며 Google과 Northernlight의 경우에는 각각 0.39와 0.42로서 전자의 검색엔진에 비해 후자의 검색엔진의 재현율이 높게 나타났다. 후자의 검색엔진인 Google과 Northernlight간에는 서로 유사한 재현율 수치를 보여주고 있는 반면에 전자의 Metacrawler와는 현저한 차이를 보였다. 앞에서도 설명하였듯이 Google과 Northernlight가 출력해낸 많은 양의 문헌중에서도 각각에서 독특하게 개발 제공하고 있는 적합성 순위방식을 채택하여 적합문헌을 출력하기 때문인 것으로 보인다.

<표 3> 재현률

	질문1	질문2	질문3	질문4	질문5	질문6	질문7	질문8	질문9	질문10	평균
Metacrawler	0.3	0.19	0.25	-	0.27	0.31	0.18	0.18	-	0.17	0.19
Google	0.35	0.38	0.44	0.4	0.36	0.31	0.45	0.45	0.25	0.5	0.39
Northernlight	0.35	0.43	0.31	0.6	0.36	0.38	0.36	0.36	0.75	0.33	0.42

4. 정도률

<표 4>에서 보는바와 같이 Metacrawler의 평균 정도률은 0.3이고, Google과 Northernlight의 평균정도률은 각각 0.53와 0.52로 나타났다. Google과 Northernlight의 경우에는 서로 유사한 평균정도률을 보이고 있는 반면에, Metacrawler와는 대략 0.5정도의 현격한 차이를 나타냈다. 본 연구에서 제시하고자 하는 연구결과가 기존의 검색엔진에 비해 새로운 색인방식 즉, 웹페이지의 로그분석을 통한 웹 마이닝기법을 이용하고 있는 Google과 텍스트 마이닝기법을 이용하고 있는 Northernlight가 보다 방대한 웹페이지를 대상으로 정확한 문헌을 검색하고 있다.

<표 4> 정도률

	질문1	질문2	질문3	질문4	질문5	질문6	질문7	질문8	질문9	질문10	평균
Metacrawler	0.6	0.4	0.4	-	0.6	0.4	0.2	0.2	-	0.2	0.3
Google	0.7	0.8	0.7	0.2	0.8	0.4	0.5	0.5	0.1	0.6	0.53
Northernlight	0.7	0.9	0.5	0.3	0.8	0.5	0.4	0.4	0.3	0.4	0.52

5. deadlinks

사이트의 연결에러를 나타내는 deadlinks항목에 대해서는 단연 Metacrawler가 높은 수치를 나타냈다. 이는 Metacrawler가 검색을 의뢰하는 개별검색엔진의 순간적인 다운여부와 가장 밀접한 관련이 있는 사항으로 비교대상인 타 검색엔진에 비해 높은 수치를 기록하는 것은 당연한 결과라 하겠다. 반면에 Google에서는 cached라는 독특한 기능을 이용하여 사이트의 Not found 404 및 505에러를 방지하며 사이트의 연결여부에 대한 한층 나은 서비스를 제공하고 있었으며 Northernlight의 경우 타 검색엔진에 비해 중복문헌과 deadlinks출력건수가 적

은 것으로 나타났다. 이는 최신성의 유지라든가 정확한 검색결과를 보여주는 성능면에서 Metacrawler와 Google에 비해 높다는 것을 의미한다.

<표 5> deadlinks의 수

	deadlinks
Metacrawler	16
Google	10
Northernlight	8

6. 각 질문에 대한 개별분석

질문 1항과 2항은 일반적인 동향분석자료를 요구하는 질문사항으로서 각 검색엔진으로부터의 결과 질문1에 대해 재현률은 Metacrawler 0.3, Google과 Northernlight은 0.35로 나타났으며 질문2에 대해서는 각각 앞의 순서대로 0.19, 0.38, 0.43으로 나타났다. 정도률은 2질문 모두 비교적 높은 것으로 나타났으며 특히 질문2항의 경우에는 Northernlight가 Google에 비해 높게 나타났는데 이는 일반적인 동향자료에 대한 검색일 경우 Northernlight이 적합하다는 반증이라고 할수 있다.

질문3,4,5항은 구체적인 수치정보를 요구하는 질문사항으로써 각각에 대해 살펴보면 먼저 질문3항의 키워드와 and연산자를 계속해서 조합한 질문식구성과 2차정보원이라 할 수 있는 디렉토리가 출력된 경우 부적합한 문헌으로 판정한 본 실험과정에서 Metacralwer는 상당수의 2차정보원이 디렉토리를 출력하고 있었으며 Google과 Northernlight의 경우에는 디렉토리보다는 해당 키워드가 존재하는 적합한 문서를 출력해냈다. 질문4항의 경우에서도 Metacralwer은 한번의 클릭과정을 요구하는 디렉토리가 출력되어 적합한 문헌을 한건도 출력해내지 못했다. 이는 여러가지 다양한 옵션기능을 제공하지 못하고 단지여러 검색엔진에 의뢰해 검색결과를 보여주는 메타검색엔진이 질문을 의뢰한 각 개별 검색엔진에서 해당 키워드에 의존한 검색결과를 제공해주고 있다는 것을 의미하는 것이다.

질문5항과 6항의 경우에는 각 검색식에서 연산자를 적용하지 않고 키워드와 띄어쓰기를 통해 검색질문을 구성하여 검색을 수행해 본 결과 2항목 모두 출력된 검색결과가 띄어쓰기를 and 연산자로 인식하여 검색결과를 출력하고 있었다. 특히 질문 5항의 경우에는 실험에서 사용된 검색질문 중 가장 정도률이 높은 결과를 출력해내고 있는 반면에 재현률은 Metacrawler가 0.27, Northernlight과 Google이 각각 0.36으로 나타났으며 여러 질문중 가장 적합여부 판정이 용이했다.

본 질문은 각 검색엔진 모두 검색질문에서 요구한 키워드와 띄워쓰기를 and연산자로 인식

하여 검색을 수행하여 적합성 기준이 된 해당 코코넛크림파이를 만들기 위한 재료가 수록된 문헌을 출력하였다.

질문 7항의 경우는 구체적인 실적수치를 요구한 질문사항으로써 적합성 판단 기준이 앞의 5항처럼 구체적인 수치자료가 나온 경우를 적합한 문헌으로 간주하였기 때문에 비교적 적합 판정이 용이했던 질문에 해당되었으며 Metacrawler검색엔진에 비해 데이터마이닝 기법을 이용한 검색엔진인 Google과 Nortemlight가 2배정도의 정확한 검색결과를 출력하였다. 이는 Google의 PageRank기법과 Northernlight의 주제전문사서에 의해 제공해주는 주제폴더검색에 기인된 결과이다.

질문8항과 9항은 새로운 신기술 및 기술사업에 대한 동향분석 자료를 요구하는 질문항목으로서 질문9항의 결과처럼 비교적 검색된 문헌의 양도 적었으며 또한 적합여부를 판단하기에 어려움을 겪은 질문항목 이었다. 또한 이 신기술이 어느 한분야에 속한 기술이나 사업이라기 보다는 여러 분야 즉, 통신과 광고가 접목되어 새롭게 출현하게 된 분야여서 2분야와 관련된 망라적인 자료를 대상으로 검색을 수행해야만 했다.

이렇게 한 주제분야보다 이와 관련된 주제분야를 브라우징(browsing)을 통해 적합문헌을 검색하기에는 Northernlight을 이용하는 것이 적합하였는데 이는 해당검색엔진이 주제전문사서에 의해 분석된 이용자 검색패턴을 고려하여 제공되는 주제폴더 검색방식이 제공되기 때문이다.

질문10항은 각 검색엔진마다 디렉토리 출력빈도가 상당히 높았는데 이는 유아전문교육기관(childhood education)이라는 키워드가 일반적인데서 기인된 결과라 할 수 있다. 질문에 적합한 문헌을 판단하기 위해 개별 검색결과를 클릭하여 해당페이지에 접속해 본 결과 Google이 타 비교검색엔진 보다 정확한 검색결과를 출력하고 있었는데 이는 10억이상의 웹페이지의 목차를 저장하였다가 이용자가 검색을 실행하면 다른 페이지의 링크횟수에 따라 순위가 매겨진 순서대로 검색결과를 보여주기 때문에 상위 10건으로 제한한 검색리스트에서 보다 적합한 문헌의 수가 많이 출력된 것이라 생각한다.

7. 데이타마이닝 기법을 이용한 검색엔진의 효율성 측정결과 해석

앞에서 살펴보았듯이 데이타마이닝 기법을 이용한 검색엔진이 일반 메타검색엔진에 비해 높은 검색효율성을 보여 주었다. 이는 데이타마이닝 기법의 특성때문이었다고 해석된다. 그 내용을 살펴보면 다음과 같다.

(1) 데이타마이닝 기법을 이용한 검색엔진들의 공통점은 검색결과의 적합성을 "인간의 판단"에 기초하고 있다는 점이다.

이는 데이타마이닝 기법의 특성이라고 할 수 있는데, 먼저 Google의 접근은 인터넷의 링크 구조에서 그 해답을 찾고 있다. 특정 키워드를 포함하고 있는 웹페이지 A와 웹페이지 B가 있다고 하자. 그런데 웹페이지 A를 많은 사이트들이 링크하고 있고, 웹페이지 B는 거의 링크되어 있지 않다고 한다. 이러한 구조에서 구글은 특정 키워드에 대해 웹페이지 A가 보다 적합하다고 판단한다는 것이다.

이렇게 하여 검색순위를 정하는 Google의 방식에서는 링크가 많은 웹페이지가 일반적으로 사람들이 우수하다고 평가한다는 가정에 기초한 것이다. 즉, 우수하니까 많은 다른 웹페이지들이 링크를 하고 있을 것이라는 생각이다. Northernlight의 경우에는 전문사서에 의해 미리 주제 분류된 폴더에 각 정보를 배열하여 검색결과를 보여주는 것이다. 또한 해당주제와 관련된 분야를 정리하여 제공함으로써 기존의 검색엔진에서는 볼 수 없었던 사용자들의 정보나 문서에서 추론하여 결합한 새로운 정보를 생성하는 이른바 지식검색이 가능해졌다는 것이다.

(2) 데이터마이닝기법을 이용한 검색엔진은 기존의 검색엔진의 단점을 보완할 수 있는 검색서비스라는 것이다.

최근의 검색엔진들에게 문제로 대두되고 있는 재현율과 정도율에 대한 문제를 보완한 검색서비스를 제공하고 있다. 즉 대부분의 검색엔진들이 자체기술개발을 통해 재현률은 좋지만 상대적으로 정도률이 떨어지는 검색서비스를 제공하고 있다. 검색엔진수가 현재 수백여개로 증가하고 있는 현재에는 재현율보다는 정도율이 무엇보다 검색엔진에 있어 중요하다고 하겠다. Google의 PageRank 기술을 도입한 검색결과의 배열과 이를 이용한 관련성 피드백(Relevance Feedback)의 랭킹구조 선정은 이러한 기존 검색엔진의 단점을 보완해주며 검색이용자에게 보다 정확한 검색내용을 이용할 수 있게끔 제공되는 검색서비스라고 할 수 있으며, Northernlight의 경우 검색과 관련된 주제의 기(旣)분류된 주제폴더를 제시함으로써 관련 주제에 대한 포괄적인 내용 검토가 가능한 검색서비스를 제공하고 있다.

또한, Google의 경우 중국어, 일본어등 전세계 25개국의 언어를 지원하는 언어옵션을 통해 기존의 검색엔진에서 간과하였던 비영어권 웹페이지까지도 검색 및 색인할수 있게 되었으며 5억6천의 웹페이지를 검색 인덱스(search index)로 구축, 제공하게 되었다.

(3) 일반적인 내용과 구체적인 내용에 따라 다른 검색결과를 보여주고 있다.

다음의 <표 6> 데이타마이닝 검색엔진의 비교는 데이터마이닝기법을 이용한 검색엔진인 Google 과 Northernlight를 중심으로 앞서 검색을 수행한 10개의 검색질문을 토대로 일반적인 내용과 구체적인 내용을 검색하는데 있어서 각 검색엔진이 얼마나 그 기능을 효과적으로 수행하고 있는 가를 비교해 본 결과이다.

<표 6>에서 나타나듯이 브라우징을 통한 주제에 관한 포괄적인 내용을 검색하고자 할 경우에는 Northernlight을 이용하여 검색을 수행하는 편이 보다 효과적이며 구체적인 수치등을 이용한 검색을 원할때는 Google을 이용한 검색이 보다 효과적이었음을 알 수 있었다. 이는 각각의 검색엔진이 전략적으로 서비스하고 있는 내용과 기술과 관련된 결과로써, Northernlight의 경우에는 검색키워드와 관련된 내용을 검색결과를 보여주는 결과화면 좌측에 주제폴더(blue foledr)형식으로 제공해줌으로써, 주제와 관련된 타 영역을 브라우징(browsing)을 통해 포괄적인 내용 검색이 가능하였다. 즉, -에 관한 동향정보, 추세, 동향분석등의 자료를 원할 경우 효과적이었다. 반면에, Google은 구체적인 수치를 나타내고, 검색자가 자신이 검색하고자 하는 내용을 구체적이고 정확하게 인지하고 있는 경우에 이용하면 효과적인 결과를 얻을 수 있었다. Google이 새롭게 개발하여 독특하게 서비스하고 있는 PageRank기술을 이용한 검색결과의 출력이 가능한 서비스를 제공하고 있으므로 검색키워드로써 자신이 찾고자하는 내용을 검색키워드로 표현하여 검색을 수행하면 만족할 만한 검색결과를 얻을 수 있는 검색서비스를 제공하고 있었다.

(4) 새로운 영역에 대한 검색서비스를 제공하고 있다.

Google과 Northernlight는 기존의 검색엔진이 배제해 왔던 아시아권의 웹문서 검색을 가능하게 하고 있고 최근 이슈화되고 있는 B2B자료를 검색할 수 있는 검색공간을 제공하고 있다. 최근 Yahoo와 Lycos등이 검색서비스로 출발하여 새로운 전자상거래 수행이 가능한 비즈니스모델을 채택하여 정보검색엔진 본연의 검색서비스 제공 보다는 컨텐츠, 커뮤니티, 커머스를 제공해주는 포탈사이트로 지향하고 있다. 그러나 Google과 Northernlight의 경우는 이상의 검색엔진들과는 약간 차별화된 접근을 시도하면서 포탈이라는 개념보다는 검색서비스에만 집중하고 있다. Google의 경우에는 그 흔한 기업광고나 주식시세정보는 볼 수 없고 유난히 빈공간이 많은 단순한 디자인과 여백이 많다. 이러한 이유로 대부분의 다른 검색엔진에 비해 자료를 내려받는 속도가 빠르며 이 때문에 Google은 검색요청을 처리하는데 걸리는 시간을 화면 상단에 제공해주고 있다.

<표 6> 데이타마이닝 검색엔진의 비교

<일반적인 내용>				
	Meta	Google	Northern	계
1. 사회내에서의 자원봉사자의 실태에 관한 자료를 검색하시오.	6	7	7	
2. 세계 통신업계에 관한 신기술 관련정보, 각국 통신정책, 대표적 통신사업체등 통신분야에 관련된 정보를 검색하시오.	4	8	9	
6. 제약산업의 규모를 알수 있는 향후 전망자료를 검색하시오.	4	4	5	
8. 이동통신을 이용한 광고현황에 관한 자료를 검색하시오.	2	5	4	
9. 캐릭터를 활용한 스포츠마케팅 사례에 관한 자료를 검색하시오.	0	1	3	
10. 미국의 유아전문교육기관에 관한 자료를 검색하시오.	2	6	4	
	18	31	32	81
<구체적인 검색내용>				
3. 미국의 2000년도 동절기 기상전망에 관한 자료를 검색하시오.	4	7	5	
4. 영국런던대학 산하교육연구소에 관한 연구정보를 검색하시오.	0	2	3	
5 .코코넛 크림파이를 만들기 위한 재료를 검색하시오.	6	8	8	
7. 자동차 리콜실적에 관한 자료를 검색하시오.	2	5	4	
	12	22	20	54
계	30	53	52	135

V. 결 론

21세기의 가치평가는 자산위주가 아닌 조직내부에 축적되어 있는 지식에 따라 평가되는 지식중심사회가 될 것이라는 예견과 더불어 이에 대한 실천적 방안으로서 지식습득, 공유 및 활용을 위한 인프라인 지식관리시스템 구축을 제시하고 있다. 지식관리시스템의 효과적인 구축 및 활용을 위하여 최근들어 데이타마이닝 기법을 많이 적용해 왔다.

지식의 개념을 노나카의 암묵지(tacit knowledg), 칼 포퍼(K.popper)의 주관적 지식의 개념에서 유추한 연구가 이미 선행되었고 이에 더욱 선행되어 새로운 경영기법이기 보다는 조직내에 이미 여러 형태로 존재하고 있는 자원을 의미한다고 주장하고 이러한 지식을 월드와이드웹, 온라인서비스, 분산하이퍼 텍스트 시스템, 인트라넷과 같은 새로운 정보기술로서 정보인프라를 구축하여 조직체 구성원들의 지식공유, 정보기술을 유통시킬수 있는 "지식창고(data warehouse)" 를 마련할 수 있는 정보인프라의 체계도가 제시되고 있다.

데이타마이닝 기법은 이러한 방대한 양의 데이타 속에서 쉽게 드러나지않는 유용한 정보를 찾아내는 과정, 데이타간의 숨겨진 관계, 혹은 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 뽑아내기에는 너무나 복잡한 관계를 찾아내고 이 관계를 바탕으로 앞날을 예측하는 기술이며 대용량의 데이타로부터 이들 데이타 내에 존재하는 관계, 패턴, 규칙등을 탐색하고 찾아내어 모형화함으로서 유용한 지식을 추출하는 일련의 과정이다.

데이타마이닝을 효율적으로 수행하기 위하여 시계열분석등 각종 통계기법과 데이타베이스 기술 뿐만 아니라 산업공학, 신경망, 인공지능, 전문가시스템, 퍼지논리, 패턴인식, 기계적 학습(machine learning), 불확실성 추론(reasoning with incertainty), 정보검색에 이르기까지 각종 정보기술과 기법들이 적용되고 있다. 그 중 정보검색분야에서 데이타마이닝 기법이 적용된 사례를 중심으로 본 연구에서는 데이타마이닝 기법을 이용한 검색엔진의 효율성과 메타검색엔진의 효율성과 비교해 보았다.

데이타마이닝 기법을 이용한 검색엔진과 기존의 검색엔진의 검색 효율성을 측정한 결과를 요약하면 다음과 같다. 첫째, 각 검색엔진이 검색해낸 문서중에 상위 10개를 기준으로 적합성 판단을 했을 경우, 각 질문당 검색된 적합문헌의 평균은 Google의 경우 5.3건이고, Northernlight의 경우 5.2건으로 그리고 Metacrawler의 경우 3.0으로 나타났다. 둘째, 재현율과 평균정도률을 측정해 본 결과 Google과 Northernlight가 Metacrawler보다 높은 것으로 나타났다. 이는 데이타마이닝 기법의 특성인 인간의 판단에 기초를 둔 테크닉이 검색엔진에 도입되어 기존의 검색엔진보다 더욱 정확한 검색 결과를 나타내고 있다고 할 수 있다. 셋째, Metacrawler와 Google, Northernlight의 중복문헌수와 deadlinks를 조사해 본 결과, 각각 8건,9건,3건과 16건,10건,8건으로 나타났는데, 이러한 결과는 데이타마이닝 기법을 이용한 검색엔진이 기존의 검색엔진보다 최신성 유지라든지, 정확한 검색결과를 보여주는 성능면에서 높은 결과를 보이고 있음을 알 수 있다. 이러한 결과 역시 데이터마이닝기법의 특성 때문인 것으로 보인다. 즉, Google의 PageRank기법을 이용한 질문에 대한 검색결과의 배열과 Northernlight의 주제별로 분류된 폴더의 배열등을 그 특성으로 들 수 있다. 이러한 특성은 기존의 검색엔진의 단점으로 제시되는 재현률을 높힘과 동시에 상대적으로 정도률을 높일 수 있는 검색서비스를 제공할 수 있을 것이다. 예를 들면, 구글의 cached기능의 활용과 GoogleScout를 이용한 출력된 결과치와 유사한 페이지를 자동적으로 보여주는 similar page

기능을 활용한 관련성 피드백(relavance feedback)은 이러한 기존 검색엔진의 단점을 보완해 주는 서비스을 제공할 수 있으리라 생각된다.

참 고 문 헌

강현철, 박태원, 임난희, 1998. Data Mining 방법론과 SAS Enterprise Miner, 한국분류학회 발표논문집.

김성회, 1997. 인터넷상의 메타검색엔진의 검색효율성 비교연구,도서관학논집,제27집

김성회, 1999. 지식관리시스템의 단계별 분석 및 구축방안에 관한 연구, 정보관리학회지 제16권, 제2호.

김현회, 배금표, 안태경공저, 1999. 정보검색론. 오름시스템.

노정란, 1999. 지식경영과 정보인프라, 정보전문가의 관계. , 한국비블리아 제9집

박민우, 2000. 인터넷 세상, 검색엔진이 주도한다 - 검색엔진 과거와 현재 그리고 미래, 마이크로 소프트웨어 2000년 3월호 특집기사

이명희, 1997. "네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 검색엔진의 검색효율에 관한 탐색적 연구", 한국문헌정보학회지, 31(1) :176-197

장남식, 홍성완, 장재호, 1999. 데이터마이닝, 대청

정영미, 김성은, 1997. "WWW탐색도구의 색인 및 탐색기능 평가에 관한 연구.", 한국문헌정보학회지, 31(1) : 153-184

진휘철, 2000. 데이터마이닝은 우리에게 어떤 이득을 주는가? , 삼성 SDS IT Review.

조재희, 박성진, 1996. 데이타 웨어하우징과 OLAP, 대청출판사.

Adriaans, P.and Zantinge,D. 1998 Data Mining, Syllogic .

Chu, H and M. Rosenthalml, 1996. "Search Engines for the World Wide Web ; A Comparative Study and Evaluation Methodology." Presented at the 96' ASIS Conference.

Courtois, M.P. et al., 1995. Cool Tools for searching the Web : A performance Evaluation, Online, 19(60 : 14-32

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, 39(11), 27-34

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. & Uthurasamy, R., 1996. Advances in Knowledge Discovery and Data Mining. Cambridge: AAAI/MIT Press.

Gordon Linoff, 1999. Data Mining *Inform*; Silver Spring.

Inmon,W.H., 1996. Building the Data WareHouse(2nd Ed.), John Wiley & Sons, Inc.

Kimmel, S., 1996. "Robot-generated Database on the World Wide Web." Database. 19(1) : 40-49.

Limb, P.R., & Meggs, G.J., 1995. "Data Mining -Tools and Techniques" British Telecom Technology Journal, 12(4), 32-41.

Lloyd-Williams, M., Jenkins, J., et al., 1995. "Knowledge Discovery in an Infertility Database Using Artificial Neural Networks" IEE Colloquium on Knowledge Discovery in Databases. IEE Digest.

Maurice D Mulvenna, 2000. Personalization on the Net using Web mining; Association for Computing Machinery. Communications of the ACM, New York, Vol. 43.

Michael Lioyd, Williams., 1997. Discovering the Hidden Secrets in Your Data--the Data Mining Aproach to Information, Department of Information Studies University of Sheffield.

Peggy Zorn et al., 1999. Mininig meets the Web, Online, 23(3).

Poe, V.,1994. "Guidelines for Warehouser Development.", Database Programing & Design, September.

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

<인터넷>

http://a-pex.co.kr/choice.htm

http://www.cio.co.kr

http://medric.chungbuk.ac.kr/bioinfo/1400_01.htm

http://human21.new21.org/dataMining/dm_2.htm

http://human21.new21.org/datamining/dm_4.htm

http://www.google.com

http://www.metacrawler.com

http://www.northernlight.com

http://my.netian.com/~kylim/interest/datamining/index.htm

http://dblab.chungbuk.ac.kr/~damine/References/dbworld9709spec.html