인터넷 홈페이지 검색시스템 구현과 검색효율 향상

박현주, 최재덕, 강상배, 박승, 박용욱, 권혁철

Implementation of an Internet Homepage Retrieval System and Improvement of Retrieval Efficiency

Hyun-Joo Park, Jae-Duck Choi, Sang-Bae Kang, Seung Park,
Yong-Uk Park, Hyuk-Chul Kwon

Dept. of Computer Science, Pusan National University

요 약

이 논문은 인터넷 홈페이지를 검색하는 정보검색시스템인 *미리내 시스템*을 제시한다. 웹문서의 특성을 고려하여 로봇의 기능을 확장하고,색인, 등록, 수정, 삭제, 분류의 자동화를 구현하여 관리효율을 높인다. 자동화에 따른 문제점과 해결방법을 제시하고, 불리언질의검색 외에 자연언어질의 검색에서 질의어 확장의 방법으로 웹페이지 링크속성검색, Relevance feedback을 통한 검색효율을 높인다.

1. 서론 [1)]

웹이 활성화하기 전부터 논문 등을 대상으로한 정보검색시스템의 연구가 진행되어왔다. 인터넷(웹)이 활성화되면서 기존의 정보검색시스템을 웹에 적용시킨 인터넷정보검색시스템의 개발이 활발해졌다.

이러한 과정에서 웹문서의 수집을 위해, 로봇에이전트의 사용이 추가되었으며, 몇몇을 제외한 대부분의 웹정보검색시스템은 인터넷로봇을 통하여 등록문서를 수집한다. 그리고 기존의 검색시스템에서와 같은 구조로 관리자가 등록을 하고, 문서를 분류한다.

그러나 웹문서는 일반적인 논문, 신문, 도서 등과는 다른 특성을 가지고 있다.

◆ 웹문서의 특성

① 분산성 : 문서의 위치가 한 곳에 모아져 있는것이 아니고 웹 전체에 흩어져있다.

② 연결성 : link를 통해 directed graph형태로, 문서단위의 내용별 연결이 되어있다.

③ 순간성 : 문서가 새로 생기고, 사라지는 작업이 빈번하다.

④ dangling URL : 참조를 하는 문서에서는 존재하고 있다고 생각하는 문서가 사라지고 없는 경우가 있다.

⑤ 비계층성 : 내용별 연결이 되어있는 구조이므로, 기존의 이름별 디렉토리 구조인 tree형태의 구조가 아니라 그래프 구조로 되어있다.

위의 특성으로 인하여 웹문서검색시스템은 기존의 정보검색시스템에 추가하여 다음과 같은 기능이 추가되어야 한다.

◆ 웹문서검색시스템에서 필요한 기능과 검색요구자에게 제공해야하는 기능.

① 분산되어있는 문서의 수집 기능.

② 순간성과 dangling URL로 인한 구축된 데이타베이스 실데이타간의 불일치를 없애는 메카니즘이 필요.

③ 비계층적인 문서의 구조를 내용별 tree형태의 구조

본 연구는 한국과학재단 산학협력연구비 (962-0100-001-2) 지원으로 수행되었으며 지원에 감사를 드립니다.

로 분류, 제시하는 기능.

④ 방대한 양의 검색문서에서의 정확도 높은 검색기능.

⑥ 방대한 데이타의 높은 압축기능과 빠른 접근기능.

이 논문에서는 위에서 언급한 웹문서의 특성을 고려하여 정보검색시스템을 구현하고, 필요한 기능들을 구현한다. 그리고 그에 따른 문제점과 해결방법을 제시한다.

2. 기존검색시스템의 문제점과 요구사항 해결

기존의 검색시스템을 웹검색시스템으로 확장시키는 데는 크게 2가지의 문제점이 대두된다. 이는 웹문서의 특성으로 인해서 발생하는데, 다음과 같이 나누어 볼 수가 있다.

① 등록상의 문제점 :

웹문서는 분산성으로 인하여 흩어져 있는 문서를 수집하는 역할이 필요하다. 이를 위하여 인터넷 로봇을 이용한다. 그러나 기존의 인터넷 로봇은 단순히 문서를 수집해오는 기능만을 담당하고, 로봇이 모아온 문서를 일정시간 후에 관리자가 등록을 한다.

이러한 메카니즘으로는 웹문서의 순간성과 dangling URL의 문제를 효과적으로 해결할 수가 없다. 따라서 없어진 문서와 바뀐 문서를 그대로 검색시스템은 저장을 하고 있게된다. 따라서 관리자의 관리효율을 떨어뜨리게 되고, 사용자의 검색요구에 최신의 정보를 제공하지 못하므로 결과적으로 검색효과를 떨어뜨리게 된다.

② 검색상의 문제점 :

웹문서는 그 양이 방대하다. 한국 내의 웹문서만 하더라도 450만건이 넘는 문서가 존재한다. 따라서 관련된 문서의 수 또한 방대하므로, 관련된 문서를 많이 찾아주는 것보다, 관련된 문서를 높은 우선순위로 찾아주는 것이 중요하다.

제시한 문제점을 해결하기 위하여 이 논문에서 제시하는 미리내시스템은 로봇을 이용한 자동등록시스템을 구현하여 등록과 관리효율을 높이고, 웹링크속성 검색과 사용자 Relevance Feedback검색을 이용하여 검색효율을 높인다.

3.미리내 시스템

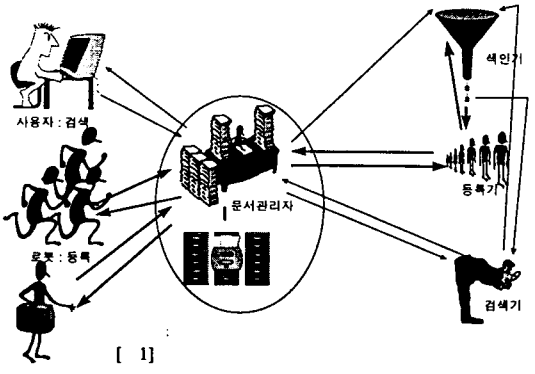
3.1. 전체 시스템 구조

이 논문에서 구현한 미리내 시스템은 인터넷로봇, 웹문서관리자, 색인기, 검색기, 등록기의 5개의 부분으로 구성되어있다[그림1]. 이들 각각은 독립된 서버로서 실행되며, 인터넷 소켓을 통하여 통신을 한다.

각 구성 요소가 독립된 서버로 동작하기때문에 시스템의 확장성이 높고, 이식성이 높아진다. 각각의 시스템이 독립된 컴퓨터에서 수행될 수도 있으므로, Computing Power가 약한 컴퓨터에서는 분리시키는 것이 좋다.

3.2. 로봇의 역할 확대

웹문서의 특성을 고려하여 로봇의 기능을 문서수집기능에서 다기능으로 확장하였다. 기능이 분리된 로봇이 다수



[1]

프로세스로 존재하며, 데이타를 공유하는 형태로 구성되어 있다. 사전관리자를 통하여 데이타를 공유하므로, 어떠한 수의 로봇이라도 수행이 가능하다. 따라서 네트워크 상황이나 시스템 상황에 따라서 프로세스 수를 조정할 수 있다.

◆ 로봇의 기능

① 웹문서의 수집

HTTP로 웹서버에 접속해서 문서를 수집해오는 로봇의 기본적인 기능을 제공한다. Gatherer로봇으로 제공된다.

② 웹문서의 변환기능

HTML Parser를 통해서 전문(Full Text)색인을 위한 Text와 속성색인(제목색인, 링크색인)을 위한 속성정보를 구하며, 로봇의 다음 접근지를 위한 링크정보를 구한다.

③ 웹 문서의 모니터링 기능

웹문서의 순간성으로 인한 등록데이타와 실데이타의 불일치를 방지하기 위하여 주기적으로 데이타 모니터링을 한다. Modify Monitor 로봇으로 제공된다.

④ 네트워크 모니터링 기능

웹문서의 분산성으로 인해 발생하는 문제점으로, 데이타 접근시 네트워크 에러나 해당문서가 존재하는 서버의 일시적 에러로 인하여 데이타 접근이 불가능해질 수 있다. 이러한 경우의 처리를 독립적인 로봇이 수행함으

로써 데이타 접근도를 높인다.Network Monitor로봇으로 제공된다.

⑤ 자동등록요구

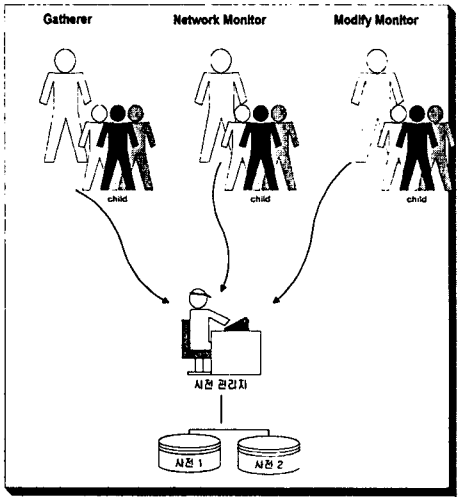
관리자의 개입이 없이 문서를 수집해와서 등록작업을 해준다. 이러한 방법으로 인하여 관리효율을 높인다.

⑥ 자동분류기능

URL의 도메인네임(.ac, .or, .re, .co, .com, .net, .nm)을 이용해서 문서를 자동으로 분류를 한다.

3.3. 등록의 자동화

다중 로봇의 작업을 실시간 유지하기 위하여 문서관리자는 다중로봇의 등록요구를 Queueing해서 로봇의 실행을 계속 진행시킨다. 로봇의 요청으로 Queueing된 등록요구는 등록기로 보내져서 등록이 된다. Modify Monitor의 경우에는 등록된 문서의 수정과 삭제를 요구한다. 이를 위하여 등록 데이타의 insert, modify, delete가 가능하여야한다. 실시간 등록 중 발생하는 문제점으로는 데이타 파손이 있다.



[그림 2]멀티로봇의 구성도

그 이유는 문서등록을 빠르게 하기위해 데이타들이 메인메모리에서 처리되기 때문이다. 언제나 데몬으로 존재하는 시스템의 자료는 정전이나 시스템 파손으로 인해 파손될 수 있으며 이를 위한 고려가 되지않는한 실시간 자동등록 정보검색시스템의 구현은 불가능한 것이다. 이를 위해 미리내 시스템에서는 등록처리를 트랜잭션으로 나누어 파손이 일어나면 트랜잭션이전의 단계로 회복절차를 거쳐서 복구작업이 일어난다.

[그림 3]의 등록절차에서 살펴보면 회복은 로봇의 회복과 등록의 회복 2부분으로 나누어진다. 로봇의 회복은 사전관리자의 Mutex Flag와 History buffer를 이용하여 이루어진다.

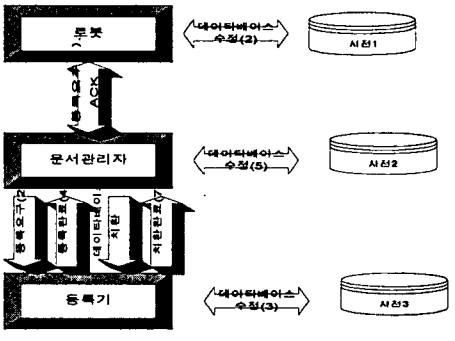
Copy Flag	Read Dic	Write Dic	Read Count	Read-Temp Count	Write-Wait

[표 1] Mutex Flag의 구조

사전관리자는 [표 1]의 Mutex Flag를 통해서 다중로봇의 read/write를 관리한다. 읽기는 동시에 여러 로봇이 가능하며, 쓰기는 한 로봇만 가능하다. Copy Flag는 한 로봇이 쓰기를 끝내고 난후 사전의 불일치가 생겼을 때 사전일치의 필요성을 알리는 역할을 한다. Read Dic, Write Dic은 현재의 읽기, 쓰기 사전을 나타낸다. Read Count는 읽기용 사전을 읽고 있는 로봇의 개수를 나타낸다. Read-Temp Count는 사전일치를 해야하는데, 기존의 Read용 사전을 읽고있는 로봇이 있어서, 일치를 하지 못했을 경우, 현재의 Write Dic으로 Read를 할 수 있게 하는 역할이다. Copy Flag가 세팅 되어있을 경우는 Write가 불가능하며, Write-Wait에 자신의 Process ID를 세팅하고 나간다.

History buffer는 각각의 로봇의 현재 접근중인 URL의 Current ID와 등록요구를 위한 등록용 임시파일을 저장하고 있어서, 회복시 로봇의 수행 시작지점과 등록시작지점을 알 수 있다.

Mutex Flag와 History buffer는 File에 쓰여지는데, File Writing시 파손을 대비해서 Duplicate File로 유지를 해서,



[그림 3]등록절차

양쪽의 데이타가 동일한지를 검사후 회복절차를 수행한다.

문서관리자와 등록기의 회복은 일치를 해야한다. 문서관리자가 요구를 시작해서 문서관리자의 데이타수정이 끝나는 시점까지가 하나의 트랜잭션단위가 된다. 기존의 연구에서 자동등록시스템이전의 단계에서는 문서 하나의 등록까지를 트랜잭션으로 처리해서 회복작업을 진행했었다[3].

이 논문에서 구현한 웹정보검색시스템에서는 문서관리자의 등록요구로 시작한 등록 시작점부터 단위등록이 끝날 때까지를 트랜잭션으로 정의했다.[그림3]에서의 (2)-(7)까지의 작업이 트랜잭션으로 처리되어 회복시 복구기점이 된다.

로봇과 같은방법으로 History buffer를 유지하며, 현재 등록중인 문서번호와 Queue의 상태를 저장하여, 회복시 이전상태로 복구하게 된다.

3.4. 검색효율향상

(1) 링크속성검색

① 기본 개념

기존의 논문을 이용한 검색시스템에서 인용문헌에 의한 검색효율향상[3]노력이 있었다. 이때의 인용검색은 질문에 관련 있는 문헌을 사용하여 그 문헌을 인용하고 있는 모든 문헌들을 검색해 낼 수 있는 방법이다. "학자들이 그들의 문헌에 참고문헌을 기입할 때 문헌을 색인하고 있는 셈이고, 그들만큼 더 지식을 가지고 색인을 맡고 있는 사람이 있다고 볼 수 없다" 라는게 인용색인의 시작점이다.

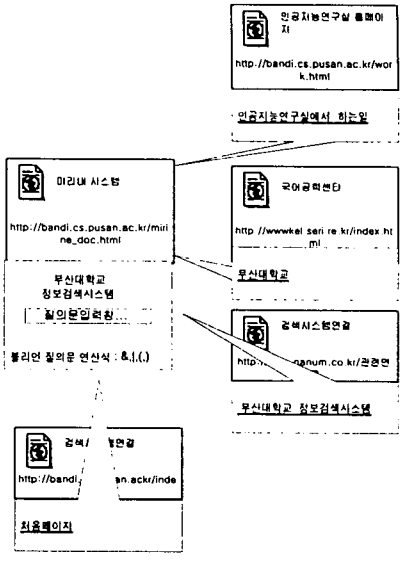
이러한 기법은 웹문서에도 적용할 수 있다. 홈페이지 작성자들이 참고로 하는 웹문서를 링크할 때 사용하는 색인어는 참고되는 문서의 제목에 상당하는 중요도를 가진다.

[그림 4]에서 보듯이 제목이 "미리내 시스템 "인 홈페이지의 본문에는 "부산대학교"와 "검색시스템"이라는 색인어가 없다. 특히 근래에 홈페이지에 그림파일을 많이 쓰기 때문에 색인어가 거의 나오지 않는 문서들도 있다. 또한 제목과 본문이 모두 영어로 쓰여져있는 국내 문서도 상당수 찾을 수 있다. "미리내 시스템" 홈페이지를 5가지 방법으로 찾게된다면 아래와 같은 결과를 가진다.

색인 \ 색인어	미리내	정보 검색	부산 대학교	인공지능 연구실
전문	O	X	X	O
제목	O	X	X	X
링크	X	O	O	O
전문 + 제목	O	O	O	O
전문 + 제목 + 링크	O	O	O	O

[표 2]속성 색인방법에 따른 검색효과

링크속성 색인어는 [그림4]의 밑줄쳐진 부분, 즉 HTML 문서내에서 *부산대학교 인공지능연구실* "부산대학교 인공지능연구실"이라는 색인어를 이용한다.



[그림 4]웹문서의 링크

링크속성색인어는 일반적으로 상호나 사이트를 찾을 때 유용하게 사용될 수 있다. in-link를 많이 가지는 문서는 링크속성검색시 높은 우선순위로 검색이 된다.

② 링크속성의 등록과정

링크정보는 자신을 참조하고 있는 홈페이지에서 자신을 참조할 때 사용하는 색인어를 이용하기 때문에, [그림4]에서 미리내시스템 문서가 full text와 제목 등의 속성이 등록될 때 링크속성 등록이 불가능하다.

이러한 문제로 링크속성등록은 문서가 등록되는 시점이 아니고, 해당문서를 참조하는 문서가 등록될 때 비로소 등록이 되게된다. 따라서 등록기는 속성등록을 지속적으로 추가할 수 있는 기능이 요구된다.

(2) Relevance Feedback

정보검색의 훈련을 거의 받지 못한 일반사용자(end-user)는 정보검색시스템을 이용하여 원하는 정보를 효과적으로 찾기 어렵다. 그 이유는 정보검색시스템의 색인어와 이용자의 탐색어가 다르고, 이용자는 자신이 원하는 정보를 정확한 용어로 표현하기 어렵기 때문에 최적의 질의문을 작성하지 못한다.

이러한 질의어 확장은 기존에는 시소러스를 구성하여 유의어, 광의어, 협의어를 확장하였다. 그러나 시소러스는 그 분야에 따라 용어의 의미가 다르기 때문에 특정한 분

야의 문서에 따라 다른 종류의 시소러스가 필요하다.

웹에서는 모든 분야를 충족시켜주는 시소러스의 작성이 불가능하다. 이러한 이유로 사용자에게 검색된 결과를 보여주고 사용자가 관련 있다고 판단한 문서에서 색인어를 추출하여 원질의문에 추가함으로써 검색효율을 높이는 작업이 Relevance Feedback작업이다. 이는 시소러스를 쓰는 것보다 효과가 좋으며 부가비용이 들지않고, 사용자별 시소러스를 작성하는 것과 같은 의미를 가진다.

4. 실험

4.1. 링크속성검색

① 실험방법

실험을 위하여 등록된 문서는 3만건이고, 10만건의 웹문서에서 링크정보를 구해서 등록된 문서에 해당되는 링크속성색인어를 구했다. 실제로 링크속성색인은 문서의 범위가 전체문서set에 근접할수록 효율이 높아진다. 그러나 웹문서의 특성상 전체 문서의 set을 알 수 없기 때문에, 좀더 높은 링크등록률을 구하기 위해서 큰set의 문서에서 일부분을 등록하고 링크속성을 구했다.

문서set의 크기가 작고, 충분한 링크정보를 이용하지 못했기 때문에 질의어는 링크정보로 등록된 색인어에서, "보라넷", "부산대학교" 등과같이 사이트를 찾기 위한 질의문을 구성했다.

질의어처리는 복합명사 처리를 하지않고, 단일명사의 AND처리를 했다. 사용자가 "부산대학교"를 찾으면 "부산" AND "대학교"라는 질의어를 처리한다. 질의문 30개에 대하여 링크검색으로 찾은 문서에 대하여 제목검색으로 찾아지는가를 봐서, 제목검색으로 불가능한 문서를 링크속성을 이용하여 검색해낼수있나를 실험한다.

② 실험결과

실험에서 링크속성으로 등록되어있는 단어를 추출하여 제목속성색인과 링크속성색인을 비교하여 검색해 보았을때 결과는 아래와 같다.

질의문 30개에 대하여 링크속성이 유용한 경우의 특성을 살펴보면 아래와 같다.

- 색인어가 거의 없는 문서 : 문서의 시작페이지는 근래들어서 이미지 파일로만 되어있는 경우가 많다.
- 제목에 영어약자로 표기되어있는 문서
- 사이트 검색은 전화번호부를 찾는 것과 마찬가지로 해당 사이트의 최상위 URL만 찾을 경우 빠른 검색을 보장한다.

속성검색은 결과로 찾아진 문서가 적지만, site를 찾기에는 효과적인 방법이다. 속성검색은 전문(Full Text)검색보다 검색속도가 빠르고, 찾아진 문서는 모두 관련된 문서를 찾아준다.

이 실험에서는 시도하지 않았지만, 질의어의 복합명사를 bi-gram으로 처리를 하게되면 좀더 많은 관련된 문서를 찾아줄것으로 기대된다. 또한 링크속성 색인어를 얻을때, <a> tag안의 문장외에, 좌우 문장을 추가하는방법도 고려하면, 더 좋은 결과를 기대할 수 있다.

4.2. Relevance Feedback

질의문	제목검색(개)	링크검색(개)
서울대학교	0	3
항공대	0	1
금오공대	0	1
국민대	0	1
제일제당	27	12
대한생명	2	1
한국전산원	5	8
국립수산진흥원	1	1
코리아제록스	0	2
대우정보시스템	2	1
순천대학교	55	1
한국표준과학연구소	0	1
경남은행	11	1
아이네트	13	17
한국밀알선교단	0	1
리크루트	1	1
광염교회	0	1
한국보건사회연구원	0	1
화성그룹	1	1
통계청	4	3
영등포구청	2	1
농수산물유통공사	2	2
건설교통부	0	1
중앙대학교	20	30
부산대학교	1	2
보라넷	30	29
대우중공업	0	3
인터넷엑스포	1	1
유니텔	2	2
넷스케이프	14	12

[표 3] 링크속성검색의 결과

① 실험방법

등록된 10만건의 문서에서, 자연언어 질의어로 초기검색을 하고, 검색된 적합문서에서 색인어를 선택하여 질의문

을 확장한다. 검색시스템은 이 문서에 있는 모든 단어를 추출한다. 그러나 모든 단어를 질의문에 첨가하는 것은 비효율적이다. 왜냐하면 이 단어들 중에는 질문 의도와 일치하지 않는 단어가 있을 수도 있고, 이러한 단어로 인하여 부적합 문서가 검색될 수 있으며, 재검색 속도가 늦어진다. 따라서 검색된 적합문서에서 추출된 색인어에 가중치를 부여하여 질의어 확장을 함으로써 부적합문서의 검색을 방지한다.

다음은 질의어확장으로 선택한 방법이다[4].

- ◆ TF : 적합문서내에서 많이 나타나는 단어들을 빈도순 상위 20개를 질의문에 추가.
- ◆ TFIDF : 적합문서내에서 빈도수가높고, 문서 내의 역문헌 빈도가 높은 단어들을 빈도순 상위 20개를 질의문에 추가.

평균 3개 단어로 구성된 질의어 30개를 사용하여 실험을 하였다. 질의문은 정보검색시스템을 사용해본 경험이 있는 부산대학교 전자계산학과 3학년 2명이 만들어낸 자연언어 질의문이다. 1차 검색방법은 본문과 제목을 모두 사용한 검색이며, 제목에 가중치 30을 준 방법을 사용하였다.

② 실험결과

평균적으로 재검색의 경우 29.1%, 45.8%의 향상이 있었지만, 문서의 종류에 따라서는 재검색의 효율이 낮아지는 경우도 있었다. 이러한 경우를 살펴보면, 관련 있다고 기입된 문서가 광범위한 내용을 담고있어서, 불필요한 질의어가 확장된 경우였다.

1차검색시 30개중 4.8개의 검색결과를 보인것은 실험한 데이타가 10만개의 문서여서 전체 데이타set에 미치지 못해서 신문기사에 비하여 검색효율이 떨어지는것으로 나타났다.

5. 결론

이 논문에서는 웹문서의 특성에 맞추어 문서의 등록과 변경, 삭제가 자동으로 가능하고 시스템 파손으로 인한 자동 복구가 가능한 정보검색시스템을 구현하였다. 또한 사용자 Relevance Feedback과 웹링크속성검색을 통하여 검색효율을 향상했다.

각 기능을 자동화함으로써 정보검색시스템의 관리효율을 높이고, 그로 인한 사용자의 검색요구만족도를 높였다.

현재 서비스 중인 대부분의 웹분류서비스는 모두 수작업분류를 하고있다. 이는 상당한 노력과 경비를 요구한다. 미리내 시스템에서는 초기단계로 URL의 도메인을 이용하여 1차적 분류시스템을 구현하였다. 앞으로는 다단계 자동문서분류를 위하여 문서클러스터링을 통한 자동분류로 확장하여야 할것이다.

	검색한갯수 (상위 30개)	재검색 향상률
1차 검색	4.8개	-
TF	6.2개	29.1%
TFIDF	7.0개	45.8%

[표 4]Relevance Feedback 실험결과

6. 참고문헌

[1]정영미, 정보검색론, 정음사, 1988.

[2]이준영,강상배,양장모,박승,박현주,김민정,권혁철, "다중색인에 의한 정보검색시스템 구현", 한글및 한국어 정보처리 학술발표 논문집, pp. 18-22, 1995.

[3]박승,강상배,박현주,권혁철," 문서내용 갱신시에도 자료검색이 가능한 데이터베이스 복구기법", '96봄 학술발표논문집, pp.163-166,1997.

[4]박세진,강상배,권혁철, "Relevance Feedback을 이용한 정보검색시스템의 검색효율 향상", HCI'97 학술대회. pp.3-7.1997.

[5]Gerard Salton and Mitchael J. McGill, Introduction to Modern Information Retrieval, New York: Mcgraw-Hill, 1983.

[6]조건준,이지철,김경수,김지하, "웹서치엔진에 대한 고찰 및 개선방향", '96 봄 학술발표 논문집, pp.557-560,1997.

[7]이란주, 한국문헌정보학회지 제27집. 1994.12.
"인용문헌에 의한 정보검색 효과에 관한 고찰"

[8]Yov Shoam,"Knowbot : an Intelligent agent", 공개토론, 스텐퍼드대학, 1992.

[9]D.K.harman,"The Proc of the second Text Retrieval(TREC)