검색엔진의 중복된 검색결과 배제 기법설계

이 서 정
동덕여자대학교 정보학부
sjlee@dongduk.ac.kr

요 약

전자상거래가 활성화되고 사회 전반의 거의 모든 활동이 인터넷으로 가능해지면서 무한이 증가하는 정보의 가치를 경쟁력 있게 만들기 위해, 기업에 종사하는 사람들이나 전자상거래의 구매자들은 강력한 검색도구를 필요로 한다. 최근 검색엔진의 성능이 급격히 좋아지고 있지만, 사용자가 완벽히 만족할 수 있는 결과를 제시하지는 못하는 것이 현실이다. 본 연구에서는 같은 서버의 같은 디렉토리 상에 존재하는 중복되는 결과로 내용의 정확성에 비해 검색 갯수가 많은 경우와 연결이 안되는 웹 페이지를 다량 포함한 경우에 대해 검색 결과를 공통 분모를 찾아 결과를 축약하는 방법과 블랙리스트를 구축하는 방법을 제안한다.

Excluding Technique Design for Duplicated Results of Search

Seo-Jeong Lee

ABSTRACT

As e-commerce has been activated and internet has been used as usual, higher efficient search engine must be used to promote the value of information and take possession of the market place. all e-commerce user seller and buyer want to competitive goods Although these needs, search results are still much to be desired. In this paper, I will suppose two ideas which are abbreviation result and making blacklist. Abbreviation result is to hide results with common factors and making blacklist is to reduce null links of search results, which makes many useless results. This routine is made of making blacklist, check list, reduce list and append list.

이 논문은 2001년도 동덕여자대학교 교비연구비 지원에 의한 것임.

Ⅰ. 서 론

2000년 7월 현재, 인터넷에 공식적으로 접속 가능한 웹 페이지 수는 20억 개가 넘는다고 보고되고 있다. 2006년까지 기업의 인트라넷 정보의 양이 1998년보다 200배 가량 증가할 것이라고 한다[1]. 이러한 단순한 증가와 더불어 정보의 추가/삭제 등 변동사항이 기하급수적으로 늘어나고 있다. 동일한 웹 페이지 내의 정보의 변경 뿐만 아니라 파일의 추가나 삭제 등의 변동도 많이 발생한다.

정확한 자료를 제공하기 위한 노력으로 각 검색엔진은 자료를 분류하는 기준을 다양화하고 검색로봇의 수를 늘리고 속도가 빠른 검색로봇을 도입하거나 고급 검색기능의 폭을 넓히고 사용자 맞춤형 검색엔진을 제공하는 등 각양의 노력을 기울이고 있다[2]. 이 때, 사용자는 추가한 파일의 경우 검색엔진에 등록하지만 파일이 삭제된 경우에는 일반적으로 무심한 편이다. 결과적으로 검색엔진은 연결되지 않는 불량링크 페이지 목록을 점점 많이 갖게되고, 사용자는 불량링크 페이지를 포함한 검색결과를 볼 수밖에 없는 상황이 발생한다. 그러나, 검색엔진의 성능이 개발되어 사용자에게 더 많은 결과를 더 빨리 보여줄 수 있다해도 사용자가 완벽히 만족할 수 있는 결과를 제시하지는 못하는 것이 현실이다[3]. 그 중 본 연구에서 해결하려는 사항은 다음의 두 경우이다.

① 같은 서버의 같은 디렉토리 상에 존재하는 중복되는 결과로 내용의 정확성에 비해 검색 갯수가 많은 경우

② 연결이 안되는 웹 페이지를 다량 포함한 경우 이러한 점은 검색엔진의 신뢰도를 떨어뜨리며, 사용자에게는 질 높은 정보를 보유한 웹 사이트에 방문할 수 있는 기회를 줄이는 결과를 초래한다. 빠르고 정확한 검색결과를 원하는 사용자의 기대에 부응하기 위해서 검색엔진은 연결이 안되는 정보를 검색결과에서 배제하고, 검색엔진에 등록된 비슷한 내용의 정보를 간추려서 사용자에게 제공하여야 한다. 사용자는 검색 결과에 대해 원하는 정보가 제대로 찾아졌는지를 빨리 판단할 수 있고, 그렇지 않은 경우, 다른 방법을 빠른 시간 내에 시도할 계획을 세우게 된다.

본 연구에서는 이러한 점을 결과 축약(abbreviation)과 블랙리스트(blacklist)관리 기법을 통해 제안하고자 한다.

Ⅱ. 관련연구

2.1 기존 검색 방식

웹 서비스의 보편화로 정보제공자들이 많아지면서 인터넷에서의 정보검색에 소요되는 시간이 점차 늘어나고 있다. 따라서, 전 세계에 퍼져있는 인터넷 정보제공자들로부터 자신에게 유효한 정보만을 선별적으로 취득할 수 있도록 함으로써 인터넷의 정보검색 시간을 줄이는 문제가 중요하게 대두되고 있다. 현재 전 세계적으로 약 300개 이상의 웹 검색엔진들이 동작하고 있다. 한편 검색엔진을 분류하는 기준은 일반적으로 동작특성에 따라 주제별 검색엔진, 키워드 검색엔진, 멀티 검색엔진으로 구분된다[4].

2.2 진보된 검색방식

기존 검색 방식의 한계를 보완하기 위해 개념기

반(content-base) 방법을 새롭게 적용하거나 불리언 모델을 적용한 지능형 정보검색엔진들을 사용한다. 개념기반 정보검색은 기존의 통계적 검색의 단점을 보완할 수 있는 차세대 검색 모형으로 간주되고 있다. 일반적으로 시스템에 입력되는 문서 데이터로부터 지식베이스를 자동으로 구축하고, 이 지식베이스를 대상으로 개념확장을 수행한 후 관련정보를 검색한다. 따라서 개념기반 검색의 성능을 좌우하는 주요 요소는 지식베이스와 개념확장 알고리즘이라고 할 수 있다[5].

문서에 대한 용어의 중요도를 가중치로 하는 벡터로 표현하여 벡터와 벡터 상이의 유사도를 측정하는 벡터모델은 이분법을 사용하는 불리언 모델의 단점을 보완하고 있다. 또한, 문서에 나타나는 용어의 빈도 수 및 희귀도를 조합한 가중치의 부여로 벡터모델은 일반적으로 검색 성능의 우수성을 보여주고 있다. 그러나, 이 모델은 인식론적 의미론의 부재와 용어와 용어 사이의 독립성을 가정하여 생기는 문제점 등을 포함하고 있다[6].

2.3 고급검색옵션

최근 검색엔진들은 기존에 제공하던 여러 가지 고급기능[7,8,9] 즉, 절단기능, 불린연산 기능들 이외에도 사용자 임의대로 검색범위를 정하거나 최근 검색한 결과에 가중치를 두는 방식으로 사용자 입장에서 검색엔진을 설계할 수 있는 커스터마이징(customizing) 기능을 추가하는 추세이다[10]. 그 외, 본 연구에서 제안하는 방법 중 하나인 축약 및 그룹핑을 포함하는 검색엔진[11,12,13,14]들이 미국에서 개발되어 있지만 시장성이 크지 않아 일반 사용자들에게는 혜택이 거의 없는 상황이다. 그 이유는 별도의 검색엔진을 사용자의 컴퓨터에 설치하여 사용해야하는데 다운로드해서 설치하는 과정이나 프로그램 설치에 따르는 메모리의 낭비로 사용자 부담이 늘어나기 때문이다.

Ⅲ. 제안하는 방법

본 연구에서는 앞에서 설명한 검색 결과의 문제점을 완화하기 위한 두 가지 방안을 제안하고, 실용 가능한 툴(tool)로 개발하기 위한 방법을 제안한다. 하나는 검색 결과를 공통 분모를 찾아 결과를 축약하는 방법으로 통합형이나 키워드 검색엔진에 적용될 수 있다. 다른 하나는 연결이 안되는 결과를 줄이기 위해 블랙리스트를 구축하는 방법이다. 키워드 검색엔진이나 주제별 검색엔진에 적용될 수 있다.

3.1 검색 결과의 축약(abbreviation)

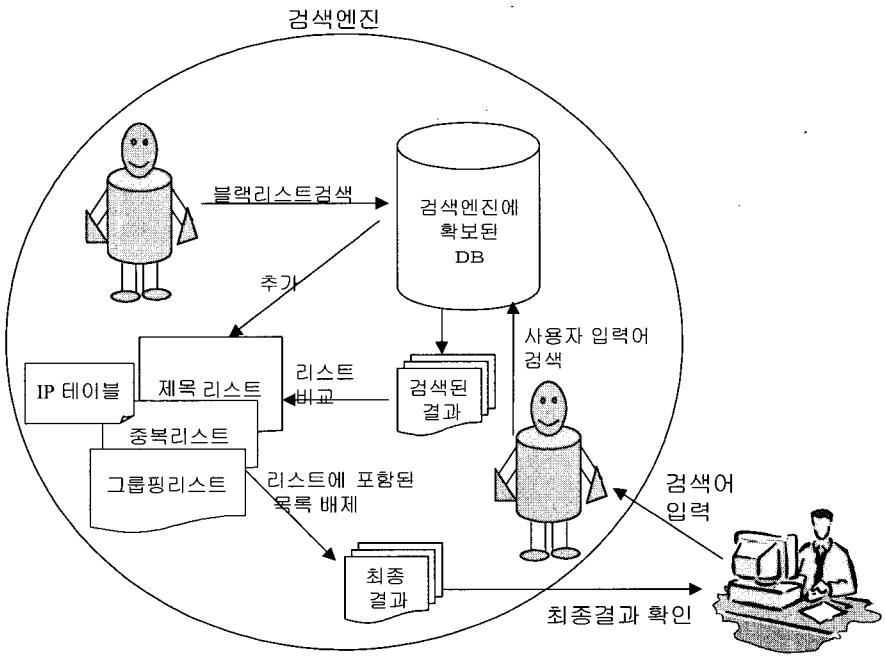
검색결과의 축약은 같은 호스트내의 정보를 그룹핑(grouping)하거나, 중복된 결과를 배제(excluding)하는 것으로 나눈다. (그림 1)은 검색결과의 축약의 개념을 표시한다.

경우1) 같은 호스트내의 정보 그룹핑

검색 결과의 공통 분모를 URL에서 찾아 같은 도메인의 같은 디렉토리에 있는 파일의 경우 그 중 하나만 보여주고, 사용자가 원하는 경우 나머지를 보여주는 방식이다. URL의 구성은 다음과 같이 구성된다.

프로토콜명://서버.호스트.기관.(국가)/디렉토리1/디렉토리2/ .. /디렉토리n/파일명

http://cs4000.dongduk.ac.kr/dir1/dir2/file1.htm



(그림 4) 검색엔진의 축약을 활용한 검색 메카니즘

http://cs4000.dongduk.ac.kr/dir1/dir2/file2.htm
:
http://cs4000.dongduk.ac.kr/dir1/dir2/filen.htm

이 경우는 첫 번째 문서로 그룹핑을 할 수 있다. 사용자가 원하면 다른 내용을 살펴볼 수 있다.

openURL(): 접속한 URL을 읽는다.
pickHost():URL중 호스트를 String으로 리턴한다.
pickDirs(): URL에 포함된 디렉토리를 리턴한다.
pickFile(): URL에 포함된 페이지의 파일 이름과 경로를 리턴한다.

addGroupinglist() : 그룹핑 리스트에 대표 URL을 등록한다.

경우2) URL을 표현하는 방식에 따른 중복 배제

다음 예는 실제로 같은 페이지를 의미한다. 이 경우에도 중복을 배제할 수 있다. 첫 번째 검색되는 페이지만 결과로 보여주고, 나머지는 중복리스트에 등록해둔다.

http://cs4000.dongduk.ac.kr
http://cs4000.dongduk.ac.kr/
http://cs4000.dongduk.ac.kr/index.html

addDuplicatelist() : 중복리스트에 추가해두고, 중복리스트에 나타나지 않는 검색 결과만을 보여준다.

경우3) IP 주소와 도메인 이름에 따른 중복 배제

도메인과 IP주소를 매치하여 일치하는 경우는 검색하지 않는다.

http://www.dongduk.ac.kr/index.html

http://202.20.108.24/indx.html

HostToIP() : host의 IP주소를 리턴한다.

IPToHost() : IP주소에 해당하는 호스트를 리턴한다.

경우4) 문서제목이 일치하는 경우 그룹핑

호스트가 다르더라도 제목이 일치하는 경우에는 같은 내용이거나 비슷한 경우가 많으므로 하나의 제목으로 그룹핑한다.

멀티미디어 기술 www.multitech.co.kr

멀티미디어 기술 multi,multicommunity.co.kr

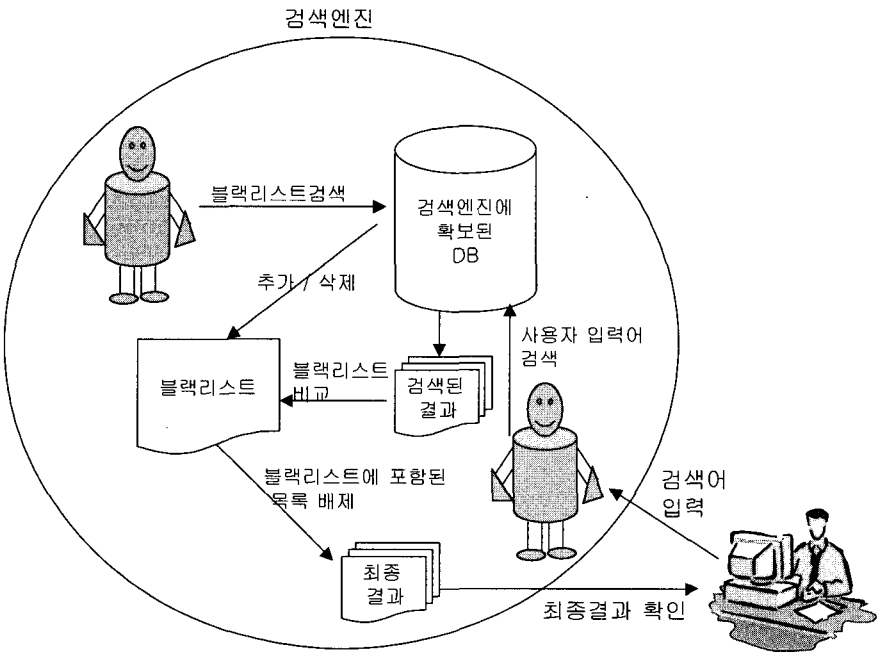
getTitle() : 검색된 웹문서의 제목을 string으로 리턴한다.

addTitlelist() : 제목리스트에 추가한다.

3.2 블랙리스트(blacklist) 관리

등록된 웹 페이지들 중 연결되지 않는 페이지 목록을 구축하는 방법으로 키워드 검색엔진이나 주제별 검색엔진에 적용할 수 있다. 키워드 검색엔진은 미리 구축된 데이터베이스에서 검색 로봇이 사용자가 입력한 검색어에 부합되는 결과를 보여준다. 이 때, 미리 준비된 블랙리스트를 비교하여 리스트에 있는 웹 페이지는 결과에서 제외한다.

블랙리스트의 작성은 검색 로봇이 검색 부하(load)가 적은 시간 또는 별도의 검색 로봇을 활



(그림 5) 검색로봇의 블랙리스트를 활용한 검색 메카니즘

용해 확보된 데이터베이스를 확인하여 블랙리스트를 구성한다. 이후, 주기적으로 블랙리스트를 검토해서 연결이 되는 웹 페이지는 블랙리스트에서 제외시키고, 새로운 리스트를 추가할 수 있다.

주제별 검색엔진의 경우에도 주기적으로 데이터베이스에 구축된 결과를 검토하여 블랙리스트를 구축할 수 있다.

(그림 2)는 키워드 검색엔진에서 검색 로봇이 어떻게 작동하는지를 개략적으로 보여준다. 주제별 검색엔진의 경우에도 검색로봇에 해당하는 프로그램으로 데이터베이스를 검색하여 블랙리스트를 구축하고, 사용자가 입력한 검색어에 대해서도 키워드 검색엔진과 마찬가지로 블랙리스트를 거쳐 수정된 결과를 사용자에게 보여준다.

Ⅳ. 결 론

본 연구는 신뢰도가 높은 검색엔진의 검색 결과를 사용자에게 제공하기 위한 방법의 하나로 결과의 축약과 블랙리스트 관리기법을 제안했다. 결과의 축약은 기존의 검색엔진으로 찾은 결과의 양이 방대한 것을 줄여 사용자에게 적당한 양으로 결과를 보여주는 의미이다. 기존의 결과 중 호스트가 같거나, 제목이 일치하는 경우에는 하나의 그룹으로 취급하고, 블랙리스트관리의 경우에는 블랙리스트를 정기적으로 관리하는 기법으로 검색엔진은 부가적인 리스트를 유지해야한다. 이 방법은 검색엔진이 보유한 웹 데이터베이스가 잘 갖춰진 검색엔진에서 가장 효과적으로 사용할 수 있으며, 통합 검색엔진의 경우에는 전체검색결과보다는 일정한 수준의 정확도 이상의 결과에 대해 적용해 볼 수 있다.

현재 본 연구에서 제안한 방법의 효과나 성능을 검토해보기 위해 간단한 웹 에이전트를 설계하고 있다.

Ⅴ. 참 고 문 헌

[1] 3Soft, "진보된 정보검색의 중요성," http://www.3soft.com/content/support

[2] Danny Sullivan, "Many Changes At Alta Vista," The Search Engine Report, Jul. 6, 1999

[3] Paul Festa, "Web search results still have human touch," CNET News.com Dec. 27, 1999

[4] "UNDERSTANDING AND COMPARING WEB SEARCH TOOLS," http://web.hamline. edu/administration/libraries/search/comparisons.html

[5] 노영희, "개념기반 정보검색 방법론," 2001, IT21

[6] 김민구, "정보검색 모델에 관한 연구," 2001, IT21

[7] "A Look at Search Engines," http://www.augsburg.edu/library/aib/sources.html

[8] "How Search Engines Work," http://www.monash.com/spidap4.html

[9] Terry A. Gray, "How to Search the Web :A Guide To Search Tools,"http://daphne.palo mar.edu/TGSEARCH/

[10] Danny Sullivan, "Many Changes At AltaVista," The Search Engine Report, Jul. 6, 1999

[11] Mata Hari, http://www.thewebtools.com/

[12] WebFerret, http://www.ferretsoft.com/

[13] BullsEye, http://www.intelliseek.com/

[14] Copernic, http://www.copernic.com/

이 서 정

1993.9-1998.2 숙명여자대학교 대학원 전산학과 박사과정 졸업

1989.3-1991.2 숙명여자대학교 대학원 전산학과 석사과정 졸업

1985.3-1989.2 숙명여자대학교 전산학과 졸업

1998.3- 현재 동덕여자대학교 정보학부 강의전임교수

2001.7-2001.12 정통부 산하 KCSC 컴포넌트 분류체계 표준

2000.7-2001.2 삼성SDS 웹마이닝 솔루션 분석

관심분야 : 소프트웨어 공학, 웹데이터베이스, 모바일응용 등