# 인트라넷 부하 평준화를 위한 분산 검색엔진 설계

고윤석° 윤희병
국방대학교 전산정보학과
{ko12647°, hbyoon}@kndu.ac.kr

# Design of the Distributed Search Engine for Intranet Load Balancing

Yunseok Ko° Heebyung Yoon
Dept. of Computer and Information Science, Korea National Defense University

**요 약**

본 논문은 인트라넷에서 검색엔진 운용시 발생되는 트래픽을 감소하여 네트워크 자원을 효율적으로 이용하기 위한 분산 웹 크롤링 에이전트와 인덱싱 에이전트를 제안한다. 일반적인 검색엔진의 구성, 인트라넷의 네트워크 구성 및 인트라넷이 인터넷과 구별되는 몇가지 특징을 제시하고 이에 적합한 분산 검색엔진을 설계하며 분산 검색엔진의 각 에이전트들이 분산 환경에 동작할 수 있도록 하기위하여 URL Sorter, URL Provider 및 분산 Indexer를 설계한다.

## 1. 서 론

1969년 ARPANET으로 시작된 인터넷은 1991년 WWW서비스가 개발되어 일반인들도 쉽게 사용할 수 있어 현재의 규모로 성장하게 되었다. 2003년 6월 국내의 경우 국민 전체 중 64.1%가 인터넷을 사용하고 있으며 인터넷 사용 목적은 71.6%가 검색엔진에 의한 자료 및 정보검색으로 조사되었다[1].

제한된 컬렉션에서 가능한 많은 결과를 사용자에게 제공하기 위한 도서관 문헌정보 검색 시스템과는 달리 인터넷 검색엔진은 방대한 양의 웹 문서를 수집, 인덱싱, 갱신하여 사용자의 정보요구에 적절할 결과를 제공하여야 한다. 검색엔진은 크게 정보검색 요구에 대하여 적절한 검색 결과를 제공하기 위하여 웹 문서를 수집하는 웹 크롤링 에이전트, 수집된 문서를 인덱싱하는 인덱싱 에이전트, 인덱싱된 문서를 사용자의 검색요구에 응답하기 위한 인터페이스 에이전트로 구성되어 있다.

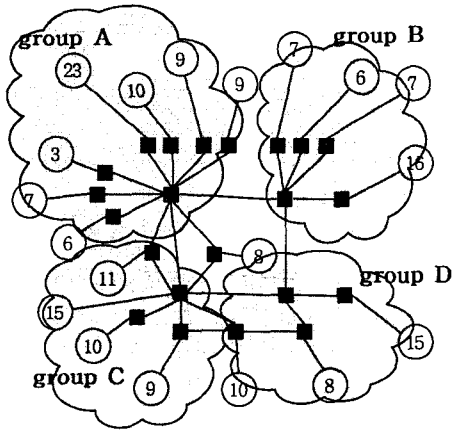
인터넷은 중앙 통제 방식이 아닌 사용자들의 규약으로 모든 네트워크가 연결되어 있기 때문에 검색엔진의 주요한 기능들을 분산하여 구성하는 것이 불가능하다. 그러나 웹과 인터넷 기술을 활용한 기업 내부 정보시스템이라고 정의되는 인트라넷에 검색엔진을 구현하려고 할 때 인트라넷이 인터넷과 구별되는 몇 가지 특징을 이용하면 네트워크 대역폭을 보다 효율적으로 사용하며 검색엔진을 경량화할 수 있는 분산 검색엔진의 구현이 가능하다.

본 논문의 구성은 인트라넷 환경과 검색엔진에 대하여 소개하고, 분산 검색엔진을 설계하며, 이를 검증하기위하여 분산 검색엔진을 라우터 홉 카운터 측면에서 평가한다.

## 2. 인트라넷 환경 및 검색엔진 분석

기업체간의 정보공유나 업무처리에 필요한 사항을 웹 브라우저와 같은 하나의 응용을 통하여 수행하기 위하여 시작된 인트라넷은 그 규모가 커질수록 능동적으로 문서를 수집하여 사용자에게 적절한 정보를 제공할 수 있는 검색엔진의 필요성이 커지고 있다. 본 논문은 인트라넷이 인터넷과 구별되는 몇가지 특징과 일반적인 검색엔진의 구성요소에 대하여 분석하여 분산 검색엔진을 설계하고, 이를 통한 인트라넷의 네트워크 자원을 효율적으로 사용하며 검색엔진을 경량화 시킬 수 있는 방법을 제시한다.

### 2.1 인트라넷 환경



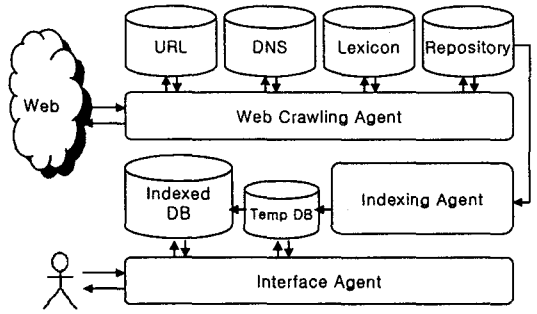

■ : Routing Server ○ : 웹서버 --- : T3 백본망

[그림 1] 인트라넷 구성 샘플

[그림 1]은 본 논문에서 사용하기 위한 인트라넷의 네트워크 구성도의 샘플로서 22개의 라우팅 서버가 백본 네트워크에 연결되어 있으며 137개 웹서버가 운용되고 있다고 가정한다. [그림 1]에서의 각 그룹은 네트워크의 특성 및 관리적인 측면을 고려하여 4개의 물리적 그룹으로 분할하였다. 이러한 그룹은 본 논문에서 제시하는 분산 검색엔진의 설계에서 하위 그룹에 인덱싱 에이전트와 웹 크롤링 에이전트를 운용하기 위하여 인위적으로 할당한 것이다.

2.2 검색엔진

[그림 2]는 검색엔진의 일반적인 시스템 구성에 대하여 나타내고 있다. 웹 크롤링 에이전트는 URL 서버로부터 URL을 제공받아 웹 문서를 수집하는 기능을 하며, 인덱싱 에이전트는 수집된 웹 문서를 분석하여 새로운 URL을 추출하며 사용자의 검색요구에 적절한 결과를 제공하기 위한 형태로 가공한다. 인터페이스 에이전트는 사용자의 질의를 분석하여 Indexed DB와 비교, 적절한 순위를 부여하여 사용자에게 제공한다[2,3,4].



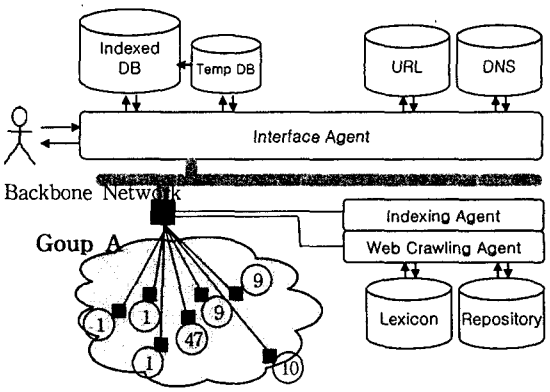

[그림 2] 검색엔진 구성도

2.3 검색엔진의 인트라넷 적용

인트라넷은 자유롭게 확장되는 인터넷과는 달리 어떤 집단의 인위적인 목적에 의하여 구성된 네트워크로 집단 내부의 관리자는 네트워크의 전반적인 구성에 대하여 알 수 있으며, 네트워크를 구성하고 있는 라우팅 서버, LAN 라우터 및 웹 서버 등에 대한 제한적인 통제가 가능하다. 또한 인트라넷은 본 논문에서 제시하는 분산 검색엔진의 도입을 위하여 지역적, 물리적 혹은 관리적 측면으로 몇 개의 영역으로 그룹화가 가능하다. 3장에서는 인트라넷이 인터넷과 구별되는 몇 가지 특성을 이용하여 분산 검색엔진을 설계한다.

## 3. 분산 검색엔진 설계

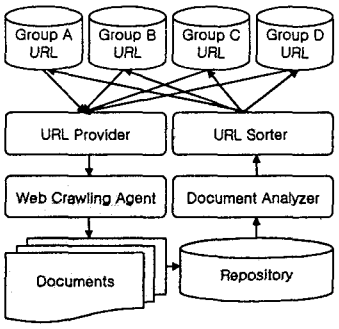
[그림 3]은 본 논문에서 제안하는 분산 검색엔진의 구성도를 나타낸다. 분산 검색엔진은 문서 수집을 위한 웹 크롤링 에이전트와 수집된 문서를 인덱싱하기 위한 인덱싱 에이전트가 인터페이스 에이전트와 분리되어 각각의 하위 그룹에 포함되어 있다. URL 서버와 Lexicon 및 DNS는 웹 크롤링 에이전트 및 인덱싱 에이전트와 일관된 정보를 교환하기 위하여 인터페이스 에이전트에 위치하게 된다. 웹 크롤링 에이전트는 해당 그룹내의 웹서버에 대하여만 문서 수집 및 갱신과정을 거치게 되며 수집된 문서는 각 그룹별 인덱싱 에이전트에 의하여 인덱싱된 후 인터페이스 에이전트에 위치한 Main Inverted Indexed DB에 종합된다[5,6].




[그림 3] 분산 검색엔진 구성도

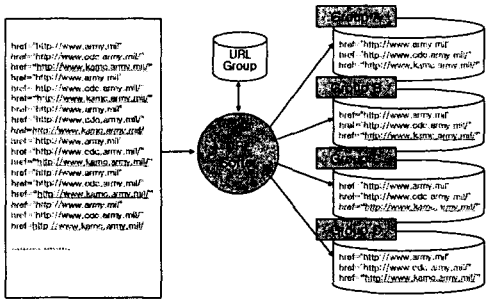
3.1 웹 크롤링 에이전트

인트라넷 환경의 분산 검색엔진에서 웹 크롤링 에이전트는 하위 그룹에 각각 위치하여 문서를 수집하게 된다. [그림 4]는 분산 검색엔진의 웹 크롤링 에이전트를 나타낸다.



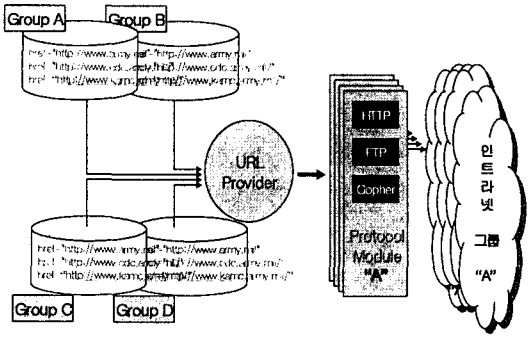

[그림 4] 분산 검색엔진의 웹 크롤링 에이전트

각 그룹별 URL 서버는 해당 그룹의 웹 크롤러가 방문해야할 URL을 QUEUE 형태로 저장한다. 각 그룹별 URL 서버는 URL Provider에 의하여 각 그룹의 웹 크롤러에게 URL을 전달한다. 웹 크롤러는 전달된 URL로 접근하여 문서를 수집한다. URL Provider는 멀티 쓰레드 프로그램 형태의 웹 크롤러가 동일한 웹서버로 접근하지 않도록 URL을 적절히 분산시킨다[5,6].




[그림 5] 분산 검색엔진의 URL Sorter

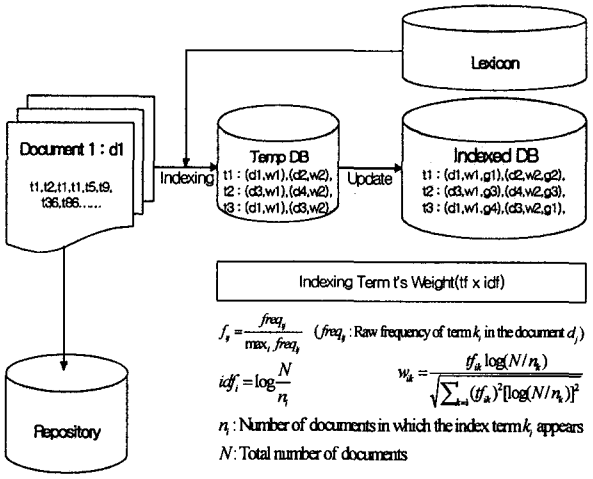
웹 크롤러에 의하여 수집된 문서는 분석 과정을 거치게 되는데 이 과정에서 발견된 새로운 URL은 웹 크롤러가 접근할 수 있는 절대 경로의 URL 주소로 변환되며 URL Sorter로 전달되며 URL Sorter는 URL 주소를 그룹별 분리 저장한다.




[그림 6] 분산 검색엔진의 URL Provider

3.2 인덱싱 에이전트

본 논문에서는 분산 웹 크롤링 에이전트와 함께 인트라넷 환경에 적합한 검색엔진을 설계하기 위하여 인덱싱 에이전트를 사용한다. [그림 7]은 검색엔진의 인덱싱 에이전트를 나타낸다.




[그림 7] 분산 검색엔진의 인덱싱 에이전트

분산 검색엔진에서 인덱싱 에이전트는 수집된 원본 문서를 역색인 형태로 변화 시킨다. 역색인 과정은 수집된 원본 문서 내에서 색인어로서의 가치가 없는 불용어를 제거과정과 한글의 경우 형태소 분석과정을 추가적으로 거치게 되며 색인어-빈도 형태의 정보는 다시 색인어-가중치-문서 ID-그룹 ID의 형태로 변환된다.

각 그룹별 인덱싱 에이전트로부터 생성된 역색인 데이터는 인터페이스 에이전트의 Main Inverted Indexed DB로 저장된다. 특히 문서에 대한 가중치 부여는 보편적으로 많이 이용되는 tf x idf 벡터 모델을 이용하며[7], 인덱싱과정에 사용자의 원본 문서 접근 시 참조를 위하여 그룹 ID가 추가된다.

## 4. 결론

본 논문에서는 인트라넷에 검색엔진을 도입하고자 할 때 인트라넷이 인터넷과 구별되는 특성을 이용한 분산 검색엔진을 설계하였다. 분산 검색엔진의 설계에서 검색엔진의 효율을 높이고 네트워크의 대역폭을 효과적으로 사용하기 위하여 분산 웹 크롤링 에이전트와 분산 인덱싱 에이전트를 설계하였으며 분산 웹 크롤링 에이전트에서 각 그룹별 효율적인 URL 배분을 위하여 URL Sorter와 URL Provider를 설계하였다. 분산 색인 에이전트는 각 그룹별 독자적으로 동일한 Lexicon을 가지고 색인과정을 거친 후 인트페이스 에이전트의 Main Inverted Indexed DB로 저장되어 각 사용자에게 서비스 된다.

## 참 고 문 헌

[1] 정보화 실태 조사, "www.nic.or.kr", 2003

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation Ranking: Bringing order to the Web. Unpublished Manuscript.

[3] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SLAM Symposium on Discrete Algorithms, 1998.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th World Wide Web Conference, 1998.

[5] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL orderring. In 7th World Wide Web Conference, Brisbane, Australia, April 1998.

[6] S. Chakrabarti, K. Punera, and M. Subramanyam, Accelerated focused crawling through online refernce feedback. WWW, pagtes 148-145, ACM, Honolulu, May 2002.

[7] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling : A new approuch to topic-specific Web resource discovery. Computer Networks, 31, pages 1623-1640, 1999. First appeared in the 8th International World Wide Web Conference, Toronto, May 1999.