디지털 도서관을 위한 분산색인 기법에 대한 연구

유 춘 식[†] · 이 종 득[††] · 김 용 성[†††]

요 약

분산환경을 기반으로 하는 디지털 도서관과 같은 정보서비스 시스템들의 성능은 분산자원에 대한 색인기법에 의해 매우 큰 영향을 받는다. 이러한 분산자원에 대한 색인기법에는 중앙집중형 색인기법, 분산형 색인기법 그리고 이 두 가지 기법을 혼합한 혼합형 색인기법이 있다.

본 논문에서는 사용자의 검색요구를 보다 빠르게 처리하면서 시스템과 네트워크에 부과되는 부담(overhead)을 줄이기 위해, 중앙집중형 색인기법과 분산형 색인기법을 혼합한 새로운 분산색인 기법을 제안한다. 즉, 질의어에 적합한 정보자원을 저장하고 있는 서버들만을 대상으로 보다 빠르게 질의를 처리하기 위해, 일반적으로 사용되는 역파일을 확장한 확장된 역파일(Extended Inverted File, EIF) 구조와 이에 대한 관리 기법 그리고 이를 이용한 검색기법을 제안한다. 또한 실험을 통하여 본 논문에서 제안한 분산색인 기법의 성능을 검증하였다.

A Study on Distributed Indexing Technique for Digital Library

Chun-Sik Yoo[†] · Chong-Deuk Lee[††] · Yong-Sung Kim[†††]

ABSTRACT

Indexing techniques for distributed resources have much effect on an information service system based on distributed environment like digital library. There is a centralized indexing technique, a distributed technique, and a mixed technique for distributed indexing techniques.

In this paper, we propose new distributed indexing technique using EIF(Extended Inverted File) structure that mix the centralized technique and the distributed technique. And we propose management techniques for EIF structure and retrieval technique using EIF structure. This distributed indexing technique proposed is able to fast process retrieval request and reduce network overload and select servers relevant to query terms. This paper investigated performance of a proposed distributed indexing technique.

1. 서 론

최근 몇 년 사이에 대량의 정보를 손쉽게 입력할 수 있는 입력장치와 이러한 정보를 저장하고 관리할 수 있는 데이터베이스 시스템, 저장된 정보에 대한 사용자의 검색요구를 보다 정확하고 편리하게 만족시키기 위한 정보검색 시스템의 성능이 많이 향상되었다. 이러한 기술들을 종합적으로 이용하여 네트워크 상에 분산된 방대한 양의 정보를 체계적으로 관리하고, 이러한 정보를 네트워크를 통해 쉽고 편리하게 접근·획득하기 위한 방안으로 제안되고 있는 것이 디지털 도서관이다[20, 21]. 일반적으로 디지털 도서관의 자료들은 다양한 형태의 멀티미디어 정보들이며, 이러한 정보들이 한 곳에서 저장·관리되는 것이 아니라 지역적으로 여러 곳에 분산되어 있다. 더구나 정보들이 표현

※ 이 논문은 1997년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.
† 준 회 원 : 전북대학교 대학원 컴퓨터과학과
†† 정 회 원 : 서남대학교 전자계산학과 교수
††† 종신회원 : 전북대학교 컴퓨터과학과 교수
논문접수 : 1998년 9월 4일, 심사완료 : 1998년 12월 18일

되는 방법과 정보들에 접근하여 검색하는 방법도 매우 다양하다. 이러한 정보의 이질성과 분산성으로 인하여 사용자가 디지털 도서관의 시스템 구성에 투명하게 디지털 도서관에 접근하여 사용할 수 있도록 하는 시스템을 개발하는 것은 쉽지 않다. 이러한 이유로 사용자가 정보의 분산성과 이질성에 무관하게 디지털 도서관을 사용하기 위해서는, 먼저 정보의 이러한 특성을 충분히 반영하여 정보들에 대한 메타 정보, 즉 색인을 작성하는 방법에 대한 연구가 필수적으로 요구되고 있으며, 이러한 분산자원에 대한 색인기법은 전체 디지털 도서관의 성능에 매우 큰 영향을 미친다[2, 7, 8, 26].

분산자원에 대한 색인기법(즉, 분산색인 기법)에는 분산되어 있는 각 서버의 색인을 그대로 사용하는 분산형 색인기법과 모든 자원에 대한 색인을 한 곳에 집중시켜 관리하는 중앙집중형 색인기법 그리고 이러한 두 가지 기법을 혼합한 혼합형 색인기법이 있다[26]. 중앙집중형 색인기법은 모든 분산자원에 대한 정보를 한 곳에 모아두는 형태로서, 질의에 대한 응답시간은 빠른 장점이 있지만 통합된 분산색인을 구축하고 유지하는 것이 어렵다는 문제점이 있다. 반면, 분산형 색인기법은 자료에 대한 색인을 각 서버에 구축한 후, 사용자가 입력한 질의를 모든 서버에 전달하여 처리하는 방법이다. 이 기법은 분산자원에 대한 색인을 구축하고 유지하는 것은 용이하나, 입력된 질의에 적합한 자료를 가지고 있지 않은 서버들에게도 질의를 전달해야 하는 문제점이 있다. 따라서 현재는 이 두 가지 기법을 혼합한 기법에 대한 연구가 주로 수행되고 있다.

본 논문에서는 사용자의 검색요구를 보다 빠르게 처리하면서 시스템과 네트워크에 부과되는 부담(overhead)을 줄이기 위해, 중앙집중형 색인기법과 분산형 색인기법을 혼합한 새로운 분산색인 기법을 제안한다. 즉, 질의어에 적합한 정보자원을 저장하고 있는 서버들만을 대상으로 보다 빠르게 질의를 처리하기 위해, 일반적으로 사용되는 역파일을 확장한 확장된 역파일(Extended Inverted File, EIF) 구조를 제안하고 이에 대한 관리 기법, 이를 이용한 검색기법을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 분산자원에 대한 색인기법들에 대해 알아보고, 3장에서 본 논문에서 제안하는 색인기법을 기술한다. 그리고 4장에서는 본 논문에서 제안한 색인기법에 대한 실험 결과를 제시하고, 마지막으로 5장에서 결론 및 향후 연구과제를 기술한다.

2. 관련연구

본 장에서는 분산자원에 대한 색인기법인 중앙집중형 색인기법과 분산형 색인기법, 그리고 이들을 혼합한 혼합형 색인기법에 대하여 설명한다.

2.1 중앙집중형 색인기법

중앙집중형 색인기법은 분산자원에 대한 통합 색인을 한 곳에서 저장·관리하는 방법으로서, 사용자가 입력한 질의에 대한 검색결과를 보다 빠르게 얻을 수 있다. 그러나 이 방법은 모든 분산자원의 표현형식에 대한 정보를 알고 있어야 하며, 새로운 정보가 추가될 때마다 통합 색인을 변경해야 하는 문제점이 있다. 또한 통합 색인에 저장되어 있는 정보와 실제로 존재하는 정보가 일치하지 않는 불일치 문제가 발생한다.

대부분의 WWW 검색엔진들(심마니, AltaVista, InfoSeek, Lycos 등)은 대표적인 중앙집중형 색인시스템들로서, 로봇(robot)이 인터넷에 존재하는 HTML 문서들을 자신의 사이트(site)로 가져와 문서들에 대한 색인을 작성한다. 비록 로봇이 필터링을 통해 가져올 문서의 수를 줄일 수는 있지만, 문서 전송 자체가 네트워크에 상당한 부담을 주게 된다. 또한 중앙집중형 색인기법에서 발생하는 문제점들(즉, 유지관리의 어려움, 불일치 문제 등)이 그대로 나타난다.

다음으로 연구개발정보센터에서 개발한 KRISTAL-II[28]는 중앙집중형 색인기법을 사용하는 시스템으로서, 모든 서버의 데이터베이스에 저장된 정보에 대한 역파일을 중앙의 서버에 구축한 후, 이를 정보검색에 이용하기 때문에 중앙집중형 색인기법에서 발생하는 문제점들이 발생한다. 즉, 각 데이터베이스의 표현 형식이 동일해야 하며, 저장된 문서정보 크기만큼의 색인(역파일)을 유지·관리해야 하는 문제점이 있다.

2.2 분산형 색인기법

분산형 색인기법은 분산되어 있는 각 서버의 색인을 그대로 사용하는 방법으로서, 각 정보에 대한 검색방법과 검색결과의 수집방법만을 고려하면 된다. 또한 실제로 저장되어 있는 정보와 색인사이의 불일치 문제가 발생하지 않는 장점이 있다. 그러나 분산형 색인기법은 각 질의에 적합한 정보가 어느 서버에 존재하는지를 미리 알 수 없기 때문에 모든 서버에 질의를 주어 검색을 수행해야 하며, 이로 인해 매우 큰 네트워

크 부하(overload)가 발생하는 문제점이 있다.

이러한 문제점을 보완하기 위한 연구들로서, UCSB (University of California at Santa Babara)의 Alexandria[19], Stanford 대학의 전자도서관[24], Michigan 대학의 UMDL(University of Michigan Digital Library) [25], 그리고 [4, 10, 15] 등과 같이 지능형 에이전트(Intelligent Agent)를 이용하는 연구들이 있다. 이러한 연구들은 사용자 프로파일(User Profile)을 이용하여 사용자의 관심분야 또는 관심주제에 적합한 정보를 검색하는 것을 목적으로 한다. 그러나 지능형 에이전트를 이용하기 위해서는 사용자가 원하는 자원들에 대한 정보를 획득하여 사용자 프로파일을 작성해야 하는데, 사용자의 관심을 가지는 주제에 대한 정보를 정확하게 획득하는 것은 상당히 어려운 작업 중의 하나이다.

2.3 혼합형 색인기법

본 절에서는 분산된 정보자원에 대한 색인기법들 중에서 중앙집중형 색인기법과 분산형 색인기법을 혼합한 기법을 사용하는 연구들에 대해 살펴본다.

먼저 WHOIS++[16], Discover[13]는 centroid(WHOIS++), content label(Discover)이라고 부르는 정보자원에 대한 요약 기술자료(description)를 사용하여 색인에 대한 계층구조(hierarchy)를 형성한다. 그리고 계층구조에 존재하는 각 서버들의 centroid(또는, content label)는 한 단계 하위의 서버들에 저장되어 있는 색인에 대한 요약정보를 저장하고 있으며, 이를 이용하여 질의를 처리한다. 그러나 이 연구들에서는 자원들에 대한 적절한 요약 기술자료를 작성하는 방법과 계층구조에서 질의에 적합한 정보자원을 선택하는 방법이 명확히 제시되지 않았다.

Harvest[1], HyPersuit[17]에서는 WHOIS++와 Discover에서 발생된 문제점을 해결하기 위해, 특수화(specialization)라는 개념을 도입하여 색인을 작성하였다. 즉, 각 서버(broker)를 구축할 때 일정한 기준에 의해 하위 단계의 서버들을 선정하고, 선정된 하위 단계의 서버들에 저장되어 있는 색인에 대한 요약정보를 작성한다. 이러한 방법에 의해 자원들에 대한 다중 계층구조를 형성한다. 그러나 이 연구들은 서버들에 대한 일반적인 분류(classification) 방법의 부재, Broker 구축의 어려움과 막대한 구축비용 그리고 실제 정보자원과의 일치성 유지의 어려움 등과 같은 문제점이 있다.

What'sHot[7]에서도 특수화 개념을 도입하여 색인을 작성하였다. 이 연구에서는 실제 정보 자원인 아닌 메타데이터를 대상으로 사용자가 원하는 관심분야 또는 주제에 적합한 정보에 대해서만 분산색인을 작성하였다.

NCSTRL 프로젝트의 Dienst 프로토콜[9]은 각 FTP 호스트에 문서를 저장하고, Dienst가 정의한 형식으로 저장된 문서들에 대해 색인을 작성한다. 그리고 Indexer가 몇 개의 FTP 호스트를 한 집단으로 묶어 이들에 저장된 모든 문서들에 대한 색인을 작성하며, 사용자의 질의는 Indexer들에 의해 처리된다. 결국 모든 Indexer를 대상으로 질의를 처리하게 된다.

서울대학교의 한울(Heterogeneously Archived and Networked Universal Library) 시스템은 가상색인(Virtual Index)[26]이라는 개념을 이용한다. 가상색인은 사용자의 질의에 대한 검색결과를 이용하여 카탈로그 관리자(Catalog Manager)가 관리한다. 그러나 가상색인에 대한 관리방법이 제시되지 않았다.

3. 분산색인의 작성 및 관리

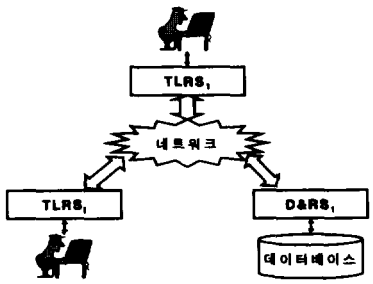
2장에서 살펴 본 바와 같이, 분산자원에 대한 기존의 기법들은 정보 표현구조의 부재, 정보 관리 기법의 부재, 그리고 분산색인 정보에 대한 유지관리의 어려움, 검색의 비효율성 등과 같은 문제점이 있다. 따라서 본 논문에서는 이러한 문제점들을 해결하기 위해, 분산색인에 대한 정보 표현구조로서의 확장된 역파일(Extended Inverted File, EIF) 구조와 이에 대한 관리기법 그리고 이를 이용한 검색 기법을 제안한다.

3.1 디지털 도서관의 구조

본 논문에서 제안하는 EIF 구조를 사용하는 디지털 도서관의 구조는 (그림 1)과 같다. (그림 1)에서 보는 바와 같이, 디지털 도서관의 사용자들은 디지털 도서관을 구성하는 최상위 인용서버(TLRS, Top-Level Reference Server)들 중에서 특정한 하나의 TLRS에 접속하여 디지털 도서관을 이용하게 된다.

사용자의 정보검색 요구는 일차적으로 사용자가 접속한 TLRS에 의해 수행된다. 사용자가 접속한 TLRS는 자신에게 속한 문서서버 및 인용서버(D&RS, Document Servers and Reference Servers)들의 EIF에서 사용자가 입력한 질의어를 검색한 후, 그 결과를 반환한다. 그러나 사용자가 검색결과에 만족하지 못하거나 D&RS에서 해당 질의를 수행하지 못하는 경우에는, 다

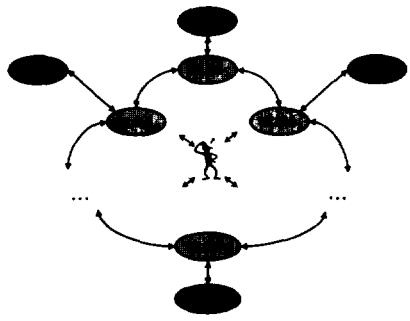
른 TLRS에 해당 질의를 전달하여 검색을 수행하도록 요청한 다. 각 TLRS가 검색결과를 반환하면 반환된 검색결과들을 병합하여 사용자에게 전달한다. 이러한 방법으로 사용자에게 디지털 도서관의 시스템 구성에 투명한 검색환경을 제공할 수 있기 때문에 사용자는 보다 편리하게 디지털 도서관을 사용할 수 있다.



TLRS : Top-Level Reference Server
D&RS : Document Servers and Reference Servers

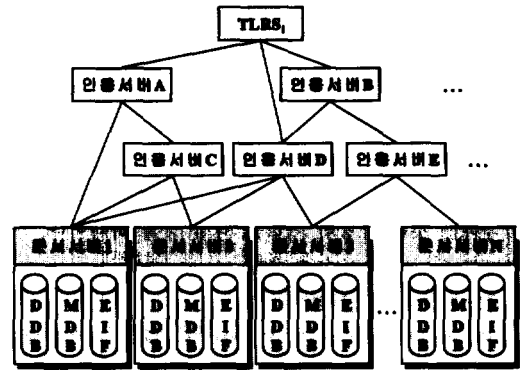
(그림 1) 디지털 도서관의 구조
(Fig. 1) Architecture of Digital Library

(그림 1)과 같은 구조를 가지는 디지털 도서관의 세부적인 구성은 (그림 2)와 같다. 결과적으로 최상위 인용서버들은 일정한 기준(분야, 지역 등)으로 특수화(specialization)되며, 사용자는 이들 중에서 특정한 하나의 TLRS에만 접속하여 검색을 수행하게 된다. 또한 사용자가 입력한 질의는 모든 인용서버에 의해 수행되는 것이 아니라, 질의어를 포함하고 있는 인용서버에 의해서만 수행되기 때문에 일반적인 분산색인 기법이 가지는 문제점, 즉 모든 서버에 질의를 전달해야 하는 문제점을 해결할 수 있다.



(그림 2) 디지털 도서관의 구성
(Fig. 2) Components of Digital Library

(그림 3)은 각 TLRS와 해당 TLRS가 참조하는 인용서버 및 문서서버 사이의 관계를 나타내고 있다. 그림에서와 같이 TLRS와 인용서버, 문서서버는 다중 계층구조(multiple hierarchy)를 형성하며 각 TLRS의 하위에는 다수의 인용서버가 있고, 계층구조의 최하위 단계에는 문서서버가 위치하고 있다.



DDB : 문서 데이터베이스(Document Database)
MDB : 메타데이터베이스(Metadatabase)
EIF : 확장된 역파일(Extended Inverted File)

(그림 3) 분산색인의 구조
(Fig. 3) Structure of Distributed Indexing

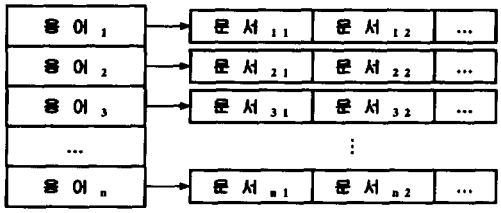
각 인용서버는 일정한 기준(분야, 지역 등)에 의해 특수화되는데, 자신의 기준에 적합한 문서서버나 인용서버에 저장되어 있는 정보(색인 정보)를 통합하여 저장하고 있다. 본 논문에서는 이러한 색인정보로서 확장된 역파일(Extended Inverted File, EIF) 구조를 사용하며, EIF의 구조에 대해서는 3.2 절에서 설명한다. 또한 분산색인의 구조가 다중 계층구조이기 때문에 각 인용서버나 문서서버는 서로 다른 다수의 TLRS, 인용서버, 문서서버에 의해 참조될 수 있다.

그리고 각 문서서버는 각종 문서 정보를 저장하고 있는 문서 데이터베이스, 문서 데이터베이스에 대한 메타 데이터(Dublin Core, MARC 등)를 저장하고 있는 메타데이터베이스, 특정 용어(질의어)를 주제어(keyword)로 포함하는 문서를 빠르게 검색하기 위한 형태의 EIF로 구성된다.

3.2 확장된 역파일 구조를 이용한 분산색인 기법

본 절에서는 분산색인 정보를 저장하기 위한 정보구조인 EIF의 구조에 대해 기술한다. 먼저 정보검색

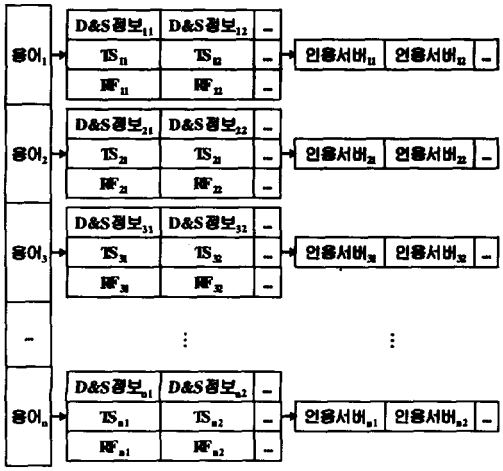
분야에서 사용되는 일반적인 역파일(Inverted File)의 구조는 (그림 4)와 같다.



(그림 4) 일반적인 역파일의 구조
(Fig. 4) Structure of common inverted file

(그림 4)에서 알 수 있듯이, 문서를 중심으로 하여 해당 문서의 내용을 대표할 수 있는 주제어에 대한 정보를 포함하는 메타 데이터(또는 색인 정보)와는 달리, 역파일은 용어를 중심으로 하여 해당 용어를 포함하고 있는 문서들의 리스트를 지정하고 있다. 이는 사용자의 검색 요구가 질의어를 중심으로 이루어지기 때문에 해당 질의어를 포함하고 있는 문서들의 리스트를 빠른 시간 내에 획득하기 위함이다.

그런데 본 논문에서는 특정 질의어를 포함하고 있는 문서서버(Document Server)들의 리스트를 빠른 시간 내에 획득하는 것이 중요하기 때문에, 이러한 일반적인 역파일의 구조를 변형, 확장하여 확장된 역파일 구조인 EIF를 제안한다. 이러한 EIF의 구조는 (그림 5)와 같다.



(그림 5) EIF의 구조
(Fig. 5) Structure of EIF

이러한 EIF의 구성과 각 구성부분에 저장되는 정보의 내용은 다음과 같다.

1) 용어 리스트
   해당 인용서버(Reference Server)의 하위 레벨에 존재하는 문서서버의 EIF 또는 다른 인용서버의 EIF에 존재하는 용어 리스트들을 병합한 용어 리스트.
2) 분산색인 정보부분
   (1) 서버 리스트
      ① D&S(문서&서버) 정보
         현재의 서버가 문서서버인 경우에는 해당 용어를 주제어로 포함하고 있는 문서의 ID, 현재의 서버가 TLRS 또는 인용서버인 경우에는 하위 서버들의 리스트
      ② TS(Time Stamp) 정보
         해당 서버에서 가장 최근에 검색결과를 반환한 시간.
      ③ RF(Return Frequency) 정보
         일정 기간동안 해당 서버가 검색결과를 반환한 횟수.
   (2) 인용서버 리스트
      해당 용어를 참조하는 인용서버(즉, 계층구조 상의 상위 서버)들의 리스트.

EIF의 구성부분 중에서 TS 정보와 RF 정보는 분산색인 정보의 관리를 위해 사용된다. 이를 이용한 분산색인 정보의 관리기법에 대해서는 다음 절에서 기술한다.

3.3 분산색인 정보의 관리

본 절에서는 EIF를 이용한 분산색인 정보의 추가, 삭제, 갱신 등의 관리기법에 대해 기술한다.

3.3.1 새로운 분산색인 정보의 추가

문서서버에 새로운 문서가 추가되는 경우에는 분산색인 정보에 대한 계층구조를 운행(traverse)하면서 해당 문서와 문서의 주제어를 EIF에 저장한다. 이때, 계층구조의 운행은 EIF에 저장되어 있는 인용서버 리스트를 이용한다. 이러한 기능을 수행하는 분산색인 정보 추가 알고리즘은 [알고리즘 1]과 같다.

[알고리즘 1] 추가 알고리즘 Insert_EIF
입력 : 문서 D_i,

문서 D_i의 주제어 리스트 KW,
문서 D_i를 저장하고 있는 서버의 EIF
출력 : 분산색인 정보가 추가된 EIF'
begin
/* 문서 D_i을 저장하고 있는 문서서버의 EIF에 정보 추가 */
for(KW에 저장되어 있는 주제어 KW_j) begin
if(EIF에 KW_j가 존재하지 않는다) then begin
EIF에 KW_j 추가
인용서버 리스트를 공리스트(empty list)로 초기화
end

/* KW_j의 서버 리스트에 정보 추가 */
KW_j의 서버 리스트의 문서&서버 정보에 D_i 추가
KW_j의 서버 리스트의 TS 정보 ← 0
KW_j의 서버 리스트의 RF 정보 ← 0
end

/* 해당 서버를 참조하는 인용서버들의 분산색인에 정보 추가 */
for(KW_j의 인용서버 리스트에 존재하는 인용 서버 RS_k) begin
분산색인 계층구조 상의 인용서버의 EIF에 정보 추가(**Insert_EIF_Hierarchy**)
end /* for(RS_k) */
end /* for(KW_j) */
end.

[알고리즘 2] 분산색인 계층구조에 대한 EIF 추가 알고리즘 Insert_EIF_Hierarchy

입력 : 서버명 SrvName,
주제어 KW_j,
현재 인용서버의 EIF
출력 : 분산색인 정보가 추가된 EIF'
/* 서버가 인용서버 또는 TLRS /*
begin
SrvName ← 정보 추가를 요청한 서버명
if(KW_j가 EIF에 존재하지 않는다) then begin
EIF에 KW_j 추가
end

KW_j의 서버 리스트의 문서&서버 정보에 SrvName 추가
KW_j의 서버 리스트의 TS 정보 ← 0
KW_j의 서버 리스트의 RF 정보 ← 0

if(현재 서버가 TLRS가 아니다) then begin
/* 해당 서버를 참조하는 인용서버들의 분산색인에 정보 추가 */
for(KW_j의 인용서버 리스트에 존재하는 인용서버 RS_k) begin
계층구조 상의 인용서버의 EIF에 정보 추가 (**Insert_EIF_Hierarchy**)
end /* for(RS_k) */
end /* if(NOT TLRS) */
end.

3.3.2 분산색인 정보의 삭제

문서서버에서 문서 또는 주제어가 삭제되는 경우에는 해당 문서서버의 EIF만을 수정하고 분산색인 정보의 수정은 갱신과정을 통해 수행한다. 이러한 기능을 수행하기 위한 삭제 알고리즘은 [알고리즘 3]과 같다.

[알고리즘 3] 삭제 알고리즘 Delete_EIF

입력 : 삭제 조건,
문서 D_i,
문서 D_i의 주제어(keyword) 리스트 KW,
문서 D_i를 저장하고 있는 문서서버의 EIF
출력 : 분산색인 정보가 삭제된 EIF'
begin
if(삭제 조건이 문서 삭제) then begin
for(KW에 저장되어 있는 주제어 KW_j) begin
KW_j의 서버 리스트의 문서&서버 정보에서 D_i를 삭제
end /* for(KW_j) */
end /* if(문서 삭제 */
else if(삭제 조건이 주제어 삭제) then begin
KW의 서버 리스트의 문서&서버 정보에서 D_i를 삭제
/* 이 경우에는 KW가 하나의 주제어만을 포함 */
end /* else if(주제어 삭제) */
end.

3.3.3 분산색인 정보의 갱신

본 논문에서 제안한 분산색인 기법은 중앙집중형 색인기법을 기본으로 하는 방법이기 때문에 문서서버

의 색인정보가 해당 문서서버를 참조하는 모든 인용서버들의 EIF에 저장된다. 그러나 특정 인용서버가 자신이 참조하는 모든 문서서버의 색인정보를 저장하게 되면 그 양이 너무 방대해지므로 각 인용서버는 일정한 기준을 두어 자신의 분산색인 정보를 관리한다. 또한 각 인용서버의 EIF에 저장되어 있는 분산색인 정보는 해당 인용서버가 참조하는 문서서버들의 색인 정보와 일치해야 하기 때문에 일정한 기준에 의해 갱신되어야 한다.

이를 위해 본 논문에서는 이러한 분산색인 정보의 관리 기준으로 사용자의 질의를 이용하는 갱신 방법, LRU(Least Recently Used) 기법과 LFU(Least Frequently Used) 기법을 혼합한 방법을 사용한다. 즉, 특정 용어를 포함하는 임의의 질의에 대한 검색결과가 상위의 서버로 전달될 때, 검색결과의 반환이 이루어진 시간을 EIF의 TS 정보에 저장하고 반환이 이루어진 횟수를 EIF의 RF 정보에 저장하여 이를 분산색인 정보의 갱신에 이용한다.

먼저 EIF의 TS 정보가 갱신 기준을 만족하지 않는다는 것은 일정 기간이 경과될 때까지 특정 용어 또는 특정 서버에 대한 재검색이 이루어지지 않았다는 의미이므로 색인 정보로서의 의미를 상실한 것으로 간주하여 해당 분산색인 정보를 EIF에서 삭제한다. 또한 EIF의 RF 정보가 갱신 기준을 만족하지 않는다는 것은 일정 기간동안 해당 용어 또는 서버에 대한 검색결과의 반환 횟수가 주어진 기준 이하임을 나타내며, 이는 해당 용어 또는 서버에 대한 검색이 매우 드물게 수행됨을 의미한다. 이 경우에도 해당 분산색인 정보가 색인으로서의 의미가 적기 때문에 이를 EIF에서 삭제한다. 이러한 기능을 수행하는 분산색인 정보 갱신 알고리즘은 [알고리즘 4]와 같다.

[알고리즘 4] 갱신 알고리즘 Modify_EIF

입력 : 갱신 조건,
　　서버의 EIF,
　　검색결과 및 질의,
　　LRU 임계값(T_{LRU}) 및 LFU 임계값(T_{LFU})
출력 : 분산색인 정보가 변경된 EIF'
begin
if(갱신 조건이 검색결과의 반환) then begin
　Modify_EIF_by_Retrieval
end
else if(갱신 조건이 일치성 보장을 위한 EIF의 갱신) then begin
　Modify_EIF_for_Consistency
end
end.

[알고리즘 5] 검색결과 반환에 의한 EIF 갱신 알고리즘 Modify_EIF_by_Retrieval

입력 : 서버의 EIF,
　　검색결과 및 질의
출력 : 분산색인 정보가 변경된 EIF'
begin
for(질의에 포함되어 있는 각 용어 T_j) begin
　SrvName ← T_j를 포함하는 검색결과를 반환한 서버명
　SrvName의 TS 정보 ← 현재 시간
　SrvName의 RF 정보의 값을 1만큼 증가
end
end.

[알고리즘 6] 일치성 보장을 위한 EIF 갱신 알고리즘 Modify_EIF_for_Consistency

입력 : 서버의 EIF,
　　LRU 임계값(T_{LRU}) 및 LFU 임계값(T_{LFU})
출력 : 분산색인 정보가 변경된 EIF'
begin
for(EIF에 포함되어 있는 용어 T_k) begin
　for(T_k의 분산정보 부분 DI_i) begin
　　if((DI_i의 TS 정보 < T_{LRU}) AND
　　(DI_i의 RF 정보 < T_{LFU})) then begin
　　T_k에 포함되어 있는 분산정보 부분 리스트에서 DI_i를 삭제

　　현재 서버로부터 TLRS까지의 계층구조 상에 존재하는 모든 서버들의 EIF에서 DI_i를 삭제
　　end　/* if(TS AND RF) */
　　if(T_k에 포함되어 있는 분산정보 부분 DI_i가 존재하지 않는다) then begin
　　EIF에서 T_k를 삭제
　　end
　end　/* for(DI_i) */

end /* for(T_k) */
end.

3.4 EIF를 이용한 정보검색

EIF 정보를 이용하여 정보를 검색하는 알고리즘은 [알고리즘 7]과 같다. 사용자가 입력한 질의는 일차적으로 현재 사용자가 접속하고 있는 TLRS에 의해 처리된다. 만약 해당 TLRS에서 질의를 처리할 수 없거나, 제공된 검색결과에 대해 사용자가 만족하지 못해 재검색을 요구하는 경우에는 다른 TLRS에 해당 질의를 전달하여 검색을 수행하도록 요청한다.

[알고리즘 7] EIF를 이용한 검색 알고리즘

```
입력 : 질의,
       EIF
출력 : 문서들의 리스트
begin
/* 현재 서버 : TLRS */
if( 질의어가 EIF에 포함 ) then begin
  SrvList ← 해당 질의어에 대한 EIF의 서버 리스트
            의 문서&서버 정보에서 획득한 서버들
            의 리스트
  SrvList에 포함되어 있는 각 서버들에게 질의 처리
    요청(QueryProcessing_by_Server)
  SrvList에 포함되어 있는 각 서버들이 전달한 검색
    결과들을 병합하여 사용자에게 전달
end
else begin
  다른 TLRS에 검색 요청
  검색결과를 병합하여 사용자에게 전달
end

if( 사용자의 재검색 요구 발생 ) then begin
  다른 TLRS에 검색 요청
  검색결과를 병합하여 사용자에게 전달
end
end.
```

TLRS를 제외한 다른 서버들(문서서버 또는 인용서버)이 질의를 처리하는 방법은 [알고리즘 8]과 같다.

[알고리즘 8] 하위 서버에서의 검색 알고리즘

QueryProcessing_by_Server

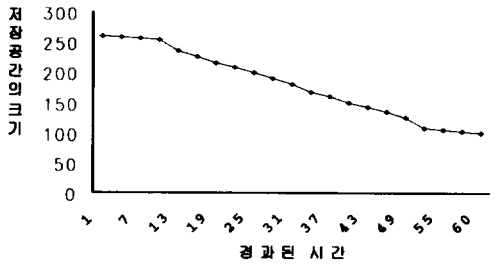
```
입력 : 질의,
       EIF
출력 : 문서들의 리스트
begin
if( 현재 서버가 문서서버 ) then begin
  DocList ← 해당 질의어에 대한 EIF의 서버 리스트
            의 문서&서버 정보에서 획득한 문서들
            의 리스트
  EIF의 서버 리스트의 TS 정보 ← 현재 시간
  EIF의 서버 리스트의 RF 정보의 값을 1만큼 증가

  검색결과인 DocList와 검색결과 반환에 의한 EIF 갱
    신 요청(Modify_EIF_by_Retrieval)을 상위 서버
    에 전달
end
else begin   /* 현재 서버가 인용서버 */
end
  SrvList ← EIF의 서버 리스트의 문서&서버 정보에
            서 획득한 서버들의 리스트
  SrvList에 포함되어 있는 각 서버들에게 질의 처리
    요청(QueryProcessing_by_Server)
  SrvList에 포함되어 있는 각 서버들이 전달한 검색
    결과들을 병합한 DocList와 검색결과 반환에 의한
    EIF 갱신 요청(Modify_EIF_by_Retrieval)을 상
    위 서버에 전달
end.
```

4. 실험 및 평가

본 논문에서 제안한 분산색인 기법의 성능을 평가하기 위해, 심마니, Altavista, Lycos 등을 비롯한 WWW 검색도구들을 사용하여 문서서버를 구축하였다. 즉, 특정 용어와 이에 대한 인터넷 검색결과 문서들을 데이터베이스로 구축한 후, 이를 문서서버로 사용하였다. 다음으로 네트워크 트래픽(Network Traffic)은 일정하다는 가정하에, 구축된 문서서버들에 대한 인용서버들을 구축하였다. 그리고 성능 평가를 위하여 시간의 경과에 따른 저장공간의 크기 변화, 인용서버 대 문서서버의 비율의 변화에 따른 응답시간의 변화, 마지막으로 시간의 경과에 따른 응답시간의 변화를 측정하여 그 결과를 이용하였다.

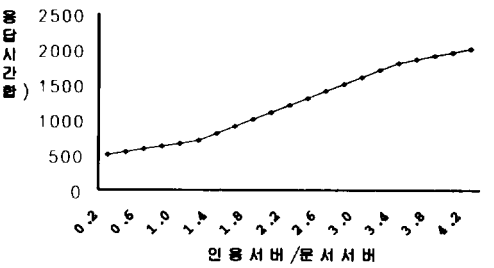
(그림 6)은 시간의 경과에 따른 저장공간의 변화 추이를 보여주고 있다. 이때 저장공간은 인용서버들에 저장되어 있는 분산색인 정보(EIF 구조에 저장된 정보)의 총합을 나타낸다.



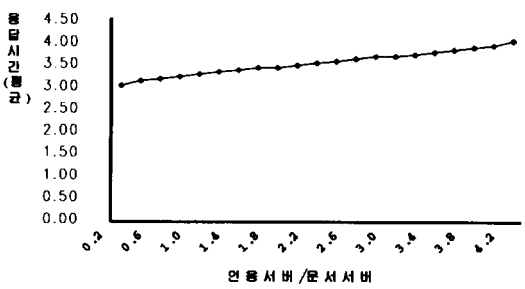
(그림 6) 시간의 경과 대 저장공간
(Fig. 6) Time elapsed vs Storage

(그림 6)에서 보는 바와 같이 시간이 지남에 따라 저장공간이 감소함을 알 수 있다. 이는 시간이 경과함에 따라 많은 질의가 입력이 되고, 그 결과 EIF 구조에 저장되어 있는 분산색인 정보에 대한 갱신이 이루어져 LFU 임계치인 정보와 LRU 임계치 이하인 정보가 삭제되었기 때문이다.

(그림 7)은 인용서버 대 문서서버의 비율에 따른 응답시간(response time)의 변화 추이를 보여주고 있다. (그림 7)을 살펴보면 인용서버 대 문서서버의 비율이 증가할수록 응답시간이 완만한 상승을 보이고 있다. 이는 인용서버가 많아질수록 동일한 문서서버나 인용서버를 참조하는 인용서버도 증가되기 때문에 질의가 수행되는 인용서버의 수도 증가되고 그 결과 전체 응답시간이 증가됨을 알 수 있다. 그러나 (그림 8)에서 알 수 있듯이 평균 응답시간은 거의 일정하다.

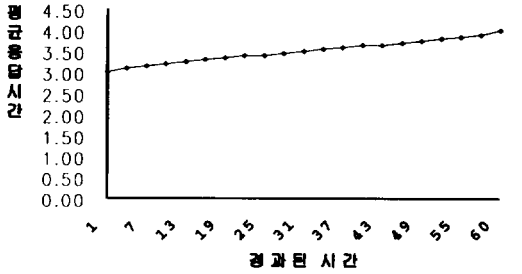


(그림 7) 서버의 비율 대 응답시간의 합
(Fig. 7) Ratio of Servers vs Sum of Response Time



(그림 8) 서버의 비율 대 평균응답시간
(Fig. 8) Ratio of Servers vs Average Response Time

(그림 9)는 시간의 경과에 따른 응답시간(response time)의 변화 추이를 보여주고 있다.



(그림 9) 시간의 경과 대 응답시간
(Fig. 9) Time elapsed vs Response Time

(그림 9)에서 보는 바와 같이 시간이 경과하여도 응답시간의 변화가 거의 없다. 이러한 결과를 통해 분산색인 정보에 대한 갱신 알고리즘에 의해 일부 분산색인 정보가 삭제되더라도 전체 시스템의 검색 성능에는 거의 영향이 없다는 것을 알 수 있다.

결국 이와 같은 여러 실험에 의해 본 논문에서 제안한 분산색인 기법은 검색성능(즉, 응답시간)에는 거의 영향을 주지 않으면서 갱신 알고리즘에 의해 일부 분산색인 정보를 삭제함으로써 전체 시스템의 저장공간을 줄일 수 있었다. 또한 사용자가 관심을 가지고 있는 분야 또는 주제에 해당하는 인용서버를 특수화 기법을 적용하여 구축할 수 있다는 장점이 있다.

5. 결론 및 향후연구과제

본 논문에서는 중앙집중형 색인기법과 분산형 색인기법을 혼합한 새로운 분산색인 기법을 제안하였다.

즉, 정보검색에서 일반적으로 사용되는 역파일을 확장한 확장된 역파일(Extended Inverted File, EIF) 구조를 제안하고, 이에 대한 관리기법과 이를 이용한 검색기법을 제안하였다. EIF에는 사용자가 입력한 질의에 적합한 정보자원을 저장하고 있는 서버들만을 선택하기 위한 정보와 각 인용서버들에 저장되어 있는 분산색인 정보들을 관리하기 위한 정보가 포함되어 있다. 또한 각 인용서버에 저장되어 있는 분산색인 정보를 문서서버들의 정보와 일치시키기 위한 방법으로는 LRU (Least Recently Used) 기법과 LFU(Least Frequently Used) 기법을 응용한 방법을 사용하였다. 즉, 사용자의 질의에 대한 검색결과가 반환된 가장 최근의 시간과 횟수를 이용하여 각 인용서버에 저장되어 있는 분산색인 정보를 갱신하였다. 그리고 본 논문에서 제안한 분산색인 기법의 성능에 대한 실험을 수행하여 전체 시스템의 성능(즉, 검색요구에 대한 응답시간)에 거의 영향을 미치지 않고, 일부 분산색인 정보를 삭제하여 전체 시스템이 요구하는 저장공간의 크기를 줄이면서도, 다양한 사용자의 관심주제를 반영한 다중 계층구조의 분산색인을 구축할 수 있음을 증명하였다.

앞으로 TLRS 사이의 효율적인 통신방법, 해당 서버에 대한 각 용어의 적합성 정보를 EIF 구조에 포함시키는 방법 그리고 인용서버를 구축하기 위한 방법론으로 에이전트를 이용하기 위한 방법 등에 대한 연구가 필요하다.

참 고 문 헌

[1] C. M. Bowman 외 6인, "Harvest: A Scalable, Customizable Discovery and Access System," Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado-Boulder, 1994.

[2] Arturo Crespo, Hector Garcia-Molina, "Distributed Differential Indexing," http://walrus.stanford.edu/diglib/pub/slides/sitevisit0497/crespo/index.htm

[3] P. B. Danzig 외 3인, "Distributed Indexing : A Scalable Mechanism for Distributed Information Retrieval," ACM SIGIR '91, pp.220-229.

[4] J. Davies, R. Weeks, M. Revett, "Jasper : Communicating Information Agent for WWW," Proc. of the 4th Int'l World Wide Web Conf., 1995.

[5] A. Dupa and M. A. Sheldon, "Content Routing in Networks of WAIS Servers," 14th IEEE Int'l Conf. on Distributed Computing Systems, 1994.

[6] Martin Hamilton, "Distributed Indexing and IP Multicast," http://www.roads.lut.ac.uk/Reports/centroid/centroid.html.

[7] Arkadi Kosmynin, "Distribution in Internet Resource Discovery," Proc. of Australian World Wide Web Technical Conf., May, 1997.

[8] Martijn Koster, "ALIWEB - Archie-Like Indexing in the WEB," http://www.bib.lu.se/elbibl/litt/paper.html.

[9] C. Lagoze and J. R. Davis, "Dienst : An Architecture for Distributed Document Librasries," CACM, Vol.38, No.4, 1995.

[10] H. Lieberman, "Letizia : an agent that assists Web browsing," Proc of the Int'l Joint Conf. on Artificial Intelligence, 1995.

[11] C. Lynch and H. Garcia-Molina, "Interoperability, Scaling, and the Digital Libraries Research Agenda," IITA Digital Libraries Workshop, 1995.

[12] Miguel Rio, Joaquim Macedo, Vasco Freitas, "A Distributed Weighted Centroid-based Indexing System," Proc. of JENC8, pp.322-331.

[13] M. A. Sheldon 외 3인, "Discover : A Resource Discovery System based on Content Routing," Proc. of the 3'rd Int'l World Wide Web Conf., 1995.

[14] S. Shen 외 7인, "An Interoperable Architecture for Digital Information Repositories," DL94, http://abgen.tamu.edu/DL94/paper/shen.html.

[15] E. M. Voorhees, "Agent Collaboration as a Resource Discovery Technique," CIKM Workshop on Intelligent Information Agents, 1994.

[16] C. Weider, J. Fullton, S. Spero, "Architecture of the WHOIS++ Index Service," RFC 1913, Feb., 1996, http://ds.internic/net/rfc/rfc1913.txt

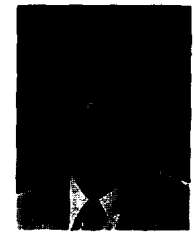
[17] R. Weiss 외 6인, "HyPursuit : A Hierarchical network search engine that exploits cintent-link hypertext clustering," Proc of the 1996 7th ACM Conf. on Hypertext, 1996.

[18] Budi Yuwono 외 3인, "A World Wide Web Re-

source Discovery," Proc. of the 4'th Int'l World Wide Web Conf., Dec., 1995.
[19] "Alexandria Digital Library," http://alexandria.sdc.ucsb.edu/descripsion/organization.html
[20] "Definition and Purposes of a Digital Library," http://sunsite.berkeley.edu/ARL/definition.html
[21] "Digital Libraries:Issues and Architectures," http://www.csdl.tamu.edu/DL95/papers/nuernberg.html
[22] "A Framework for Distributed Digital Object Services," http://www.cnri.reston.va.us/home/cstr/k-w.html
[23] "Network Information Discovery and Retrieval," http://www.finchcomputer.com/~phil/class/cs586/presentation/
[24] "Stanford Digital Library Project," http://www.computer.org/pubs/dli/r50061/r50061.html
[25] "University of Michigan Digital Library Initiatives Overview," http://www.lib.umich.edu/libhome/digitalprojects.html
[26] 김동규 외 2인, "분산자원 검색을 위한 색인기법 연구", 한국정보과학회지, 제15권, 제2호, 1997.
[27] 맹성현, "디지털 도서관 관련 기반 기술 및 고급 기술", 한국정보과학회 데이터베이스연구회, '96 추계 튜토리얼, pp.111-139, 1996.
[28] 이준호, 안정수, "정보검색 시스템 KRISTAL-II", 한국정보과학회지, 제15권, 제2호, 1997.

유 춘 식

e-mail : csyoo@cs.chonbuk.ac.kr
1991년 8월 전북대학교 전산통계학과 졸업(이학사)
1994년 전북대학교 대학원 전산통계학과(이학석사)
1994년~현재 전북대학교 대학원 전산통계학과 박사과정
관심분야 : 디지털 도서관, 정보검색, 자동색인, 분산색인, 인공지능, 멀티에이전트 시스템 등



이 종 득

e-mail : cdlee@tiger.seonam.ac.kr
1983년 전북대학교 전산통계학과 졸업(이학사)
1989년 전북대학교 전산통계학과 (이학석사)
1998년 전북대학교 전산통계학과 (이학박사)
1992년~현재 서남대학교 전자계산학과 조교수
관심분야 : 디지털 도서관, 정보검색, 지식공학, 인공지능, 개념 클러스터링 등

김 용 성

e-mail : yskim@moak.chonbuk.ac.kr
1978년 고려대학교 수학과 졸업 (이학사)
1984년 광운대학교 전산학과(이학석사)
1992년 광운대학교 전산학과(이학박사)
1985년~현재 전북대학교 컴퓨터과학과 교수
1996년~1998년 1월 한국학술진흥재단 전문위원
관심분야 : 디지털 도서관, 정보검색, 인터넷 기반 정보검색, 멀티미디어 시스템, 인공지능 등