□ 기술해설 □

분산자원 검색을 위한 색인기법 연구

서울대학교 김동규*·이상구**
명지대학교 전종훈**

● 목 차 ●

1. 서 론

인터넷의 확장과 보급은 하드웨적인 요구뿐만 아니라 소프트웨어적으로 새로운 요구들을 만들어 내고 있다. 그 중에서도 우리가 가장 기본적이면서도 쉽게 접하게 되는 부분이 정보검색이다. 단일 공간내의 소규모 정보로부터 확장 공간내의 대규모 정보로의 이동은 대상이 되는 정보를 검색하기 위한 새로운 기술들을 요구하고 있다. 이미 웹상에서 서비스되고 있는 다양한 정보검색 서버들은 이러한 요구사항에 맞게 발전해 나가고 있고 하나의 분야로 자리잡아 가고 있음이 사실이다. 이와같은 사실들은 크게 두가지 측면에서 특징적인 환경의 변화를 암시하고 있다[1].

첫째는, 정보원천 측면에서의 분산성이다. 여기서의 분산성은 결함내성(fault tolerance)의 관점이라기보다는 하나 이상의 정보제공자 혹은 대규모 정보량의 효율적 관리를 위한 분산이라고 할 수 있다. 웹상의 여러 검색엔진들은 정보제공자가 아무런 제한이 없다는 점에서 분산성의 극단적인 예라고 할 수 있다.

두번째는 이렇게 분산되어 있는 정보들의 사용자 측면에서의 투명성이다. 사용자들은 다양한 정보를 원하기 때문에 가능하면 많은 정보원천을 검색하고 싶어하지만 그 모든 것이 하나의 인터페이스로 이루어지기를 바란다.

만일 전자 도서관 연구와 관련된 기술보고서나 논문들을 찾으려고 한다면 가장 먼저 웹의 검색엔진들을 생각해 볼 수 있지만 이들은 검색의 대상이 무제한적이기 때문에 필요로하는 정보보다는 불필요한 정보를 더욱 많이 검색한다. 따라서 각 대학들이 제공하고 있는 대학별 검색 서비스를 사용하는 것이 정보의 질적 측면을 고려할 때 오히려 효과적일 수 있다. 하지만, 이러한 대학별 검색서비스의 사용은 일일이 새로운 서버에 접속하여 검색을 해야하기 때문에 각기 특성에 맞는 검색서버의 사용법을 익혀야 하고, 검색해야 할 대상이 많으면 많을수록 불편을 겪게 된다. 이때 필요한 것이 하나의 인터페이스를 통한 통합검색이다. 결과적으로 한번의 검색요청을 통합검색 서버가 받아서 제어가 가능한 분산되어 있는 여러 정보원천들에 대해 검색을 수행하게 되고 이를 통해 사용자에게는 마치 하나의 정보원천에서 추출된 것처럼 결과를 보여주게 된다. 이것이 바로 위에서 살펴본 정보원천의 분산성과 사용자 검

*비 회 원
**종신회원

색의 투명성이 요구되는 하나의 예라고 할 수 있다[2].

이를 위해서는 기존의 정보검색 시스템들에 대해 몇가지 새로운 기능성들이 요구되는데 본 글에서는 그중에서도 분산된 자원에 대한 색인 방법에 대하여 알아보고 새로운 접근방식을 제안할 것이다. 본 글은 다음과 같이 구성된다. 2장에서는 전자 도서관에서의 정보검색 기술들을 간단히 살펴보고, 3장에서는 다양한 시스템에 적용되고 색인관리 방법을 분류함으로써 어떤 방식으로의 접근이 이루어지고 있는지를 살피는 동시에 각각의 장단점을 설명한다. 끝으로 4장에서는 가장 효율적인 분산색인 관리를 위해 필요한 일들과 새로운 접근방식을 제시한다.

2. 전자 도서관(Digital Library)과 정보검색

전자 도서관에 대한 관점은 무척 다양하다[1][2]. 전자 도서관은 일반적인 정보검색 시스템과의 차이점이 불분명하고 최근 유행하고 있는 웹의 검색 시스템들과의 구분도 애매한 면이 있다. 분명 최근의 전자 도서관 연구 및 개발 프로젝트들은 여러 가지 면에서 기존 검색 시스템들과는 차별화된 전략들을 밝히고 있지만, 그들간에는 정보검색이라는 기능적 공통성이 존재하고 있다. 즉, 기존의 검색 시스템들과 현재의 전자 도서관 시스템간에는 병렬적이고 근본적인 차이점이 존재한다기 보다 일반 검색 시스템이 전자 도서관으로 그 기능이 확장된 것이라는 발전적 선상에서, 또는 일반 검색 시스템이 포함되는 또하나의 새로운 응용형태로서 전자 도서관을 논하는 것이 올바르다. 이러한 관점으로 볼 때, 웹상의 여러 검색 시스템들을 굳이 전자 도서관과 분리해 낼 필요는 없다. 다만 분명히 짚고 넘어가야 할 점은 웹이라는 특징적인 형태인데 이는 단지 폭넓은 사용자층을 가진 인터페이스의 활용일 뿐이라는 점이다. 즉, 많은 검색 시스템들이 각각 고유의 클라이언트 프로그램들을 가지고 있는 반면 웹서버의 HTTP에 의한 모든 서비스들은 웹브라우저라는 일반적인 인터페이스로 사용이 가능하다는 막강한 장점을 지니고 있다는 점에서 고려의 대상이 될 수는 있지만, 그것이 모든 전자 도서관의 필수적인 요소는 아니다. 또한 웹상의 검색 시스템들에서는 검색의 대상이 다른 웹서버에 의한 브라우징이 가능한 문서들에 국한된다는 사실을 알아야 한다. 즉, Yahoo나 infor-seek등의 일반적 웹 검색 시스템들은 검색대상을 웹상의 홈페이지들로 하고 Netscape와 같은 웹 브라우저를 인터페이스로 하며, 로봇을 통해 만들어진 색인을 사용하는 분산 전자 도서관으로 분류가 가능하다.

이렇게 전자 도서관 연구의 주요관점은 단순한 도서관 자료의 전산화가 아니라 보다 효율적이고 편리한 자료열람을 위한 전자화된 자료의 수집, 배치 그리고 운영을 위한 동적 기능이다. 사실 이러한 연구 관점은 기존의 정보검색 시스템의 관점과 별반 다르지 않으며 결국 정보 검색이라는 큰 테두리를 중심으로 여러 가지로 분화되고 있다.

2.1 전자 도서관의 검색과정

일반적으로 검색은 크게 두 단계를 거치게 된다. 첫번째 단계에서는 사용자의 질의에 의해 자료의 목록을 검색한다. 그런다음 두번째 단계에서는 추출된 목록들 중에서 사용자에 의해 요구되는 특정 자료를 추출한다. 이 두가지 단계에서 검색이란 첫번째 단계에서 이루어지는 목록의 추출을 말한다. 즉, 사용자의 질의어에 적합한 문서들을 찾아내어 그 목록을 사용자에게 다시 보여주는 일련의 과정이 바로 검색과정이라고 할 수 있다. 이 검색과정은 다음과 같이 이루어진다.

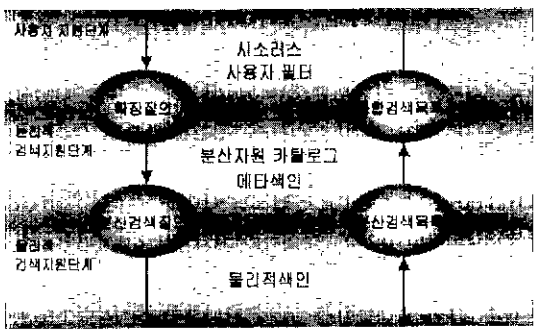

그림 1 전자 도서관 검색과정과 각 단계별 사용자원

사용자 지원단계에서는, 시소러스등을 사용하여 사용자의 질의어를 유사어, 상위어, 하위어등으로 확장함으로써 검색의 폭을 넓게하는 동시에 검색 시스템이 지원하는 검색모델에 맞추어 검색된 결과에 대해 가중치를 정한다.

논리적 검색 지원단계는, 사용자 지원단계에서 구성된 최종 질의를 실제 검색을 위한 물리적 검색 엔진에 전달하기 전에 정보원천의 분산성과 이질성을 고려하여 중간 해석단계를 거치는 과정으로서 특히 분산된 자원에 대해서는 별도의 메타색인을 활용하여 검색의 효율을 높이고 마찬가지로 물리적 검색의 결과를 다시 논리적으로 통합하여 사용자에게 보여주는 역할을 맡는다.

물리적 검색 지원단계는, 실제로 최종적인 검색이 이루어지는 단계로서 색인을 사용하여 내용기반검색등을 지원하고 정보저장공간의 효율적 사용할 수 있게 하는 여러 가지 기술들이 사용된다.

이와 같은 구분이 모든 시스템에서 곧바로 독립적인 모듈들로 연결되지는 않는데, 특히 두 번째의 논리적 검색지원 단계 없이도 전자 도서관의 검색시스템은 완결될 수가 있다. 하지만, 분산 자원 내지는 분산 전자 도서관의 상위 통합 서버를 구축한다면 대부분의 시스템이 각 단계의 기능성을 유지한다는 면에서 위와 같은 공통적인 구조를 갖게된다. 그중에서도 논리적, 물리적 검색지원에서 보여지듯 실제 자료와 검색기능을 이어주는 역할을 하는 색인은 모든 검색 시스템에 필수적인 요소이다[3][4].

2.2 자료구조(data structure)로서의 색인과 카탈로그(catalog)로서의 색인

일반적으로 색인은 특정 필드에 대해 정의되며 검색속도의 복잡도(complexity)를 O(n)이하로 낮추기 위한 일종의 도구로서 몇가지 자료구조로 표현될 수 있다. 대표적인 것으로는 B-tree나 hashing기법 등이 있고, 공간자료(spatial data)를 위한 grid file, kd tree, R-tree등이 있다. 이런 자료구조로서의 색인의 가장 기본적 목적은 무수히 많은 단어나 모양들 중에서 원하는 것을 빨리 찾아내는 것이다.

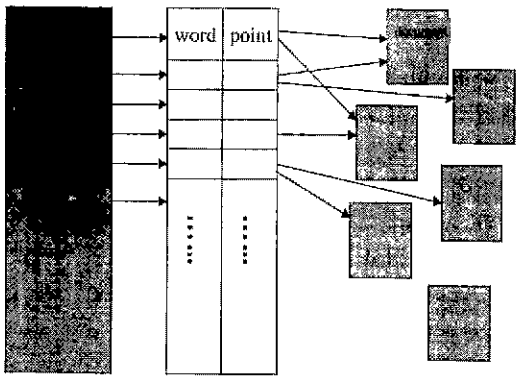

그림 2 역화일(inverted file)로 구성한 검색시스템의 색인구현의 예

하지만, 전자 도서관과 같은 검색 시스템에서는 자료구조 이상의 동작을 하는 색인을 요구한다. 단지 단어를 빨리 찾는 것이 아니라 그 단어와 관련이 있거나 정확히 일치하는 항목을 가진 문서들의 목록을 원한다는 측면에서 자료구조보다는 한 차원 높은 수준의 카탈로그로서의 색인이 요구된다. 결국 전자 도서관의 색인은 자료구조로서의 색인을 사용하여 해당 단어를 빨리 찾아내고 그와 연관된 별도의 문서목록을 최종적으로 반환할 수 있는 카탈로그로서의 색인이라고 할 수 있다. 이런 카탈로그로서의 색인으로 많이 사용이되는 것들에는 역화일(inverted file)이나 시그너쳐 파일(signature file)등이 있다[5][6].

3. 분산자원 검색과 색인

3.1 분산환경과 색인

분산환경은 그림 1의 단일 검색과정을 몇가지 가능성으로 분화할 수 있게 한다. 위에서도 밝혔듯이 분산환경을 고려할 때 두번째 단계인 논리적 검색지원 단계의 기능이 기능적으로 강화될 필요가 있다[7]. 이 단계의 기능적 강화란, 상위 단계에서 요청하는 사용자 질의를 하위 단계에 맞도록 재해석하여 요청을 하기 위한 중간매체적 기능의 강화를 말하는데, 하위의 실제자료들이 분산되어 있는 방식에 따라 의존적으로 구성하게 된다. 이러한 기능을 담

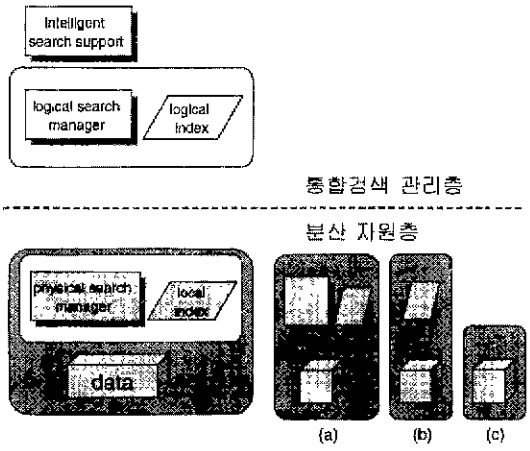

그림 3 분산의 형태

당하기 위하여 필요한 메타정보들은 실제자료의 분산형태를 충분히 반영할 수 있어야 한다.

그림 3은 분산 형태를 분류한 것이다.

그림 1에서 보여지는 논리적 검색지원 단계는 그림 3의 통합검색 관리층의 핵심과정으로서 통합검색을 주도하는 모듈로 구현이 된다. 분산의 대상이 되는 부분은 순수한 자료만일 수도 있고, 그 이상일 수도 있다. 그림 3의 분산 자원층으로 분류되는 층에서의 분산 대상은 기본적으로 문서 및 멀티미디어 자료이다. 하지만, 분산 시스템의 검색 효율을 위하여 분산의 대상을 넓힐 수 있다. 각각의 분산된 자료에 대한 직접적인 정보를 갖는 지역 색인, 그리고 그 색인을 사용하여 통합 검색과는 별도로 검색이 가능하게 하는 최하위 검색관리자가 새로운 분산 대상이 된다. 이중에서도 색인의 분산은 전체 통합검색 시스템의 색인을 구성하는 방식에 중요한 영향을 미친다.

이와 같은 분산 대상에 따라 분산형태의 가능성은 다음과 같은 세가지 방식으로 나누어진다.

(a) 완전히 독립적인 검색시스템으로의 분산이다. 이 경우에는 통합검색 관리층의 구현 방향이 다양해 질 수 있다. 즉, 독립적인 분산 검색시스템의 색인과 검색모듈이 활용가능하기 때문에 별도의 전체 색인이 없이도 통합검색이 가능하다. 이에 대해서는 다음 절에서 자세히 설명할 것이다.

(b) 독립적인 검색모듈이 없이 자료와 색인만의 분산된 형태이다. 이 경우에는 통합검색 관리층의 검색엔진에 의해 각 분산 색인들이 사용되는 구조로 통합검색 시스템이 구축될 수 있다.

(c) 단순한 자료만이 분산되어 있는 형태로서 전체 분산 자료에 대한 색인과 검색을 통합검색 관리층이 담당해야 한다. 이러한 방식 역시 다른 형태와 비교할 때 몇가지 장단점을 지니게 된다.

3.2 분산된 자원에서의 색인관리

분산 자원에 대한 색인관리 기법을 논하기 위해서는 한 가지 가정이 필요하다. 일단 기존의 자료가 방대해짐에 따라 그 자료를 나눔으로써 행해지는 자료의 분산화도 물론 이 논의의 대상에 포함되지만 그보다는 이미 구축되어 있는 다른 자료 및 검색 시스템의 통합 관점에서의 분산 색인관리를 가정한다. 사실 최근의 이 분야에 대한 연구에서는 대부분 이러한 이질성의 고려가 필수적으로 요구되고 있다.

3.2.1 분산 색인방식과 중앙집중식 색인방식

최근 분산 전자 도서관의 통합관리에서 사용되는 가장 일반적인 방법은 분산되어 있는 색인을 그대로 사용하는 것이다. 그림 3에서 본다면 (a), (b)의 형태와 같이 분산자료들에 대한 독립적인 색인이 존재할 때 쓸 수 있는 방법으로서 단순히 검색결과의 수집만을 생각하면 된다는 장점을 지니고 있다. 즉, 각 분산되어 있는 색인에 대한 고려의 부담을 최소화할 수 있기때문에 통합검색 관리층의 구현이 상대적으로 쉬워진다. 그뿐 아니라 색인과 실제 정보사이에 생겨날 수 있는 불일치성(inconsistency)에 대한 가능성도 줄어든다는 장점이 있다. 뒤에 언급하겠지만, 통합검색 관리층에 미리 수집하여 두는 중앙집중식 색인관리 기법은 정보의 변화에 민감히 반응하지 못하는 반면 독립적인 분산색인의 활용은 이 문제를 고려하지 않아도 된다. 하지만, 이러한 방식은 결정적인 문제점을 가지고 있다. 매 질의마다 관리되고 있는 모든 분산 자료에 대한 질의를 수행해

표 1 중앙집중식 색인(Centralized Index)와 분산 색인(Distributed Index)의 비교

	Centralized Index	Distributed Index
response time for user's query	fast and efficient	slow and useless search
consistency between index and data	inconsistent need periodic index update	consistent
resource capability for data expansion	impossible to control the central index	no problem

야 한다는 점이다. 이같은 상황은 검색이 요구될 때 이 질의에 대해 어느 위치에 어떤 적당한 결과가 있을 것이라는 것을 실제 검색이 이루어지기 전에는 알 수 없다는 점에서 발생되는 문제로, 적당한 결과를 주지 못할 분산 자료에 대해서도 질의를 주어야만 하는 부담이 발생한다.

이러한 문제는 한 곳에서 전체 분산된 자료들에 대한 색인을 한꺼번에 관리하지 않는 한 해결이 불가능한 문제이기도 하다. 반면 이런 문제를 해결하기 위해 분산되어 있는 모든 자료에 대한 색인을 통합 관리층에 모아 둔다면 새로운 문제가 야기된다. 색인을 통합관리함으로써 검색의 반응속도는 빨라지지만, 그 색인을 작성하기 위해 많은 노력이 필요하게 된다. 일단 분산되어 있는 독립적인 색인들의 형식이 다를 수 있기 때문에 이들을 통합한다는 것은 모든 색인의 형식을 알아야 한다는 것을 의미할 뿐만아니라 새로운 자료의 추가시에는 매번 새로운 노력이 필요하다. 더욱이 대부분의 색인은 앞장에서 설명했듯이 단순한 텍스트가 아니라 그 색인을 사용하는 검색 엔진에 의존적으로 작성되는 특수한 구조이므로 여간 복잡한 처리를 거치지 않고는 색인의 통합은 불가능하다. 물론, 그림 3의 (c) 형태와 같이 아예 각 분산 자료의 독립적 색인은 고려하지 않고 자체적으로 통합관리층의 주도하에 새로운 전체 색인을 만들 수는 있지만, 이 역시 각 분산 자료가 다양한 저장방식으로 이루어져 있을 수 있기 때문에 색인 작성의 어려움은 마찬가지이다. 거기다가 최근의 검색시스템들은 전문검색을 지원하기 위해 색인을 카탈로그화하고 있기 때문에 색인의 크기는 거의 문서의 크기만큼 만들어 진다. 이 때문에 통합된 색인의 크기는 관리하기에는 너무나 큰 자원을 차지하게 되고 이를 또다시 분산화해야 하는 새로운 문제가 생겨난다. 그리고, 앞에서 말했듯이 한 곳에 미리 작성된 색인과 실제 자료간에는 정기적인 갱신주기를 사이로 불일치 상태가 생겨날 수밖에 없다.

3.2.2 실제 전자도서관 적용을 통해 본 다양한 변형들

중앙집중식 색인방식을 사용하는 대표적인 시스템은 대다수 웹상의 검색엔진들이다[8][9]. 이들은 앞서 설명했듯이 HTTP를 사용해서 추출이 가능한 문서들을 대상으로 한 비교적 폭넓은 영역에 대한 검색 시스템들이다. 결국 HTML로 만들어진 문서들이 주 대상이 되는데 인터넷에 퍼져있는 많은 웹서버들이 분산되어 있는 자료관리층이라 할 수 있고, 검색서비스를 담당하는 별도의 웹서버가 통합검색 관리층이 된다. 이런 구조는 웹의 특성상 분산된 독립적 색인이 존재하지 않기 때문에 통합관리층이 색인에 대한 모든 부담을 안게 된다. 이때 모든 문서에 대한 색인을 미리 수집하여 작성할 필요가 있는데 이 일을 담당하는 것이 로봇이다. 로봇은 몇가지 단계의 필터링을 하긴 하지만 웹상의 모든 문서를 일일이 통합관리층으로 가지고 와서 전문을 검색하여 색인을 만들기 때문에 네트워크에 상당한 부담을 주는 등 위에서 이야기 한 중앙집중식 색인의 특징

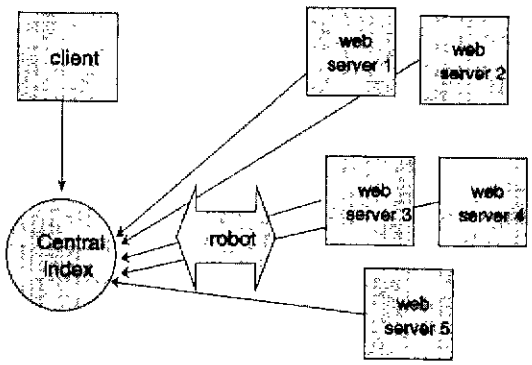

그림 4 robot을 사용하는 중앙집중식 색인방식 예

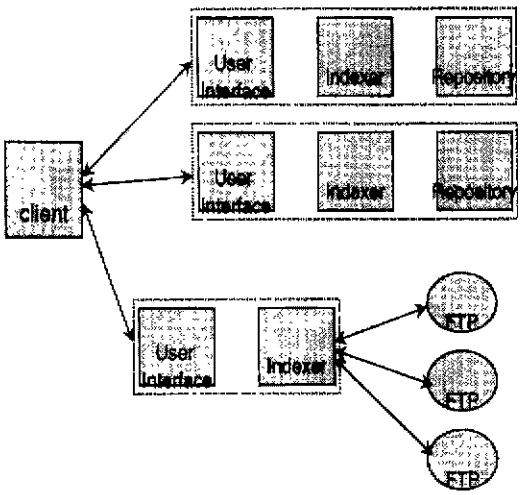

그림 5 분산 색인을 사용하는 Dienst protocol

을 그대로 가지고 있다.

반면, NCSTRL 프로젝트에서 사용하는 Dienst protocol은 분산색인 방식을 쓰고 있다[10][11][12]. 엄밀히 이야기하자면 중앙집중식 색인방식을 함께 사용하는데, 일단 별도의 검색모듈이 없이 FTP가 가능한 호스트에 검색대상이 되는 문서들을 저장하고 Dienst가 정하는 형식의 색인을 작성하여 문서들과 함께 두면 이러한 몇몇 FTP 호스트를 관리하는 중간단계의 Indexer가 있어서 하위의 FTP 호스트들에 있는 모든 문서들에 대한 색인을 작성해 놓는다.

이 Indexer와 같은 수준의 다른 Indexer들이 여러 개가 있을 수 있는데 이들에 대해서는 별도의 전체 색인 없이 분산되어 있는 형태로 검색을 하게 되어있다. 이렇게 이중적인 구조를 갖는 이유는, 자기의 호스트에 특별한 검색 소프트웨어를 추가하지 않고도 다른 검색서버를 통해 검색이 가능하도록 할 수 있게 하기 위해서이다.

물론, 정보제공자가 원한다면 중간단계의 Indexer 서버와 User Interface 서버를 설치함으로써 자체적으로 검색이 가능하게 할 수도 있다. 결과적으로 이 시스템은 전체 문서를 검색하기 위해서는 모든 Indexer를 검색해야 하기 때문에 분산 색인 방식이 갖는 장단점을 갖게된다.

4. 새로운 경향과 앞으로의 전망

4.1 기존 방식들의 문제점 해결의 노력과 새로운 방식들

분산환경하에서 색인을 처리하는 문제는 지금까지 살펴 본 방식들에서 약간씩의 변형으로 이루어지고 있다. 중앙집중식 색인 방식의 전형이라고 할 수 있는 로봇 시스템들은 색인의 작성시 비용이 많이 든다. 이 문제는 전체 네크워크 차원에서도 자원낭비라는 큰 문제를 야기시키고 있다. 물론, 웹과 같은 큰 영역이 아니라 필요한 자료를 가지고 있는 몇군데의 정보저장소를 대상으로 한다면 작성 비용이 많이 줄어들 수는 있지만, 다른 방식과 비교되는 상대적인 비용은 여전히 클 수밖에 없다. 왜냐하면, 로봇은 모든 문서를 검색서버로 가지고 와서 색인을 작성하기 때문이다. 그래서, 최근에는 색인을 작성시에 여러 가지 필터링 기법을 동원하여 가져와야 할 문서의 양을 최소화하려는 노력들이 있다. 색인작성을 위해 검색서버에 가지고 온 문서내에서 또다른 문서를 가져오기 위해 하이퍼링크를 조사할 때 필터링을 해서 꼭 필요한 문서만을 색인작성 문서목록에 추가시킨다. 이렇게 중앙집중식 색인 방식을 쓰는 시스템들은 색인작성시에 드는 비용을 줄이고 꼭 필요한 내용에 대해서만 색인을 작성함으로써 색인자체의 크기를 최소화하려는 방향으로 이동하고 있다.

이와는 달리 이미 중앙색인에 대한 부담이 없는 분산색인 방식의 시스템들은 질의 응답시간을 최소화하기 위해 노력하고 있다. 이 방식은 자료의 분산형태 뿐만 아니라 분산된 자료내용에도 큰 영향을 받게 된다. 만일 자료들이 내용별로 각기 다른 곳에 분산되어 있는 환경이라면 무조건적인 분산 색인 방식의 사용은 거의 항상 불필요한 질의를 유발하게 된다. 왜냐하면, 질의내용이 아주 일반적이라면 상관없지만 많은 경우에 특정 분야에 속하게 되고 그렇다면 이 질의에 대한 결과는 그 분야에 대한 정보를 소장하고 있는 저장소에서만 찾아볼 수 있기 때문이다. 이러한 관점에서 분산 색인 방식은 다양한 새로운 시도들이 필요하다.

4.2 서울대 전자 도서관 한울(Heterogeneously Archived and Networked Universal Library)에서의 카탈로그 관리자

서울대 전자 도서관 한울은 카탈로그 관리자(Catalog Manager)라는 이름으로 분산된 자료에 대해서 색인을 처리한다. 한울은 이름에서 보여지듯이 이질적이고 분산되어 있는 정보저장소들을 대상으로 사용자들에게는 한 전자도서관으로 투명한 서비스를 제공하는 시스템이다.

그림 6은 한울의 시스템 아키텍처이다. 카탈로그 관리자는 사용자 질의가 있을 때 연결 관리자(Connection Manager)에 의한 색인정보 문의를 처리하는 부분으로써 기존과는 다른 가상색인 기법을 사용하고 있다. 카탈로그 관리자는 전체 분산되어 있는 문서들에 대한 색인을 부분적으로 가지고 있게 되는데 만일 사용자 질의가 그 가상색인내에 이미 작성되어 있다면 별도의 분산 질의 없이 바로 사용자에게 검색 목록을 보여주게 된다. 이러한 과정은 그림 7에 나와있다. 그림에서 실선으로 표시된 부분이 방금 설명한 과정으로 가상색인 제어기(Virtual Index Controller)는 사용자 질의를 가상색인을 사용해 1차 검색을 하고 그 결과를 연결관리자에게 넘겨주면 연결관리자는 그 결과를 그대로 사용자에게 보여준다. 하지만, 가상색인으로부터 아무런 결과가 반환되지 않았다면 2차로 분산되어 있는 모든 자료에 대해

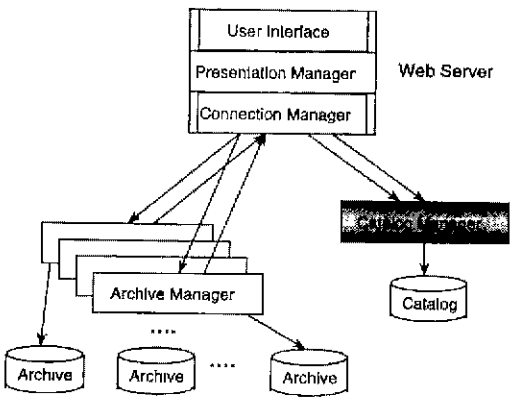

그림 6 전자 도서관 한울의 시스템 아키텍처

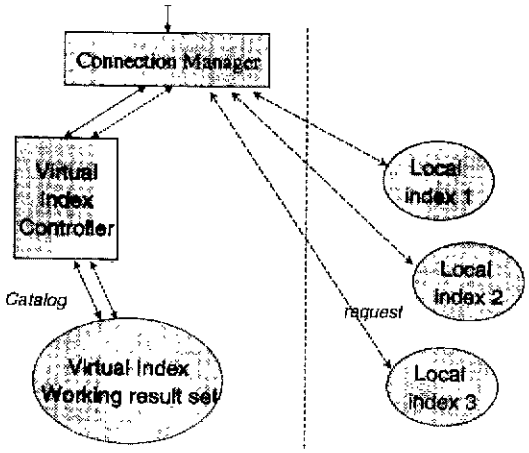

그림 7 카탈로그 관리자의 가상색인 기법 과정

직접 검색을 시도한다. 그림에서 점선으로 표시된 부분이다.

가상색인 기법은 기본적으로 분산 색인방식을 사용한다. 최초의 질의가 있을 때는 가상색인은 아무런 정보도 가지고 있지 않기 때문에 분산 색인방식에 따라 모든 분산 자료에 대한 검색을 한다. 대신 그 결과를 수집하여 가상색인을 작성한다. 즉, 이 기법에서는 사용자 질의를 색인작성의 기초적 근거로 하여 한번 이상 질의된 내용에 대해서만 색인을 작성한다. 이렇게 하면 사용의 빈도수가 높아질수록 시스템의 색인 방식은 최소량의 전체 색인을 갖는 중앙집중식 색인방식으로 전환된다. 물론, 분산색인 방식은 지속적으로 유지된다. 일반적인 사용자들의 질의는 통계적 수치로 볼 때 매우 높은 집중성을 보인다. 다시 말하면, 비슷한 질의를 다수의 사람들이 사용한다는 사실이다. 뿐만아니라 문서 자동추출에 의한 색인은 가능한 모든 질의어에 대해 색인을 만들기 때문에 문서크기에 맞먹는 양이 생성되지만, 실제로 사용자에 의해 쓰이는 부분은 극히 작다. 이러한 통계적 근거가 가상색인 기법의 타당성을 증명한다.

4.3 결론 및 전망

분산 자원을 위한 색인은 첫째, 빠른 검색 응답시간을 지원할 수 있어야 한다. 이 요구사항은 모든 검색 시스템이 가져야할 첫 번째 특징이다. 두 번째로는, 자료와의 일치성을 유지

할 수 있어야 한다. 어떠한 색인이든 본질적으로 자료와는 분리되어 있기 때문에 어느 정도 자료와의 불일치 상태는 감수를 해야 한다. 하지만, 분산되어 있는 자료일 경우에는 이 정도가 더욱 심해질 수 있고, 검색 시스템의 성능과 밀접한 관련을 맺게 된다. 그리고, 마지막으로 자료의 양적팽창에 따른 이질성, 비용의 문제에 동적으로 대처할 수 있어야 한다. 이미 존재하고 있는 자료와 구축할 자료들은 각기 이질적인 구조를 지니고 있을 수밖에 없다. 이 문제에 대해 항상 새로운 방식의 색인이 작성되어야 한다면 큰 비용 손실이 생기게 된다.

이와 같은 요구사항에 대해 중앙집중식 방식과 분산 방식은 서로간에 장단점을 가지고 있음을 밝혔다. 구축하려는 전자 도서관의 특성에 따라 꼭 필요한 기능을 지원할 수 있도록 둘 중에 어느 한가지 방법을 택할 수도 있지만 근래의 전자 도서관들은 이 두가지 방법을 혼용하여 모든 요구사항을 수용할 수 있는 방향으로 나아가고 있다.

현재는 서울대 전자 도서관 한울과 같이 두가지 방식의 장점을 최대한 살릴 수 있도록 유동적인 방식을 택하는 경향이 주류를 이룰 전망이다. 중앙집중식 색인 방식은 전체 색인을 작성하되 꼭 필요한 문서에 대해서만 작성함으로써 비용을 최소화 하려는 경향으로, 분산 색인 방식은 반응 속도를 빠르게 하기 위해 어느 정도의 전체 색인을 허용하는 방향으로 나아가고 있다. 이러한 경향은 분산 색인과 전체 색인을 동시에 사용하는 기법에 대한 논의로 모아지고 있다. 이러한 기존 방식의 변형으로서의 접근이 아니라 완전히 새로운 방식의 접근도 있다. 에이전트에 의한 방식이 그것인데, 에이전트는 색인의 역할을 하는 능동적인 모듈이라고 할 수 있다. 색인의 역할을 하는 에이전트는 자료를 관리하는 에이전트와 메시지를 주고 받음으로써 문서에 대한 메타정보를 지식(knowledge)로 갖게 되고, 사용자와의 상호작용으로 그 지식을 확장, 변화시켜 나가는 방식을 택하고 있다[13]. 집에 가기전에 사용자가 몇가지 자신의 요구를 입력해 놓으면 에이전트가 밤새 그 요구에 맞는 문서들을 찾아서 사용자의 영역에 저장해 놓는다. 사용자는 다음날 그 문서들을 읽고 간단한 평가를 해두면 에이전트는 그날 입력된 사용자 선호도를 이용해 사용자가 원하는 것이 무엇인지를 판단하고 다음에 문서를 찾을 때 그 지식을 사용하게 된다. 이와 같은 에이전트 기반 아키텍처는 이미 오래전부터 응용되어 왔었고, 전자 도서관과 같은 검색 시스템에 많이 활용될 전망이다.

참고문헌

[1] William Y. Arms, "Key Concepts in Architecutre of the Digital Library", D-Lib Magazine, July 1995.

[2] R. R. Larson., "Design and development of a network-based electronic library.", Proc. ASIS Mid-Year Meeting, Portland, Oregon, May 1994.

[3] B rtschi, M., "An overview of information retrieval subjects", IEEE Computer, May 1985.

[4] Hersh W, Buckley C, Leone T, and Hickam D. "OHSUMED : An Interactive Retrieval Evaluation and New Large Test Collection for Research", SIGIR 1994.

[5] Gerad Salton., "Introduction to Modern Information Retrieval", McGraw-Hill, 1983.

[6] C. J. Crouth. "An approach to the automatic construction of global thesauri", Information Processing and Management, 1990.

[7] Cliffaord Lynch and Hector Garcia-Molina, "Interoperability, Scaling, and the Digital Libraries Research Agenda", IITA Digital Libraries Workshop, August 22, 1995.

[8] A. Dupa, M. A. Sheldon, "Content Routing in Networks of WAIS Servers", 14th IEEE International Conference on Distributed Computing System, Poznan, Poland, June 1994.

[9] Martijn Koster, "Aliweb-Archie-Like

Indexing in the Web", WWW94, 1994.

[10] VanHeyningen, M., "The Unified Computer Science Technical Report Index : Lessons in indexing diverse resources", 2nd International World Wide Web Conference, WWW94, 1994.

[11] James R. Davis and Carl Lagoze, "The Networked Computer Science Technical Report Library", IEEE Computer special issue on Building Large-scale Digital Libraries", May 1996.

[12] Lagoze, C. and J.R. Davis, "Dienst : An Architecture for Distributed Document Libraries", Communications of the ACM, vol 38, no 4, April 1995.

[13] Finin. T., R.Fritzon, D. McKay, et al., "KQML as an agent communication language", Third International Conference on Information and Knowledge Management, ACM Press, 1995.

김 동 규

1995 서울대학교 계산통계학과 전산과학 전공 학사
1995~현재 서울대학교 전산과학과 석사과정 재학중
관심분야 : 정보검색, 전자 도서관

이 상 구

1985 서울대학교 계산통계학과 계산학 전공 학사
1987 University of North-Western 전산과학 석사
1990 University of North-Western 전산과학 석사
1990~1992 Electronic Data System(EDS) R&D USA연구원
1992~현재 서울대학교 전산과학과 조교수
관심분야 : 질의 최적화, 정보검색, 데이터웨어하우스, 데이터 마이닝

전 종 훈

1983~1986 덴버대학 전산과학과 학사
1986~1988 노스웨스턴대학 전산과학과 석사
1988~1992 노스웨스턴대학 전산과학과 박사
1992~1995 셀프렐 오클라호마 주립대학 전산과학과 조교수
1995~현재 명지대학 컴퓨터공학과 조교수
관심분야 : 디지탈 라이브러리, 멀티미디어 데이타베이스, 논리데이타베이스, 정보검색