오디세우스 객체관계형 DBMS를 사용한 웹 검색 시스템에서의 사이트, 도메인, 커뮤니티 제한 검색 및 홈서치

김민수[O] 이재길 김민수 황규영
한국과학기술원 전산학과/첨단정보기술연구센터
{mskim2[O], jglee, mskim, kywhang}@mozart.kaist.ac.kr

Site-, Domain-, Community- Limited Search, and Home Search in a Web Search System Using the ODYSSEUS Object-Relational DBMS

Min S. Kim[O] Jae-Gil lee Min-Soo Kim Kyu-Young Whang
Department of Computer Science and
Advanced Information Technology Research Center
Korea Advanced Institute of Science and Technology

요 약

인터넷이 널리 활용되면서 웹 사이트의 수가 기하급수적으로 증가하는 동시에, 각각의 웹 사이트에 저장된 정보의 양도 급속히 증가하고 있다. 사용자가 이러한 웹 환경에서 원하는 정보를 효과적으로 찾을 수 있도록 하기 위해서는 크게 두 가지 요소가 중요한데, 첫 번째는 웹 검색 사이트에서 다양한 제한 검색 기능들을 제공하는 것이고, 두 번째는 일반적인 웹 사이트에서 홈서치 서비스를 제공하는 것이다. 제한 검색은 각 웹 사이트의 정보를 중앙 데이타베이스에 저장하고, 검색 범위를 특정 범위로 한정시켜 검색 결과를 제공하는 방법으로, 이를 활용하면 일반적인 웹 사이트들도 저렴한 비용으로 방문자들에게 홈서치 서비스를 제공할 수 있다. 본 논문에서는 이러한 제한 검색 기능들과 홈서치 기능을 오디세우스 정보검색용 객체관계형 DBMS를 사용하여 실제로 어떻게 구현할 수 있는지에 대해 SQL 및 HTML 레벨에서 설명한다. 따라서, 본 논문에서 제안하는 구현 방법은 Oracle, MySQL 등의 정보검색 기능이 제공되는 관계형 DBMS들에 대해 모두 적용될 수 있다.

1. 서 론

인터넷이 널리 활용되면서 웹 사이트(site)의 수가 기하급수적으로 증가하는 동시에, 각각의 웹 사이트에 저장된 정보의 양도 급속히 증가하고 있다. 사용자가 이러한 웹 환경에서 원하는 정보를 보다 효과적으로 찾을 수 있도록 하기 위해서는 크게 두 가지 요소가 중요한데, 첫 번째는 웹 검색 사이트에서 다양한 제한 검색 기능들을 제공하는 것이고, 두 번째는 일반적인 웹 사이트에서 홈서치(Home Search) 서비스를 제공하는 것이다.

질의에 대한 검색 결과의 질을 높이기 위한 효과적인 방법으로 본 저자들이 제한 검색 개념을 제안한 바 있다[1]. 제한 검색은 인터넷 상의 웹 문서들에 대한 정보를 중앙 데이타베이스에 저장하고 검색 범위를 특정 범위로 한정시켜 검색 결과를 제공하는 방법으로서, 사용자의 의도에 따라 검색 범위를 제한함으로써 검색 결과의 질을 높이게 된다.

일반적인 웹 사이트들에서 홈서치 서비스를 제공하기 위해서는 전용의 검색 엔진을 설치해야 하지만, 비용 및 인력 문제로 소규모 웹 사이트에서는 홈서치 서비스를 제공하기 힘들다. 이러한 문제를 해결하기 위해, 본 논문의 저자들이 제한 검색 기능을 이용한 홈서치 서비스에 대한 개념을 제안한 바 있다[1][3]. 이는 웹 사이트를 중앙 검색 시스템에 등록하고, 검색 범위를 해당 웹 사이트로 한정시킴으로써 저렴하고 손쉽게 홈서치 서비스를 제공하는 개념이다[1].

본 논문에서는 한국과학기술원 첨단정보기술연구센터에서 개발한 정보검색용 객체관계형 DBMS 오디세우스[4]를 사용하여 제한 검색 및 홈서치 서비스를 실제로 어떻게 구현할 수 있는지에 대해 SQL 및 HTML 레벨에서 총체적으로 설명한다. 오디세우스는 정보검색 엔진과 DBMS 엔진이 밀결합[5]되어 있기 때문에 제한 검색 및 홈서치 서비스를 빠르게 처리한다[1]. 또한, 본 논문에서 제안하는 구현 방법은 SQL 및 HTML 레벨에서 설명되어 있기 때문에 Oracle, MySQL 등의 정보검색 기능이 제공되는 관계형 DBMS들에 대해 모두 적용될 수 있다.

본 논문의 구성은 다음과 같다. 제 2 절에서는 제한 검색의 구현 방법에 대해 설명하고, 제 3절에서는 홈서치 서비스의 구현 방법에 대해 설명한다. 마지막으로 제 4 절에서는 결론을 내린다.

2. 제한 검색의 구현 방법

본 절에서는 다양한 제한 검색 기능들에 대해 오디세우스 객체관계형 DBMS를 사용하여 구현하는 방법들을 SQL 레벨에서 설명한다. 제 2.1 절에서는 사이트(site) 제한 검색 구현 방법을 설명하고, 제 2.2 절에서는 도메인(domain) 제한 검색 구현 방법을 설명한다. 그리고, 제 2.3 절에서는 커뮤니티(community) 제한 검색 구현 방법을 설명한다.

그림 1은 제한 검색 기능을 위해 DBMS에 저장할 주요 테이블들의 스키마이다. 그림 1에서 ***siteInfo*** 는 사이트 정보를 저장하는 테이블이고, ***pageInfo***는 웹 페이지 정보를 저장하는 테이블이다. siteInfo 테이블에는 사이트 식별자, 사이트가 속한 도메인의 식별자, 사이트가 속한 커뮤니티의 식별자와 사이트의 URL, 제목, 설명 등이 저장된다. 사이트 식별자는 각각의 웹 사이트마다 고유하게 부여되는 정수 타입의 식별자이고, 사이트가 속한 도메인의 식별자는 각각의 도메인마다 고유하게 부여되는 정수 타입의 식별자이다. 또한, 사이트가 속한 커뮤니티의 식별자는 각각의 커뮤니티마다 고유하게 부여되는 정수 타입의 식별자이다. pageInfo 테이블에는 웹 페이지가 속한 사이트의 식별자, 웹 페이지가 속한 도메인의 식별자, 웹 페이지가 속한 커뮤니티의 식별자와 웹 페이지의 식별자, 제목, URL, 본문 등이 저

* 본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

장된다. 웹 페이지 테이블에 저장된 사이트 식별자는 특정 웹 페이지가 어떤 사이트에 속한 웹 페이지인지 판별하는데 사용되며, 정수 타입과 text 타입의 두 가지 타입으로 저장된다. 마찬가지로 도메인 식별자와 커뮤니티 식별자도 사이트 식별자와 동일한 형태와 의미를 갖는다.

사이트 테이블(siteInfo) 스키마

컬럼 이름	컬럼 타입	설명
siteId	integer	사이트 식별자
domainId	integer	사이트가 속한 도메인 식별자
communityId	integer	사이트가 속한 커뮤니티 식별자
URL	varchar	사이트 URL
title	text	사이트 제목
description	text	사이트 설명

웹 페이지 테이블(pageInfo) 스키마

컬럼 이름	컬럼 타입	설명
siteId	integer	사이트 식별자
siteIdText	text	사이트 식별자
domainIdText	text	사이트가 속한 도메인 식별자
communityIdText	text	사이트가 속한 커뮤니티 식별자
pageId	integer	웹 페이지 식별자
title	text	웹 페이지 제목
URL	varchar	웹 페이지 URL
content	text	웹 페이지 본문

그림 1. 검색 시스템의 주요 테이블 스키마.

그림 1의 스키마에서 text 타입의 컬럼은 키워드 검색을 위한 컬럼이다. 오디세우스 DBMS에서 text 타입의 컬럼에 저장된 내용은 자동적으로 텍스트 인덱스에 의해 관리된다.

2.1. 사이트 제한 검색

본 논문의 저자들은 중앙 데이타베이스에 저장된 전체 데이타 중에서 지정된 웹 사이트만을 검색 범위로 제한하는 검색 기능을 *사이트 제한 검색* 이라는 개념으로 정의한 바 있다[1][3]. 구글(Google) 웹 검색 시스템도 "Site Search"라는 유사한 개념의 서비스를 제공하고 있으나[6], 이 "Site Search"는 구글에 특화되어 구현되어 있고 구현 방법이 공개되지 않았기 때문에 일반적인 DBMS를 이용하여 구현하기가 어렵다.

사이트 제한 검색의 예는 다음과 같다. 삼성정밀화학(sfc.samsung.co.kr) 사이트를 검색 범위로 제한하고, "채용"이라는 키워드로 검색을 하면 삼성정밀화학 사이트 내의 웹 페이지들 중에서 "채용"이라는 키워드를 포함하는 웹 페이지들을 찾게 된다. 그림 2는 앞에서 설명한 사이트 제한 검색 예의 SQL 질의식이다. 여기서 삼성정밀화학의 사이트 식별자는 "228931"이라고 가정한다. 질의식의 형태는 표준 SQL3[7]와 동일하게 SELECT, FROM, WHERE 절로 구성되어 있다.

```
SELECT  p.oid
FROM    pageInfo p
WHERE   MATCH(p.content, "채용")>0 AND
        MATCH(p.siteIdText, "228931")>0;
```

그림 2. 사이트 제한 검색 SQL 질의.

WHERE 절에 명시된 MATCH() 함수는 키워드 검색을 제공하기 위해 오디세우스에서 제공되는 함수이며, Oracle DBMS에서도 이와 유사한 CONTAINS()라는 함수를 제공하고 있다.

2.2. 도메인 제한 검색

본 절에서는 중앙 데이타베이스에 저장된 전체 데이타 중에서 지정된 도메인만을 검색 범위로 제한하는 검색 개념을 *도메인 제한 검색*이라고 정의하고, 이에 대한 구현 방법을 SQL문으로 설명한다.

도메인 제한 검색의 예는 다음과 같다. "samsung.co.kr" 도메인을 검색 범위로 제한하고, "채용"이라는 키워드로 검색을 하면 "samsung.co.kr" 도메인에 속하는 삼성정밀화학(sfc.samsung.co.kr), 삼성중공업(shi.samsung.co.kr), 삼성전기(sem.samsung.co.kr) 등의 사이트에서 "채용"이라는 키워드를 포함하는 웹 페이지들을 찾게 된다. 그림 3은 앞에서 설명한 도메인 제한 검색 예의 SQL 질의식이다. 여기서 "samsung.co.kr" 의 도메인 식별자는 "7966"이라고 가정한다.

```
SELECT  p.oid
FROM    pageInfo p
WHERE   MATCH(p.content, "채용")>0 AND
        MATCH(p.domainIdText, "7966")>0;
```

그림 3. 도메인 제한 검색 SQL 질의.

2.3. 커뮤니티 제한 검색

본 논문의 저자들은 중앙 데이타베이스에 저장된 전체 데이타 중에서 지정된 커뮤니티만을 검색 범위로 제한하는 검색 기능을 *커뮤니티 제한 검색* 이라는 개념으로 제안하 바 있다[2]. 커뮤니티는 의미적으로 관련된 사이트들의 집합으로서, 커뮤니티 제한 검색은 URL에 의해 명시되는 사이트나 도메인은 다르지만 의미적으로 관련된 사이트들에 대해서도 범위를 제한하여 검색할 수 있다는 특징을 가진다. 참고로 구글에서는 URL에 의해 명시되는 사이트 또는 도메인에 대해서만 제한 검색 기능을 제공하고 있다[6].

커뮤니티 제한 검색의 예는 다음과 같다. 삼성 계열사들을 나타내는 커뮤니티를 검색 범위로 제한하고, "채용"이라는 키워드로 검색을 하면 삼성정밀화학(sfc.samsung.co.kr), 삼성전자(www.sec.co.kr), 삼성SDI(www.samsungsdi.com), 제일기획(www.cheil.com) 등의 삼성 계열사 사이트에서 "채용"이라는 키워드를 포함하는 웹 페이지들을 찾게 된다. 그림 4는 앞에서 설명한 커뮤니티 제한 검색 예의 SQL 질의식이다. 여기서 삼성 계열사들을 나타내는 커뮤니티의 커뮤니티 식별자는 "376"이라고 가정한다.

```
SELECT  p.oid
FROM    pageInfo p
WHERE   MATCH(p.content, "채용")>0 AND
        MATCH(p.communityIdText, "376")>0;
```

그림 4. 커뮤니티 제한 검색 SQL 질의.

3. 홈서치 서비스의 구현 방법

웹 사이트에서 자신이 보유하고 있는 데이타에 대한 홈서치 서비스를 제공하기 위해서는 전용 검색 엔진을 설치하고 운영해야 한다. 이를 위해서는 검색 엔진의 구입 및 설치 비용과 지속적인 데이타베이스 관리 등의 유지 보수 비용이 필요한데, 규모가 작은 웹 사이트에서는 비용의 부담으로 인해 홈서치 서비스를 방문자들에게 제공하는 것이 힘들다.

이러한 문제점을 해결하기 위해 본 논문의 저자들이 제한 검색 기능을 이용한 홈서치 서비스에 대한 개념을 제안한 바 있다[1][3]. 홈서치 서비스는 사이트 제한 검색을 활용하여 각 웹 사이트에 대한 검색 기능을 제공하는 ASP(Application Service Provider) 서비스이다. 즉, 그림 5에서와 같이 각 웹 사이트는 자신을 중앙 검색 시스템에 등록한 후, 웹 사이트에 추가한 검색 창을 통해 중앙 검색 시스템에 사이트 내 검색을 요청하고, 중앙 검색 시스템은 중앙 데이타베이스에 저장된 전체 데이타 중에서 요청 받은 웹 사이트만으로 검색 범위로 제한하여 질의를 처리한다. 참고로 구글 웹 검색 시스템에서도 "사이트 검색 기능이 포함된 무료 웹 검색(Google Free Web Search with Site Search)[8]"을 통해 최근 들어 홈서치 서비스를 제공하기 시작하였다. 그러나, 그 구현 방법이 공개되어 있지 않기 때문에 일반적인 DBMS를 이용하여 구현하기가 어렵다.

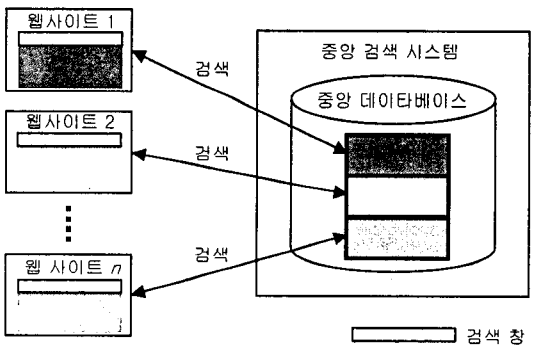

그림 5. 홈서치 아키텍쳐.

그림 6은 오디세우스 객체관계형 DBMS를 사용하여 구현한 홈서치 기능을 제공하는 웹 사이트에 접속한 화면이다. 화면에서 페이지 프레임이 위, 아래로 분리되어, 위 프레임에는 키워드를 입력할 수 있는 홈서치 검색창이, 아래 프레임에는 실제 웹 페이지 내용이 표시된다.

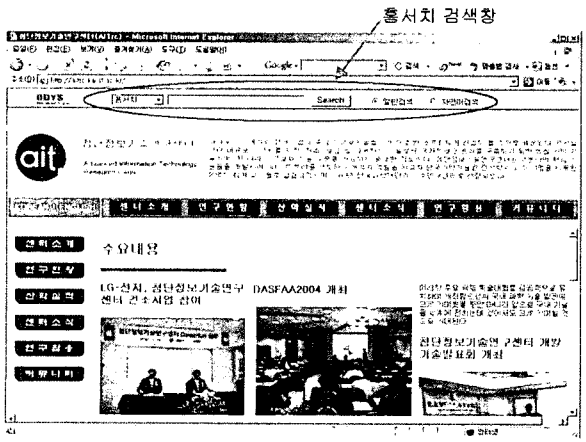

그림 6. 홈서치 화면.

그림 7은 일반적인 웹 사이트에서 홈서치 서비스를 제공하기 위해 메인 웹 페이지에서 추가해야 하는 HTML 코드의 예를 나타낸다. 이 코드에서 FORM 태그(tag) 내의 action 항목에는 중앙 검색 시스템의 URL을 입력하며, 세 번째 input 태그 내의 value 항목에는 홈서치 서비스 사용하고자 하는 웹 사이트의 사이트 식별자를 입력한다. 제 2 절에서 설명한 바대로 제한 검색 기능이 구현되어 있는 검색 시스템과 연동하여 이러한 HTML 코드를 삽입하기만 하면 홈서치 서비스를 사용할 수 있게 된다.

```
<FORM method=GET action= "http://odys-search.kaist.
.ac.kr:7779/homesearch.html">
   <input type=text name="query" size="32">
   <input type=submit name="search" value="Search">
   <input type=hidden name="siteId" value=2000001>
</FORM>
```

그림 7. 홈서치 검색창을 추가하기 위한 HTML 코드 예.

4. 결론

본 논문에서는 사이트, 도메인, 커뮤니티의 제한 검색 기능들과 홈서치 기능을 오디세우스 정보검색용 객체관계형 DBMS를 사용하여 실제로 구현하는 방법들을 SQL 및 HTML 레벨에서 설명하였다. 따라서, Oracle, MySQL과 같은 정보검색 기능이 제공되는 관계형 DBMS들에 대해 본 논문에서 설명한 구현 방법들을 적용하여 제한 검색 및 홈서치 서비스를 손쉽게 구현하고 또 서비스할 수 있을 것이다.

참고문헌

[1] 이재길, 이민재, 김민수, 황규영, "오디세우스 객체관계형 DBMS를 사용한 사이트 제한 검색의 구현," 한국정보과학회 춘계학술발표회 논문집, pp. 755-757, 2003년 4월.

[2] 김계정, 김민수, 김이른, 황규영, "커뮤니티 제한 검색을 위한 웹 크롤링 및 PageRank 계산," 한국컴퓨터종합학술대회 2005 논문집, pp. 1024-1026, 2005년 7월.

[3] Myseek 웹 검색엔진: 내 집 검색기, ㈜마이씨크, 2001.

[4] 한욱신, 이민재, 이재길, 박상영, 황규영, "오디세우스 객체관계형 멀티미디어 DBMS의 아키텍쳐," 한국정보과학회 추계학술발표 논문집, pp. 45-47, 2000년 10월.

[5] Kyu-Young Whang, Min-Jae Lee, Jae-Gil Lee, Min-Soo Kim, and Wook-Shin Han, "Odysseus: a High-Performance ORDBMS Tightly-Coupled with IR Features," In *Proc. IEEE 21st Int'l Conf. on Data Engineering (ICDE)*, pp. 1104-1105, Tokyo, Japan, Apr. 5-8, 2005.

[6] Google Site Search, http://www.google.com/intl/en/help/features.html#sitesearch, 2005.

[7] Melton, J. and Simon, A.R., SQL: A Complete Guide, Morgan Kaufmann Publishers, 1993.

[8] Google Free, http://services.google.com/searchcode2.html, 2005.