오디세우스 객체관계형 DBMS를 사용한 사이트 제한 검색의 구현

이재길 [0] 이민재 김민수 황규영

한국과학기술원 전산학과/첨단정보기술연구센터

{jglee, mjlee, mskim, kywhang}@mozart.kaist.ac.kr

Implementation of a Site-limited Search using the ODYSSEUS Object-Relational DBMS

Jae-Gil Lee [0] Min-Jae 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 일반인들에게 널리 활용되면서 웹 사이트의 수가 기하급수적으로 증가하고 있으며, 각각의 웹 사이트에 저장된 정보의 양도 급속히 증가하고 있다. 따라서, 웹 사이트에서 자신이 보유한 정보를 방문자들이 쉽게 검색할 수 있도록 자체 검색 서비스를 제공해야 하는 필요성이 점차 커지고 있다. 이를 위해, 각 웹 사이트의 정보를 중앙 데이타베이스에 저장하고, 검색 범위를 자신의 사이트에서 제공한 데이타에 제한하여 검색하는 사이트 제한 검색이 널리 사용되고 있다. 본 논문에서는 오디세우스 정보검색용 객체관계형 DBMS 를 사용하여 사이트 제한 검색을 효율적으로 구현하는 두 가지 방법을 제안한다. 다음으로, 본 논문에서 제안한 사이트 제한 검색의 구현 방법은 대용량 데이타베이스에서 사이트 제한 검색을 수행하더라도 검색 속도가 전혀 저하되지 않음을 실험을 통해 입증한다.

1. 서 론

인터넷이 일반인들에게 널리 활용되면서 웹 사이트(site)의 수가 기하급수적으로 증가하고 있으며, 각각의 웹 사이트에 저장된 정보의 양도 급속히 증가하고 있다. 따라서, 웹 사이트에서 자신이 보유한 정보를 방문자들이 쉽게 검색할 수 있도록 자체 검색 서비스를 제공해야 하는 필요성이 점차 커지고 있다.

자체 검색 서비스를 제공하기 위해서는 웹 사이트에 검색 엔진을 설치하고 운영해야 한다. 즉, 검색 엔진을 구입하여 검색 서버에 설치하고 관리자가 지속적으로 데이타베이스를 유지보수 해야 한다. 그러나, 비용 및 인력 문제로 이와 같이 검색 엔진을 설치하고 운영하기에 부담이 따르기 때문에, 소규모 개인 웹 사이트에서는 자체 검색 서비스를 제공하기 힘들다.

이러한 문제를 해결하기 위해, 사이트 제한 검색이 자체 검색 서비스의 제공 방법으로 널리 사용되고 있다. 사이트 제한 검색은 각 웹 사이트의 정보를 중앙 데이타베이스에 저장하고, 검색 범위를 자신의 사이트에서 제공한 데이타에 제한하는 검색 방법이다. 웹 사이트 관리자는 자신의 사이트를 검색 시스템에 등록함으로써 쉽게 자체 검색 서비스를 제공할 수 있다. 현재, 구글(Google) 웹 검색 시스템에서 이러한 사이트 제한 검색 기능을 제공하고 있다[1].

일반적인 웹 정보검색은 텍스트 인덱스만으로 처리되지만, 사이트 제한 검색을 수행하기 위해서는 속성 검색이 병행되어야 한다. 즉, 텍스트 인덱스를 통해 찾아진 결과 투플들 중에서 속성 검색으로 사이트 제한 조건을 만족하는지를 확인해야 한다. 이때, 각 결과 투플들을 액세스해야 하는데, 그 액세스 비용은 매우 크다. 따라서, 이러한 속성 검색을 효율적으로 처리하지 못한다면, 사이트 제한 검색은 일반적인 웹 정보검색에 비해 매우 느리게 처리된다.

본 논문에서는 한국과학기술원 첨단정보기술연구센터에서 개발한 정보검색용 객체관계형 DBMS 인 오디세우스[2]를 사용하여 사이트 제한 검색을 효율적으로 구현하는 두 가지 방법을 제안한다. 제안하는 방법은 정보검색 엔진과 DBMS 엔진이 밀결합된[4] 오디세우스를 활용하여 키워드 검색과 속성 검색을 텍스트 인덱스만으로 처리하기 때문에 사이트 제한 검색 질의를 빠르게 수행한다. 실험 결과, 제안하는 구현 방법은 대용량 데이타베이스에서도 검색 속도가 전혀 저하되지 않는다.

본 논문의 구성은 다음과 같다. 제 2 절에서는 사이트 제한 검색의 개념을 설명한다. 제 3 절에서는 오디세우스 객체관계형 DBMS 를 사용하여 사이트 제한 검색을 구현하는 방법을 설명한다. 제 4 절에서는 사이트 제한 검색의 수행 성능에 대한 실험 결과를 제시한다. 마지막으로, 제 5 절에서는 결론을 내린다.

2. 사이트 제한 검색

사이트 제한 검색은 그림 1 과 같이 중앙 데이타베이스에 저장된 전체 데이타 중에서 지정된 사이트의 데이타만을 검색하는 기능이다. 제한 검색을 사용하기 위하여 웹 사이트 관리자가 자신의 사이트를 검색 시스템에 등록하면, 웹 로봇[5]이 등록된 사이트에 포함되어 있는 웹 페이지를 수집하여 중앙 데이타베이스에 저장한다. 이때 어떤 사이트로

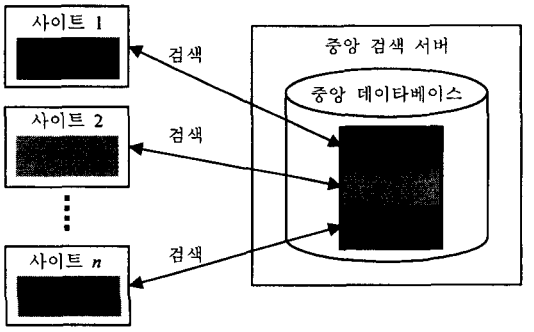

그림 1. 사이트 제한 검색의 개념.

* 본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

부터 수집된 웹 페이지인지 나타내는 정보를 함께 저장하며, 사이트 제한 검색 요청이 들어오면 이 정보를 사용하여 해당 사이트로부터 수집된 웹 페이지에 대해서만 검색을 수행한다.

사이트 제한 검색 기능을 활용하면 웹 사이트에 검색 엔진을 설치하지 않고도, 마치 웹 사이트에 검색 엔진을 설치하여 운영하는 효과를 볼 수 있다. 그림 2 는 사이트 제한 검색을 위해 등록된 웹 사이트에 접속한 화면이다. 화면에서 페이지 프레임이 위 아래로 분리되어 위 프레임에는 키워드를 입력할 수 있는 사이트 제한 검색창이, 아래 프레임에는 실제 웹 페이지 내용이 표시된다. 검색창에 키워드를 입력하고 검색을 수행하면 현재 사이트에서만 검색을 수행하여 검색 결과를 출력한다. 이와 같이 실제 사이트에는 검색창을 제공하지 않고서도 자체 검색 서비스를 제공할 수 있다.

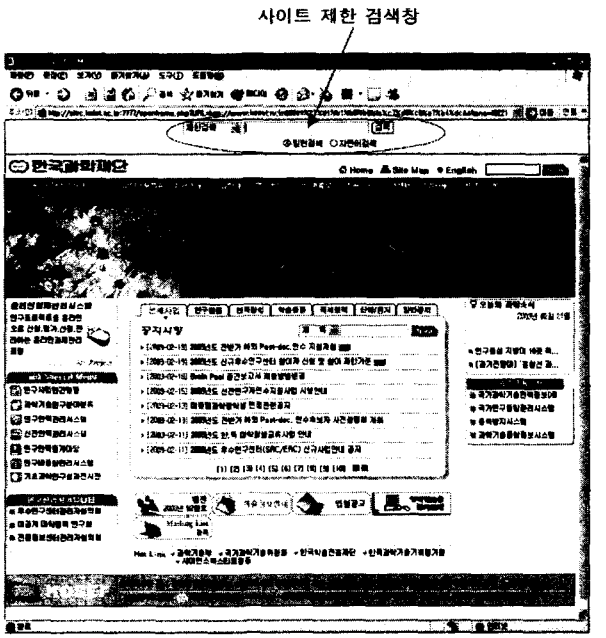

그림 2. 사이트 제한 검색의 수행 화면.

3. 사이트 제한 검색의 구현 방법

본 절에서는 오디세우스 객체관계형 DBMS 를 사용하여 사이트 제한 검색을 구현하는 두 가지 방법을 설명한다. 제 3.1 절에서는 첫 번째 구현 방법으로 사용자 정의 포스팅[2]을 사용하는 구현 방법을 설명하고, 제 3.2 절에서는 두 번째 구현 방법으로 텍스트 인덱스의 조인을 사용하는 구현 방법을 설명한다.

구현 방법을 설명하기 전에, 우선 설명에 사용되는 예제 스키마를 소개한다. 그림 3 은 사용된 예제 스키마로, 사이트 정보를 저장하는 테이블 ***siteInfo*** 와 웹 페이지 정보를 저장하는 테이블 ***pageInfo*** 로 구성된다. ***siteInfo*** 테이블에는 사이트 식별자와 사이트의 타이틀, 설명, URL 등이 저장된다. 사이트 식별자는 각각의 사이트마다 고유하게 부여되는 정수 타입의 식별자이다. ***pageInfo*** 테이블에는 웹 페이지가 속한 사이트의 사이트 식별자, 웹 페이지 식별자, 웹 페이지의 타이틀, 본문 내용, URL 등이 저장된다. 웹 페이지 테이블에 저장된 사이트 식별자는 특정 웹 페이지가 어떤 사이트에 속한 웹 페이지인지 판별하는 데 사용되며, 정수 타입과 **text** 타입의 두 가지 타입으로 저장된다.

사이트 테이블(siteInfo) 스키마

컬럼 이름	컬럼 타입	설명
siteId	integer	사이트 식별자
URL	varchar	사이트 URL
title	text	사이트 타이틀
description	text	사이트 설명

웹 페이지 테이블(pageInfo) 스키마

컬럼 이름	컬럼 타입	설명
siteId	integer	사이트 식별자
siteIdText	*text*	사이트 식별자
pageId	integer	웹 페이지 식별자
title	*text*	웹 페이지 타이틀
URL	varchar	웹 페이지 URL
content	*text*	웹 페이지 본문

그림 3. 검색 시스템의 주요 테이블 스키마.

그림 3 의 스키마에서 **text** 타입의 컬럼은 키워드 검색을 위한 컬럼이다. **text** 타입의 컬럼에 저장된 내용은 자동적으로 텍스트 인덱스에 의해 관리된다.

3.1. 사용자 정의 포스팅을 사용하는 방법

첫 번째 구현 방법에서는 사이트 제한 검색을 수행하는 질의식을 처리하기 위해 사용자 정의 포스팅을 사용한다. 사용자 정의 포스팅이란 일반 속성의 값을 텍스트 인덱스에 직접 내포시키도록 사용자가 포스팅의 구조를 정의하는 기능이다. 따라서, 텍스트 인덱스만으로 키워드 검색을 하면서 동시에 속성 검색을 처리할 수 있다. 그러므로, 이를 활용하여 사이트 제한 검색을 매우 빠르게 처리할 수 있다.

사용자 정의 포스팅을 사용한 사이트 제한 검색을 수행하기 위해서는 ***pageInfo*** 테이블의 ***content*** 컬럼에 생성된 텍스트 인덱스가 ***siteId*** 컬럼의 값을 내포하도록 선언한 뒤에, 질의식에 ***pageInfo*** 테이블의 ***siteId*** 컬럼에 대한 조건을 추가하면 된다. 그림 4 는 사용자 정의 포스팅을 사용한 사이트 제한 검색 질의식의 예이다. 질의식의 형태는 표준 SQL[3]과 동일하게 **SELECT**, **FROM**, **WHERE** 절로 구성되어 있다. 단, **WHERE** 절에 명시된 **MATCH()** 함수는 키워드 검색을 제공하기 위해 오디세우스의 질의어에 추가된 함수이다. 그림 4 의 질의식을 수행하면 사이트 식별자가 "69883"인 사이트에 포함된 웹 페이지의 본문 중에서 "정보"라는 키워드를 포함한 웹 페이지를 검색하여 결과로 반환한다.

```
SELECT pageInfo
FROM pageInfo
WHERE MATCH(content, "정보")>0 AND siteId = 69883;
```

그림 4. 사용자 정의 포스팅을 사용하는 사이트 제한 검색 질의.

3.2. 텍스트 인덱스의 조인을 사용하는 방법

두 번째 구현 방법에서는 사이트 제한 검색을 수행하는 질의식을 처리하기 위해 오디세우스가 제공하는 텍스트 인덱스의 조인을 사용한다. 이 방법에서는 속성 값을 키워드로 취급하여 **text** 타입의 컬럼에 저장함으로써 속성 검색을 위한 텍스트 인덱스로 속성 검색이 처리되도록 한 후, 기존 키워드 검색을 위한 텍스트 인덱스와 빠르게 조인하여 질의를 처리한다.

텍스트 인덱스의 조인을 사용한 사이트 제한 검색을 수행하기 위해서는 질의식에 ***pageInfo*** 테이블의 ***siteId*** 컬럼 대신 ***siteIdText*** 컬럼에 대한 조건을 추가하면 된다. 그림 5 는 텍스트 인덱스의 조인을 사용한 사이트 제한 검색 질의식의 예이다. 질의식의 의미는 그림 4 의 질의식과 동일하다. 그림 5 의 질의식을 수행하면 ***content*** 컬럼의 텍스트 인덱스와 ***siteIdText*** 컬럼의 텍스트 인덱스를 조인하여 두 개의 키워드 검색 조건을 모두 만족하는 웹 페이지를 결과로 반환한다.

```
SELECT pageInfo
FROM pageInfo
WHERE MATCH(content, "정보")>0 AND
      MATCH(siteIdText, "69883")>0;
```

그림 5. 텍스트 인덱스 조인을 사용하는 사이트 제한 검색 질의.

4. 성능 평가

본 절에서는 사이트 제한 검색의 수행 성능에 대한 실험 결과를 제시한다. 우선 실험 환경에 대해서 설명하고, 그 다음으로 실험에 사용한 질의식과 실행 시간을 제시한다.

■ 실험 환경

본 실험에서 사용한 실험 데이타, 사용 장비, 데이타베이스 크기는 각각 아래와 같다.

- 실험 데이타: 웹 로봇에 의해 수집된 약 12 만개의 사이트의 약 800 만건의 웹 페이지
- 사용 장비: SUN E3500 서버 (400MHz CPU * 2, 1GB RAM)
  SUN T3 디스크 어레이 (70GB * 9, RAID 레벨 5)
- 데이타베이스 크기: 약 120GB

■ 수행 질의식

본 실험에서 사용한 질의식은 아래와 같이 4 종류이다.

- 질의식 1: 한국과학재단 사이트에서 "한국과학기술원"이 포함된 웹 페이지를 검색하는 질의
- 질의식 2: 이화여자대학교 사이트에서 "공과대학"이 포함된 웹 페이지를 검색하는 질의
- 질의식 3: CBS 사이트에서 "헌금"이 포함된 웹 페이지를 검색하는 질의
- 질의식 4: 교보증권 사이트에서 "주가지수"가 포함된 웹 페이지를 검색하는 질의

■ 질의 실행 시간

위 질의식 1~4 를 텍스트 인덱스를 조인하는 방법으로 실행하여 질의 실행 시간을 측정한다. 제안하는 두 가지 구현 방법은 데이타베이스의 상태에 따라 각각 유리한 경우가 있으며, 분석을 통해 어느 구현 방법이 더 유리한지 알아낼 수 있다. 본 실험에 사용한 데이타베이스에서는 텍스트 인덱스를 조인하는 방법을 사용하는 편이 보다 더 유리하다.

질의식 1~4 의 질의 실행 시간은 표 1 과 같다. 표 1 에서 볼 수 있듯이 오디세우스를 사용하여 구현한 검색 시스템은 매우 빠르게 사이트 제한 검색을 실행한다. 참고로, 질의식 1~4 에서 사용한 키워드에 대하여 사이트 제한 조건을 추가하지 않은 질의식의 실행 시간은 각각 67ms, 84ms, 58ms, 41ms 이다. 사이트 제한 조건에 의해 검색 결과가 줄어들어 오히려 질의 실행 시간이 더 짧아진다. 따라서, 본 논문에서 구현한 검색 시스템은 실행 시간의 저하 없이 사이트 제한 검색을 수행할 수 있음을 입증한다.

표 1. 사이트 제한 검색 실행 시간.

(단위: ms)

	질의식 1	질의식 2	질의식 3	질의식 4
실행시간	36	41	37	32

5. 결 론

본 논문에서는 오디세우스 정보검색용 객체관계형 DBMS 를 사용하여 사이트 제한 검색을 효율적으로 구현할 수 있는 두 가지 방법을 제안하였다. 제안하는 구현 방법은 텍스트 인덱스만을 액세스하므로 매우 빠르게 사이트 제한 검색을 처리한다. 첫 번째 구현 방법은 사용자 정의 포스팅으로 속성 값을 텍스트 인덱스에 내포함으로써 텍스트 인덱스만을 액세스하여 사이트 제한 검색을 처리한다. 두 번째 구현 방법은 속성 값을 키워드로 취급함으로써 텍스트 인덱스간의 조인을 통하여 사이트 제한 검색을 처리한다. 실제 검색 시스템을 구현하여 사이트 제한 검색의 성능을 측정해본 결과, 사이트 제한 검색으로 인해 질의 실행 시간이 전혀 저하되지 않음을 입증하였다.

참고 문헌

[1] Google, Hosted SiteSearch, http://www.google.com/services/, 2003.

[2] 한욱신, 이민재, 이재길, 박상영, 황규영, "오디세우스 객체관계형 멀티미디어 DBMS 의 아키텍처," 한국정보과학회 추계학술발표 논문집, pp. 45-47, 2000 년 10 월.

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

[4] 박병권, 정보 검색과 데이타베이스 관리 시스템의 밀결합을 위한 역색인 구조와 질의 최적화, 박사 학위 논문, KAIST 전산학과, 1998.

[5] Shkapenyuk, V. and Suel, T., "Design and Implementation of a High-Performance Distributed Web Crawler," In *Proc. of the 18th Int'l Conf. on Data Engineering*, San Jose, California, Feb. 2002.

[6] 우준호, ODYSSEUS 객체지향 데이타베이스 시스템을 위한 질의처리기의 설계 및 구현, 석사 학위 논문, KAIST 전산학과, 1995 년.