한국컴퓨터정보학회 동계학술대회 논문집 제22권 제1호 (2014. 1)

HashMap 기반의 트라이를 이용한 파일 내용 검색 프로그램

김성완*, 이우순°
*삼육대학교 컴퓨터학부
°삼육대학교 컴퓨터학부
e-mail:swkim@swkim@syu.ac.kr*, comnet100@naver.com°

File Content Retrieval Program Using HashMap-based Trie

Sung Wan Kim*, Woosoon Lee°
*Division of Computer, Sahmyook University
°Division of Computer, Sahmyook University

● 요 약 ●

본 논문에서는 파일 내용 기반 검색 프로그램을 설계하고 구현하였다. 역 인덱스 구조를 이용하여 설계하였으며 별도의 정보 검색 라이브러리 사용 없이 구현하였다. 인덱스 파일은 트라이 자료 구조를 직접 설계 및 구현 하였으며 자바 언어의 HashMap 구조를 중첩 형태로 구현하였다. 개발 시스템의 유용성을 테스트하기 위해 GRE 단어집에 수록된 약 3,300개의 단어를 사용하여 임의 생성한 텍스트 파일 집합을 사용하였다.

키워드: 파일 내용 검색(file content retrieval), 역 인덱스(inverted index), 트라이(trie)

I. 서 론

컴퓨터 저장 용량의 대형화에 따라 개인 컴퓨터에 저장된 정보를 체계적으로 찾아 주는 데스크탑 검색 기술이 사용 중에 있다. 데스크탑 검색은 파일이름 뿐만 아니라 파일 내용 기반의 검색 서비스를 제공한다. 이를 위해 파일 내용을 사전에 인덱스로 구축한 뒤 사용자의 검색 요구에 대해 신속하게 검색 조건에 일치되는 문서를 찾아준다[1]. 본 논문에서는 별도의 정보검색 라이브러리 사용 없이 파일 내용 기반 검색 프로그램을 자바 언어의 HashMap을 사용하여 개발하였다. 개발된 프로그램은 단순 키워드 검색 및 불리언 연산을 지원한다. 실험용 데이터를 통해 개발 프로그램의 유용성을 확인하였다.

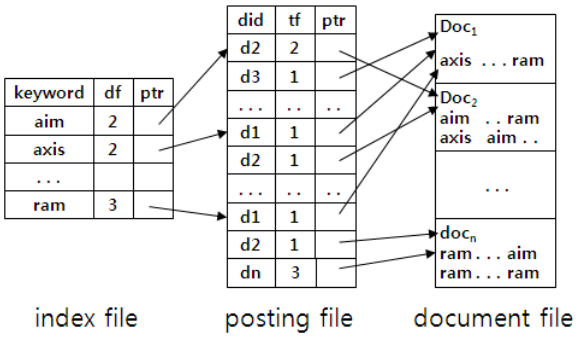

그림 1. 역 인덱스 구조
Fig. 1. Inverted Index Structure

II. 시스템 설계 및 구현

파일 내용 기반 검색을 위해서는 역 인덱스 구조를 사용한다. 역 인덱스는 <그림 1>과 같이 실제 내용이 저장되어 있는 문서 파일, 각 문서에서 추출된 키워드와 문서 파일과의 연결 정보를 유지하는 포스팅 파일, 그리고 키워들을 사전 순으로 정렬하여 저장한 인덱스 파일로 구성된다[2].

본 논문에서 인덱스 파일은 트라이를 이용하여 구현하였다. 트라이[2]는 가변 길이의 키워드를 탐색하는 용도로 널리 사용되고 있는 자료구조이다. 트라이 구현을 위해 자바 언어의 HashMap를 중첩하여 사용하였다. 트라이의 각 노드를 위한 클래스의 주요 구조는 <그림 3>과 같다. val 필드는 한 트라이 노드에 포함되는 문자 값을 의미하며, isTerminal 필드는 단말 노드 여부를 나타내며, ptrPost는 단말 노드일 경우 첫 번째 포스팅 엔트리를 참조하는 필드이다. child 필드는 하위 노드를 나타내며 자기 참조를 이용한 중첩형태이다.

```
class TrieNode {
   char val;
   HashMap<char, TrieNode> child;
   boolean isTerminal;
   int ptrPost;
   public TrieNode(char ch) {
       val = ch;
       child = new HashMap<>();
       isTerminal = false;
       ptrPost = -1;
   }
   . . .
}
```

그림 2. 트라이 노드 위한 클래스
Fig. 2. Class for Trie Node

<그림 3>은 <그림 1>의 인덱스 파일을 <그림 2>의 정의에 따라 개념적으로 나타낸 트라이 구조이다.

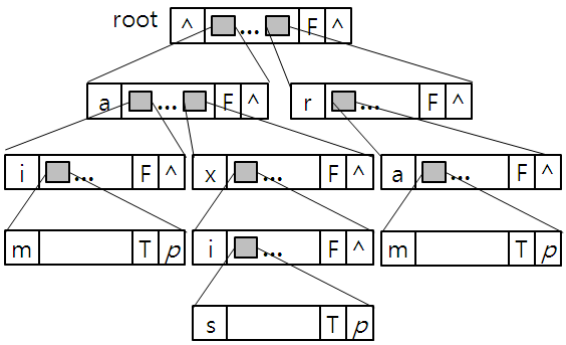

그림 3. 트라이 구조도
Fig. 3. Trie Structure

시스템 구현은 자바 언어를 사용하였으며, 개발도구는 Eclipse Kepler Service Release 1과 Window Builder 1.6.0을 사용했다. 테스트 데이터는 GRE 단어집에 수록된 약 3,300개의 단어를 랜덤 추출하여 생성한 100개의 문서 파일을 사용하였으며, 총 32,900개의 단어가 포함되었다. 생성된 포스팅 파일의 엔트리는 약 31,000개이다.

개발 시스템에서 제공되는 검색 기능은 첫째, 하나의 키워드를 입력 받아 일치되는 문서를 검색하는 단순 키워드 검색이다. <그림 4>는 단순 검색을 위한 사용자 인터페이스를 나타낸 것이다. 둘째, 불리언 검색으로 a AND b, a OR b, a AND NOT b 등 3개의 불리언 연산을 지원한다. <그림 5>는 불리언 검색 인터페이스를 나타낸 것 이다. 좌측 리스트 박스에서 AND, OR, NOT AND 중 하나를 선택하여 검색 조건을 지정 할 수 있다.

그림 4. 단순 키워드 검색 인터페이스
Fig. 4. Interface for Simple Keyword Retrieval

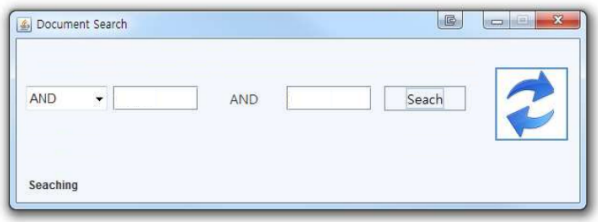

그림 5. 불리언 검색 인터페이스
Fig. 5. Interface for Boolean Retrieval

III. 결 론

본 논문에서는 자바 언어의 HashMap 기반의 트라이 를 이용한 역 인덱스를 설계하여 파일 내용 검색 프로그램을 구현하였다. 특히, 별도의 정보 검색 라이브러리 사용 없이 중첩 구조를 갖는 HashMap 클래스 구조를 정의하여 구현하였다. 또한, 실험용 데이터 셋을 사용하여 개발 시스템의 유용성을 확인하였다.

참고문헌

[1] Wiki, http://en.wikipedia.org/wiki/Desktop_search
[2] B. Lee, "Information Retrieval,", Green Press, 2012
[3] Y. Choi, "Novel Java," 2nd Ed., JABOOK, 2005