2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

양방향 검색을 지원하는 전자사전 구조의 설계 및 구현

김철수*, 박인철**
*서남대학교 컴퓨터영상정보통신학부
**호원대학교 컴퓨터과학부
*chskim@tiger.seonam.ac.kr
**icpark@sunny.howon.ac.kr

A Desgn and Implementation of Electronic Dictionary for support bidirectional searching

Cheol-su Kim○, in-chul Park
Dept. of Computer Science and Information, Seonam University
Dept. of Computer Science, Howon University

요 약

본 논문에서는 빠른 검색 시간을 가지면서 단어의 역문자열도 검색할 수 있는 사전 구조를 설계하고 구현한다. 빠른 검색 시간을 지원하고, 역문자열 검색을 효율적으로하기 위해 트라이 구조를 이용하였으며 트라이 성질 잘 표현하는 배열을 이용한 구현 방법을 사용하였다. 이 사전 구조는 형태소분석, 정보검색, 음성인식 및 문자 인식 과정 등 다양한 분야에서 유용하게 이용할 수 있다.

1. 서론

정보검색이나 자연어 처리를 위해서는 전자사전은 필수적인 요소이다. 전자사전은 검색 대상이 되는 단어의 관련 정보를 얻기 위해서 자주 참조될 뿐만 아니라 사용할 분야에 따라 검색 방법 및 지원되는 기능이 다양하다. 형태소 분석 과정에서는 분석 후보들의 진위 여부를 판별하기 위해 사전을 매우 많이 참조하며 필요에 따라 단어를 구성하는 문자열의 역순으로 된 역문자열 일부를 찾을 수 있어야 한다. 정보검색을 위한 사전이라면 질의어의 유연성을 제공하기 위해 와일드카드 지원해 주어야 한다. 사전 구조의 검색시간과 기억공간에 대한 연구는 많이 진행되었으나 단순히 검색을 위한 용도로 여러 가지 기능들을 동시에 지원해 주지 못하는 단점이 있다.

본 논문에서는 빠른 검색 시간을 가지면서 일반적인 단어 검색뿐만 아니라 단어의 역문자열로 구성된 단어도 검색할 수 있는 사전 구조를 설계하고 구현한다. 이 방법은 단어의 빠른 검색 시간을 지원하고, 역문자열 검색을 효율적으로하기 위해 트라이 구조를 이용하였다. 이 사전 구조는 형태소분석, 정보검색, 음성인식 및 문자 인식 과정 등 다양한 분야에 이용할 수 있다.

2. 관련연구

검색 과정은 색인어의 저장 구조와 밀접하게 연관되어 있으므로 검색어를 효율적으로 검색하기 위해서는 단어를 효율적으로 저장하여야 한다. 검색 방법은 순차 탐색 방법과 이진 탐색 방법 법, 트리 구조, 해싱 구조[1], 트라이 구조[2]등을 이용할 수 있다.

순차 탐색 방법은 검색 대상이 되는 단어를 찾

기 위해 소요되는 시간이 삽입된 단어 수에 비례하여 증가하는 단점이 있다. 트리 구조는 순차 구조에 비해 검색 시간을 줄일 수 있는 좋은 방법이지만 역문자열로 구성된 단어를 검색하기 위해 동일 단어를 중복하여 저장하는 단점을 가진다. B^*트리는 트리 구조의 특징을 유지하면서 순차 구조의 특징을 모두 가지고 있으나 순차 검색 과정이 비효율적이다. 해싱 구조는 검색 시간은 빠르지만 단어를 중복 저장해야 하는 단점이 있다. 빠른 검색 시간을 지원하면서 중복을 최소화할 수 있는 방법으로 트라이 구조가 있다.

2.1 트라이 구조

트라이는 트리 구조의 특수한 경우로 트리 구조가 다음 레벨로의 분할 및 이동이 단어 전체에 의해 이루어지지만 트라이는 단어 전체가 아닌 단어를 구성하는 일부 문자에 따라 이동할 위치가 달라진다. 트라이 구조는 삽입 단어의 개수에 관계없이 검색할 단어의 길이에 의해서 검색 시간이 결정되므로 빠른 검색이 가능하다. 단어 집합 S1={보정, 정보, 정보검색, 정보검색시스템, 정보처리, 정보통신, 통신}의 트라이 구조를 상태 전이도로 표현하면 그림 1과 같다.

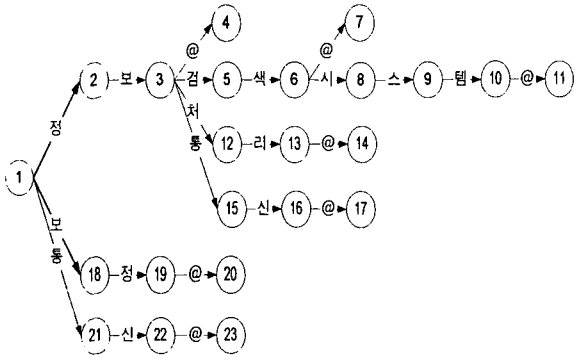

그림 1 상태전이도로 표현된 트라이 구조

트라이 구조를 구현하기 위한 일반적인 방법은 리스트 구조를 이용하는 방법이다. 그러나 이 방법은 삽입단어 수가 증가함에 따라 검색 시간이 길어지는 단점이 있다. 이런 단점을 해결하기 위한 방법중의 한가지는 배열을 이용하는 방법이다[3]. 이 방법은 삽입 단어 숫자에 관계없이 단어의 길이에만 검색 시간이 결정되므로 검색 속도가 빠르다.

[3]는 트라이 성질을 최대로 발휘할 수 있는 방법을 보여주고 있다. 이 방법은 트라이 구조를 상태 전이도 형태로 표현하고 상태 전이를 간단한 산술 연산을 이용한 함수를 사용한다. 이 방법은 검색 시간이 삽입 단어 수에 무관하게 단어의 문자열 길이에만 의존하여 검색 시간이 결정되는 특징을 가진다. 또한 트라이 구조 특성을 이용하여 문자열 확장도 가능하다.

3. 설계

그림 1과 같이 상태전이도로 표현된 트라이 구조는 순방향 문자열로 구성된 단어 검색은 가능하지만 역문자열로 구성된 단어 검색은 불가능하다. 따라서 역문자열로 구성된 단어 검색을 위한 사전 구조가 필요하다. 단어 집합 S1의 단어들에 대한 역문자열로 구성된 단어들의 집합 S2 = {정보, 보정, 색검보정, 템스시색검보정, 리처보정, 신통보정, 신통신}에 대한 검색을 위하여 그림 2와 같은 트라이 구조를 구성할 수 있다.

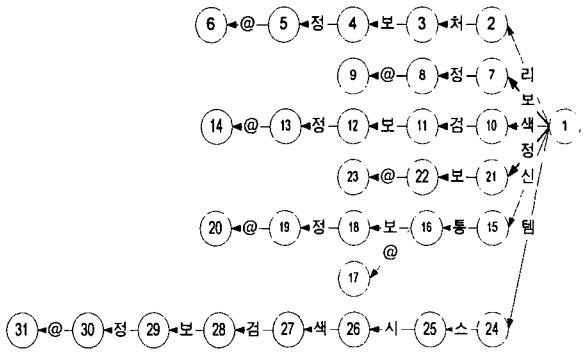

그림 2 역문자열로 구성된 상태 전이도

그림 2의 트라이 구조는 단어의 역문자열로 구성된 단어의 검색이 가능하다. 그러나 순방향 문자열로 구성된 단어들에 대한 검색은 불가능하다. 따라서 순방향 문자열 단어 검색과 역방향 문자열 단어 검색을 위해서는 2개의 사전을 유지하거나 새로운 방법이 필요하다. 순방향 문자열 단어 검색과 역방향 문자열 단어 검색을 동시에 지원하기 위해서는 새로운 사전 구조가 필요하다.

순방향 문자열 단어와 역방향 문자열 단어 검색을 동시에 지원하기 위해 그림 1과 2를 통합한 구조를 고려할 수 있다. 집합 S1의 단어 개수가 7개이므로 7개의 단어를 모두 저장하면 14개의 엔트리를

가져야 한다. 그러나 "정보"와 "보정"의 서로 다른 단어를 저장하는 과정에서 단어 "정보"와 단어와 "보정"의 역문자열 "정보"가 동일 위치에 저장되는 문제점이 발생한다. 역시 "보정"의 역문자열 "정보"를 저장하는 과정에서 같은 문제점이 발생한다. 따라서 14개의 엔트리가 아닌 12개의 엔트리를 가진다. 이는 결과적으로 검색 과정에서 "정보"라는 정방향 문자열과 "보정"의 역방향 문자열 "정보"의 관련 정보가 동일한 위치에 저장되어 잘못된 정보를 가져오는 결과를 가져온다. 따라서 그림 1과 2를 단순 통합한 사전 구조는 비효율적임을 알 수 있다. 지금까지 지적한 단점을 해결하기 위해서는 새로운 방법이 필요하다. 그림 1의 트라이 구조는 단어의 prefix를 공유하고, 그림 2의 구조는 단어의 suffix를 공유하는 특징을 가지므로 prefix와 suffix를 최대로 공유하는 것이 바람직하다. 그림 1과 2의 특징을 최대로 활용하여 새로 설계한 사전 구조는 그림 3과 같다. [4]에서는 단어의 suffix를 공유하도록하는 연구가 있었으나 이 구조는 역방향 단어 검색이 불가능한 단점이 있다.

본 논문에서는 이 단점을 해결하기 위하여 [4,5]을 개선한 새로운 사전 구조를 제안한다. 제안한 구조는 순방향 문자열 단어 검색은 물론 단어의 역방향 문자열로 구성된 단어 검색을 동시할 수 있다. 그림 3에서 기호 "@"는 단어의 순방향 검색 과정에서 1개의 단어가 다른 단어에 내포되어 나타나는 부분 문자열일 때 두 개의 단어(정보, 정보처리)를 구분하기 위한 단어의 종료 기호이고, 기호 "&"는 임의 단어의 역문자열 검색 과정에서 1개의 역문자열이 다른 역문자열에 내포되어 나타나는 부분 문자일 때 두 개의 서로 다른 단어(신통-통신 역문자열, 신통보정-정보통신 역문자열)를 구분하기 위한 단어의 시작 기호이다. 단어의 검색 과정은 순방향 검색일 때는 왼쪽 트라이의 시작 노드 1에서 시작하여 오른쪽 트라이의 마지막 노드 1에 도달하면 검색에 성공하고, 역문자열 검색은 오른쪽 트라이 시작노드 1에서 시작하여 역문자열에 의해 왼쪽 트라이 시작노드 1에 도달하면 역방향 단어 검색은 성공한다. 이 구조는 순방향 문자열의 prefix가 역방향 문자열의 suffix를 포함하고 역방향 문자열의 prefix가 순방향 문자열의 suffix를 포함하여 양방향 검색을 위해 필요한 노드 수를 최소화 하였다.

4. 실험 평가

기억장소 실험을 위하여 2개의 트라이 구조에 순방향 단어와 역방향 단어를 별도로 저장하는 방법(방법 1), 순방향 문자열과 역방향 문자열이 동일한

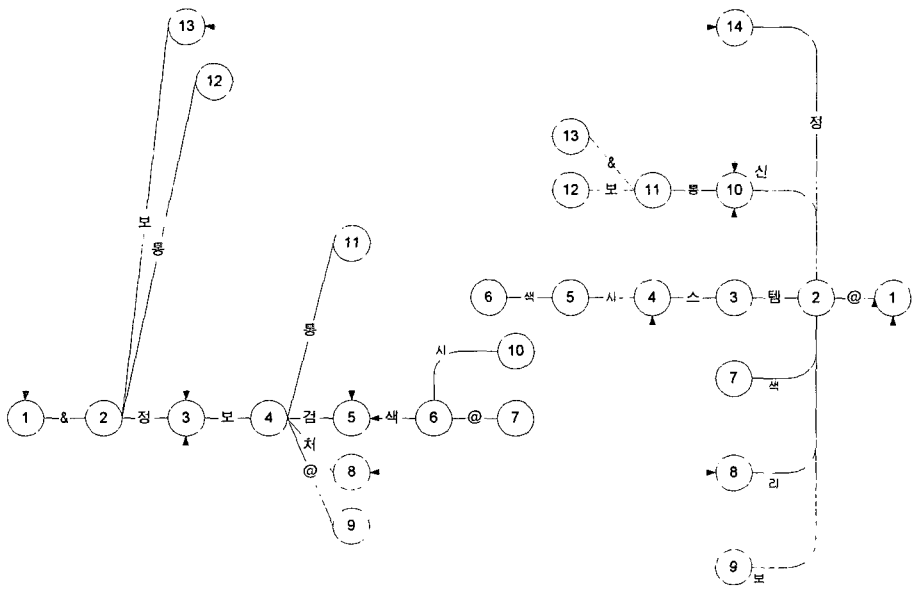

그림 3 제안된 양방향 트라이 구조

경우를 고려하지 않고 통합된 트라이에 저장하는 방법(방법 2)과 본 논문에서 설계한 트라이 구조에 저장한 방법(방법 3)을 각각 이용했을 때 필요한 기억장소를 보면 그림 4와 같다. 그림 4에서 보는 것처럼 단어 1만개를 삽입했을 때 방법 1과 방법 2가 필요한 기억장소는 200KB이고 방법 3은 170KB를 요구하여 비슷하지만 8만개의 단어를 삽입했을 때 방법 1은 1,380KB, 방법 2는 1,300KB, 방법 3은 1,180KB가 필요하였다. 이는 논 논문에서 설계 구현한 방법이 기억장소를 작게 차지함을 알 수 있다.

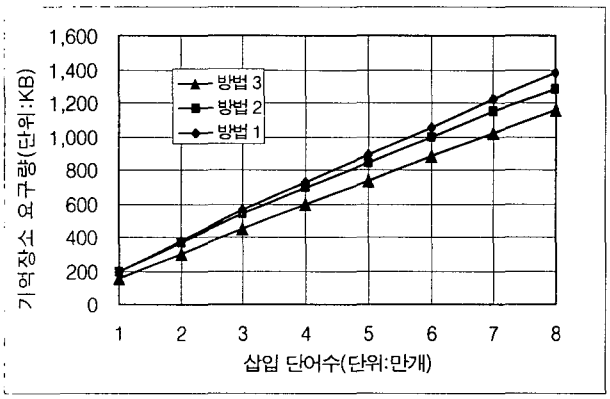

그림 4 기억 장소 요구량

검색 시간은 삽입 단어 수에 관계없이 단어 길이에만 의존하여 검색 시간이 결정되므로 1개의 단어 검색 시간이나, 와일드카드 지원을 위한 검색 시간은 모두 방법 1, 2, 3 모두 같으므로 비교하지 않았다.

5. 결 론

본 논문에서는 순방향 문자열 단어와 역방향 문자열 단어를 효율적으로 검색할 수 있는 전자 사전 구조를 설계하고 구현하였다. 구현된 전자 사전은 기억장소 공간을 최소화하면서 1개의 단어를 검색하기 위해 소요되는 검색 시간이 빠를 뿐만 아니라 삽입 단어 수에 관계없이 단어 길이에만 의존하는 검색 시간이 정해지는 특징을 가진다.

형태소 분석 과정에서 전자 사전 참조가 매우 많이 발생한다. 분석 방향에 관계없이 이용할 수 있으며, 분석 과정에서 분석의 대상이 되는 어간과 어미에서 어간은 순방향 단어 검색을 통하여 어미는 역방향 단어 검색을 통하여 효율적으로 이용할 수 있다.

본 사전 구조는 형태소 분석 환경뿐만 아니라 정보 검색을 위한 사전으로 유용하게 이용할 수 있다. 정보검색 시스템에서 질의어가 전절단된 환경의 앞부분 확장, 후절단된 경우의 뒤부분 확장, 중간 절단된 경우의 중간 부분 확장을 지원하도록 검색 루틴을 변경하여 절단 검색을 지원해 주고자 한다. 절단 검색은 정보 검색 환경뿐만 아니라 문자열 인식 후처리, 음성인식등 처리 과정에서도 매우 유용하게 사용할 수 있다.

참고 문헌

[1] T. G. Lewis and C. R. Cook, "Hashing for dynamic and static internal tables,: IEEE Computer, pp. 45-56, Oct. 1988.

[2] E. Fredkin, B. Beranek and Newman, "Trie memory", CACM, Vol3, pp. 490-499, 1960.

[3] H. I. Aoe, "An Efficient Digital Search Algorithm by Using Double-array Structure," IEEE Trans. on S/W Eng., Vol. 15, No. 9, pp.1066-1077,1989.

[4] 김철수외 3인, "이중배열 트라이 구조를 이용한 한국어 전자 사전 구축, 한국정보과학회 논문지, Vol. 23, pp. 85-94, 1996.

[5] K. Morimoto, H. Iroguchi and J. I. Aoe, A Retrieval Algorithm of Dictionaries by Using Two Trie Structures, 일본 전자공학회 논문집D-II Vol. J76-D-II No.11, pp.2374-2383, 1994.