길이 비트맵 해시 기반 형태소 분석 시스템

선충녕° 민경구 서정연
°다이퀘스트
서강대학교 컴퓨터학과
wiliwisp@diquest.com°, mingk24@gmail.com, seojy@sogang.ac.kr

Length Bitmap HASH Based POS Tagging System

Choong-Nyoung Seon°, Kyungkoo Min, Jungyun Seo
°Diquest Inc., Department of Computer Science, Sogang University

요 약

인터넷의 확장에 따라 형태소 분석기에서 사용하는 사전의 규모도 커지고 있다. 이러한 상황은 사전의 증가를 가져옴으로써 기존 형태소 분석기의 자료 구조에 대한 새로운 요구를 발생시켰다. 기존의 트라이를 이용한 방법은 노드의 과다 생성과 데이터 부족문제로 발생하는 메모리 낭비의 문제를 가지고 있다. 효율적인 메모리 사용을 위해서는 해시 구조가 적절하다. 하지만 이 경우 트라이에 비해 검색 횟수의 복잡도가 비약적으로 증가되는 문제점을 안고 있다. 본 논문에서는 해시를 위한 길이 비트맵을 이용하여 검색 횟수를 제한할 수 있는 방법을 제안하였다. 실험을 통해 제안된 자료 구조와 해시와 트라이의 형태소 사전 검색 횟수를 비교하였으며, 비문 사용이 많은 영역에서 효율적임을 입증하였다.

1. 서론

인터넷의 이용이 활발해짐에 전자문서의 수가 비약적으로 증가하고 있고, 신조어, 전문 용어 등의 다양한 단어가 출현하였다. 이에 따라 정보 검색 등의 문장 분석을 위해 사용해야 하는 사전의 크기도 꾸준히 증가하고 있다[8].

많은 한국어 형태소 분석기는 트라이 사전을 이용한 Tabular parsing 방법을 사용한다[4][7]. 트라이 구조는 하나의 단어를 찾기 위해 철자(음소 혹은 음절)를 이용한다. 이 때문에 불필요한 탐색 시간을 소모하기도 한다.. 뿐만 아니라 링크 구조를 유지하기 위해 필요한 메모리도 다른 구조에 비해 큰 편이다. 따라서 단어 수와 길이가 길어지는 인터넷 상의 요구에 부응하기 위해 형태소 분석기의 사전 탐색 구조의 변화가 요구된다.

본 논문에서는 구조적으로 단순한 해시를 이용한 형태소 분석 방법을 제안한다. 해시는 구조가 단순하면서도 성능과 메모리 사용에 있어서 다른 자료 구조에 비해 효율적임이 증명되어왔다[1][3]. 하지만 해시 검색은 트라이에 비해 검색 횟수를 많이 요구하는 단점 때문에 형태소 분석에는 많이 사용되지 않았다. 이 논문에서는 해시 형태소 사전의 탐색 횟수를 줄이기 위한 메타 자료 구조를 제안한다. 제안한 메타 자료 구조는 단어들의 길이 정보를 이용하여 검색횟수를 줄여주어 형태소 분석 영역에서 해시가 가지는 단점을 보완해 줄 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 형태소 분석 작업에서 트라이의 장단점을 살펴보고, 3장에서는 본 논문에서 제안한 길이 비트맵 자료 구조에 대해 소개한다. 4장에서 제안한 구조를 검증하기 위한 실험 결과를 보이고 분석하며, 마지막으로 5장에서 결론을 맺는다.

2. 형태소 분석에서 트라이 구조의 장단점

트라이는 각 단어를 실제 단어가 출현한 문장에서 철자별 비교를 통해 사전에 등록된 단어를 찾아내는 자료구조다[2]. 사전의 문자열은 중복적으로 발생할 수 있는 특징이 있기 때문에 트라이는 Tabular Table에서의 사전 탐색 횟수를 줄여준다. 하지만 단어 혹은 구조 단위의 형태가 아닌 트리의 노드를 구성하는 것이 철자단위가 되면서 하나의 단어를 표현하기 위해 많은 리소스를 요구하게 된다. 예를 들어, '남산'이라는 단어를 표현하기 위해 일반적인 자료구조에서는 하나의 정보단위를 활용하는데 비해 트라이는 음소 단위인 경우 6 노드를, 음절 단위인 경우에는 2개의 노드를 이용해야 한다. 사전에 등록된 단어가 서로 정보를 공유하기는

하지만 길이가 긴 단어가 등록되고, 하위 노드로 갈수록 데이터의 희소성 문제가 심각하게 발생하게 된다. 이러한 희소성 문제는 탐색의 속도와 리소스의 사용 측면에서 부정적인 영향을 미치고 있다. 이를 해결하기 위해 트라이의 변형이 있어왔고 그것을 통해 탐색의 효율성을 확보할 수 있었다[1][3]. 하지만 그럼에도 불구하고 여전히 다른 사전 구조에 비해 많은 리소스를 요구하고 있으며 더욱 복잡한 구조를 요구하게 되었다.

3. 길이 비트맵 해시

해시는 트라이와 같은 구조상의 문제를 가지지 않는다. 따라서 대규모 사전이나 긴 항목에 대해서는 해시가 더 유리하다. 하지만 앞 절에서 설명한 바와 같이 트라이는 한번의 검색으로 중복된 여러 단어를 한꺼번에 검색하기 문에 모든 단어를 찾는데 입력 길이에 비례한 검색 횟수만을 요구한다. 이에 비해 해시는 모든 후보를 고려해야 하므로 길이의 제곱에 비례한 사전 탐색이 요구된다. 다행히 이와 같은 차이는 어절을 대상으로 하는 경우 길이가 짧기 때문에 크게 부각되지 않는다. 하지만 어절 분리가 명확하지 않은 영역에서 이러한 검색 횟수 복잡도 차이는 심각한 문제를 불러올 수 있다.

따라서 본 논문에서는 해시를 이용하면서 검색의 횟수를 최소화하는 자료 구조인 길이 비트맵을 제안한다. 길이 비트맵은 등록 단어의 첫 글자를 색인으로 가지는 비트맵 배열로 각 항목의 비트들은 그 색인으로 시작하는 해당 길이의 단어가 있는지를 의미한다. 이를 이용하여 사전 탐색 전에 찾고자 하는 단어의 길이로 길이 비트맵에서 조건을 검사하고, 없는 경우 검색을 생략할 수 있게 된다. 예를 들어 '김치', '김밥말이', '김', '김장시즌', '나들이꽃', '나라', '나랏말', '나루터지기' 등과 같은 단어가 포함된 사전이 있을 때 음절 기반의 사전 생성시 다음과 같은 길이 비트맵이 함께 생성되게 된다. 음절 '김'으로 시작하는 단어는 길이가 1인 '김', 길이가 2인 '김치', 길이가 4인 '김밥말이', '김장시즌' 등이 사전에 등록되어 있다. 위의 단어들만 등록된 사전에서는 단어의 최대 길이가 5가 되므로 각 첫 음절에 해당하는 길이 비트맵은 다음과 같이 구성된다.

길이 비트맵
김 : 11010
나 : 01111

그림 1 길이 비트맵 예

이렇게 생성된 길이 비트맵은 형태소 분석에서 사전 탐색할 때 이용되게 된다. '김'으로 시작하는 어절을 분석할 때 어절의 길이가 몇이건 항상 사전에 존재하는 1, 2, 4의 길이에 대해서만 탐색하여 검색 횟수를 줄이는 효과를 가져온다.

4. 실험

본 논문에서 제안한 길이 비트맵을 검증하기 위해 Tabular Table을 가정하고 실제 영역에서 검색 횟수를 비교하였다. 이를 통해 제안된 자료 구조가 해시의 사전 탐색 횟수를 어느 정도 제한해 줄 수 있는지 실험하였다. 실험 대상으로는 신문 기사 코퍼스, 웹 커뮤니티 게시판의 제목 코퍼스, 웹 커뮤니티 게시판의 본문 코퍼스로 각각의 성격은 아래의 표1 과 같다.

표 1 사정등록어의 길이 정보

	신문	게시판 제목	게시판 본문
어절 수	20,519,681	18,066,667	23,129,653
평균 어절 길이	3.90	4.36	3.43
최대 어절 길이	143	194	10,064

위와 같이 어절의 평균 길이는 영역에 따라 큰 차이를 보이지는 않는다. 하지만 최대 어절의 길이는 영역에 따라 크게 달라지는 것을 볼 수 있다. 사전에 등록된 단어 길이에 해당하는 단어들의 수와 최대 길이를 파악할 수 있는 그래프는 그림 2와 같다

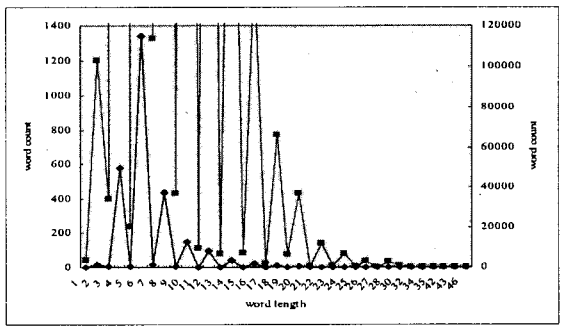

그림 2 사전에 등록된 단어 수와 길이에 대한 분포

왼쪽의 Y축은 빈도가 작은 항목을 나타내기 위해, 그리고 오른쪽의 Y축은 빈도가 큰 항목들을 나타내기 위한 축이다. 빈도가 많은 것과 적은 것이 번갈아

출현하는 이유는 한국어가 2바이트 문자이기 때문에 상대적으로 길이가 짝수인 단어의 비율이 크기 때문이다. 단어의 길이가 3글자(길이 6)인 단어들이 가장 많이 등록되었고, 그 이후 급격히 줄어 8글자(길이 16) 이내의 단어들이 사전의 99%이상을 차지한다. 즉, 길이 17이상의 탐색은 대부분이 실패하고, 최대 길이인 48 이상의 탐색은 의미가 없음을 보여주고 있다. 아래의 그림 2는 사전 내용들로 표현된 길이 비트맵의 일부를 보여준다.

	길이 비트맵
!	10000000000000000000000000000000000000000000000000
"	10000000000000000000000000000000000000000000000000
…	…
대	01011111110111011101010001110000000000000000000
댁	01010100000000000000000000000000000000000000000000
…	…

그림 3 형태소 사전 길이 비트맵의 예

표2 길이 비트맵의 자료 희소성

첫 음절 수	길이있음	길이없음	%
1782	8301	73671	10.13%

위의 표에서 보여주듯 길이 비트맵은 전체의 10%만 채워지는 수준으로 매우 성기게 구성된다. 빠른 속도로 길이 비트맵을 탐색할 수 있다면 이와 같은 희박한 구성은 해시를 검색하는 것에 비해 효율적으로 사전탐색을 수행할 수 있다는 것을 의미한다. 트라이, 해시, 길이 비트맵+해시의 검색 횟수를 비교하면 아래 그림4와 같다. 그림4에서 보듯이 문서에 비문이 많은 경우, 트라이에 비해 해시의 탐색 횟수가 크게 증가됨을 보여주고 있다. 제안한 길이 비트맵을 이용하는 경우 해시와는 달리 비문의 사용이 많은 영역에서 트라이와 유사한 형태의 변화를 보여주고 있다.

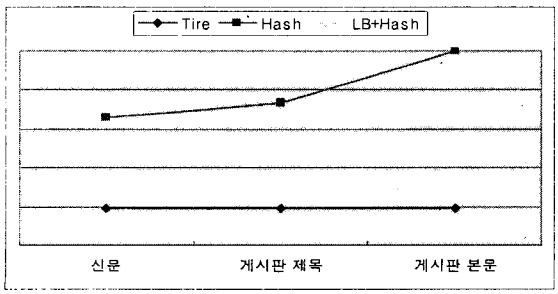

그림 4 사전구조와 영역에 따른 검색횟수 비교

5. 결론

기존의 형태소 분석기들은 트라이에 의존한 경우가 많았으며, 이 경우 복잡한 구조로 인해 사전 용량이 커지는 경우 메모리 요구나 처리 상의 문제가 발생하는 경우가 있었다. 본 논문에서는 해시를 보완할 수 있는 자료구조를 제안하였다. 해시가 트라이에 비해 가지는 단점인 검색횟수 문제를 길이 비트맵을 사용하여 줄여줄 수 있었다. 또한 실험을 통해 제안한 시스템이 띄어쓰기가 잘 지켜지는 영역뿐만 아니라 그렇지 못한 영역에서는 더욱 기여할 수 있음을 보여주었다.

6. 참고 문헌

1. A. Andersson, S. Nilsson, "Imporved Behavior of 트라0ls by Adaptive Branching", Information Processing Letters 46, 6(June), 295-300, 1993.
2. E. Fredkin. " Trie Memory" . Comm. ACM 3(9), 490-499, Sept. 1960
3. S. Heinz, J. Zobel, H. E. Williams, " Burst tries: A Fast, Efficient data structure for string keys" , ACM Trans-actions on Information Systems 20(2), 192-223, 2002.
4. Kyung-il Lee, Tae-sung Ahn., " Development of POS Tagging System independent to Word Spacing" , Proc. of the 15th conference on Hangeul and Korean Information Processing (Korean), 2003
5. NIST., http://www.nist.gov/dads/HTML/trie.html
6. C. Seon, Y. Ko, J. Kim, J. Seo, " Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules" , NLPRS, 2000
7. S. Kim, " A morphological Analyzer for Korean Language with Tabular Parsing Method and Connectivity Information" , 1986.
8. H. E. Williams, J. Zobel, "Searchable Words on the Web", 200