후위부자유 정규언어의 복합 연산 상계 상태복잡도 [1)]

엄해성 ○, 한요섭
연세대학교 컴퓨터과학과
{haesung21, emmous}@cs.yonsei.ac.kr

State Complexity Upper Bound of Combined Operations for Suffix-Free Regular Languages

Hae-Sung Eom ○, Yo-Sub Han
Department of Computer Science, Yonsei University

정규언어는 현재 패턴매칭[1], 검색엔진[2] 그리고 여러 응용프로그램[3]에 널리 사용된다. 응용프로그램들의 크기가 커짐에 따라서 사용되는 정규언어의 크기도 커지고 있다. 상태복잡도(state complexity)는 정규언어의 크기를 나타낼 수 있는 표현 중 하나로써 현재 많은 연구가 진행되고 있다[4-7]. 우리는 기존 연구 결과를 바탕으로 후위부자유 정규언어에 대한 몇 가지 복합연산(합집합의 스타, 교집합의 스타, 역의 스타 그리고 연쇄의 스타)의 상계(Upper bound) 상태복잡도를 살펴본다.

네 가지 복합연산의 상계 상태복잡도를 알아보기 위해 상태 개수가 m 개인 임의의 최소 결정적 유한 인식기 A_1 을 $A_1 = (Q_1, \Sigma, \delta_1, q_{0,1}, F_1)$, 상태 개수가 n 개인 임의의 최소 결정적 유한 인식기 A_2 을 $A_2 = (Q_2, \Sigma, \delta_2, q_{0,2}, F_2)$라고 정의하고, $L_1 = L(A_1)$, $L_2 = L(A_2)$라 정의하자. 먼저 합집합의 스타 $(L_1 \cup L_2)^*$를 승인하는 최소 결정적 유한 인식기 $A = (Q, \Sigma, \delta, q_0, F)$를 설계하면, $Q = P \cup R \cup \{ q_{0,1}, q_{0,2} \} \cup \{ d_{0,1} , d_{0,2} \}$이고, 여기서 $P = \{ P_1 \cup P_2 \mid \emptyset \neq P_i \subseteq Q_i - F_i - Q_{0,i} , i = 1 , 2 \}$, $R = \{ R \subseteq Q_1 \cup Q_2 \mid q_{0,1}, q_{0,2} \in R , R \cap (F_1 \cup F_2) \neq \emptyset \}$이다. 최소 결정적 유한 인식기 A 의 상태 개수는 가장 많이 나오는 경우 $2^{m+n-4} - 2^{m-3} - 2^{n-3} + 3$ 개이다. 다음으로 교집합의 스타 $(L_1 \cap L_2)^*$를 승인하는 최소 결정적 유한 인식기 $A = (Q, \Sigma, \delta, q_0, F)$를 설계하면, $Q = P \cup R \cup \{ q_{0,1}, q_{0,2} \} \cup \{ d_{0,1} , d_{0,2} \}$이다. 여기서 $d_{0,1}$, $d_{0,2}$은 A_1 , A_2 각각의 막힘상태(sink state)이고, $P = \{ P_1 \cup P_2 \mid \emptyset \neq P_i \subseteq Q_i - F_i - Q_{0,i} , i = 1 , 2 \}$, $R = \{ R \subseteq Q_1 \cup Q_2 \mid q_{0,1}, q_{0,2} \in R , R \cap F_1 \neq \emptyset , R \cap F_2 \neq \emptyset \}$이다. 최소 결정적 유한 인식기 A 의 상태 개수는 가장 많이 나오는 경우 $2^{m+n-5} - 2^{m-3} - 2^{n-3} + 3$ 개이다. 다음으로 역의 스타 $(L_1^R)^*$의 상계 상태복잡도를 계산하기 위해 우선 $(L_1^R)^*$ 를 승인하는 비결정적 유한 인식기(NFA) $A' = (Q', \Sigma, \delta', q_0', \{q_{0,1}, q_0'\})$를 설계한다. 여기서 q_0'는 Q 에 있지 않은 새로운 초기상태이다. 비결정적 유한 인식기 A'를 A'와 같은 것을 승인하는 최소 결정적 유한 인식기 $A = (Q, \Sigma, \delta, q_0, F)$로 변형한다. A 의 상태 개수는 Q_1 의 전체 상태들에서 막힘상태를 제외한 m-1 개의 상태들의 부분집합의 개수이므로 2^{m-1} 개이다. 최소 결정적 유한 인식기 A 는 후위부자유를 역으로 한 전위부자유이기 때문에, 2^{m-1} 개의 부분집합들에서 승인상태인 $q_{0,1}$ 가 전이에 영향을 미치지 않으므로 $q_{0,1}$ 을

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0009168)

포함한 부분집합은 $q_{0,1}$ 를 포함하지 않은 부분집합과 같은 입력 값에 같은 전이를 하게 되므로 총 2^{m-1} 개의 부분집합에서 2^{m-2} - 1 개의 부분집합을 빼면 2^{m-2} + 1 개가 남는다. 그러므로 상계 상태복잡도는 2^{m-2} + 1 이 된다. 마지막으로 연쇄의 스타 $(L_1 \cdot L_2)^*$를 승인하는 최소 결정적 유한 인식기 A = (Q, Σ, δ, q_0, F)를 설계한다. q_0는 빈 문자열(empty string)을 승인하기 위해 새롭게 만든 초기상태이며, Q = ({q_0} ∪ P) – (B ∪ C ∪ D ∪ E)이다. 여기서, P는 (Q_1 ∪ Q_2) – {$d_{0,1}$, $d_{0,2}$}의 모든 부분집합들을 요소로 가진 집합이고, $d_{0,1}$, $d_{0,2}$는 각각 A_1와 A_2의 닫힘상태이다. B 는 Q_2 – {$d_{0,2}$}의 부분집합이고, C = {X ⊆ Q_1 – {$d_{0,1}$} | X ≠ Ø, X ≠ {q}, q ∈ Q_1 – F_1 – {$d_{0,1}$}}, D = {X ⊆ Q_1 ∪ Q_2 - {$d_{0,1}$, $d_{0,2}$} | X ∩ F_1 ≠ Ø, X ∩ Q_2 ≠ Ø, $q_{0,2}$ ∉ X}, E = {X ⊆ Q_1 ∪ Q_2 - {$d_{0,1}$, $d_{0,2}$} | X ∩ F_2 ≠ Ø, X ∩ Q_1 ≠ Ø, $q_{0,1}$ ∉ X}이다. 최소 결정적 유한 인식기 A 의 상태 개수는 가장 많이 나오는 경우 $7 \cdot 2^{m+n-6} – 2^{m-2} – 2^{n-2}$+ m 개이다.

우리는 후위부자유 정규언어라는 특수한 경우에 대하여 몇 가지 복합연산(합집합의 스타, 교집합의 스타, 역의 스타, 연쇄의 스타)에 대해 상계 상태복잡도를 살펴보았고 각각의 복합연산에 대해 다음과 같은 결과를 얻을 수 있었다.

표 1. L_1 은 m 개의 상태를 가진 정규언어, L_2 는 n 개의 상태를 가진 정규언어

연산자	상계 상태복잡도
$(L_1 \cup L_2)^*$	$2^{m+n-4} – 2^{m-3} – 2^{n-3} + 3$
$(L_1 \cap L_2)^*$	$2^{m+n-5} – 2^{m-3} – 2^{n-3} + 3$
$(L_1^R)^*$	$2^{m-2} + 1$
$(L_1 \cdot L_2)^*$	$7 \cdot 2^{m+n-6} – 2^{m-2} – 2^{n-2}$+ m

위와 같은 결과로 상계 상태복잡도가 기본 연산의 합성에 비해서 복합 연산으로 인해 더 적어진 결과를 얻을 수 있었으며, 후위부자유라는 특수한 경우를 적용했을 때 그렇지 않을 때보다 더 적어진 결과를 얻을 수 있었다.

참고문헌

[1] A. V. Aho.: Algorithms for Finding Patterns in Strings. Handbook of Theoretical Computer Science: Volume A: Algorithms and Complexity, pp. 255-300 (1990)

[2] Ye Fei-yue, Bian Li-ya, Li Hang.: Application of Information Extraction in Oil Search Engine. Web Information Systems and Mining, WISM 2009, pp 98 – 102 (2009)

[3] Mulder. M, Nezlek. G.S.: Creating protein sequence patterns using efficient regular expressions in bioinformatics research. Information Technology Interfaces, pp 207-212 (2006)

[4] Narad Pampersad.: The state complexity of L^2 and L^k. Information Processing Letters 98 , pp. 231-234 (2006)

[5] S. Yu, Q. Zhuang, K. Salomaa.: The State Complexities of some basic operations on regular languages, Theoretical Computer Science 125 (2), pp.315-328 (1994)

[6] Yo-Sub Han, Kai Salomaa, Sheng Yu.: State Complexity of Combined Operations for Prefix-Free Regular Languages. LATA 2009, LNCS 5457, pp. 398-409 (2009)

[7] Yo-Sub Han, Kai Salomaa.: State complexity of basic operations on suffix-free regular languages. Theoretical Computer Science 410, pp. 2537-2548 (2009)