# New Composition Method of TRIE's Key

*Kwak Son Il, Jo Chol Min* and *Kwon O Chol*

**Abstract** In this paper we proposed a new method for composition of TRIE's key in order to extend the coverage of application of TRIE, high speed search structure and verified its effectiveness.

**Key words** TRIE, search structure

## Introduction

The precedent studies have the restriction that the end mark symbol has to be attached behind each key in composing the key of TRIE structure and this leads TRIE not to correspond any key set. Namely, as the end mark symbol has to be attached to the end of key in the precedent TRIE structure, the key doesn't include the same character to the end mark symbol, thus the coverage of using TRIE is narrower than other structures.

TRIE structure has been used most widely in the programs for natural language processing including Chinese-Korean machine translation and so on because of high speed search time without dependency on number of key count, providing the several functions necessary in natural language processing and effectiveness in manipulation of dictionary [1].

In the TRIE structure of the precedent studies[2, 3] several branches are generated by each character of key and the end mark symbol '#' for unique identification of each key is attached to the state that is the end of key.

The whole keyword corresponding to search key becomes the unit of comparison in Hash search, binary tree search and B/B+ tree search, but each character consisting of key becomes the unit of comparison in TRIE structure [4].

In the precedent concept of TRIE structure, the limitation that the end mark symbol has to be attached to the end of each key makes it impossible for TRIE to correspond to arbitrary key set. In fact, not only normal string key but also arbitrary data such as numeric key, Unicode string key and so on can be key in other excellent search structures: B/B+ tree structure and Hash structure. Because in the precedent TRIE structure the end mark symbol has to be attached to the end of key, key cannot include the character same as the end mark symbol, thus the coverage of TRIE's usage is more limited than other search structures.

## 1. Some Definitions for New Composition of Key

TRIE is a species of string pattern matching machine and is formally defined as follows.

$$I=\{a_1,\ a_2,\ \cdots,\ a_e\}:\ \text{finite set of input symbols.}$$

$$I^{+} = \{a_1 a_2 \cdots a_n,\ a_i \in I,\ n \geq 1\},$$
$$I^{*} = I^{+} \cup \varepsilon$$

where $\varepsilon$ is the blank string(its length is 0).

$K(\subseteq I^{*})$: Key set

**Definition 1** String Pattern Matching Machine

If $M = \{S,\ I^{*},\ g,\ S_1,\ U\}$, $S$ is the finite set of states, $S_1$ the finite set of initial states, $U$ the set of final states, and $g$ state transition function $S * I^{*} \to S \cup \{fail\}$:

$$\exists x \in I^{*},\ \exists q \in S_1,\ g(q,\ x) := f,\ f \in U$$

it is said that $M$ accepts sequence $x$, the set of all the sequences that $M$ accepts is represented by $L(M)$.

Now, when $K = L(M)$, $M$ is called as String Pattern Matching Machine of key set $K$.

The branch that is marked with $a \in I$ from state $r$ to state $t$ is represented by $g(r,\ a) := t$ and if the branch is not defined, $g(r,\ a) :=$ fail or $g(r,\ a) := 0$.

**Definition 2** Input dimension $\text{in}\deg(r)$ and output dimension $\text{out}\deg(r)$ of state $r \in S$ are defined as follows.

$$\text{in}\deg(r) = |\{(t,\ a) := r,\ a \in I,\ t \in S\}|,\ \ \text{out}\deg(r) = |\{(t,\ a) \mid g(r,\ a) := t,\ a \in I,\ t \in S\}|$$

State with output dimension 0 is called as final state, the whole set of final states is represented by $F$.

**Definition 3** TRIE

If string pattern matching machine $M$ satisfies the following conditions on the state $r \in S$, $M$ is called as TRIE.

① $\exists r \in S,\ r = 1 \to \text{in}\deg(r) = 0$

② $r \neq 1,\ \forall r \in S \to \text{in}\deg(r) = 1$

③ $\forall r \in S,\ \exists y \in I^{*} \to g(1,\ y) := r$

**Definition 4** New Key of TRIE

When $k^{*} = ls,\ s \in I^{*},\ l = |s| + 1,\ k^{*}$ is defined as new key of TRIE.

From definition 4 the set of final states can be defined again as follows.

**Definition 5** Set of final states of TRIE

For the initial state $r \in S$ and key of TRIE $k^{*} = ls$

$$U^{*} := \{u \mid g(r,\ l) := q,\ g(q,\ s) := u,\ \text{out}\deg(u) = 0,\ q,\ u \in S\}$$

is defined as new set of final states of TRIE.

**Theorem** For the set of key $K,\ K^{*}(K \subset I^{*},\ K^{*}$ is new set of TRIE's key corresponding to $K$ ) and the set of final states $U^{*}$, each key of $K$ is uniquely identified in TRIE.

**Proof** If each key of the set of key constructed into TRIE is corresponded to only one another final state, this theorem is proved.

$k_1^* = l_1 a_1 a_2 \cdots a_n,\ k_2^* = l_2 b_1 b_2 \cdots b_m$ are respectively new keys corresponding to

$\forall k_1,\ k_2 \in K(k_1 \ne k_2,\ k_1 = a_1 a_2 \cdots a_n,\ k_2 = b_1 b_2 \cdots b_m,\ l_1 = n+1,\ l_2 = m+1)$.

① In case of $n \ne m$

Because $l_1 \ne l_2$, for the initial state $r$, $\exists t,\ q \in S$, $g(r,\ l_1) := t$, $g(r,\ l_2) := q$, $t \ne q$ is concluded. Thus each other final states are responded to $k_1^*$, $k_2^*$.

② In case $n = m$

Because $l_1 = l_2$, for the initial state $r$, $\exists t \in S$, $g(r,\ l_1) := t$, $g(r,\ l_2) := t$. At that time $k_1 \ne k_2$, thus $\mathrm{Index_{diff}} = \{i | a_i \ne b_i,\ 1 \le i \le n\}$, $|\mathrm{Index_{diff}}| > 0$, and there is $i_1 = \min\limits_{i \in \mathrm{Index_{diff}}} i$.

Supposing that the prefix string is

$$s_{front\_share} = \begin{cases} a_1 \cdots a_{i_1 - 1}, & i_1 > 1 \\ \varepsilon, & i_1 = 1 \end{cases},$$

For $\exists q \in S$

$$g(t,\ s_{front\_share}) := \begin{cases} q, & i_1 > 1 \\ t, & i_1 = 1 \end{cases}$$

and there are states $z_1,\ z_2 \in S$ with $z_1 \ne z_2$. Where

$$z_1 = g(t,\ s_{front\_share} a_{i_1}),\quad z_2 = g(t,\ s_{front\_share} b_{i_1}).$$

Thus $k_1^*$, $k_2^*$ respond to each other final states.

From ① and ②, $\forall k_1,\ k_2 \in K(k_1 \ne k_2)$ are uniquely identified by new composition method of key.

## 2. New Composition Method of TRIE's Key

In the practical implementation of new key, the length of key (count of bytes) enters into the first byte of key and the original key puts from the second byte(Fig.).

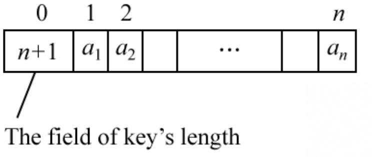

Fig. New composition method of TRIE's key

If key is composed by this method, the first byte of key becomes a field that contains the length of key and the maximum that one byte is stored is 255, thus the length of key can not exceed 255.

In the practical dictionary construction, there is no case that the length of key is over 255, but in the special case that the length of key is over 255 the first two bytes may be used to store the length of key, thus coverage of the length of key is extended to 65 536.

If key is composed by new method, the keys of TRIE can be identified clearly and uniquely without attaching the end mark symbols to the end of key.

## Conclusion

TRIE structure can be applied to any key set by new key composition method.

Because the length of new key equals to the length of key that is composed by the original method, there is no indirect time consumption in search, insertion and deletion by the new composition method of key.

The TRIE structure with new key composition method has been effectively using in developing several applications for natural language processing: Chinese-Korean machine translator "Amrokgang", Korean spellchecker and so on with Unicode version.

## References

[1] 김일성종합대학학보(자연과학), 54, 8, 23, 주체97(2008).

[2] J. Aoe et al.; IEEE Trans. Know. & Data Eng., 8, 3, 1321, 1996.

[3] Kurt Maly; Communications of the ACM, 19, 7, 7, 1976.

[4] D. E. Zegour; Elsevier Information and Software Technology, 46, 923, 2004.