บทที่ 2

หลักการของโครงสร้างพจนานุกรม

ในการแก้ไขโครงสร้างข้อมูลสำหรับพจนานุกรมที่ใช้ในการตัดคำ ต้องคำนึงถึงหลักการสองประการ ที่สำคัญคือ โครงสร้างข้อมูลแบบทราย ที่ใช้กับระบบพจนานุกรมประเภทนี้ และระบบหน่วยความจำเสมือน ซึ่งจำเป็นต่อการขยายขอบเขตและปริมาณของคำในพจนานุกรม

2.1 โครงสร้างข้อมูลแบบทราย (Trie Structure)

โครงสร้างข้อมูลแบบทราย (Fredkin and Edward , 1960; Liang and Franklin M , 1983) มีลักษณะเป็นต้นไม้ (tree) โดยที่แต่ละโหนดมีการเชื่อมโยงไปยังโหนดถัดไปเท่ากับจำนวนตัวอักษรที่มีอยู่ในภาษานั้นๆ ซึ่งโครงสร้างแบบทราย จะแตกต่างจาก โครงสร้างข้อมูลแบบอื่น คือ ข้อมูลจะถูกพิจารณาเป็นตัวอักษรเรียงตามลำดับ ในเซ็ตของคำโดยตรง แทนที่จะมองแต่ละคำเป็นข้อมูลหนึ่งชิ้น ลักษณะเช่นนี้ทำให้สามารถจัดการกับข้อมูลที่มีความยาวแปรผันได้ดี นอกจากนี้ยังสามารถทำอัดคำ (compression) ได้โดยจะจัดการกับ prefix ร่วมของคำในพจนานุกรมโดยการยุบรวมกันหมด ทำให้ลดขนาดพจนานุกรมลงได้ในทันที

ทรายมีโครงสร้างเป็นต้นไม้แบบ N มิติ (N-ary tree) โดยที่ N เป็นจำนวนตัวอักษรในแต่ละคำ แต่ละโหนดของทราย คือ สถานะ (state) ในไฟไนออโตเมตอล (FA :Finite Automaton) แต่ละ edge ซึ่งมีตัวอักษรหนึ่งตัวกำกับ (label) อยู่คือ transitionใน FA

2.2 โครงสร้างข้อมูลแบบทรายในภาษาซี

โครงสร้างของโหนดในทรายประกอบด้วย เขตข้อมูล (field) ต่างๆ ดังนี้

```
typedef struct {
            unsigned  short  chr;
            unsigned  short  next;
            unsigned  short  link;
}
```

โดยที่ chr เก็บตัวอักษร

next เก็บตัวชี้ (pointer) ที่ชี้ไปยังตัวอักษรในลำดับชั้นเดียวกันตัวถัดไปที่มี โหนดแม่ (parent node) เดียวกัน หรือ เรียกว่าเป็นโหนดที่เป็นพี่น้องกัน(Sibling)

link เก็บตัวชี้ที่ชี้ไปยัง ตัวอักษรในลำดับชั้น ถัดไป หรือที่เรียกว่าเป็นโหนดลูก(Child)

ตัวอย่างของโครงสร้างแบบทราย แสดงได้โดยใช้ศัพท์ต่อไปนี้

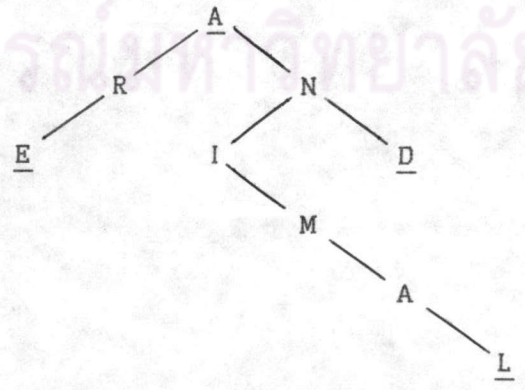
A

AND

ANIMAL

ARE

ศัพท์ทั้ง 4 คำนี้สามารถสร้างเป็นทรายได้ดังต่อไปนี้



รูปที่ 2.1 แสดงโครงสร้างแบบทราย

หรือมองอีกแง่หนึ่ง ทราย คือ การแปลงจากต้นไม้แบบ N มิติ ให้เป็นโครงสร้างต้นไม้แบบไบนารี (binary tree) ซึ่งในแต่ละ edge จะมี เครื่องหมายที่เป็นตัวบอกว่าเป็นอักษรตัวสุดท้ายของคำจากรูป คือ ตัวอักษรที่ขีดเส้นใต้

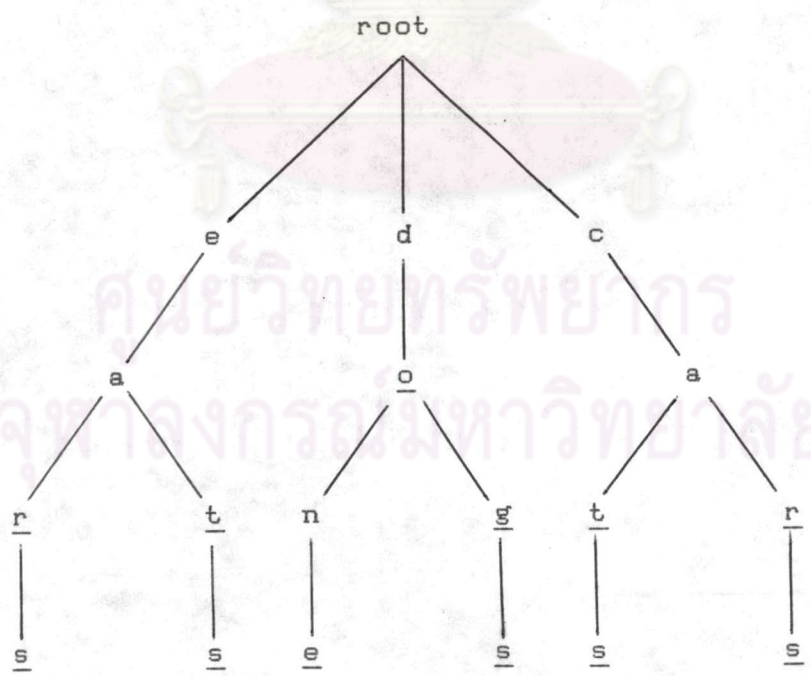
2.3 การอัดคำของโครงสร้างข้อมูลแบบทราย

เนื่องจากในพจนานุกรมใดๆจะประกอบด้วยคำศัพท์ภาษาไทย ซึ่งอาจมีตัวอักษรที่ซ้ำกันซึ่งสามารถใช้ร่วมกันได้ ดังนั้นจึงได้มีการทำ การอัดคำ เพื่อประหยัดเนื้อที่ในการจัดเก็บพจนานุกรม ซึ่งสามารถอธิบายได้ดังตัวอย่างต่อไปนี้

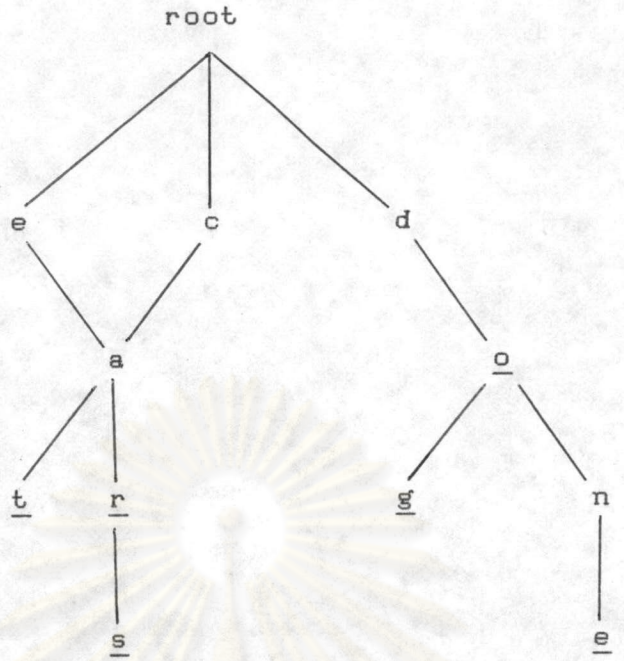
สมมุติว่าพจนานุกรมประกอบด้วยคำศัพท์ 12 คำดังนี้

car	do	ear
cars	dog	ears
cat	dogs	eat
cats	done	eats

สามารถนำมาสร้างเป็นทรายได้ดังนี้



จากทรายข้างต้นสามารถนำมาทำการอัดคำได้ดังนี้



เมื่อจำนวนคำศัพท์ในพจนานุกรมเพิ่มขึ้น ปัญหาที่พบสำหรับ โครงสร้างแบบทรายคือ ถ้าทรายมีขนาดใหญ่ขึ้น จะทำให้ไม่สามารถนำข้อมูลในพจนานุกรมทั้งหมดเข้าไปเก็บในหน่วยความจำหลักได้ จึงจำเป็นต้องมีโครงสร้างใหม่ที่สามารถแบ่งทรายออกเป็นส่วนๆ เมื่อต้องการใช้โครงสร้างส่วนใด จะนำเฉพาะส่วนดังกล่าวเข้ามาในหน่วยความจำหลัก ส่วนที่เหลือจะคงอยู่ในหน่วยความจำสำรอง บทที่ 4 จะได้กล่าวถึงการออกแบบโครงสร้างดังกล่าว