บทที่ 3

การออกแบบโครงสร้างของแฟ้มข้อมูลสำหรับระบบ

แฟ้มข้อมูลของระบบการเก็บและการค้นคืนสารเทศ แบ่งได้เป็น 2 ประเภท คือ

1) แฟ้มข้อมูลนำเข้า (input file)
2) แฟ้มข้อมูลที่สร้างขึ้นจากระบบ

1. แฟ้มข้อมูลนำเข้า

แฟ้มข้อมูลนำเข้า ของระบบการเก็บและการค้นคืนสารสนเทศ โดยใช้แนวความคิดของแฟ้มข้อมูลผกผัน เป็นแฟ้มข้อมูลที่ประกอบด้วย สารสนเทศที่เก็บอยู่ในรูปของเอกสาร ซึ่งสามารถใช้ได้ ทั้งเอกสารที่เป็นภาษาไทย และเอกสารที่เป็นภาษาอังกฤษ โดยที่เอกสารที่เป็นภาษาไทย จะต้องมีการแบ่งข้อความในเอกสาร ออกเป็นคำๆ คั่นด้วยช่องว่าง แต่ละคำต้องมีความยาว ไม่เกิน 15 ตัวอักษร และจะต้องจบภายในบรรทัดเดียวกัน แต่ละเอกสารจะต้องเริ่มต้นด้วย .dh และตามด้วย ชื่อเรื่องของเอกสาร มีความยาวไม่เกิน 80 ตัวอักษร และต้องอยู่ภายในบรรทัดเดียวกัน แต่ละตอน หรือแต่ละย่อหน้าในเอกสารจะต้องเริ่มต้นด้วย .p

ในการเตรียมแฟ้มข้อมูลนำเข้า เตรียมโดยใช้ โปรแกรมเวิร์ดโปรเซสเซอร์ทั่วไป เช่น ชีฮูไรท์เตอร์ หรือ เวิร์ดราชวิถี เป็นต้น โดยชื่อแฟ้มข้อมูลมีความยาวไม่เกิน 8 ตัวอักษร และในส่วนของเอกสารภาษาไทย จะต้องพิมพ์ช่องว่างแทรกระหว่างคำ ที่ต้องการแบ่ง แต่สำหรับในการวิจัยนี้ได้รับความอนุเคราะห์ ให้ใช้โปรแกรมสำหรับตัดคำ จากบริษัทซัมมิตคอมพิวเตอร์

ตัวอย่างเอกสารที่เก็บในแฟ้มข้อมูลข้อความ

.dh วัฏจักร ศุกร์ 5 สิงหาคม 2534 "จับ ประเด็น ต่างประเทศ"
.p "สหรัฐ ห้าม วัยรุ่น สูบ บุหรี่"
การ สูบ บุหรี่ เป็น เรื่อง ที่ ผู้ใหญ่ สูบ เพราะ ความ เคยชิน คน ทำงาน สูบ เพราะ ต้องการ คลาย เครียด จาก การ ทำงาน วัยรุ่น สูบ เพราะ เห็น ว่า เป็น เรื่อง ที่ โก้เก๋ เป็น แมน เป็น เกิร์ล แล้ว แต่ จะ คิดว่า เป็น อะไร แต่ ไม่เคย มี ใคร ที่ สูบ บุหรี่ แล้ว คิดว่า ได้ วิตามิน
.p " บุหรี่ " เป็น สินค้า ชนิด เดียว ใน เมืองไทย ที่มี การ โฆษณา ไม่ให้ สูบ แต่ ก็ ยัง ขาย ดี เป็น เทน้ำเทท่า สถิติ การ สูบ บุหรี่ ที่มี ข่าว ว่า ลดลง เป็น การ ลวงตา แทบ ทั้งสิ้น เพราะ การ จำหน่าย บุหรี่ ยัง เป็น ธุรกิจ ที่ ทำ รายได้ ดี ให้ แก่ พ่อค้า ติด อันดับ โลก อยู่ จนถึง ปัจจุบัน
.p ความจริง ความคิด เรื่อง สูบ บุหรี่ แล้ว โก้ เป็น เรื่อง ที่ บ้านเรา เอง ก็ รับ มาจาก เมืองนอก เช่น มาจาก สหรัฐ อเมริกา แล้ว ก็ พัฒนา มา โดยลำดับ จน ปัจจุบัน บ้านเรา ก็ ชัก จะ ติด อันดับ คน สูบ บุหรี่ มากที่สุด ใน โลก แล้ว
.p เมื่อ มองเห็น โทษภัย ของ บุหรี่ ทั่ว โลก ต่าง ก็ หาทาง ป้องกัน ไม่ให้ มี การ สูบ บุหรี่ มากขึ้น บาง ประเทศ ออก กฎหมาย ห้าม สูบ บุหรี่ ใน ที่ สาธารณะ บาง ประเทศ หาทาง ปิดกั้น ตลาด บุหรี่ ไม่ให้ แพร่หลาย
.p ที่ สหรัฐ เอง เมื่อ วันที่ 4 กรกฎาคม ที่ ผ่านมา กฎหมาย ที่ ทาง มลรัฐ ไอโอวา ออกมา เพื่อ ห้าม ไม่ให้ "วัยรุ่น" สูบ บุหรี่ มี ผล บังคับ ทั้งนี้ เนื่องจาก ปรากฏ ว่า วัยรุ่น สูบ บุหรี่ กัน มาก เหลือเกิน จน ทำให้ มี โทษ ทั้ง ที่ เกิด กับ ผู้ สูบ และ บุคคล ข้างเคียง จึง ต้อง ออก กฎหมาย ห้าม โดยเฉพาะ อย่างยิ่ง ใน ที่ สาธารณะ
.p โดย กำหนด โทษ ไว้ ว่า หาก มี การ ฝ่าฝืน ต้อง โทษ ปรับ ครั้งละ 100 ดอลล่าร์ (ประมาณ 2,500 บาท) หรือ ไม่ ก็ ต้อง ถูก ปรับ ให้ ทำ สาธารณะ ประโยชน์ ตาม ระยะ เวลา ที่ กำหนด ซึ่ง เป็น มาตรการ หนึ่ง ที่ ทาง สหรัฐ ต้องการ ปราบ พวก วัยรุ่น ที่ ชอบ สูบ บุหรี่ ไทย เรา ก็ น่า เอาอย่าง บ้าง
.p ผู้นำ มลรัฐ ดังกล่าว ให้ ความเห็น ว่า เนื่องจาก การ สูบ บุหรี่ มี มากขึ้น ใน กลุ่ม วัยรุ่น จน น่า เป็นห่วง เพราะ วัยรุ่น ที่ ติด บุหรี่ มัก จะ เป็น พาหะ ที่ นำ มา ซึ่ง การ เสพยา เสพติด ชนิด ต่างๆ เนื่องจาก การ สูบ บุหรี่ ทำให้ วัยรุ่น หลง คิด ว่า เป็น การ โก้หรู

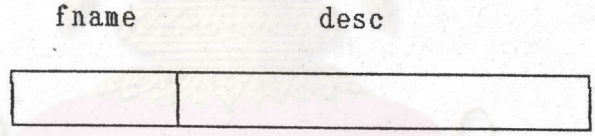
2. แฟ้มข้อมูลที่สร้างขึ้นจากระบบ

แฟ้มข้อมูลที่สร้างขึ้นจากการประมวลผล ระบบการเก็บและการค้นคืนสารสนเทศแบ่งออกเป็น 2 ประเภท คือ

1) แฟ้มข้อมูลรวม
2) แฟ้มข้อมูลเฉพาะ

2.1 แฟ้มข้อมูลรวม

ระบบการเก็บและการค้นคืนสารสนเทศโดยใช้แนวความคิดของแฟ้มข้อมูลผกผัน สามารถจะประมวลผลกับแฟ้มข้อมูลนำเข้า ได้หลายแฟ้มข้อมูล ดังนั้นเพื่อเป็นการอำนวยความสะดวกกับผู้ใช้ในการจดจำชื่อแฟ้มข้อมูล ที่ผ่านการประมวลผลแล้ว จึงได้ออกแบบแฟ้มข้อมูลสารบัญ (DIRECTORY) ขึ้นมา เพื่อเก็บรายชื่อของแฟ้มข้อมูล พร้อมรายละเอียดเกี่ยวกับแฟ้มข้อมูลที่ผ่านการประมวลผลแล้ว โดยใช้โครงสร้างของแฟ้มข้อมูลแบบลำดับ ซึ่งแต่ละระเบียนประกอบด้วย



- fname เป็นชื่อแฟ้มข้อมูลนำเข้า ที่ผ่านการประมวลผลแล้ว มีความยาวไม่เกิน 8 ตัวอักษร

- desc เป็นคำอธิบายรายละเอียดเกี่ยวกับแฟ้มข้อมูลนำเข้า มีความยาวไม่เกิน 80 ตัวอักษร

2.2 แฟ้มข้อมูลเฉพาะ

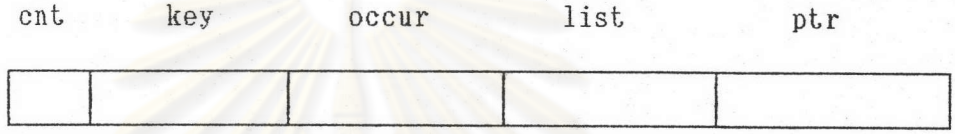
ในการประมวลผลแฟ้มข้อมูลนำเข้าแต่ละแฟ้ม โปรแกรมจะทำการสร้างชุดของแฟ้มข้อมูลที่เกี่ยวข้องขึ้นมา 1 ชุด ประกอบด้วย 2 แฟ้มข้อมูลคือ filename.DIC และ filename.INX โดยที่ filename คือชื่อของแฟ้มข้อมูลนำเข้า

2.2.1 **แฟ้มข้อมูล filename.DIC**

เป็นแฟ้มข้อมูลที่ประกอบด้วย 2 ส่วนย่อย คือ

2.2.1.1 **ส่วนของดิคชันนารี**

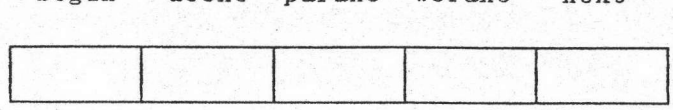
ดิคชันนารี เป็นส่วนที่ใช้สำหรับเก็บคำทุกคำที่เป็นคีย์ จำนวนซ้ำของคีย์ และตัวชี้รายการตำแหน่งของคีย์ ในการออกแบบได้ใช้โครงสร้างข้อมูลแบบบีทรี (B-Tree) ที่มีลำดับ 2M+1 โดยที่ M เป็นจำนวนคีย์ที่น้อยที่สุดที่สามารถมีได้ในแต่ละโหนด โครงสร้างของแต่ละโหนดประกอบด้วย



- cnt เป็นตัวนับจำนวนคีย์ที่มีในแต่ละโหนด
- key ประกอบด้วย เขตข้อมูลย่อยจำนวน 2M เขต แต่ละเขตข้อมูลย่อยจะเก็บค่าคีย์ ที่มีความยาวไม่เกิน 15 ตัวอักษร
- occur ประกอบด้วย เขตข้อมูลย่อยจำนวน 2M เขต แต่ละเขตข้อมูลย่อย จะเป็นตัวนับจำนวนซ้ำของคีย์แต่ละตัว ที่มีลำดับสอดคล้องกัน
- list ประกอบด้วย เขตข้อมูลย่อยจำนวน 2M เขต แต่ละเขตข้อมูลย่อย จะเป็นตัวชี้ตำแหน่งเริ่มต้น ของลิงค์ลิสในส่วนของข้อมูลผกผัน ที่มีลำดับสอดคล้องกับคีย์แต่ละตัว
- ptr ประกอบด้วย เขตข้อมูลย่อยจำนวน 2M+1 เขต แต่ละเขตข้อมูลย่อย จะเป็นตัวชี้ตำแหน่งของโหนดลูก ที่สัมพันธ์กับคีย์แต่ละตัว

2.2.1.2 **ส่วนของข้อมูลผกผัน**

ข้อมูลผกผัน เป็นส่วนที่ใช้สำหรับเก็บรายการตำแหน่งของค่าแต่ละค่าที่เป็นคีย์ ในการออกแบบ ได้ใช้โครงสร้างข้อมูลแบบลิงค์ลิส (linked list) ซึ่งแต่ละโหนดประกอบด้วย



- begin เป็นตำแหน่งเริ่มต้น ของเอกสาร หรือตำแหน่งเริ่มต้น ของตอนในเอกสาร ที่คีย์ปรากฏอยู่

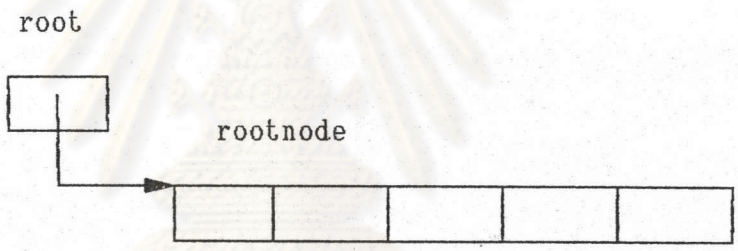
- docno เป็นหมายเลขเอกสารที่คีย์ปรากฏอยู่

- parano เป็นหมายเลขของตอนในแต่ละเอกสารที่คีย์ปรากฏอยู่

- wordno เป็นหมายเลขของคำที่เป็นคีย์ ในแต่ละตอนของเอกสาร

- next เป็นตัวชี้ตำแหน่งของโหนดถัดไปในลิงค์ลิส

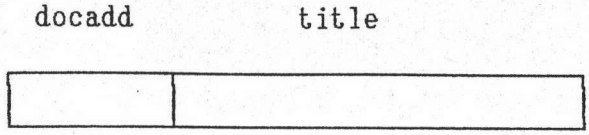
ระเบียนแรกของแฟ้มข้อมูล filename.DIC จะเก็บสารสนเทศพิเศษ เรียกว่า ราก (root) ซึ่งเป็นตัวชี้ตำแหน่งของโหนดราก (rootnode) ของบีทรี ดังนี้



ในการค้นหาคีย์ที่ต้องการ จะทำการค้นหาในแฟ้มข้อมูล filename.DIC โดยการอ่านระเบียนแรกของแฟ้มข้อมูล เพื่อนำค่าตัวชี้ตำแหน่งของโหนดราก มาใช้ในการเริ่มต้นค้นหาคีย์ในโหนดของบีทรี และค่าใน list ที่สัมพันธ์กับคีย์ที่ค้นหา จะเป็นตัวชี้รายการตำแหน่งของคีย์ในลิงส์ลิส จากนั้น ข้อความในแฟ้มข้อมูลข้อความ ที่มีตำแหน่งสอดคล้องกับค่าในลิงส์ลิส ก็จะถูกอ่านออกมา

2.2.2 แฟ้มข้อมูล filename.INX

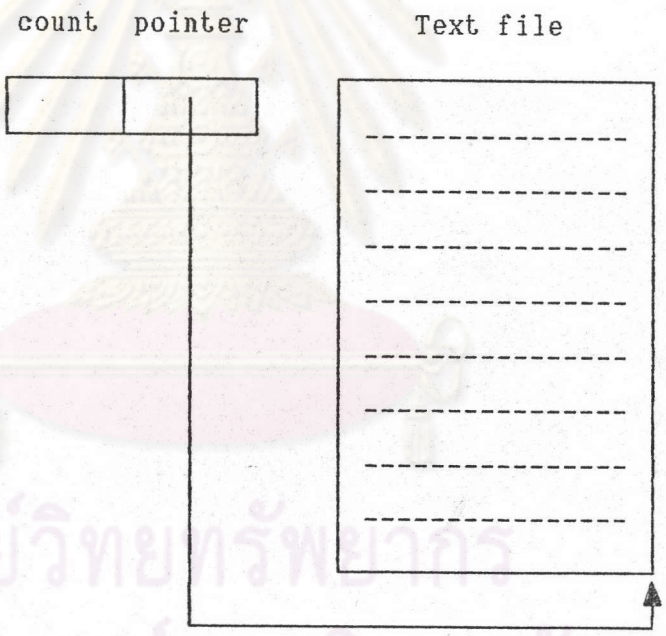
เป็นแฟ้มข้อมูลดัชนี ที่ใช้สำหรับเก็บตำแหน่งเริ่มต้นของแต่ละเอกสารในแฟ้มข้อมูลข้อความ ในการออกแบบได้ใช้โครงสร้างของแฟ้มข้อมูลแบบลำดับ แต่ละระเบียนประกอบด้วย



- docadd เป็นตำแหน่งเริ่มต้นของแต่ละเอกสาร ในแฟ้มข้อมูลข้อความ

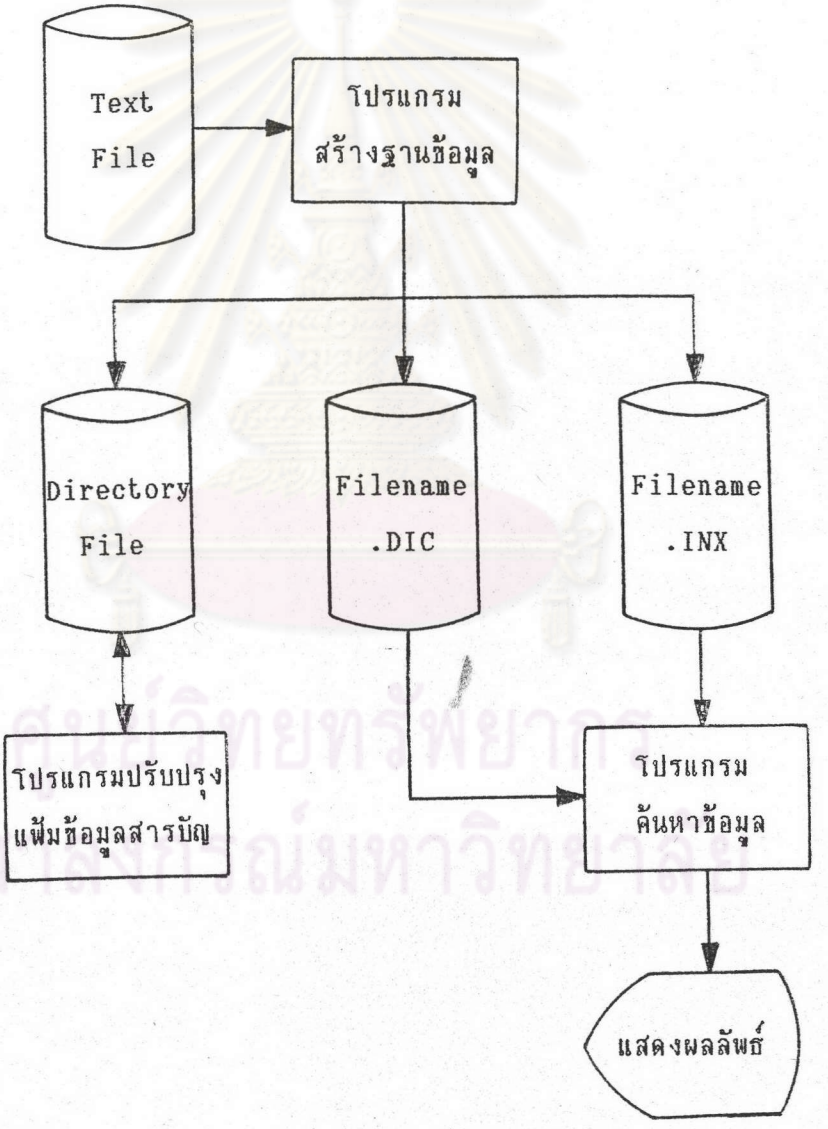
- title เป็นชื่อเรื่องของแต่ละเอกสาร มีความยาวไม่เกิน 80 ตัวอักษร

ระเบียนแรกของแฟ้มข้อมูลดัชนีจะเก็บสารสนเทศพิเศษ ซึ่งประกอบด้วยตัวนับ (count) จำนวนเอกสารในแฟ้มข้อมูลข้อความ และตัวชี้ (pointer) ตำแหน่งสุดท้ายของแฟ้มข้อมูลข้อความ เพื่อใช้ในการตรวจสอบแฟ้มข้อมูลข้อความ ในกรณีที่มีการเพิ่มข้อมูลลงไปในแฟ้มข้อมูลข้อความ ต่อท้ายข้อมูลเดิม เมื่อทำการประมวลผลใหม่อีกครั้งหนึ่ง จะทำการประมวลผลเฉพาะส่วนของข้อมูลที่เพิ่มเข้าไป และนับจำนวนของเอกสาร ต่อจากจำนวนเดิมที่มีอยู่แล้ว ดังนี้



รูปที่ 3.1ก และรูปที่ 3.1ข แสดงขั้นตอนการทำงานของระบบ (System Flowchart) โดยรูปที่ 3.1ก แสดงการเพิ่มข้อมูลในแฟ้มข้อมูลข้อความ และรูปที่ 3.1ข แสดงการสร้างฐานข้อมูล การค้นหาข้อมูล และการปรับปรุงแฟ้มข้อมูลสารบัญ

รูปที่ 3.1ก



รูปที่ 3.1ข

แผนภาพแสดงการทำงานของระบบ (System Flowchart)