โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ในประเทศไทย

(Internet Search Engine in Thailand)

โดย

น.ส.ชาลินี ชยุตมาพงศา

นายสมนึก พฤกษ์พิจารณ์

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง

ปีการศึกษา 2539

ปีการศึกษา 2539

โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ในประเทศไทย

โดย

นางสาวชาลินี ชยุตมาพงศา 36014115

นายสมนึก พฤกษ์พิจารณ์ 36014456

อาจารย์ที่ปรึกษา

อ.อภิเนตร อูนากูล

ปริญญานิพนธ์ปีการศึกษา 2539

ภาควิชา วิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง หุ่นยนต์ค้นหาข้อมูลบนระบบเครือข่ายคอมพิวเตอร์ในประเทศไทย

ผู้จัดทำ

1. นางสาวชาลินี ชยุตมาพงศา
2. นายสมนึก พฤกษ์พิจารณ์

..... อาจารย์ที่ปรึกษา

(.....)

โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ในประเทศไทย

ชาลินี ชยุตมาพงศา
สมนึก พฤกษณ์พิจารณ์

อาจารย์ อภิเนตร อูนากูล
ปีการศึกษา 2539

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้ เรียบเรียงขึ้นตามรายละเอียดของโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ ซึ่งเป็นโครงงานที่จัดทำขึ้นเพื่อพัฒนาแนวทางในการค้นหาข้อมูลบน เครือข่ายคอมพิวเตอร์ โดยมุ่งเน้นพัฒนาการค้นหาข้อมูลที่มีขอบเขตอยู่ภายในประเทศไทย อันจะมีส่วนทำให้การค้นหาข้อมูลภายในประเทศไทยสามารถทำได้สะดวกขึ้น โครงสร้างของ โครงงานนี้แบ่งออกได้เป็น 3 ส่วน คือ ส่วนแรกคือส่วนดึงข้อมูลและวิเคราะห์ข้อมูล ซึ่งจะเป็น ส่วนที่ทำงานโดยโรบอท (Robot) ซึ่งเป็นเอเจนท์ (Agent) ประเภทหนึ่งบนเครือข่าย คอมพิวเตอร์ ในการดึงข้อมูลจากเว็บไซท์ (Website) ต่างๆ ที่อยู่ในประเทศไทยแล้วนำข้อมูลที่ ได้มาวิเคราะห์จัดให้อยู่ในรูปแบบที่เหมาะสม ส่วนที่สองคือส่วนที่ทำการติดต่อระหว่างผู้ใช้ บริการกับเซิร์พเวอร์ การทำงานในส่วนนี้จะทำผ่าน โปรแกรมที่เรียกว่า โปรแกรมซีจีไอมีหน้าที่ ในการส่ง และ รับข้อมูลระหว่างผู้ใช้บริการและเซิร์พเวอร์ส่วนสุดท้ายคือส่วนจัดเก็บและค้นหา ข้อมูลซึ่งในส่วนนี้จะเป็นส่วนที่จัดเก็บข้อมูลที่ ได้จากโรบอทในรูปแบบที่พร้อมที่จะทำการค้นหา ตามเงื่อนไขที่ผู้ใช้บริการต้องการ

INTERNET SEARCH ENGINE IN THAILAND

Chalinee Chayutamapongsa

Somnuek Phurkpijarn

Aprinetr Unakul

1996

Abstract

This thesis cover all information about Internet Search Engine in Thailand Project that has a purpose to develope performance of finding information espectially about Thailand . The structure of this project has three parts , First part is Search and Analyse Information that has important responsibility to find the information from website in Thailand and later covert these information into proper form . Second part is Communication between User and Server , this part is use to pass and receive data between user and server via program called cgi . Data passing in this part are words for search and options in searching . The last part of this project is Database and Query Information , Database in this search engine is designed by using ER-Model and using in Interbase server in Delphi to create database system , so server can query data in demand of user from Interbase server .

สารบัญ

สารบัญรูปภาพและตาราง

สารบัญรูปภาพและตาราง

บทที่ 1
โครงงานหุ่นยนต์ค้นหาข้อมูล
บนเครือข่ายคอมพิวเตอร์

บทนำ

ในปัจจุบันนี้ การใช้งานระบบเครือข่ายอินเทอร์เน็ตได้รับความนิยมอย่างแพร่หลายภายในระยะเวลาอันรวดเร็ว สาเหตุ เนื่องมาจากระบบเครือข่ายอินเทอร์เน็ต กลายเป็นแหล่งข้อมูลขนาดใหญ่ สำหรับผู้ที่เข้ามาใช้บริการระบบเครือข่ายอินเทอร์เน็ต ทั้งนี้เพราะแต่ละไซท์จัดตั้งขึ้นเพื่อให้บริการข้อมูล ตามจุดประสงค์ที่แตกต่างกัน โดยอาจจะเป็นแหล่งที่ให้ข้อมูลเกี่ยวกับวิชาการ กีฬา ความบันเทิง การเมือง เศรษฐกิจ ไปจนถึงการโฆษณาขายสินค้า แต่การที่ค้นหาข้อมูลบนระบบเครือข่ายอินเทอร์เน็ตก็ไม่อาจทำได้ง่าย เพราะผู้ใช้บริการไม่สามารถทราบ ได้ว่าข้อมูลที่ต้องการอยู่ที่ใด และเป็นข้อมูลที่ถูกต้องเชื่อถือได้มากน้อยแค่ไหน (ข้อมูลข่าวสารจะมีประโยชน์ได้ก็ต่อเมื่อข้อมูลนั้นเป็นข้อมูลที่ทันต่อเหตุการณ์และมีความถูกต้อง)

ต่อมาเมื่อปัญหาที่เกิดขึ้นในการค้นหาข้อมูลเพิ่มมากขึ้น จึงมีนักศึกษา 2 คนในสาขาวิศวกรรมไฟฟ้า ชื่อ David Filo และ Jerry Yang ที่มหาวิทยาลัย Standford เริ่มมีแนวคิดที่จะทำการจัดหมวดหมู่ของข้อมูลที่มีอยู่มากมายขึ้นเป็นระบบ อันเป็นที่มาของโครงงานที่มีชื่อว่า David and Jerry's Guide to the Web ต่อมาก็มี แนวคิดที่จะเปลี่ยนชื่อเพื่อความเหมาะสมจึงได้มีการเปลี่ยนชื่อ ระบบจากเดิมเป็น Yahoo (ปัจจุบันเป็นหนึ่งในหลายเว็บไซท์ที่ได้รับความนิยมอย่างสูงจากผู้ใช้บริการบนระบบเครือข่ายอินเทอร์เน็ต) แต่ถึงแม้ว่าจะได้มีการพัฒนาระบบที่ใช้ในการค้นหาข้อมูลเพิ่มขึ้นอีกหลายเว็บไซท์ เพื่อ เป็นการอำนวยความสะดวกในการค้นหาข้อมูล และสามารถค้นหาข้อมูลได้มากที่สุด ซึ่งจากการ ที่ระบบค้นหาข้อมูลต่างๆ สามารถที่จะทำการค้นหาข้อมูลได้เป็นจำนวนมาก ก็ทำให้เกิดปัญหา จากการใช้งานระบบค้นหาข้อมูลเหล่านี้ เนื่องจากในการทำการค้นหาข้อมูลใดๆ ก็ตามมักจะได้ ผลลัพธ์ออกมาเป็นเว็บไซท์จำนวนมากซึ่งเป็นการยากที่ผู้ใช้บริการจะสามารถเข้าไปศึกษาหรือนำ ไปใช้งานได้หมดหรือเป็นการยากที่จะทำการเลือกว่าจะเอาข้อมูลจากเว็บไซท์ใด

ปัญหาที่เกิดจากการที่มีผลลัพธ์ของการค้นหาข้อมูลเป็นจำนวนมาก ก็พบในกรณีที่ผู้ใช้บริการทำการค้นหาข้อมูลที่เกี่ยวกับประเทศไทยโดยเน้นค้นหาจากเว็บไซท์ที่ตั้งอยู่ในประเทศไทย ทางผู้จัดทำโครงงานได้เห็นความสำคัญของปัญหาเกิดขึ้น จึงมีแนวคิดที่จะทำหุ่นยนต์ค้นหาข้อมูลบนระบบเครือข่ายอินเทอร์เน็ตขึ้นมาเอง โดยมีขอบเขตของข้อมูลที่รวบรวมว่าเป็นข้อมูลของ เว็บ

ไซท์ที่ตั้งอยู่ในประเทศไทย เพื่อช่วยอำนวยความสะดวกต่อผู้ใช้บริการที่ต้องการเข้ามาหาข้อมูลที่เกี่ยวกับประเทศไทย อันอาจจะเป็นการช่วยส่งเสริมข้อมูลที่เกี่ยวกับประเทศไทย ทางด้านต่างๆ เช่น การท่องเที่ยวของประเทศ เทคโนโลยี วัฒนธรรม ศาสนา หรืออาจมีส่วนช่วยสร้างภาพพจน์ที่ดีให้กับประเทศไทยได้

นอกจากนี้โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์นี้ยังมีทฤษฎีและการทำงานที่ตรงต่อความสนใจของผู้จัดทำโครงงาน ผู้จัดทำโครงงานจึงได้เริ่มต้นทำการศึกษา ทฤษฎีต่าง ๆที่เกี่ยวข้องกับโครงงาน เช่น การทำงานของระบบอินเตอร์เน็ต คุณสมบัติและการ ทำงานของโรบอท การจัดเก็บข้อมูลบนระบบฐานข้อมูล เป็นต้น

1.1 วัตถุประสงค์ของโครงงาน

1 สามารถจัดทำฐานข้อมูลที่รวบรวมข้อมูลบนเครือข่ายอินเทอร์เน็ตของเว็บไซท์ที่ตั้งอยู่ในประเทศไทยไว้เป็นหมวดหมู่ได้

2 สามารถคำนวณและจัดลำดับความสำคัญของคำที่เก็บในฐานข้อมูลได้

3 สามารถจัดทำหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายอินเทอร์เน็ตของเว็บไซท์ ที่ตั้งอยู่ในประเทศไทยได้ตามเงื่อนไขที่กำหนด

4 ทำการศึกษาทฤษฎีต่างๆ ที่เกี่ยวข้องกับการทำงานของหุ่นยนต์ค้นหาข้อมูลและสามารถนำมาประยุกต์ใช้งานได้จริง

1.2 ขอบเขตการทำงานของโครงงาน

โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ เป็นโครงงานที่จัดทำการรวบรวมข้อมูลต่างๆ ที่อยู่บนเว็บไซท์ ที่ตั้งอยู่ภายในประเทศไทย ทั้งนี้เพื่อเป็นการอำนวยความสะดวกต่อผู้ใช้บริการที่ต้องการ ทราบข้อมูลที่เกี่ยวกับประเทศไทย

ในส่วนของการทำงานของหุ่นยนต์จะทำการวิ่งไปตามเว็บไซท์ ต่างๆ ทำการตรวจสอบว่าอยู่ในขอบเขตที่กำหนดหรือไม่ (เฉพาะเว็บไซท์ที่อยู่ในประเทศไทย) เมื่อตรวจสอบพบว่าอยู่ในขอบเขตที่กำหนดไว้ ก็จะทำการดึงข้อมูลกลับมา ข้อมูลที่ถูกดึงกลับมาจะถูกนำมาวิเคราะห์และเก็บ เฉพาะคำที่มีความสำคัญตามเงื่อนไขที่โครงงานกำหนด

หลังจากนั้นทำการสร้างส่วนที่ใช้ในการค้นหาข้อมูลบนระบบฐานข้อมูล ที่ใช้จัดเก็บข้อมูล โดยมีการกำหนดเงื่อนไขในการค้นหา เช่น รูปแบบตัวอักษรเหมือนต้นแบบ เป็นต้นเมื่อทำการค้นหาข้อมูลเรียบร้อยแล้ว ก็ทำการแสดงผลลัพธ์ออกมาทางหน้าจอ ในรูปแบบที่ผู้ใช้งานสามารถทำงานได้

การออกแบบการทำงานของส่วนต่างๆของโครงงาน จะทำการออกแบบจาก ทฤษฎีที่เกี่ยวข้องโดยจะทำการเลือก วิธีที่เหมาะสมต่อการทำงานในระดับการศึกษามากที่สุด ทั้งนี้เนื่องจาก การมีข้อจำกัด ทางด้าน งบประมาณ ทรัพยากรเครื่องคอมพิวเตอร์และอุปกรณ์เสริมต่างๆ ที่มีอยู่อย่างจำกัด ทฤษฎี ที่เกี่ยวข้องกับโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่าย คอมพิวเตอร์ นั้นสามารถแบ่งออกได้เป็นหลายส่วน

1. ทฤษฎีที่เกี่ยวข้องกับส่วนค้นหาและวิเคราะห์ข้อมูล อันได้แก่ ทฤษฎีเกี่ยวกับ การทำงานของเอเจนท์ , โรบอท , การวิเคราะห์คำและคำนวณค่าความสำคัญของคำ

2. ทฤษฎีที่เกี่ยวข้องกับส่วนที่ใช้ติดต่อระหว่างผู้ใช้บริการและระบบจัดเก็บข้อมูลอันได้แก่ทฤษฎีเกี่ยวกับการเขียนโปรแกรมเชื่อมต่อ (โปรแกรมซีจีไอ) รวมทั้งภาษาที่ใช้เขียนโปรแกรมซีจีไอ

3. ทฤษฎีที่เกี่ยวข้องกับส่วนจัดเก็บและค้นหาข้อมูล อันได้แก่ ทฤษฎีเกี่ยวกับ ระบบฐานข้อมูล การออกแบบฐานข้อมูล โปรแกรมที่ใช้สร้างระบบฐานข้อมูล รวมทั้งภาษาที่ใช้ในการดึงข้อมูล

4. ทฤษฎีที่เกี่ยวข้องกับการรับและแสดงผลลัพธ์ออกทางหน้าจอผู้ใช้งาน อันได้แก่ การศึกษาทฤษฎีเกี่ยวกับภาษาเอ็ชทีเอ็มแอล (HTML)

1.3 ผลที่คาดจะได้รับ

ในการจัดทำโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ ทางผู้จัดทำโครงงานก็มีความหวังที่จะได้รับความรู้เกี่ยวกับที่ทฤษฎีและหลักการที่สนใจ รวมทั้งคาดหมายถึงประสิทธิภาพในการทำงานของโครงงานที่สร้างขึ้น ดังต่อไปนี้

1. สามารถทำความเข้าใจทฤษฎีเกี่ยวกับการทำงานของโรบอทบนเครือข่ายคอมพิวเตอร์
2. สามารถสร้างโรบอทขึ้นมาทำงานตามวัตถุประสงค์ที่ตั้งไว้ได้จริง
3. สามารถออกแบบฐานข้อมูลเพื่อใช้ในการจัดเก็บข้อมูลได้
4. สามารถทำการจัดสร้างระบบฐานข้อมูลที่ทำงานได้ตามวัตถุประสงค์
5. สามารถเขียนโปรแกรมที่ทำหน้าที่ในการติดต่อระหว่างผู้ใช้งานบนระบบเครือข่ายคอมพิวเตอร์ กับ ระบบฐานข้อมูลที่จัดสร้างไว้ได้
6. โครงงานที่เสร็จสมบูรณ์แล้ว สามารถที่จะทำการค้นหาข้อมูลตามที่กำหนดขอบเขต ไว้ได้จริง

1.4 การประเมินผลโครงงาน

เมื่อทำโครงงานหุ่นยนต์ค้นหาข้อมูลเสร็จสิ้นแล้ว ก็จำเป็นที่จะต้องทำประเมินผลเพื่อเป็นการวัดประสิทธิภาพของงาน , รวบรวมปัญหาต่าง ๆที่เกิดขึ้นในการทำงาน เพื่อเป็นแนวทางในการ พัฒนาประสิทธิภาพของโครงงานให้ดียิ่งขึ้นไปในอนาคต

หัวข้อที่ใช้ในการประเมินผลโครงงาน

1. ประสิทธิภาพในการดึงข้อมูลและวิเคราะห์ข้อมูลของโรบอท ซึ่งเป็นการพิจารณาทั้งในแง่ของความเร็วในการทำงาน ความถูกต้องของข้อมูล และความสามารถในการครอบคลุมข้อมูลตามขอบเขตที่กำหนด

2. ความเหมาะสมของการออกแบบระบบฐานข้อมูล รวมทั้งการเลือกใช้ระบบฐานข้อมูลจริงให้เหมาะสมกับการทำงานของโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์

3.ประสิทธิภาพในการค้นหาข้อมูลตามที่ผู้ใช้บริการต้องการ ซึ่งเป็นการพิจารณาทั้งในแง่ของความเร็วและความถูกต้อง

4. ความสวยงามของหน้าจอที่ใช้ติดต่อกับผู้ใช้งานรวมทั้งความยากง่ายในการใช้งาน

บทที่ 2

ทฤษฎีและหลักการ

ในการจัดทำโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ ทางผู้วิจัยได้ทำการศึกษาทฤษฎีต่างๆ ที่มีความเกี่ยวข้องกับโครงงาน เช่น โรบอท , ซีจีไอ , ระบบฐานข้อมูล ในแง่ของหลักการ , การทำงาน เพื่อนำมาใช้อ้างอิง เปรียบเทียบข้อดีข้อเสียในขั้นตอนของการออกแบบ นอกจากนี้ยังได้ทำการศึกษา ภาษาเดลไฟในส่วนที่เกี่ยวข้องกับโครงงาน เพื่อที่จะนำมาใช้ใน การเขียนโปรแกรม

ระบบค้นหาข้อมูล (Search Engine)

ระบบค้นหาข้อมูล คือ ระบบค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ (อินเทอร์เน็ต) ที่ทำการรวบรวมข้อมูลจากเว็บไซท์ต่างๆ นำข้อมูลที่รวบรวมได้มาทำการวิเคราะห์เลือกข้อมูลที่มีความสำคัญ นำมามาจัดเก็บในรูปแบบที่เหมาะสม พร้อมที่จะทำการค้นหาข้อมูลตามเงื่อนไข ที่กำหนด

ระบบค้นหาข้อมูล ประกอบด้วย 4 ส่วนหลักดังนี้

2.1 ส่วนดึงข้อมูลและวิเคราะห์ข้อมูล

2.2 ส่วนเชื่อมต่อระหว่างผู้ใช้บริการและส่วนจัดเก็บข้อมูล

2.3 ส่วนจัดเก็บและค้นหาข้อมูล

2.4 ส่วนหน้าจอรับข้อมูลและแสดงผล

2.1 ส่วนดึงข้อมูลและวิเคราะห์ข้อมูล

ส่วนดึงข้อมูลและวิเคราะห์ข้อมูล เป็นส่วนของเซิร์สเอเจนท์ ที่ทำหน้าที่ต่อไปนี้

1. ทำการท่องไปบนเว็บเพจต่างๆ
2. ตรวจสอบเว็บเพจว่าอยู่ในขอบเขตที่กำหนดหรือไม่
3. ทำการดึงข้อมูลกลับมา
4. ทำการคำนวณและวิเคราะห์ข้อมูลที่ดึงกลับมา

ส่วนดึงข้อมูลและวิเคราะห์ข้อมูล จะเป็นการทำงานของโปรแกรมประยุกต์ที่เรียกว่า โรบอท ซึ่งเป็นประเภทหนึ่งของเอเจนท์ ดังนั้นจึงควรศึกษาถึงทฤษฎีและหลักการของเอเจนท์ประเภทโรบอท

2.I.1 เอเจนท์ (Agent)

เอเจนท์ คือโปรแกรมคอมพิวเตอร์ที่จัดสร้างความสัมพันธ์ระหว่างมนุษย์โดยจะทำหน้าที่เปรียบเสมือนคนอีกคนหนึ่งที่ทำงานแทนเรา

ประเภทของเอเจนท์

เอเจนท์สามารถแบ่งเป็นประเภทได้ 4 ประเภท ตามที่มีอยู่บนอินเตอร์เน็ต

1. เว็บโรบอท (Web Robot),สไปเดอร์ (Spider),วันเดอเรอร์ (Wanderer)

เป็นโปรแกรมที่ท่องไปในตามข้อมูลใน เวิร์ลไวด์เว็บ (www)โปรแกรมเหล่านี้จะย้ายจากเว็บ (Web) หนึ่งไปยังอีกเว็บหนึ่งโดยอาศัยลักษณะของไฮเปอร์ลิงค์ (Hyper Link) ไม่มีอยู่ใน เว็บเพจ เว็บโรบอทเหล่านี้จะอาศัย เอ็ชทีทีพี โปรโตคอล (HTTP Protocol) ของเวิร์ลไวด์เว็บใน การที่จะดึงเอกสารจากเซิร์พเวอร์ (Server) ต่าง ๆ โปรแกรมเหล่านี้ค่อย ๆ ผ่านไปยังเว็บต่าง ๆ เพื่อค้นหาแหล่งข้อมูลใหม่ ๆ เพื่อนำมาจัดทำดรรชนีในการค้นหาข้อมูล

2. เว็บคอมเมอร์ส (Web Commerce Agent)

เป็นโรบอทซึ่งจะทำหน้าที่ช่วยในการตัดสินใจเลือกบริการต่างๆ หรือซื้อสินค้าบน ระบบออนไลท์ (On-line) โดยทำการเลือกสิ่งที่ดีที่สุด เหมาะสมที่สุดจะมีเอกสารและรายละเอียดสินค้าและบริการต่าง ๆ ของบริษัทและร้านค้าต่าง ๆ เอเจนท์จะทำหน้าที่ติดต่อเสมือนพ่อค้าคนกลางในการเลือกใช้บริการหรือซื้อสินค้าบนระบบออนไลน์นี้

3. วอร์ม (Worm),ไวรัส (Virus)

เป็นเอเจนท์ที่มีเจตนาไม่ดีและไม่เป็นที่ปรารถนาวอร์มและไวรัสจะจำลองตัวเองใน รูปแบบที่หลีกเลี่ยงจากการตรวจสอบจากเครื่องหนึ่งไปยังอีกเครื่องหนึ่งจากระบบเครือข่ายหนึ่งไปยังอีกเครือข่ายหนึ่งในสมัยก่อนมักจะพบโปรแกรมเหล่านี้บนดิสก์แต่ทุกวันนี้แม้แต่ในอินเตอร์เน็ต ก็ไม่อาจรอดพ้นจากโปรแกรมเหล่านี้ ซึ่งโปรแกรมเหล่านี้มักจะนำความเสียหาย มาสู่ระบบและข้อมูลต่าง ๆ

4. เอ็มยูดีเอเจนท์ (Mud Agent),แชตเตอร์บอท (Chatter Bot)

เอ็มยูดีเอเจนท์ เป็นโปรแกรมที่มีประโยชน์ในการให้บริการต่าง ๆ สำหรับมนุษย์ เช่น ถามที่ผู้ใช้ต้องการหรือให้คำแนะนำ โดยผ่านทางยูสเซอร์อินเตอร์เฟส (User Interface) ที่ใช้ภาษาเฉพาะ ส่วนแชตเตอร์บอท เป็นเอเจนท์ที่สามารถสนทนา โดยหน้าที่หลักคือ สื่อสารกับคนที่ใช้โปรแกรมอยู่โดย แชตเตอร์บอท ไม่เหมือนกับเอ็มยูดีเอเจนท์ตรงที่ แชตเตอร์บอท จะกว้างกว่าเอ็มยูดีเอเจนท์ที่เป็นเฉพาะทางมากกว่า

เอเจนท์มีอยู่หลายประเภท แต่เอเจนท์ที่นำมาใช้ในโครงงานหุ่นยนต์ค้นหาข้อมูล คือ โรบอท ประเภทเว็บโรบอทดังนั้นทางผู้วิจัยจึงทำการศึกษาคุณสมบัติ,การทำงาน รวมทั้งข้อกำหนดต่างๆ ที่เกี่ยวข้องกับการทำงานของโรบอทประเภทเว็บโรบอท

2.1.2 คุณสมบัติและการทำงานของเว็บโรบอท

การนำเว็บโรบอท ไปใช้งาน

เว็บโรบอทในปัจจุบันได้มีการนำมาใช้งานกันอย่างแพร่หลายมาก โดยในระยะเริ่มแรกนั้น ได้มีการนำมาใช้เพื่อตรวจวัด อัตราการเจริญเติบโตของอินเตอร์เน็ท โดยใช้ โรบอทที่สร้างขึ้นทำการนับจำนวนเว็บไซท์ (Web Site) ที่เพิ่มขึ้นมา ต่อมาได้มีการนำโรบอทมาใช้งานทางด้านอื่นๆ มากขึ้น

เราสามารถแบ่งการใช้งานเว็บโรบอทออกได้เป็น 3 อย่าง

1. เว็บรีซอร์สดิสโคเวอรี่ (Web Resource Discovery)
2. เว็บเมนเทนแนนซ์ (Web Maintenance)
3. เว็บไมร์เรอริ่ง (Web Mirroring)

เว็บรีซอร์สดิสโคเวอรี่

เป็นตัวที่เข้ามาเกี่ยวข้องกับงานในการแก้ไขปัญหาในการทำการค้นหาข้อมูลอันเป็นประโยชน์ภายในเว็บที่มีอยู่เป็นจำนวนมาก เนื่องจากค่าใช้จ่าย , การเป็นข้อมูลแบบกระจายการเปลี่ยนแปลงที่เกิดขึ้นบ่อยๆ และลักษณะเฉพาะของเว็บทำให้ผู้ที่เข้ามาใช้งานมีความสนุกสนานมาก แต่ก็ต้องได้รับความลำบากอย่างมากในการที่จะต้องทำการค้นหาข้อมูล เนื่องจากจะทำการค้นหาได้ยาก

ดังนั้นจึงได้มีการคิดค้นการทำ เซิร์สเอ็นจินขึ้นมากเพื่อใช้ในการช่วย หาข้อมูล ซึ่ง เซิร์สเอ็นจิน นี้ก็จำเป็นที่จะต้องใช้ โรบอทเข้ามาช่วย ซึ่งโดยทั่วไปแล้วมักจะเรียกว่า สไปเดอร์ หรือเว็บโรบอท ซึ่งจะเป็นโปรแกรมที่ทำการท่องไปตามเว็บต่างๆ อย่างอัตโนมัติ แล้วนำข้อมูล กลับมาทำเป็นดรรชนีเพื่อใช้ในการค้นหา

เว็บเมนเทนแนนซ์

สิ่งหนึ่งซึ่งเป็นการยากในการที่จะดูแลข้อมูลบนเว็บ เนื่องจากต้องมีการทำการ ตรวจ สอบอยู่ตลอดเวลาว่าเว็บ ที่เราทำการเชื่อมต่อไปนั้น ได้มีการเปลี่ยนแปลงข้อมูลหรือมีการลบ ทิ้งไปหรือไม่ ซึ่งหากมีการทำการเคลื่อนย้ายหรือทำการลบเว็บที่มีการเชื่อมต่อ ถึงทิ้งไปก็อาจ ทำให้เกิดกรณีของการขาดการเชื่อมต่อ(Dead Link)

ในปัจจุบันนี้ยังไม่มีการทำการตรวจสอบ เรื่องดังกล่าวโดยอาศัยการทำงานแบบอัตโนมัติ ดังนั้นบางเซิร์พเวอร์ จะหยุดการทำงาน เอ็ชทีทีพี รีเควส (HTTP Request) เนื่องจากการเกิดการขาดการเชื่อมต่อขึ้นที่ ยูอาร์แอล (URL)ตัวแรกที่เราจะทำการติดต่อคือในขณะที่มี การคืนค่ากลับมาด้วย ดังนั้นข้อมูลที่อยู่ภายในล็อกไฟล์ของ เซิร์พเวอร์ ก็สามารถที่จะนำมาใช้ในการบอกว่ามีรายการของเว็บใดบ้างที่มีการขาดการเชื่อมต่ออยู่แต่วิธีดังกล่าวนี้ก็ไม่ใช่วิธีที่ดี เนื่องจากเว็บต่างๆ ค่อนข้างที่จะมีความเป็นอิสระ จึงเป็นการยากที่จะทำการตรวจสอบการติดต่อระหว่างเว็บไซท์ต่างๆ ได้ทั้งหมดจากการใช้ โรบอทในการทำการตรวจสอบ

ดังนั้นการทำงานทางด้านนี้จึงมักเป็นที่นิยมในการที่จะใช้งานโรบอทในประเภท ที่เรียกว่า เว็บเมนเทนแนนท์ สไปเดอร์ ซึ่งจะเป็นโปรแกรมที่ทำหน้าที่ช่วยผู้ที่เป็นเจ้าของเว็บ และ เว็บมาสเตอร์ (Web Master) ในการที่จะทำการดูแลรักษา โครงสร้างต่างๆของเว็บโดยจะทำ การท่องโดยอัตโนมัติไปตามการเชื่อมต่อ ต่างๆ ที่เกี่ยวข้อง แล้วทำการตรวจสอบหา การขาด การเชื่อมต่อว่ามีปรากฏหรือไม่

เว็บไมร์เรอริ่ง

ไมร์เรอริ่ง เป็นเทคนิคทั่วไปที่ใช้การที่จะทำการติดตั้ง สำหรับโครงสร้างของข้อมูล ตัวอย่างเช่น ไมร์เรอริง เอฟทีพี ไซท์ (Mirroring FTP Site) ที่เกี่ยวข้องกับการทำการคัดลอกเอฟทีพีไฟล์ ทั้งหมดโดยอาศัยการทำงานแบบการทำซ้ำ (Recursive) และทำการสร้างตัวของ มันเองขึ้นมาอีกในเครื่องอีกเครื่องหนึ่งภายในระบบเครือข่ายเอฟทีพีไซท์ ที่เป็นที่รู้จักกันโดย ทั่วไปมักจะถูกทำการจำลอง (Mirror)ไปยังส่วนต่างๆของระบบเครือข่าย ซ้ำแล้วซ้ำเล่าเพื่อทำ การร่วมใช้ข้อมูลไปยังส่วนที่ต้องการ แต่ก็อาจเป็นการเสี่ยงที่จะทำให้เกิดการหยุดการทำงาน ได้ง่าย

โรบอทที่ทำหน้าที่ตามหน้าที่ของเว็บไมร์เรอริ่ง นั้นในปัจจุบันมีอยู่แล้วซึ่งจะเป็นโปรแกรม ที่ทำการคัดลอกได้ดีกว่าตัวที่มีอยู่แล้วในการที่จะทำการคัดลอก ข้อมูลต่างๆ จาก เอฟทีพีไซท์แต่อย่างไรก็ตามโรบอท ที่ทำหน้าที่เป็นเว็บเมนเทนแนนซ์ ก็ไม่สามารถที่จะทำการ ตรวจสอบการเปลี่ยนแปลง ข้อมูลต่างๆบนเว็บไซท์ ได้

2.1.3 กฎสี่ข้อของเว็บโรบอท

ในการที่จะทำการสร้างโรบอท ขึ้นมาใช้งานสักตัวหนึ่งเราควรที่จะทำตามข้อกำหนดต่างๆ ต่อไปนี้ ซึ่งได้รวบรวมข้อควรระวังและข้อควรปฏิบัติสำหรับ โรบอทโดยทำการรวบรวมข้อมูลโดยเว็บมาสเตอร์ ที่มีความชำนาญและผู้เชี่ยวชาญทางด้านโรบอท ซึ่งกฎทั้ง 4 ข้อมีดังตาราง ต่อไปนี้

1. เว็บโรบอทต้องประกาศตัวเองอย่างชัดเจน
2. เว็บโรบอทต้องปฏิบัติตามข้อกำหนดมาตรฐาน
3. เว็บโรบอทต้องไม่ยึดครองแหล่งข้อมูลแบบถาวร
4. เว็บโรบอทต้องแสดงข้อผิดพลาดเมื่อเกิดความผิดพลาด

1. เว็บโรบอทต้องแสดงตัวเองอย่างชัดเจน

เว็บมาสเตอร์ ผู้ที่ทำการดูแลเซิร์พเวอร์ต้องการที่จะรู้ว่าโรบอทที่เข้ามาติดต่อด้วยนี้ มีหน้าที่ในการทำการเข้าถึงข้อมูลแบบใด และใครเป็นผู้เขียนเพื่อในกรณีที่เกิดปัญหาขึ้นจะได้ สามารถติดต่อกลับไปได้ และนอกจากเว็บมาสเตอร์มักจะอยากรู้ว่า โรบอท เหล่านี้รู้จักไซท์ ของพวกเขาได้อย่างไร ดังนั้น เว็บโรบอท สามารถที่จะทำการบอกแก่ เว็บมาสเตอร์ ได้โดยการ กำหนดค่าต่างๆในฟิลด์ต่อไปนี้ ยูสเซอร์เอเจนท์ ฟิลด์ , ฟอร์มฟิลด์และเว็บเพจ

การแสดงตัวของเว็บโรบอท

เว็บไคลเอนท์สามารถที่จะกำหนดในการแสดงค่าเกี่ยวกับตัวเองได้โดยใช้ ยูสเซอร์เอเจนท์ ฟิลด์ ที่สนับสนุนในเอ็ชทีทีพี รีเควส อย่างเช่นภายในเน็ตสเคพ (netscape) สามารถที่จะเรียกใช้ โมซิลลา (Mozilla) ที่มีอยู่ของมันเองได้ดังนี้

User-agent : Mozilla /1 .1 N

ดังนั้น เว็บโรบอทสามารถที่จะใช้ฟิลด์ ยูสเซอร์- เอเจนท์ ในการที่จะแสดงชื่อและบอกเกี่ยวกับเวอร์ชั่นต่างๆ ดังตัวอย่าง

User-agent : Terminator/1.0

ซึ่ง ยูสเซอร์-เอเจนท์ ฟิลด์ นี้จะช่วยให้เว็บมาสเตอร์ ในการกำหนดให้ เว็บโรบอท ให้สามารถถูกใช้งานได้โดยคนที่ใช้งานเว็บบราวเซอร์

การแสดงคำสั่งการทำงานของโรบอท

เอ็ชทีทีพี จะมี ฟิลด์ฟอร์มอยู่ภายในส่วนของรีเควสเฮดเดอร์ เพื่อจะใช้ในการกำหนด หน้าที่ต่างๆที่ได้รับมา ซึ่งโดยทั่วไปมักจะใช้อีเมลแอดเดรส (e-mail address) ในการกำหนด ค่าที่ฟิลด์นี้ ดังตัวอย่างต่อไปนี้

Form: joe.robomas@roboland.com

ซึ่งค่าดังกล่าวนี้จะช่วยให้ เว็บมาสเตอร์ สามารถติดต่อกับผู้ที่ใช้งาน โรบอทตัวดังกล่าวอยู่ได้ในกรณีที่เกิดปัญหาขึ้น นั่นทำให้ผู้ที่สร้างโรบอทหรือใช้งานโรบอทอยู่สามารถที่จะติดต่อกลับไปยังเว็บมาสเตอร์ได้

การอ้างอิงเว็บเพจ

เว็บมาสเตอร์ มักจะแปลกใจว่า คนโดยทั่วไปสามารถเข้ามาเรียนรู้เกี่ยวกับเว็บไซท์ของพวกเขาได้อย่างไร เมื่อโรบอท มีการเข้าถึงข้อมูลใน เว็บเพจ ต่างๆมันจะเป็นการดีมากสำหรับเว็บโรบอทที่จะทำการบอกแก่เว็บ เซิร์พเวอร์ ซึ่งเป็นผู้ดูแลเอกสารดังกล่าวอยู่ซึ่งเอกสาร นี้เราเรียกว่า เว็บเพจ รีเฟอเรน เอ็ชทีทีพี จะมี ฟิลด์ ที่ชื่อ รีเฟอเรน ฟิลด์ สำหรับใช้ในการกำหนดเอกสาร ซึ่งข้อมูลเหล่านี้จะแจ้งแก่ เว็บมาสเตอร์ ให้รู้ว่าเว็บเพจที่กำลังถูกดึงข้อมูลอยู่นี้ได้รับ การรีเฟอเรนมาจากตัวอย่าง

referer :HTTP://WWW.referRus.com/launchpad.html

2. เว็บโรบอทต้องปฏิบัติตามข้อตกลง

มาตราฐานสำหรับโรบอท เอ็กซ์คลูชั่น นี้ถูกกำหนดขึ้นโดย มาร์ติน คอสเตอร์ ในเรื่องที่เกี่ยวกับการติดต่อระหว่าง เว็บเซิร์พเวอร์ กับ เว็บโรบอท ซึ่งมีข้อกำหนดว่าโรบอทตัวใดจะมีสิทธิบ้างและตัวใดมีสิทธิในการเข้าถึงข้อมูลมากน้อยเพียงใด ดังตัวอย่างข้างต้นมีการจำกัดไม่ให้โรบอท เข้าไปเข้าถึ งข้อมูลในเว็บต้องห้าม ซึ่งเป็นสิ่งที่เว็บโรบอทจะต้องปฏิบัติตาม

3. เว็บโรบอทต้องยึดครองแหล่งข้อมูลแบบถาวร

เว็บโรบอท เป็นตัวที่ช่วยในการหาแหล่งข้อมูลโดยจะทำให้ อินเตอร์เน็ต ขนาดใหญ่ดู เล็กลง ดังนั้น การที่เว็บโรบอทจะทำงานได้ดีควรมีการกำหนดสิ่งที่มัน ต้องการไว้อย่างชัดเจนดัง ต่อไปนี้

- เอ็ชทีทีพี จะมี เฮดรีเควสเมธอทที่จะดึงข้อมูลเฉพาะเฮดเดอร์ ในรูปแบบแอคชั่นจากเว็บเอกสาร โดยไม่นำส่วนของตัว เอ็ชทีเอ็มแอล กลับมาด้วย ซึ่งการใช้สิ่งนี้จะทำให้เราสามารถ ลดโอเวอร์เฮด (overhead) ที่เกิดขึ้นได้แตกต่างจากได้รับรีเควสที่จะทำการดึงข้อมูลทั้ง เอกสาร การใช้สิ่งนี้จะเป็นประโยชน์ต่อโรบอท .ในการที่จะงานกับเฉพาะ ที่มีอยู่โดยไม่ ต้องทำไปตามการเชื่อมต่อทั้งหมดที่มีอยู่ ใน ไฮเปอร์ลิงค์ คอนเทนท์ (hyperlink content)

- เอ็ชทีทีพี จะมีฟิลด์ที่กำหนดใน รีเควส เฮดเดอร์ สำหรับเว็บโรบอทในการที่จะกำหนดค่าให้แก่ เซิร์พเวอร์ ว่ามีข้อมูลชนิดใดบ้างที่มันสามารถเข้าถึงสำหรับโรบอท ที่ถูกออกแบบมาเพื่อที่จะทำการวิเคราะห์ในรูปแบบแอคชั่นเท่านั้นจะกำหนดได้ดังนี้

Accept : x-text

การกำหนดสิ่งที่ต้องการจะช่วยให้ลดช่วงกว้างของระบบเครือข่าย ที่จะต้องพิจารณา เนื่องจากเว็บเซิร์พเวอร์ จะไม่จัดส่งข้อมูลในประเภทที่ โรบอท ไม่สามารถเข้าถึงได้

- ยูอาร์แอล ที่อยู่ในตอนท้ายจะเป็นตัวที่ช่วยบอกว่ามีข้อมูลชนิดใดที่อยู่ที่สุดท้ายของ การเชื่อมต่อ ถ้าไฟล์ที่มีส่วนนามสกุล ที่เป็น ps , zip , z , gif ซึ่งจะเป็นไฟล์ที่ไม่ได้รับความ สนใจ เนื่องจากโรบอทสามารถที่จะติดต่อได้เฉพาะข้อมูลที่เป็นเท็กซ์

เว็บโรบอท มักจะเสี่ยงกับการที่จะต้องเข้าไปอยู่ในส่วนต้องห้าม ดังนั้น เว็บโรบอท จะต้องมีการกำหนดรายการ ที่ควรจะหลีกเลี่ยงดังตัวอย่าง ยูอาร์แอล ที่ขึ้นต้นด้วย news และ WAIS ควรจะถูกตัดออกจากส่วนที่ เว็บโรบอท จะผ่านเข้าไป นอกจากนี้ยังต้องระวังในเรื่อง ของซับเพจ รีเฟอเรนท์ เพื่อเป็นการหลีกเลี่ยงไม่ให้มีการดึงข้อมูลจากเอกสารเดียวมากกว่า หนึ่งครั้ง

- บางระบบจะมีช่วงเวลาที่เหมาะสมในการทำการดึงข้อมูล เพราะจะเป็นช่วงที่มีดึงข้อมูลน้อย ดังนั้นโรบอทจึงควรที่จะต้องคำนึงถึงสิ่งนี้ด้วย

- โรบอท ไม่ควรที่จะพิจารณาว่า เอ็ชทีเอ็มแอล เอกสารทั้งหมดไม่มีข้อผิดพลาดโดยเฉพาะในขณะที่มีการตรวจหา ยูอาร์แอล จะต้องทำการระวังสิ่งต่างๆอย่างเช่นสิ่งต่อไปนี้

A HREF=" http://somehost.somedom/doc "

ซึ่งเป็นการอ้างถึงสองครั้ง แต่ก็มีเว็บไซท์จำนวนมากที่ไม่ใช้เครื่องหมายคั่นใน ยูอาร์แอล สำหรับไดเร็กทอรี ซึ่งหมายความว่า วิธีในการหาชื่อของยูอาร์แอลใช้การอ้างอิงจากชื่อ

- โรบอทควรจะตรวจสอบผลลัพธ์ที่ได้รับมาทั้งหมด โดยอาจจะตรวจสอบจากโปรแกรมคำสั่งและถ้าพบว่าเซิร์พเวอร์ กำลังยุ่งอยู่อาจจะไม่สามารถที่จะทำการตอบสนอง ในการให้บริการในส่วนของเอกสารที่เราต้องการได้ เราก็ต้องคอยฟังว่าเซิร์พเวอร์จะว่าอย่างไรแล้วก็ปฏิบัติตาม

- เป็นการเสี่ยงต่ออันตรายอย่างมาก หากเกิดกรณีที่มีการวนลูปแบบไม่รู้จบขึ้นบนเว็บโดยที่ไม่เคยได้คาดเดาไว้ก่อนว่าจะเกิดอะไรขึ้น และไม่สามารถคาดเดาได้ว่าจะเกิดอะไรขึ้น ดังนั้นเพื่อป้องกันปัญหาดังกล่าวนี้ เราจำเป็นที่จะต้องมีการเก็บค่าเอาไว้ว่า เว็บ ใดที่โรบอทได้ ผ่านไปแล้ว นอกจากนี้ยังต้องมีการตรวจสอบถึงกรณีที่อาจจะเกิดการมีชื่อของเซิร์พเวอร์มากกว่า 1 เซิร์พเวอร์ อยู่บนเครื่องเดียวกันนั่นคือมี ไอพี แอดเดรส เดียวกัน

- ถึงแม้ว่า เว็บโรบอท จะสามารถที่จะเข้าถึงข้อมูลได้คราวละหลายร้อยเอกสารต่อนาที แต่ปัญหาในการใช้งานอย่างหนัก และการมีการเข้าถึงหลาย ๆ อย่างพร้อมๆกันทำให้เซิร์พเวอร์ ดึงข้อมูลมาก ทำให้โรบอทต้องใช้การทำการค้นตามหลักการแบบ ราวนด์-โรบิน (round -robin) และอาจจะต้องหยุดทำงานในช่วงเวลาสั้นๆในบางครั้ง ดังนั้นการที่เราจะให้โรบอทสามารถที่ จะทำการดึงข้อมูลเพียง 1 เอกสารต่อ 1 นาที ในแต่ละเซิร์พเวอร์จึงเป็นวิธีการที่น่าสนใจมากกว่า การทำการเข้าถึงทีละหลายๆเอกสาร

- บางเว็บสามารถที่จะทำการค้นหาได้โดยใช้ ไอเอสอินเด็กซ์ (isindex) ที่มีอยู่ใน เอ็ชทีเอ็มแอล ในขณะที่บางเว็บไม่สามารถค้นหาได้ รวมทั้งยังมีการเปลี่ยนแปลงอยู่บ่อยๆดังนั้น จึงไม่

เป็นการง่ายและมีข้อมูลมากเพียงพอ ที่จะทำการวิ่งตามการเชื่อมต่อและทำการค้นหาว่าจะ อยู่ที่ใด การวิเคราะห์ไฟล์บนเอ็ชทีเอ็มแอล ของเว็บเอกสารจะถูกกระทำโดยโรบอทซึ่งสามารถ ที่จะช่วยกำหนดว่าจะหาข้อมูลที่ใด

4. เว็บโรบอทต้องมีการแสดงข้อผิดพลาด

เมื่อโรบอทมีการวิ่งไปตามเว็บต่างๆ อาจจะเกิดกรณีที่มีการวิ่งไปตามการขาดการเชื่อมต่อ ซึ่งชี้ไปยังเว็บที่ไม่มีอยู่หรือไม่สามารถที่เข้าไปได้ ซึ่งสาเหตุเหล่านี้อาจเกิดจากการที่ เว็บมาสเตอร์ทำการเคลื่อนย้ายเว็บเพจไปยังที่ใหม่ โดยการย้ายไปยังเครื่องใหม่หรือทำการย้าย/ เปลี่ยน ไดเร็กทอรีไป นอกจากนี้ยังมีสาเหตุมาจากการที่มีการเปลี่ยนชื่อของ เอกสารรวมทั้งการเปลี่ยน โดเมนเนม (Domain name) ของเซิร์พเวอร์

เมื่อเป็นเช่นนี้โรบอทควรที่จะทำการส่งข่าวสารแสดงถึงข้อผิดพลาดไปยังแอดเดรส ที่ระบุไว้ตรงเมล์ทู (Mailto) เพื่อบอกแก่เว็บมาสเตอร์ผู้ดูแลไซท์ดังกล่าว

2.1.4 สิ่งที่ควรปฏิบัติในการสร้างโรบอท

1. เพื่อเป็นการดีในการทำการติดต่อสื่อสาร ดังนั้นจึงเป็นการเหมาะสมที่จะมีการประกาศโรบอทก่อนที่จะมีการส่งโรบอทวิ่งออกไปใช้งานจริง

- ถ้าหากว่าเราได้ทำการประกาศโรบอทออกไปก่อนที่จะมีการใช้งานจริงจะทำให้ เว็บมาสเตอร์ต่างๆรู้ว่าจะมี โรบอท เข้ามาก็ไม่ทำการสังเกตหรือแปลกประหลาดใจต่อการทำงาน ของโรบอท ซึ่งจะเป็นการทำให้โรบอทเป็นที่ยอมรับและยินดีให้เข้าใช้บริการข้อมูลต่าง ๆ ของเซิร์พเวอร์มากกว่าที่ไม่ได้มีการประกาศ ซึ่งการทำการประกาศการใช้งานโรบอทนี้เราสามารถ ทำได้โดยการส่งข่าวสารไปยัง ยูสเน็ต นิวกรุ๊ป (USENET newgroup)

comp.infosystem.www.providers

หรืออาจจะทำได้โดยการส่งอีเมล ไปยังแอดเดรสข้างล่างนี้

robots@nexor.co.uk

ทางที่ดีควรจะมีการรวมข้อมูลเกี่ยวกับปัญหาต่างๆ ที่อาจจะเกิดขึ้นได้จากโรบอทส่งไปด้วย

- พิจารณาไซท์เป้าหมายหากว่าโรบอทมีการทำงานกับไซท์เพียงเล็กน้อยเราควรที่จะทำการส่งข่าวสารไปยังไซท์ปลายทาง เหล่านั้นเพื่อเป็นการติดต่อกับ เว็บมาสเตอร์โดยตรง

- ทำการรายงานการทำงานของโรบอท ที่จะเกิดขึ้นหรือบริการต่างๆที่โรบอทต้องการใช้แก่ผู้ดูแลระบบทีไซท์ที่เราจะติดต่อด้วยรับรู้ โดยเน้นเกี่ยวกับเรื่องของการเพิ่มการจราจรบน ระบบเครือข่าย หรือความจำเป็นในการที่จะต้องใช้เนื้อที่ (space) ในการทำงานเมื่อมีการทำงานของ โรบอทเกิดขึ้น ซึ่งถ้าหากมีข้อผิดพลาดเกิดขึ้นผู้ดูแลระบบจะทำการเตือนและทำการแก้ปัญหา ต่างๆ

2. ในการทดสอบการทำงานของโรบอท เราควรที่จะทำการตรวจสอบการทำงาน จากเว็บเซิร์พเวอร์ในระดับโลคอล (locally)ก่อน ไม่ควรที่จะทำการทดลองกับเซิร์พเวอร์ที่ไกล ก่อนที่จะมีการทดสอบจนกระทั่งไม่มีข้อผิดพลาดอยู่ ดังนั้นเมื่อมีการทำการทดสอบในครั้งแรก เราก็ควรที่จะทำการทดสอบกับเซิร์พเวอร์ใกล้ ๆโดยอาศัยโลคอล ยูอาร์แอล

เมื่อทำการทดสอบในระดับใกล้ๆ แล้วควรที่จะทำการวิเคราะห์ถึงประสิทธิภาพ ในการทำงานของโรบอท รวมทั้งผลลัพธ์ที่ได้มาว่ามีความถูกต้องมากเพียงใด เพื่อนำสิ่งที่ได้จากการ วิเคราะห์นี้ไปใช้เป็นแนวทางในการแก้ไขปรับปรุงสำหรับการทำงานกับหลายพันเอกสาร โดยเฉพาะต้องมีการคำนึงถึงการดึงข้อมูลที่จะเกิดขึ้นแล้วทำการแก้ไขให้เรียบร้อย

3. ผู้ทำการเขียนโรบอทจะต้องมีความรู้เกี่ยวกับโรบอทที่เขียนขึ้นว่าจะทำงานอะไรอย่างไร รวมทั้งสามารถที่จะสามารถทำการควบคุมได้ ซึ่งสามารถทำได้จากแนวทางต่อไปนี้

- มีการจัดทำล็อก (log) เพื่อทำการดูแลว่า ขณะนี้ โรบอทอยู่ที่ใดรวมทั้งดูแลการทำงานต่างๆ ของโรบอทว่าทำงานคืบหน้าไปถึงไหน รวมทั้งเพื่อดูแลให้สามารถที่จะทำการควบคุมการทำงานของโรบอทได้ รวมทั้งโรบอทก็ต้องมีการควบคุมไม่ให้เกิดการทำงานแบบวนรอบและต้องดูแลการใช้เนื้อที่ต่างๆ (Disk space requirement) ในแต่ละช่วงเวลาต่างๆ เพื่อเป็นประโยชน์ ในการดูแลไม่ให้เกิด ดิสก์ สเปส ครันช์ (Disk space crunch)

- ต้องออกแบบโรบอทให้สามารถทำงานได้ง่าย รวมทั้งมีการทำคำสั่งที่สามารถกำหนดเส้นทางของโรบอท ,การยกเลิกการทำงานของโรบอท , การข้ามเซิร์พเวอร์ที่ไม่ต้องการและต้องมีการติดตามการทำงานของโรบอทว่าหยุดการทำงานหรือผิดพลาดหรือไม่

4. เมื่อทำการรันการทำงานของโรบอท เราควรจะแน่ใจว่า เว็บมาสเตอร์ ของไซท์ ที่เราติดต่อด้วยสามารถที่จะส่งข่าวสารหรือพูดคุยกับเราได้ ในกรณีที่จำเป็นเนื่องจากหากมี ความผิดพลาดต่างๆ เกิดขึ้นจากการทำงานของโรบอท จะมีผู้เขียนโรบอทเพียงคนเดียวที่สามารถ ที่จะทำการแก้ไขได้อย่างรวดเร็วที่สุดถ้าเป็นไปได้จึงควรที่จะทำการลอกออน (log on) เครื่องที่ ใช้ในการทำการรัน โรบอทเอาไว้ตลอดในช่วงที่ทำการรันโรบอทควรจะทำการรันโรบอทเฉพาะ ในช่วงที่เราอยู่เท่านั้น ไม่ควรทำการรันในกรณีที่เราจะไม่อยู่ในเวลานาน เช่นไม่อยู่เป็นเวลาหนึ่ง อาทิตย์

5. ระหว่างที่มีการทำงานของโรบอทอยู่โรบอทอาจจะผ่านไปยังเว็บไซท์ จำนวนมาก อาจจะไปมีผลทำให้เว็บมาสเตอร์ มีความไม่พอใจในการทำงานของโรบอท ดังนั้นผู้ที่ทำการรัน โรบอท จะต้องเตรียมพร้อมในการตอบสนองและพยายามอธิบายว่า โรบอท กำลังทำอะไรอยู่ เพื่อเป็นการป้องกันไม่ให้ โรบอทมีการทำความไม่พอใจให้แก่ เว็บมาสเตอร์ เราควรที่จะทำการกำหนดให้โรบอททำงานเฉพาะโฮมเพจของไซท์ดังกล่าว

5. การทำงานควรจะมีการรวบรวม เว็บ ไซท์ ให้มากที่สุดเท่าที่จะมากได้ โดยการนำ โรบอทไปทำงานบนอินเตอร์เน็ตที่มีขนาดใหญ่ แล้วนำข้อมูลต่อไปนี้ไปใช้ให้เกิดประโยชน์ ในการใช้เป็นสิ่งที่ช่วยในการค้นหา ซึ่งสามารถทำได้ดังนี้

- ร่วมใช้ข้อมูลที่ได้รับมาจะประกอบด้วยเว็บเพจ ที่ดึงกลับมาไม่ว่าจะมาจาก เอฟทีพี เวิร์ลไวท์เว็บ ซึ่งข้อมูลเหล่านี้จะถูกเก็บและถูกใช้โดย ยูสเซอร์ คนอื่นเพื่อที่จะไม่ต้องมีการทำงานซ้ำอีกครั้งในการหาข้อมูลจากโรบอทตัวเดียวกัน

- เว็บโรบอทถูกนำมาใช้งานในงานเฉพาะทาง อาจจะถูกนำมาใช้ในการช่วยทำฐานข้อมูลหรือทำการรวบรวมสิ่งต่างๆ เพื่อนำมาใช้ประโยชน์ซึ่งจะช่วยทำให้โรบอทได้รับการยอมรับมากขึ้น มากกว่าการเป็นเพียงสิ่งที่ทำงานแล้วทำให้เซิร์พเวอร์มีการดึงข้อมูลมากขึ้น

2.1.5 การทำการป้องกันการเข้าถึงข้อมูลของโรบอท

การที่เราจะทำการป้องกันการเข้าถึงข้อมูลของโรบอท นั้นเป็นหน้าที่ของเว็บมาสเตอร์ ในการที่จะทำการสร้างไฟล์ บนเซิร์พเวอร์เพื่อที่จะเป็นตัวกำหนดแนวทางต่างๆของโรบอท ซึ่ง ไฟล์ดังกล่าวนี้มีชื่อว่า โรบอท เอ็กซ์คลูชั่นไฟล์ (Robot Exclusion File) ซึ่งเป็นไฟล์ที่สามารถจะ ทำการเข้าถึงได้จากเอ็ชทีทีพี ที่ยูอาร์แอล ที่มี /robot.เท็กซ์ ซึ่งภายในโรบอท เอ็กซ์คลูชั่น ไฟล์ นั้นจะประกอบด้วยข้อกำหนดต่างๆดังจะได้กล่าวต่อไปวิธีการดังกล่าวนี้เป็นวิธีที่ได้รับความนิยม อย่างสูง เนื่องจากเป็นวิธีที่สามารถทำได้อย่างง่ายดาย และสามารถทำได้บนเซิร์พเวอร์ทุกตัวด้วย โดยเว็บโรบอทนั้นสามารถที่จะรู้ถึงข้อกำหนดต่างๆ ได้จากการที่ทำการเข้าถึงไฟล์ที่มียูอาร์แอล / robot.txt ซึ่งจะมีอยู่เพียงเอกสารเดียวที่ เว็บโรบอทจะได้รับข้อมูลกลับไป ถึงแม้ว่าเว็บมาสเตอร์สามารถที่จะทำการกำหนดข้อกำหนดต่างๆ ได้ภายในไฟล์ โรบอท เอ็กซ์คลูชั่นไฟล์ ก็ยังต้องพิจารณาถึงโรบอท แต่ละตัวว่าได้มีการตรวจสอบไฟล์ ที่มีอยู่แล้วนี้หรือไม่ในที่แรกที่มันทำการเข้าถึงว่าจะต้องทำอะไรบ้างหรือมีสิทธิอย่างไรบ้างตามที่ได้กำหนดไว้

แต่อย่างไรก็ตามปัญหาที่น่าจะต้องพิจารณาก็คือ โรบอท เอ็กซ์คลูชั่นไฟล์ ตัวนี้มีเพียง เว็บมาสเตอร์ที่สามารถทำการแก้ไขได้ และก็ไม่ได้เป็นเอกสารอิสระในแต่ละการแก้ไขปัญหาสามารถทำได้ดังนี้คือมีการใช้โลคอลเมนเทนแนนซ์โพรซีเดอร์ (Local maintenance procedure) ซึ่งจะทำการสร้างไฟล์ robot.txt เพียงไฟล์เดียว จากจำนวนของไฟล์ทั้งหมด

จากหลักการเบื้องต้นเราจะนำไปใช้อ้างอิง ในการออกแบบโรบอทที่จะนำมาใช้งานในโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ ซึ่งมีหน้าที่ในการทำการค้นหาข้อมูลและทำการวิเคราะห์ข้อมูลที่จะเลือกนำมาจัดเก็บ โดยในโครงงานได้ทำการเลือกที่จะใช้ ภาษาเดลไฟเป็นภาษาที่ใช้ในการเขียนโรบอท เนื่องจาก การทำงานหลักของโครงงานเขียนด้วย ภาษาเดลไฟ

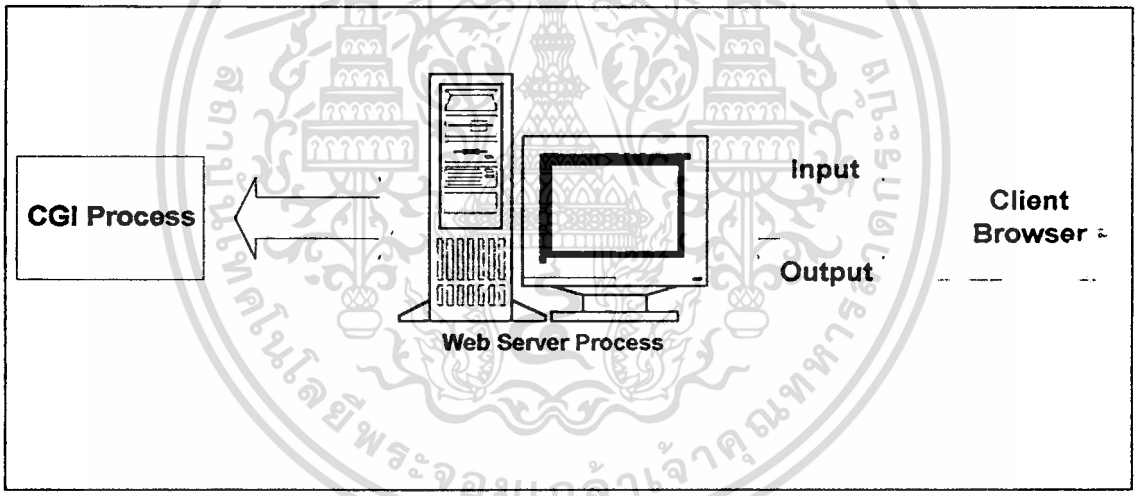
จึงเป็นการง่ายที่จะทำการเชื่อมต่อโปรแกรม นอกจากนี้ภาษาเดลไฟเป็นภาษาที่ค่อนข้าง มีความสมบูรณ์ในตัวเอง และมีการทำงานแบบวิชวลทำให้ส่วนที่เป็น ยูสเซอร์อินเฟส สามารถที่จะสร้างได้ง่ายและมีความสวยงาม รวมทั้งมีคอมโพเนนท์ให้ใช้งานเยอะ

2.2 ส่วนเชื่อมต่อระหว่างผู้ใช้บริการและส่วนจัดเก็บข้อมูล

ส่วนเชื่อมต่อระหว่างผู้ใช้บริการกับส่วนจัดเก็บข้อมูล เป็นส่วนที่เชื่อมต่อการทำงานของโฮมเพจของเซิร์สเอนจิน กับผู้ใช้บริการ โดยจะทำหน้าที่ดังต่อไปนี้

- ทำการรับตัวแปรและเงื่อนไขการทำงานของผู้ใช้บริการจากโฮมเพจ
- ส่งค่าตัวแปรและเงื่อนไขไปยังเซิร์พเวอร์ที่ให้บริการ
- รับค่าผลลัพธ์จากเซิร์พเวอร์ที่ให้บริการ แล้วส่งออกมาทางหน้าจอให้ผู้บริการ

การทำงานของส่วนเชื่อมต่อนี้จะอาศัยการทำงาน ของโปรแกรมที่เรียกว่า CGI ซึ่งมีหลักการและลักษณะการทำงานดังต่อไปนี้



รูปที่ 2.1 ภาพแสดงหลักการทำงานของโปรแกรมซีจีไอ

2.2.1 คอมมอนเกทเวย์อินเตอร์เฟส (Common Gateway Interface)

คอมมอนเกทเวย์อินเตอร์เฟสเป็นมาตรฐานในการติดต่อกันระหว่างแอพพลิเคชั่นภายนอกกับอินฟอร์เมชั่นเซิร์พเวอร์ เช่น เอ็ชทีทีพี หรือเว็บเซิร์พเวอร์ โดยปกติเอกสารแบบ เอ็ชทีเอ็มแอล ที่เว็บบราวเซอร์ได้รับจะเป็นแบบสแตติค (static) นั่นคือมันอยู่ในสถานะที่คงที่ เท็กซ์ไฟล์ ที่ได้ไม่มีการเปลี่ยนแปลงค่าใด ๆ โปรแกรมซีจีไอจะทำงานในแบบเรียลไทม์ (real-time) ดังนั้นมันจึงสามารถให้ผลลัพธ์เป็นข้อมูลชนิดไดนามิก (dynamic)ได้

ตัวอย่าง

เมื่อคุณต้องการติดต่อฐานข้อมูลบนระบบยูนิกซ์ของคุณไปยังเวิร์ลไวด์เว็บเพื่อให้ คน ทั่วไปสามารถค้นหาข้อมูลได้ วิธีพื้นฐานที่สุด คุณต้องสร้างโปรแกรมซีจีไอที่ เว็บไดมอน สามารถทำการเอ็กซ์เซคคิว (execute) ได้โดยการส่งข้อมูลไปยังระบบฐานข้อมูลรับผลที่ถูกส่ง กลับมาและนำผลที่ได้ส่งไปแสดงผลยัง ไคลเอนท์ นี่คือตัวอย่างของเกทเวย์และนี่คือสิ่งที่ทำซีจีไอ เกิดขึ้น

ข้อกำหนด (Specifics)

เนื่องจากการเอ็กซ์เซคคิวของโปรแกรมซีจีไอ เป็นการเปิดโอกาสให้บุคคลอื่นสามารถ ทำการรันโปรแกรมบน เซิร์พเวอร์ ดังนั้นปัญหาในเรื่องความปลอดภัยจึงเป็นปัญหาสำคัญจึงต้อง มีการป้องกันเมื่อต้องใช้โปรแกรมซีจีไอ วิธีการที่พอจะเป็นไปได้สำหรับ เว็บ ยูสเซอร์ คือการนำโปรแกรมซีจีไอแยกเก็บไว้ยังไดเรคทอรี เฉพาะเพื่อให้เว็บเซิร์พเวอร์สามารถทำการเอ็กซ์เซคคิว ได้ที่ไดเรคทอรีนี้เท่านั้น ซึ่งไดเรคทอรีนี้อยู่ภายใต้การควบคุมของ เว็บมาสเตอร์ ถ้าเว็บเซิร์พเวอร์ ที่ใช้เป็นเวอร์ชั่นของเอ็นซีเอสเอ เอ็ชทีทีพี เซิร์พเวอร์ (NCSA HTTP Server) จะปรากฏ ไดเรคทอรีที่ชื่อว่า /ซีจีไอ-วิน (Cgi-win) ซึ่งเป็นไดเรคทอรีที่ใช้เก็บโปรแกรมซีจีไอและไดเรคทอรี ชื่อ /ซีจีไอ-เอสซีอาร์ (Cgi-scr) ใช้เก็บตัวโปรแกรมคำสั่งของโปรแกรม

โปรแกรมซีจีไอสามารถเขียนด้วยภาษาต่างๆ มากมายที่สามารถทำการเอ็กซ์เซคคิว บนระบบได้ เช่น

ภาษาซี, ซีพลัสพลัส (C/C++)

ฟอร์แทน (Fortan)

เพิร์ล (Perl)

ทีซีแอล (TCL)

ยูนิกซ์เชลล์ (Any Unix shell)

วิชวลเบสิก (Visual Basic)

แอพเปิลสคริปต์ (AppleScript)

เดลไฟ (Delphi)

ในการเลือกภาษาที่จะใช้ในการเขียนโปรแกรมซีจีไอ นั้นขึ้นอยู่กับระบบที่ใช้อยู่ เช่น ภาษาซี หรือ ฟอร์แทน ต้องทำการคอมไพล์โปรแกรมแล้วนำไปใส่ไว้ใน /ซีจีไอ-บิน ไดเรคทอรี จะนำมาใช้งานได้และเก็บโปรแกรมไว้ในไดเรคทอรี ที่ชื่อ /ซีจีไอ-เอสซีอาร์ แต่ถ้าใช้ภาษา สคริปต์แทนเช่น เพิร์ล,ทีซีแอล หรือยูนิกซ์ เชลล์ เมื่อเขียนซีจีไอ สคริปต์ เสร็จแล้วนำเข้าไปไว้ใน /ซีจีไอ-บินไดเรคทอรี แล้วใช้ได้เลยไม่ต้องยุ่งเกี่ยวกับโปรแกรมอีก

จากข้อได้เปรียบข้างต้นจึงทำให้หลายคนเลือกที่จะเขียนซีจีไอ สคริปต์ แทนการเขียน ซีจีไอโปรแกรมเนื่องจากทำการตรวจสอบ,เปลี่ยนแปลงและดูแลรักษาง่ายกว่า

วิธีการรับข้อมูลจากเซิร์พเวอร์ของซีจีไอ

ในแต่ละครั้งที่มีรีเควสมายังยูอาร์แอลของซีจีไอโปรแกรมทางเซิร์พเวอร์จะทำการ เอ็กซ์เซคคิว ในแบบเรียลไทม์และส่งผลลัพธ์ของโปรแกรมกลับไปยังไคลเอนท์จึงเกิดความเข้าใจ ผิดเกี่ยวกับซีจีไอ นั่นคือ สามารถทำการเรียกใช้และส่งออปชั่น (options) ต่าง ๆ ให้กับโปรแกรม โดยใช้คอมมอนไลน์(command-line)ได้ เช่น

command% myprog -qa blorf

ซีจีไอจะใช้คอมมอนไลน์สำหรับวัตถุประสงค์อื่นและวิธีนี้ไม่สามารถทำได้โดยตรง นั่นคือ ซีจีไอ จะใช้ตัวแปรภายนอกต่างๆในการส่งค่าให้กับโปรแกรมตัวแปรที่สามารถใช้ได้หลัก ๆ มีอยู่ 2 ตัว คือ

1 คิวรีสตริง (Query_String)

คิวรีสตริงถูกกำหนดให้ตามหลังเครื่องหมาย “?” ใน ยูอาร์แอล ข้อมูลนี้ถูกเพิ่มเข้ามาโดย ไอเอสอินเด็กซ์ หรือเอ็ชทีเอ็มแอล ฟอร์ม (โดยใช้ GET แอคชั่น) เราสามารถทำการ ฝังเอาไว้ใน เอ็ชทีเอ็มแอล อันชอร์ (anchor) ได้ด้วยสตริงนี้มักจะเป็นอินฟอร์เมชั่น คิวรี

สตริงที่ได้นี้จะถูกเข้ารหัสในรูปแบบ ยูอาร์แอล มาตรฐาน นั่นคือเนื้อที่จะถูกแทนด้วยเครื่องหมาย ‘+’ และตัวอักษรพิเศษแทนด้วย %xx เลขฐานสิบหก ดังนั้นจึงต้องทำการถอดรหัสก่อนที่จะนำมาใช้

2 พาร์ทอินโฟ (PATH_INFO)

ซีจีไอยอมให้มีข้อมูลพิเศษถูกส่งมาพร้อมกับ ยูอาร์แอล สำหรับ เกทเวย์ ของคุณที่จะใช้ในการส่ง เอ็กซ์ทรา คอนเท็กซ์ สเปคซิฟิก (extra context-specific) อินฟอร์เมชั่นให้กับ สคริปต์ ข้อมูลนี้จะถูกเพิ่มเข้ามาในส่วนหลังของเกทเวย์ ในส่วนยูอาร์แอล โดยไม่มีการเข้ารหัส จากเซิร์พเวอร์

วิธีการส่งข้อมูลกลับไปยังไคลเอนท์

ปัญหาของผู้เริ่มต้นเขียนซีจีไอที่มักจะพบอยู่เสมอ นั่นคือไม่ได้กำหนดรูปแบบของผลลัพธ์ที่จะส่งกลับไปยังไคลเอนท์

ซีจีไอโปรแกรมสามารถส่งข้อมูลชนิดต่าง ๆ กลับไปยังไคลเอนท์ได้ เช่น image, audio, zip file หรือ อื่นๆโดยซีจีไอโปรแกรม จะต้องบอกเซิร์พเวอร์ว่าข้อมูลที่จะทำการส่งเป็นข้อมูลชนิดใด โดยการเพิ่มเฮดเดอร์ให้กับผลลัพธ์ซึ่งเฮดเดอร์ประกอบด้วยข้อมูลชนิด เท็กซ์ มีการแบ่งบรรทัดโดยใช้ไลน์ฟีด (linefeeds) หรือแคร์เรจรีเทิร์น (carriage returns) หรือทั้ง 2 อย่าง ตามด้วยบรรทัดว่าง ๆ 1 บรรทัดก่อนที่จะเป็นข้อมูลที่มีรูปแบบตามที่กำหนด

ตัวอย่าง

การส่งเอกสารแบบเอ็ชทีเอ็มแอลกลับไปยังไคลเอนท์, ผลลัพธ์ ที่ ซีจีไอจะต้องทำส่ง จะมีรูปแบบตามข้างล่างนี้

```
Content-type: text/html
<HTML><HEAD>
<TITLE>output of html from cgi script</TITLE>
</HEAD><BODY>
<H1>Sample output</H1>
What do you think of <String>this?</String>
</BODY></HTML>
```

ตัวอย่าง การอ้างอิงไปยังเอกสารอื่นสามารถทำได้โดยการเพิ่มโลเคชั่น:ยูอาร์แอล เข้าไปในส่วนเฮดเดอร์หลังคอนเทน-ไทพ์: เท็กซ์/เอ็ชทีเอ็มแอล

```
Content-type: text/html
Location: gopher://httprules.foobar.org/0
<HTML><HEAD>
<TITLE>Sorry...it moved</TITLE>
</HEAD><BODY>
<H1>Go to gopher instead</H1>
Now available at
<A HREF="gopher://httprules.foobar.org/0">a new location</A>
on our gopher server.
</BODY></HTML>
```

การรับข้อมูลจากฟอร์มบนเอกสารแบบเอ็ชทีเอ็มแอล

เนื่องจากเมธอท (Method) ที่ใช้ในฟอร์มมีอยู่ 2 วิธี คือ GET และ POST ดังนั้นการ รับข้อมูลจากฟอร์มจึงขึ้นอยู่กับเมธอทที่เลือกใช้

- เก็ท เมธอท รับข้อมูลที่ถูกเอ็นโค้ด (encode) ผ่านตัวแปรคิวรีสตริง

- โพส เมธอท รับข้อมูลที่ถูกเอ็นโค้ดผ่าน สแตนดาร์ดอินพุท (stdin)โดยดูความยาวของอินพุตจากตัวแปรคอนเทนเลจน์ (Content_Length)

ในการเขียนฟอร์มในเอกสารแบบเอ็ชทีเอ็มแอลจะมีแอททริบิว(attribute) 2 ตัวคือ เนม (name) และแวลูล (value)

- เนม คือชื่อของตัวแปรที่ใช้แทนแต่ละไอเทม (items)ในฟอร์ม

- แวลูล คือค่าของตัวแปรแต่ละตัว

ฟอร์มของข้อมูลคือ เนม=คู่ของแวลูล ที่ต่อกันโดยใช้เครื่องหมาย “&” คั่นกลาง ระหว่างคู่และเนม=คู่ของแวลูล คือสตริงถูกเอ็นโค้ดแล้ว เช่นการแทนสเปสด้วย “+” เป็นต้น

เนื่องจากการดีโค้ดข้อมูลจากฟอร์มเป็นเรื่องที่ยุ่งยากในบางภาษาจึงได้มีโปรแกรมย่อย (routine) หรือโปรแกรมที่ใช้สำหรับการดีโค้ดมากมาย เช่น

- บรอนด์เชลล์ใช้คำสั่ง sed และ awk ในการแปลงสตริงที่ได้จากฟอร์มออกเป็นตัวแปรต่าง ๆ

- ภาษาซีมีโปรแกรมย่อยที่ใช้ในการแปลงคิวรี สตริงเป็นกลุ่มของตัวแปร

- เพิร์ล ประกอบด้วยเพิร์ล ซีจีไอ-ลิบ ซึ่งมีโปรแกรมย่อยมากมายในการดีโค้ดฟอร์ม

- ทีซีแอล อาร์กูเมนท์ โพรเซสเซอร์ (TCL argument processor) คือกลุ่มของทีซีแอลโปรแกรมย่อยที่ใช้ในการดึงข้อมูลจากฟอร์มและแทนค่าเป็นตัวแปรของทีซีแอล.

การรักษาความปลอดภัยเมื่อมีการใช้ซีจีไอสคริปต์

ทุกครั้งที่โปรแกรมที่ติดต่อกับไคลเอนท์บนเครือข่ายเพิ่มมากขึ้น โอกาสที่ไคลเอนท์ จะสามารถเข้าถึง ส่วนที่ไม่ได้รับอนุญาติก็มีมากขึ้น แม้ว่าบางสคริปต์จะดูเหมือนธรรมดา แต่ก็สามารถทำให้เกิดอันตรายกับระบบได้เช่นกัน ดังนั้นจึงมีข้อแนะนำบางประการในการป้องกันไม่ให้โปรแกรมถูกโจมตี

- ภาษาเช่นเพิร์ลและบอร์นเชลล์ มีคำสั่ง eval ที่ใช้ในการสร้างและจัดการกับสตริงซึ่งเป็นสิ่งที่อันตราย

- ระวังเรื่องการแปรความหมายผิดเนื่องจากการใช้ตัวอักษรพิเศษที่ถูกป้อนเข้ามาเป็น อินพุตจากไคลเอนท์

- ถ้ามีการใช้ข้อมูลจากไคลเอนท์ เพื่อสร้างคอมมานด์ไลน์สำหรับเรียก popen หรือ system ต้องแน่ใจว่าใส่เครื่องหมาย "\" หน้าตัวอักษรที่มีความหมายพิเศษกับบอร์นเชลล์ก่อน ที่จะเรียกใช้ฟังก์ชันวิธีนี้เป็นวิธีที่ใช้ได้ง่ายในฟังก์ชันภาษาซี

- บางกรณีเซิร์พเวอร์ทำการปิด เซิร์พเวอร์-ไซด์ อินคลูดส์เพื่อป้องกันสคริปต์ ไดเรคทอรี เนื่องจากเซิร์พเวอร์-ไซด์ อินคลูดส์ สามารถถูกรบกวนโดยไคลเอนท์ที่รบกวนสคริปต์ที่ส่ง ผลลัพธ์ออกไปโดยตรง

ตัวแปรของซีจีไอ

เพื่อการส่งค่าข้อมูลจากอินฟอร์เมชั่นรีเควส จากเซิร์พเวอร์ให้กับสคริปต์, เซิร์พเวอร์ จะใช้คอมมานด์ไลน์เหมือนกับตัวแปรซึ่งตัวแปรนี้จะถูกกำหนดค่าเมื่อเซิร์พเวอร์ทำการเอ็กซ์เซคคิว โปรแกรมเกทเวย์

1. ข้อกำหนดเกี่ยวกับตัวแปร

ตัวแปรต่อไปนี้เป็นตัวแปรทั่ว ๆ ไปที่ถูกกำหนดค่าทุกครั้งที่มี รีเควส

- เซิร์พเวอร์_ซอฟท์แวร์ (Server_Software) ชื่อและเวอร์ชันของอินฟอร์เมชั่น เซิร์พเวอร์ซอฟท์แวร์ที่ตอบรับรีเควสและทำการรันเกทเวย์

Format : name/เวอร์ชัน

- เซิร์พเวอร์_เนม (Server_Name) ชื่อของเซิร์พเวอร์ (hostname), ดีเอ็นเอส (DNS alias) หรือไอพีแอดเดรสที่ปรากฏเมื่อมีการอ้างอิงถึงยูอาร์แอลของตัวเอง

- เกทเวย์_อินเตอร์เฟส (Gateway Interface) ของซีจีไอสเปคซิฟิเคชันที่เซิร์พเวอร์คอมไพล์

Format: cgi/revision

ตัวแปรต่อไปนี้เป็นตัวแปรเฉพาะที่ถูกกำหนดขึ้นมาตาม รีเควส เพื่อให้บรรลุผลตามโปรแกรมเกทเวย์

- เซิร์พเวอร์_โปรโตคอล (Server_Protocol) ชื่อและเวอร์ชันของโปรโตคอลที่ รีเควสใช้

Format : protocol/revision

- เซิร์พเวอร์_พอร์ต (Server_Port) หมายเลขพอร์ตของรีเควส

- รีเควส_เมธอท (Request_Method) เมธอทที่ รีเควสส่งมาสำหรับเอ็ชทีทีพี

- พาร์ท_อินโฟ อินฟอร์เมชั่นที่ส่งมาจากไคลเอนท์เป็นข้อมูลที่จะถูกดีโค้ดโดย เซิร์พเวอร์ถ้ามันมาจาก ยูอาร์แอล ก่อนที่จะถูกส่งไปยังซีจีไอ สคริปต์

- พาร์ท_ทรานสเลต (Path_Translaed) เซิร์พเวอร์จะแปลงพาร์ทอินโฟและทำการแปลงจากเวอร์ชวลพาร์ทเป็นฟิสิคอล พาร์ท

- สคริปต์_เนม เวอร์ชวลพาร์ทที่ชี้ไปยังสคริปต์จะทำการเอ็กซ์เซคคิวใช้สำหรับการอ้างอิงยูอาร์แอล ด้วยตัวเอง

- คิวรี_สตริง ข้อมูลที่ตามหลังเครื่องหมาย "?" ใน ยูอาร์แอล ที่อ้างอิงไปยังสคริปต์ จัดว่าเป็นคิวรีอินฟอร์เมชั่นตัวแปรนี้จะถูกกำหนดค่าเมื่อมคิวรี อินฟอร์เมชั่น

- รีโมท_โฮส (Remote_Host)ชื่อของโฮสที่ทำการ รีเควส ถ้าเซิร์พเวอร์ไม่มีข้อมูลนี้ จะไม่มีการกำหนดค่าตัวแปรนี้แต่ไปกำหนดค่ารีโมทแอดเดรสแทน

- รีโมท_แอดเดรส (Remote_Addr) ไอพีแอดเดรสของรีโมทโฮสทำการรีเควส

- รีโมท_ไอเดนท์ (Remote_Ident) ถ้าเอ็ชทีทีพี เซิร์พเวอร์ สนับสนุนตามข้อกำหนด RFC 931 ตัวแปรนี้จะถูกกำหนดค่าให้เป็นรีโมทยูสเซอร์ที่ดึงมาจากเซิร์พเวอร์ การใช้ตัวแปรนี้ จะกำหนดให้ใช้ในการลอกกิงเท่านั้น

- คอนเทนท์_ไทพ์ (Content_Type) สำหรับคิวรีที่ประกอบด้วยข้อมูล เช่น เอ็ชทีทีพี POST และ PUT นี่คือชนิดของข้อมูล

- คอนเทนท์_เลจท์ (Content_Length) ความยาวของข้อมูลที่ถูกส่งมาจากไคลเอนท์

นอกจากนี้ เฮดเดอร์ ไลน์ ที่ได้รับจากไคลเอนท์จะถูกเก็บไว้ในตัวแปรที่ขึ้นต้นด้วย เอ็ชทีทีพี_ ตามด้วย เฮดเดอร์เนม โดย "-" จะถูกเปลี่ยนเป็น "_" เซิร์พเวอร์ จะเอาเฮดเดอร์ที่ได้ ทำตามกระบวนการแล้วออก ถ้าจำเป็นเซิร์พเวอร์จะเลือก เฮดเดอร์ บางส่วนหรือทั้งหมดออก ถ้าหากว่าการเพิ่มเข้าไปจะทำให้มากเกินกว่าที่ระบบกำหนดไว้

2 <u>ซีจีไอสคริปต์อินพุต</u>

<u>ข้อกำหนดของซีจีไอสคริปต์อินพุต</u>

สำหรับ รีเควส ที่มีข้อมูลประกอบเข้ามาหลังจากส่วนเฮดเดอร์ เช่น เอ็ชทีทีพี โพส หรือใส่ข้อมูลจะถูกส่งให้กับสคริปต์ทางสแตนดาร์ดอินพุต

เซิร์พเวอร์จะส่งคอนเทนท์เลจท์ที่ใช้กำหนดความยาวของข้อมูลที่ส่งมา และต้องจำไว้เสมอว่า จะต้องให้คอนเทนท์ไทพ์ของข้อมูลดีแล้ว เซิร์พเวอร์ไม่มีหน้าที่ในการส่งเอ็นออฟไฟล์หลังจากที่ สคริปต์ อ่านค่าคอนเทนท์เลจน์แล้ว

3 <u>ผลลัพธ์ของซีจีไอสคริปต์</u>

<u>ผลลัพธ์ของสคริปต์</u>

สคริปต์ จะส่งผลลัพธ์ไปยังสแตนดาร์ดเอาท์พุต ซึ่งผลลัพธ์นี้เป็นได้ทั้งเอกสารที่ถูกสร้างโดยสคริปต์หรือคำสั่งที่ส่งให้ เซิร์พเวอร์ เพื่อแสดงผลลัพธ์

โดยปกติสคริปต์จะสร้างผลลัพธ์และส่งกลับไปยังไคลเอนท์ ข้อได้เปรียบของวิธีนี้คือ สคริปต์ไม่ต้องส่ง เอ็ชทีทีพี/1.0 เฮดเดอร์แบบเต็มรูปแบบสำหรับทุกรีเควส

บางสคริปต์ต้องการที่จะหลีกเลี่ยงการเพิ่มโอเวอร์เฮดให้กับเซิร์พเวอร์ จากการส่งผ่าน ค่าผลลัพธ์และติดต่อกับไคลเอนท์ในการแยกสคริปต์ เหล่านี้ออกจากสคริปต์ อื่น ซีจีไอ ต้องการให้ชื่อของสคริปต์ ขึ้นต้นด้วยเอ็นพีเฮ็ช ถ้าสคริปต์ไม่ต้องการให้เซิร์พเวอร์ ส่งผ่าน เฮดเดอร์ของสคริปต์เหล่านั้น ในกรณีนี้สคริปต์จะตอบสนองโดยส่งค่าของ เอ็ชทีทีพี/1.0 กลับ ไปยังไคลเอนท์

การส่งส่วนเฮดเดอร์ (Parsed headers)

ผลลัพธ์ของสคริปต์จะเริ่มด้วยส่วน เฮดเดอร์ ซึ่งประกอบด้วยเท็กซ์ไลน์ตามรูปแบบ ของเอ็ชทีทีพี เฮดเดอร์ และปิดท้ายด้วยบรรทัดว่าง

เซิร์พเวอร์ไดเรคทอรี ที่กำหนดขึ้นมา 3 ตัว ได้แก่

- คอนเทนท์_ไทพ์ ใช้กำหนด เอ็มไอเอ็มอีไทม์ของเอกสารที่จะส่งกลับ

- โลเคชัน ใช้กำหนดเซิร์พเวอร์ที่สามารถอ้างอิงของเอกสารมากกว่าที่จะส่งเอกสารจริง ถ้าอาร์กูเมนท์ที่กำหนดเป็น ยูอาร์แอล, เซิร์พเวอร์ จะย้อนกลับไปยังไคลเอนท์ ถ้าอาร์กูเมนท์ที่กำหนดเป็นเวอร์ชวลพาร์ท เซิร์พเวอร์ จะทำการดึงเอกสารที่ไคลเอนท์ต้องการ ซึ่งดูต้นฉบับของเอกสารจาก -- ? ไดเรคทอรีจะทำงานเซิร์พเวอร์แต่ถ้าเป็น # ไดเรคทอรี จะต้องย้อนกลับไปยัง ไคลเอนท์

- สถานะ ให้เอ็ชทีทีพี/1.0 สเตตัสไลน์กับเซิร์พเวอร์ เพื่อส่งต่อไปให้ไคลเอนท์

หลักการของซีจีไอจะถูกนำมาใช้ในการออกแบบซีจีไอที่ ทำหน้าที่ของซีจีไอ คือ เป็นตัวที่ทำหน้าที่ในการเชื่อมต่อระหว่าง ส่วนของยูสเซอร์ อินเตอร์เฟสกับส่วนที่ใช้ในการเก็บข้อมูล เมื่อเข้าใจหลักการทำงานของ ซีจีไอแล้วก็จำเป็นที่จะต้องเลือกภาษาที่จะนำมาเขียนภาษา ซีจีไอ โดยในโครงงานนี้เราได้ทำการศึกษาโปรแกรมภาษาต่างๆ ที่จะนำมาใช้ในการเขียนโปรแกรม ซีจีไอ ไม่ว่าจะเป็น ภาษาสคริปต์ ภาษาเพิร์ล ภาษาวิชวลเบสิก ภาษาเดลไฟ ภาษาวิชวลซีพลัส พลัส ซึ่งแต่ละภาษาก็มีข้อดีและก็ข้อเสียต่างกัน โดยเหตุผลแรกที่นำมาใช้ในการตัดสินใจคือแพลทฟอร์มที่จะใช้ เนื่องจากภาษาสคริปต์ และ ภาษาเพิร์ล นั้น จะใช้แพลทฟอร์ม ที่เป็นยูนิกซ์ ซึ่งไม่เหมาะสมที่จะนำมาเป็นแพลทฟอร์ม ของโครงงาน ส่วนภาษาซีพลัสพลัส เป็นภาษาที่มีความซับซ้อนมาก จึงไม่เหมาะสมที่จะศึกษาเนื่องจากข้อจำกัดทางเวลา ทางผู้จัดทำโครงงานจึงเลือกภาษาเดลไฟในการเขียนซีจีไอ เนื่องด้วยสาเหตุต่อไปนี้

1 ภาษาเดลไฟเป็นภาษาที่มีความสะดวกในการเขียนโปรแกรม

2 ภาษาเดลไฟที่เลือกใช้สามารถเขียนโปรแกรมซีจีไอ ที่สามารถติดต่อกับส่วนจัดเก็บข้อมูลได้ง่าย

3 ภาษาเดลไฟมีลักษณะการทำงานแบบวิชวล ดังนั้นจึงง่ายในการที่ทำส่วนยูสเซอร์ อินเทอร์เฟส

ดังนั้นในส่วนต่อไปนี้จะทำการกล่าวถึงภาษาเคลไฟที่เกี่ยวข้องกับการทำงานของซีจีไอ

2.2.2 ภาษาเคลไฟกับซีจีไอ

ภาษาเคลไฟเป็นภาษาที่เป็นวิชวล คือ มีการทำออปเจค ไว้สำหรับทำงาน และมีส่วนที่ใช้ในการสร้างยูสเซอร์อินเทอร์เฟสได้ง่าย นอกจากนี้ภาษาเคลไฟยังมี คอมโพเนนท์ที่ได้รับการพัฒนาจากโปรแกรมเมอร์ต่างๆ เป็นจำนวนมาก จึงสามารถที่จะนำมาประยุกต์ใช้งานตามที่ต้องการได้ง่าย

การใช้ภาษาเคลไฟในการเขียนซีจีไอ

ในการเลือกที่จะใช้ภาษาเคลไฟ เขียนภาษาเคลไฟ เราจำเป็นที่จะต้องใช้แพลทฟอร์ม ของเว็บ เซิร์พเวอร์ เป็นไมโครซอร์ฟวินโดว์ ซึ่งแพลทฟอร์มที่เห็นที่นิยมกันก็คือ วินโดว์เอ็นที , วินโดว์ 95

เว็บเซิร์ฟเวอร์ เป็นตัวที่เริ่มการทำงานของโปรแกรม ซีจีไอ ด้วยเหตุผลนี้เราจึงไม่มีความจำเป็นและความสามารถที่จะควบคุมการทำงานของ ซีจีไอได้ ดังนั้นการเขียนโปรแกรมซีจีไอโดยใช้เคลไฟจึงไม่จำเป็นต้อง สร้างยูสเซอร์อินเทอร์เฟส

ก่อนที่จะทำการเขียนโปรแกรมซีจีไอโดยใช้ ภาษาเคลไฟเราจะเริ่มทบทวนการส่ง ผ่านค่าของซีจีไอ และการเขียนโปรแกรมซีจีไอบนวินโดว์ก่อนเพื่อเป็นแนวทาง แล้วจึงอธิบายถึงการเขียนโปรแกรมซีจีไอโดยใช้ภาษาเคลไฟ

เว็บเซิร์พเวอร์มีการส่งค่าและรับค่า จากโปรแกรมซีจีไอ 2 วิธี

- คอมมานด์ไลน์ พารามิเตอร์ (command line paramiter)
- ดาต้าไฟล์ (data file)

ในกรณีที่เราทำงานโดยใช้ คอมมานด์ไลน์ พารามิเตอร์ จะมีคำสั่งรูปแบบต่อไปนี้

WinCGI-exe cgi-data-file

โดย WinCGI.exe เป็นชื่อของโปรแกรม ซีจีไอที่เขียนด้วยเคลไฟ เช่น c:\wibsite\cgi-win \ checkbal.exe ส่วน cgi-data-file เป็นค่าพารามิเตอร์ที่อยู่ในไฟล์ Window.INI ซึ่งเป็นไฟล์ที่เก็บค่าอินฟอร์เมชันต่างๆเกี่ยวกับการทำงานของ ยูสเซอร์และค่าต่าง ๆ ที่ต้องใช้สำหรับแพลทฟอร์ม สาเหตุที่เราต้องใช้ไฟล์ .INI เนื่องจากบนวินโดว์ ไม่มีส่วนที่เป็น stdin และ stdout เหมือนบน ยูนิกซ์

ตัวอย่าง .INI ไฟล์

[CGI]

Request Protocol=HTTP/1.0

Request Method=POST

Excutable Path=/cgi-win/quecgi1.exe

...

[System]

GMT Offset=-28800

Debug Mode=No

Output File=c:\temp\4cws.out

Content File=c:\temp\4cws.inp

...

บนวินโดว์โปรแกรมเมอร์สามารถที่จะเขียนและอ่านไฟล์ .INI ได้โดยใช้ Window API ฟังก์ชันดังต่อไปนี้

- GetPrivateProfileString( )
- GetPrivateProfileInt ( )
- WritePrivateProfileString ( )
- WritePrivateProfileInt ( )

จากการที่เราจะติดต่อไฟล์ .INI โดยผ่านทาง Window API นั้นเป็นการยาก แต่หากเราใช้ภาษาเดลไฟเราสามารถที่จะทำการติดต่อกับ .INI ไฟล์ได้โดยผ่านทาง Object ที่มีอยู่แล้วในภาษาเดลไฟ ซึ่งจะเป็นการง่ายกว่า ต่อไปเราจะทำการศึกษาเกี่ยวกับส่วนต่างๆ ภายในไฟล์ .INI

ภายใน cgi-data-file จะประกอบด้วย 8 ส่วน ซึ่งแต่ละส่วนก็จะมีค่าตัวแปรต่างกันไป ส่วนที่ 1 [CGI] Section ซึ่งภายในส่วนนี้จะประกอบด้วยค่าดังต่อไปนี้

- Request Protocol
- Request Method
- Executable Path
- Logical Path
- Physical Path
- Query String
- Request Range

- Referer
- From
- User Agent
- Content Type
- Content Length
- Content File
- Server Software
- Server Name
- Server Port
- CGI Version
- Remote Host
- Remote Address
- Authentication method
- Authentication Realm
- Authenticated Username

ส่วนที่ 2 [Accept] Section

ส่วนนี้จะประกอบด้วยส่วนของ เอ็ชทีทีพี รีเควส เฮดเดอร์ , ชนิดของข้อมูลที่รับ

ส่วนที่ 3 [System] Section ประกอบด้วยค่าต่อไปนี้

- GMT
- Debug Mode
- Output File
- Content File

ส่วนที่ 4 [Extra Header] Section

บราวน์เซอร์จะมีการส่งเอ็กซ์ตราอินฟอร์เมชั่นจากส่วนเฮดเดอร์ของเอ็ชทีทีพี มายังส่วนนี้

ส่วนที่ 5 [Form Literal] Section

จะเก็บค่าตัวแปรต่างๆ ที่ใช้ในการทำงานของซีจีไอและใช้ในการทำงาน

ส่วนที่ 6 [Form External] , ส่วนที่ 7 [Frm Huge] , ส่วนที่ 8 [Form File] เป็นส่วน ที่ไม่เก็บค่า

การอ่านค่าจาก cgi-data-file โดยใช้ Delphi's TiniFile Object

ภายในภาษาเดลไฟเราสามารถ ที่จะทำการอ่านข้อมูลจาก cgi-data-file ได้โดยใช้ TiniFile Oject โดยการทำงานร่วมกับ Method ReadString ( ) ดังตัวอย่างโปรแกรมต่อไปนี้

```
procedure TForm1.GetOutputFile;
var
  CgiIniFile : TIniFile;
  OutputFileName : String ;
begin
      CgiIniFile := TIniFile.Create( ParamStr (1) );
      OutputFileName := CgiIniFile.ReadString (' System ' , ' Output file ' , ' ' );
      CgiIniFile.Free;
end;
```

จากตัวอย่างโปรแกรมด้านบนเป็นการทำการอ่านค่าจาก cgi-data - file เพื่อนำค่ามาใช้งานต่อไปเราจะทำการเขียนโปรแกรม Cgi อย่างง่าย คือจะทำการรับชื่อและนามสกุลมาจาก โฮมเพจแล้ว ทำการส่งค่ากลับออกไปให้ผู้ใช้บริการ ในรูปของการแสดงชื่อแล้วนามสกุล :แต่ในการอธิบายนี้จะทำการเพียงบางส่วนที่สำคัญของโปรแกรม โดยอ้างอิงว่าผู้ที่ทำการศึกษาวิทยานิพนธ์เล่มนี้ มีพื้นฐานในการเขียนภาษาเดลไฟ อยู่แล้ว

ส่วนคำสั่งในการอ่านค่าตัวแปรจากฟอร์ม สมมติว่าตัวแปรที่รับค่าของชื่อ และ นามสกุล มาจากโฮมเพจ มีชื่อว่า firstname และ lastname ตามลำดับ ภายในโปรแกรมซีจีไอต้องมีการกำหนด ค่าตัวแปรขึ้นมาใหม่เพื่อจะรองรับค่าตัวแปร fisrtname และ lastname ที่ส่งมา ในที่นี้เรากำหนดค่าตัวแปร FirstName2 และ LastName2 ขึ้นมาเพื่อรับค่าที่ส่งผ่านมา ดังคำสั่งต่อไปนี้

```
FirstName2 := CgiIniFile.ReadString ( ' Form Literal ' , ' firstname' , ' ');
LastName2 := CgiIniFile.ReadString ( ' Form Literal ' , ' lastname ' , ' ' );
```

ต่อไปก็จะทำการอธิบายถึงคำสั่งที่จะใช้ในการส่ง ผลลัพธ์ในรูปของภาษา เอ็ชทีเอ็มแอล ไปยังบราวเซอร์ของยูสเซอร์โดยผ่านทาง outputfile โดยใช้คำสั่งต่อไปนี้

```
AssignFile ( outputfile , outputfilename );
Rewrite (outputfile);
Writeln (outputfile, ' <HTML> ' );
```

```
Writeln (outputfile, ' <HEAD><TITLE> Example </TITLE></HEAD>' );
Writeln (outputfile, ' <BODY> Hello ' + FirstName2 + ' ' + LastName2 + ' </BODY> ' );
Writeln (outputfile, ' </HTML>' );
Close (outputfile);
```

จากตัวอย่างข้างต้นเป็นเพียงการยกตัวอย่างที่สำคัญในการเขียนโปรแกรมซีจีไอเท่านั้น หากผู้ที่ทำการศึกษาต้องการทราบรายละเอียดมากขึ้นก็สามารถ ศึกษาได้จากหนังสืออ้างอิง ท้ายวิทยานิพนธ์ และหากต้องการตัวอย่างโปรแกรมซีจีไอที่เขียนด้วย ภาษาเดลไฟ ที่มีการทำงาน ที่ซับซ้อนมากกว่าภายในตัวอย่างก็สามารถที่จะ ทำการศึกษาได้จากรายงาน source code ของกลุ่ม โครงงานค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์

2.3 ส่วนจัดเก็บและค้นหาข้อมูล

การทำงานของเซิร์สเอนจิน จำเป็นที่จะต้องมีการสร้างส่วนที่ใช้จัดเก็บข้อมูล ที่ได้จาก โรบอท โดยข้อมูลที่จะนำมาเก็บนี้จะต้องเป็นข้อมูลที่ได้รับการวิเคราะห์ แล้ว และมีความสำคัญมากพอที่จะนำมาเก็บไว้

การจัดเก็บข้อมูลสามารถทำได้หลายวิธี ไม่ว่าจะเป็นการจัดสร้างโครงสร้างของข้อมูล เพื่อใช้ในการจัดเก็บขึ้นมาเองหรือ จะใช้งานระบบฐานข้อมูล อยู่แล้วก็ได้ ทั้งนี้ขึ้นอยู่กับความ เหมาะสมของการใช้งาน สำหรับโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ เลือก การจัดเก็บข้อมูลบนระบบฐานข้อมูล ดังจะอธิบายถึงทฤษฎี หลักการและ เหตุผลต่อไป

2.3.1 ระบบฐานข้อมูล (Database System)

ความหมายและที่มา

ฐานข้อมูล คือแหล่งเก็บข้อมูล และความสัมพันธ์ระหว่างฐานข้อมูลเหล่านี้

ในอดีตนั้นการจัดเก็บและค้นหาข้อมูลยังเก็บอยู่ในรูปแบบไฟล์ ซึ่งก็ได้งานได้ในระบบงานขนาดเล็ก มีผู้ใช้งานอยู่เพียงหนึ่งคนหรือไม่กี่คน แต่ในระบบที่มีขนาดใหญ่ขึ้นมีผู้ใช้มากขึ้น การค้นหาและจัดเก็บข้อมูลจึงต้องการประสิทธิภาพที่สูงขึ้น สาเหตุที่ทำให้มีการนำระบบฐานข้อมูลมาใช้แทนระบบไฟล์แบบเก่ามีดังนี้

- ลดความซ้ำซ้อนของข้อมูล

การที่แผนกต่างๆ ต้องการข้อมูลชนิดเดียวกัน แต่ต่างก็แยกกันเก็บทำให้สิ้นเปลืองเนื้อที่ในการจัดเก็บข้อมูลหน่วยความจำ

- หลีกเลี่ยงการขัดแย้งกันเองของข้อมูล

เมื่อมีการเปลี่ยนแปลงข้อมูลที่จัดเก็บในหลายๆที่หลาย ๆ ไฟล์ จะต้องตามทำการเปลี่ยนแปลงทุกไฟล์ที่จัดเก็บ ถ้าทำไม่ครบทุกไฟล์ก็จะทำให้ข้อมูลมีความผิดพลาดไปจากความจริง และขัดแย้งกัน

- แก้ปัญหาการรักษาความปลอดภัยของข้อมูล

เนื่องจากมีการเก็บข้อมูลกระจัดกระจาย ทำให้ยากต่อการรักษาความปลอดภัยเสี่ยงต่อการรั่วไหลและดัดแปลงข้อมูล

- ทำให้ข้อมูลมีบูรณภาพมากที่สุด หรือการควบคุมความถูกต้องของข้อมูล
- ทำให้สามารถควบคุมการใช้งานโดยส่วนกลางได้
- สามารถใช้ข้อมูลร่วมกันได้
- มีความคล่องตัวและความยืดหยุ่นในการใช้งาน
- มีการควบคุมมาตรฐานร่วมกัน

ระบบจัดการฐานข้อมูล (Database Management System)

ระบบจัดการฐานข้อมูล คือ ซอฟต์แวร์ที่เปรียบเสมือนสื่อกลางระหว่างผู้ใช้และโปรแกรมต่าง ๆ ที่เกี่ยวข้องกับการใช้ฐานข้อมูล หรือ

ระบบจัดการฐานข้อมูล คือ ซอฟต์แวร์ที่ทำหน้าที่จัดการการเข้าถึงทุกอย่างสู่ฐานข้อมูล

เนื่องจากการใช้และควบคุมดูแลฐานข้อมูลเป็นเรื่องที่ซับซ้อนยุ่งยาก ระบบจัดการฐานข้อมูลจึงเป็นเครื่องมือสำคัญในการที่จะลดภาระของผู้ใช้ไปอย่างมาก ทำให้การใช้งานระบบฐานข้อมูลมีประสิทธิภาพสูงขึ้น

โมเดลของข้อมูล

โมเดลของข้อมูล ได้แก่ประเภทของระบบการจัดการฐานข้อมูล แบ่งออกโดยจำแนกตามการแสดงความสัมพันธ์ระหว่างข้อมูลได้ 3 โมเดล ดังนี้

1 แบบโครงข่าย (Network Model)

เป็นการแสดงความสัมพันธ์ด้วยลิงก์สิสต์เป็นโครงข่าย โดยมีตัวชี้ ชี้ระหว่างข้อมูลเชื่อมกันเป็นชุด มีความสัมพันธ์แบบแมนี่ทูแมนี่

2. แบบแผนภูมิต้นไม้ (Hierachical Model)

ทำการแทนความสัมพันธ์ระหว่างข้อมูลในลักษณะของแผนภูมิต้นไม้ โดยมีความสัมพันธ์ในลักษณะวันทูแมนี่

3. แบบสัมพันธ์ (Relational Model)

เป็นการเก็บข้อมูลที่มีลักษณะการเก็บในรูปตาราง 2 มิติธรรมดา คือ มีแถวและคอลัมน์

สถาปัตยกรรมของระบบฐานข้อมูล (Architecture for a Database System)

ลักษณะสถาปัตยกรรมของระบบฐานข้อมูล ได้ถูกกำหนดให้เป็นมาตรฐานจาก 3 องค์กร หลัก คือ ไอเอสโอ , ไดเอฟไอพี , เอเอ็นเอสไอ โดยสามารถจัดแบ่งออกเป็น 3 ระดับดังนี้

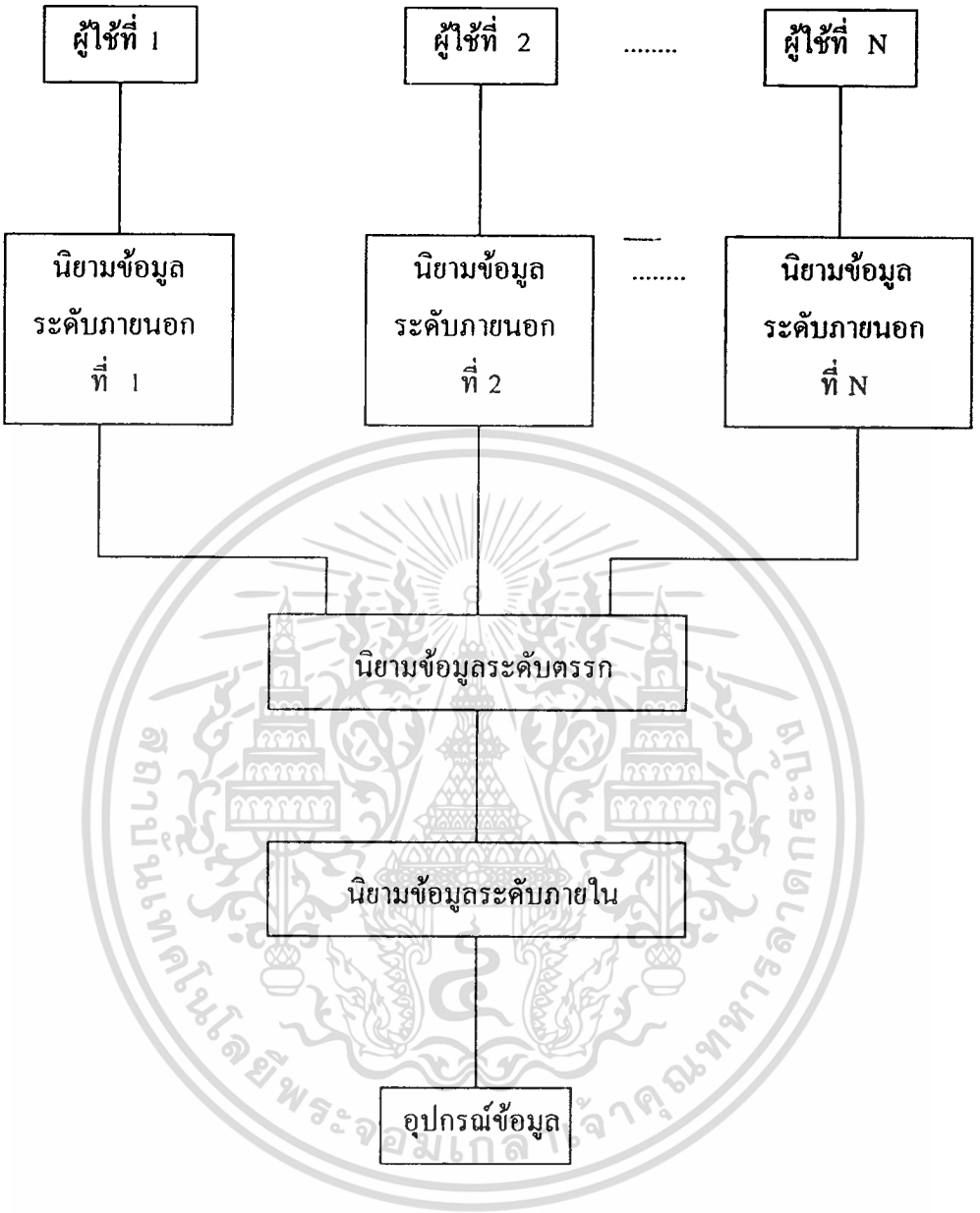
1 นิยามข้อมูลระดับภายนอก (External Schema)

2 นิยามข้อมูลระดับแนวคิด (Conceptual Schema)

3 นิยามข้อมูลระดับภายใน (Internal Schema)

ทั้งสามระดับจะมีความสัมพันธ์กัน ดังภาพที่จะแสดงในรูปที่ 2.2 โดยรายละเอียดของส่วนต่าง ๆ มีดังนี้

1. ผู้ใช้ : ผู้ใช้ขั้นสุดท้าย , คนเขียนโปรแกรม , โปรแกรมใช้งาน
2. นิยามข้อมูลระดับภายนอก : โครงสร้างข้อมูลที่ผู้ใช้แต่ละคนมองเห็น
3. นิยามข้อมูลระดับแนวความคิด : เป็นส่วนกำหนดลักษณะ , ขนาดและโครงสร้างของข้อมูล และความสัมพันธ์ระหว่างข้อมูลทั้งหมดที่อยู่ในขอบเขต
4. นิยามข้อมูลระดับภายใน : โครงสร้างข้อมูลที่จัดเก็บในอุปกรณ์เก็บข้อมูล , ลักษณะการเก็บข้อมูล
5. ฐานข้อมูลทางกายภาค : อุปกรณ์เก็บข้อมูล เช่น ฮาร์ดดิสก์



รูปที่ 2.2 สถาปัตยกรรมของระบบฐานข้อมูล

หรืออาจมองสถาปัตยกรรมดังกล่าวออกเป็นระดับ ๆ ดังนี้

1. ระดับกายภาพ หรือ ระดับภายใน (Physical level , Internal level)
2. ระดับแนวคิด หรือ ระดับตรรก (Conceptual level , Logical lever)
3. ระดับภายนอก หรือ ระดับผู้ใช้ (External level ,User level)

ผู้ใช้ระบบฐานข้อมูล

ผู้ใช้ในระบบฐานข้อมูลสามารถแบ่งได้เป็น 3 กลุ่มดังนี้

1 ผู้ใช้งานขั้นสุดท้าย (End-user)

ได้แก่ ผู้ที่จะได้รับข่าวสารที่เหมาะสมตามชนิดของงาน และความต้องการของตนจากฐานข้อมูล โดยทั่วไปจะเป็นผู้ที่มีความรู้ทางคอมพิวเตอร์และฐานข้อมูลน้อยมาก

2 คนเขียนโปรแกรมใช้งานฐานข้อมูล (Application Programmer)

จะเป็นผู้เขียนโปรแกรมใช้งานฐานข้อมูล ให้เป็นไปตามความต้องการของผู้ใช้งานขั้นสุดท้าย

3 ผู้บริหารฐานข้อมูล (Database Administrator)

เป็นผู้ที่ทำหน้าที่รับผิดชอบควบคุมระบบ ฐานข้อมูลทั้งหมด โดยมีคุณสมบัติหรือหน้าที่ดังนี้

- เป็นผู้เชี่ยวชาญทางด้านเทคนิคระดับสูงและใช้ระบบจัดการฐานข้อมูลเป็น
- เป็นผู้ออกแบบนิยามข้อมูลระดับแนวความคิดของทั้งระบบงาน
- เป็นผู้ออกแบบนิยามข้อมูลระดับแนวความคิดของทั้งระบบงาน
- เป็นผู้จัดการนิยามข้อมูลระดับภายนอกให้แก่ผู้ใช้ขั้นสุดท้ายแต่ละคน รวมทั้งการให้อำนาจที่เหมาะสมแก่ผู้ใช้ขั้นสุดท้าย
- เป็นผู้พิจารณาเลือกทฤษฎีการเข้าถึงข้อมูล ที่เหมาะสมรวมทั้งอุปกรณ์ที่จะใช้ในการจัดเก็บข้อมูลด้วย
- เป็นผู้จัดการปรับปรุงการทำงานของระบบ
- เป็นผู้กำหนดรูปแบบในการตรวจสอบความถูกต้องแน่นอนของข้อมูล
- เป็นผู้กำหนดวิธีการในการเก็บข้อมูลสำรองและการนำกลับมาใช้หม่
- เป็นผู้คอยติดต่อผู้ใช้ขั้นสุดท้าย เพื่อให้การทำงานของผู้ใช้ขั้นสุดท้ายทำไปได้อย่างมีประสิทธิภาพครงตามความต้องการ

2.3.2 ข้อดีของการจัดเก็บข้อมูลแบบฐานข้อมูล

การจัดเก็บข้อมูลโดยใช้ระบบฐานข้อมูลเข้ามาช่วยนั้นเป็นวิธีการ ที่ได้รับความนิยม อย่างแพร่หลาย เนื่องจากมีประสิทธิภาพและการใช้งานที่เหมาะสมดังมีข้อดีต่อไปนี้

1 สามารถลดความซ้ำซ้อนของข้อมูล

การจัดเก็บข้อมูลโดยวิธีอื่น เช่น การจัดเก็บข้อมูลเป็นไฟล์ อาจทำให้เกิดปัญหาการจัดเก็บข้อมูลซ้ำซ้อนกัน เช่น การที่ผู้ใช้งานทุกคนต่างก็ทำการคัดลอกไฟล์ดังกล่าวเก็บไว้เพื่อใช้งาน ซึ่ง

การซ้ำซ้อนในการเก็บข้อมูลนี้ จะทำให้เกิดการเสียเนื้อที่ในการจัดเก็บข้อมูล โดยไม่จำเป็นซึ่ง หากใช้การเก็บข้อมูลแบบฐานข้อมูล สามารถที่จะช่วยลดปัญหาดังกล่าวได้ เนื่องจากการจัดเก็บ ข้อมูลโดยใช้ฐานข้อมูล จะมีการออกแบบและควบคุมข้อมูลที่จัดเก็บเพื่อลดความซ้ำซ้อน และ นอกจากนี้จุดประสงค์หลักของการใช้ ฐานข้อมูลก็คือ การลดความซ้ำซ้อนของข้อมูลให้มากที่สุด

2 สามารถควบคุมความถูกต้องของข้อมูลได้

การจัดเก็บข้อมูลโดยวิธีทั่วไป อาจเกิดปัญหาในการทำการควบคุมความถูกต้องของข้อมูล อันอาจจะมีผลเนื่องจากการเก็บข้อมูล ซ้ำซ้อนทำให้ผู้ใช้งาน แต่ละคนสามารถทำการเปลี่ยนแปลงข้อมูลได้ ซึ่งทำให้ข้อมูลในแต่ชุดที่คัดลอกไปอาจจะไม่เหมือนกัน จนไม่อาจจะแยกแยะได้ว่าข้อมูลใดเป็นข้อมูลที่ถูกต้อง แต่หากมีการใช้การเก็บข้อมูลโดยใช้ ฐานข้อมูลจะมีการควบคุมการเปลี่ยนแปลงข้อมูล โดยจะอนุญาตให้มีเพียงผู้ใช้งาน เพียงคนเดียวในการทำการเปลี่ยนแปลงแต่ละครั้ง ทำให้สามารถควบคุมความถูกต้องของข้อมูลได้

3 สามารถใช้ข้อมูลร่วมกันได้

การที่ทำการจัดเก็บข้อมูลโดยใช้ ฐานข้อมูลทำให้ข้อมูลที่จัดเก็บ สามารถที่จะใช้ร่วมกันได้ระหว่างผู้ใช้หลายคน โดยที่สามารถควบคุมความถูกต้องของข้อมูลได้ ในขณะที่หากเราใช้ การเก็บข้อมูลแบบอื่น อาจมีปัญหาในการใช้ข้อมูลร่วมกันในเรื่องของความถูกต้องของข้อมูล

4 สามารถตรวจสอบความขัดแย้งกันของข้อมูลที่จัดเก็บได้

5 การรักษาความปลอดภัยของข้อมูลสามารถทำได้อย่างมีประสิทธิภาพ

2.3.3 เปรียบเทียบการเก็บข้อมูลโดยใช้โครงสร้างข้อมูลที่สร้างเองกับใช้ฐานข้อมูล

การจัดเก็บข้อมูลของโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ นี้ทางผู้วิจัยเลือกที่จะใช้การเก็บโดยระบบฐานข้อมูล เนื่องจากสาเหตุต่อไปนี้

1 การจัดสร้างโครงสร้างข้อมูลเพื่อใช้ในการจัดเก็บข้อมูลสามารถทำได้ยาก เนื่อง จากต้องออกแบบให้เหมาะสมกับการใช้งาน ต่างจากการใช้ระบบฐานข้อมูลที่มีโปรแกรมสำเร็จ รูปอยู่แล้วจึงเป็นการสะดวกที่จะนำมาใช้งาน

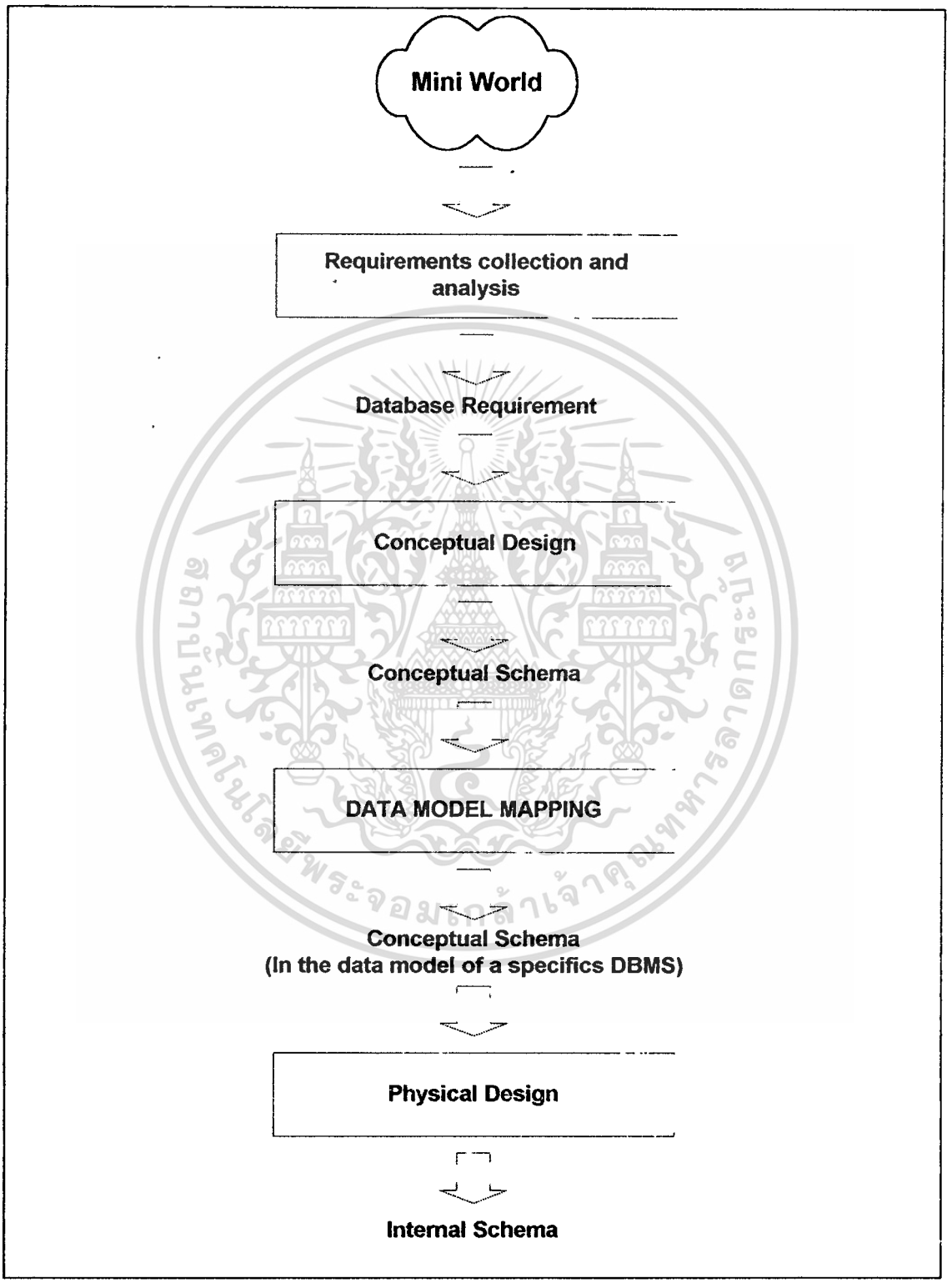
2 โครงสร้างข้อมูลที่จัดสร้างขึ้นเองไม่สามารถที่ยืนยัน ความถูกต้องของข้อมูลที่ เกิดจากการเปลี่ยนแปลงโดยผู้ใช้ได้ หากว่าการออกแบบไม่เหมาะสม ในขณะที่ระบบฐานข้อมูล จะมีการตรวจสอบความถูกต้องของข้อมูลอยู่แล้ว

3 การใช้โครงสร้างข้อมูลที่จัดสร้างขึ้นเอง มีปัญหาเรื่องของประสิทธิภาพในการ ทำงานที่ไม่สามารถควบคุมได้ หรือหมายความว่ายังมีการทำงานได้ไม่ดี ในเรื่องของความเร็ว ในการค้นหาข้อมูล,ความยากง่ายในการเปลี่ยนแปลงข้อมูล

4 เวลาที่ใช้ในการจัดสร้างโครงสร้างข้อมูลขึ้นเองนั้น จะใช้เวลามากซึ่งอาจจะเป็นอุปสรรคทางด้านเวลาสำหรับการจัดทำ โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์

ในการที่จะจัดสร้างฐานข้อมูลขึ้นมาใช้งาน เราจำเป็นที่จะต้องรวบรวมความต้องการของผู้ใช้งานแล้วทำการออกแบบฐานข้อมูล แต่การออกแบบจำเป็นที่จะต้องมีหลักในการออกแบบ ซึ่งหลักในการออกแบบ มีอยู่หลายวิธีไม่ว่าจะเป็นไนแอม (niam) , นอร์มอลไลซ์ (normalize) อีอาร์โมเดล (er-model) ซึ่งในการทำโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ นั้นใช้วิธีการออกแบบที่เรียกว่า อีอาร์โมเดล

2.3.4 การออกแบบโดยใช้ อีอาร์ โมเดล



รูปที่ 2.3 ขั้นตอนในการออกแบบโดยอาศัยหลักการของ อีอาร์โมเดล

การออกแบบโดยอาศัยหลักการดาต้าโมเดลจะมีหลักการทำงาน ตามขั้นตอนต่อไปนี้

การรวบรวมและวิเคราะห์ความต้องการของผู้ใช้

ขั้นตอนการรวบรวมและวิเคราะห์ความต้องการของผู้ใช้นี้ ผู้ออกแบบระบบฐานข้อมูล จะทำการสัมภาษณ์ลักษณะของฐานข้อมูลที่ต้องการจากผู้ที่เกี่ยวข้องและจัดทำรายงานสรุป ลักษณะฐานข้อมูลที่ต้องการ

การสร้างแบบจำลองของระบบฐานข้อมูล

ขั้นตอนการสร้างแบบจำลองของระบบฐานข้อมูลจะทำการสร้างแบบจำลองของระบบฐานข้อมูลโดยใช้ ดาต้าโมเดลระดับสูง ซึ่งแบบจำลองที่สร้างขึ้นจะอธิบายรายละเอียดต่างๆ ของความต้องการที่ได้จากขั้นตอนแรก การสร้างแบบจำลองนี้จะเป็นประโยชน์ในการที่จะใช้ช่วยอธิบายการจัดเก็บข้อมูลให้แก่ผู้ที่ไม่มีความรู้ทางเทคนิคได้เข้าใจมากยิ่งขึ้น

การจัดสร้างระบบฐานข้อมูล

ขั้นตอนการจัดสร้างระบบฐานข้อมูล เป็นการจัดสร้างระบบฐานข้อมูลขึ้นมาเพื่อใช้งานจริง โดยจัดสร้างจากโปรแกรมสำเร็จรูปทางด้านฐานข้อมูลที่มีอยู่โดยทั่วไป โดยผู้ที่ทำการจัดสร้างฐานข้อมูลจะใช้แบบจำลองฐานข้อมูลที่ได้จัดทำไว้เป็นต้นแบบ

การออกแบบในระดับฟิสิคอล

ขั้นตอนการออกแบบในระดับฟิสิคอล จะมีการทำการกำหนดโครงสร้างภายในของระบบฐานข้อมูลที่ใช้ในการจัดเก็บข้อมูลอาจรวมถึงการจัดการไฟล์ที่ใช้ภายในระบบฐานข้อมูล

หลักการของอีอาร์โมเดล

ในการออกแบบตามหลักของอีอาร์โมเดล จะต้องเข้าใจส่วนประกอบต่างๆ ของ อีอาร์โมเดล ไดอะแกรมว่ามีความหมายว่าอย่างไร

เอนติตี้ (Entities) และ แอททริบิว (Attribute)

ภายในการออกแบบโดยอาศัย อีอาร์โมเดล จะมีการแสดงสิ่งต่างๆ ที่มีอยู่บนโลก บนไดอะแกรม โดยเรียกว่า เอนติตี้ เช่น คน บ้าน รถ หรืออาจจะเป็นสิ่งที่ถูกสร้างขึ้น เช่น บริษัท มหาวิทยาลัย , งาน แต่ละเอนติตี้จะมีคุณลักษณะของตัวเอง โดยสิ่งที่ใช้แสดงคุณสมบัติดังกล่าวนี้ เรียกว่า แอททริบิว เช่น เอนติตี้ คน อาจจะมีแอททริบิวของ ชื่อ , ที่อยู่ , อายุ เป็นต้น บางแอททริบิว ยังสามารถที่จะแบ่งต่อออกไปได้อีก ซึ่งแอททริบิวที่มีลักษณะดังกล่าวนี้ เราเรียก ว่า กลุ่มแอททริบิว (composite attribute) ดังแสดงตัวอย่างต่อไปนี้

ตัวอย่าง composite attribute

ประเภทของเอนติตี้ (Entities Type)

ในระบบฐานข้อมูล มักจะประกอบด้วยกลุ่มของแอททริบิว ที่คล้าย ๆ กัน เช่นภายในบริษัท อาจจะมีพนักงานจำนวนมาก โดยแต่ละคนจะมีข้อมูลคล้าย ๆ กัน นั่นคือพนักงานเหล่านี้เป็นเอนติตี้ ประเภทเดียวกันมี แอททริบิว เหมือนกันจะต่างกันก็แต่เพียงค่าของ แอททริบิว ของพนักงานแต่ละคน ดังนั้นเราจึงมีการกำหนดให้กลุ่มของแอททริบิวที่คล้ายกันรวมกันอยู่ในเอนติตี้ไทพ์เดียวกัน เช่น

พนักงาน	บริษัท
ชื่อ , อายุ , เงินเดือน	ชื่อ , ที่อยู่ , ประธาน
e1	e1
(นาย ก , 21 , 3000)	(ปูนไทย , บางพลี , นายสมชาย)
(นาย ค , 31 , 5600)	(กระเบื้องไทย , กรุงเทพ , นายปาน)

คีย์แอททริบิว (Key Attribute)

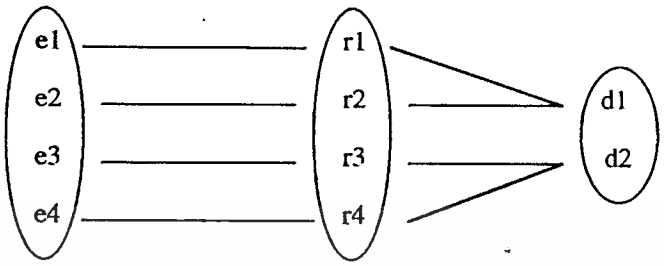
คีย์แอททริบิวเป็นส่วนที่สำคัญมาก สำหรับเอนติตี้เพราะคีย์ แอททริบิว เป็นตัวที่สามารถแยกแยะ เอนติตี้ แต่ละตัวภายในเอนติตี้ไทพ์เดียวกันได้ เช่น เอนติตี้ไทพ์ บริษัทอาจจะมีชื่อเป็นคีย์แอททริบิว

ความสัมพันธ์ (Relationship)

ในระบบฐานข้อมูลจะประกอบด้วยหลาย เอนติตี้ไทพ์ ซึ่งแต่ละเอนติตี้ในระบบ ฐานข้อมูล ก็จะมีความสัมพันธ์กับเอนติตี้ตัวอื่น เช่น เอนติตี้พนักงาน มีความสัมพันธ์กับเอนติตี้ บริษัท โดยจะต้องมีการกำหนดความสัมพันธ์ระหว่างเอนติตี้ออกมาให้ชัดเจน เช่น ความสัมพันธ์ ระหว่างเอนติตี้พนักงานและเอนติตี้บริษัทจะอยู่ในรูปของความสัมพันธ์ ที่เรียกว่า ทำงานให้ (work_for) ซึ่งความสัมพันธ์นี้ อาจจะเป็นแบบหนึ่งต่อหนึ่ง , หนึ่งต่อหลาย หรือ หลายต่อหลาย ก็ได้แล้วแต่ความสัมพันธ์

ตัวอย่าง ความสัมพันธ์

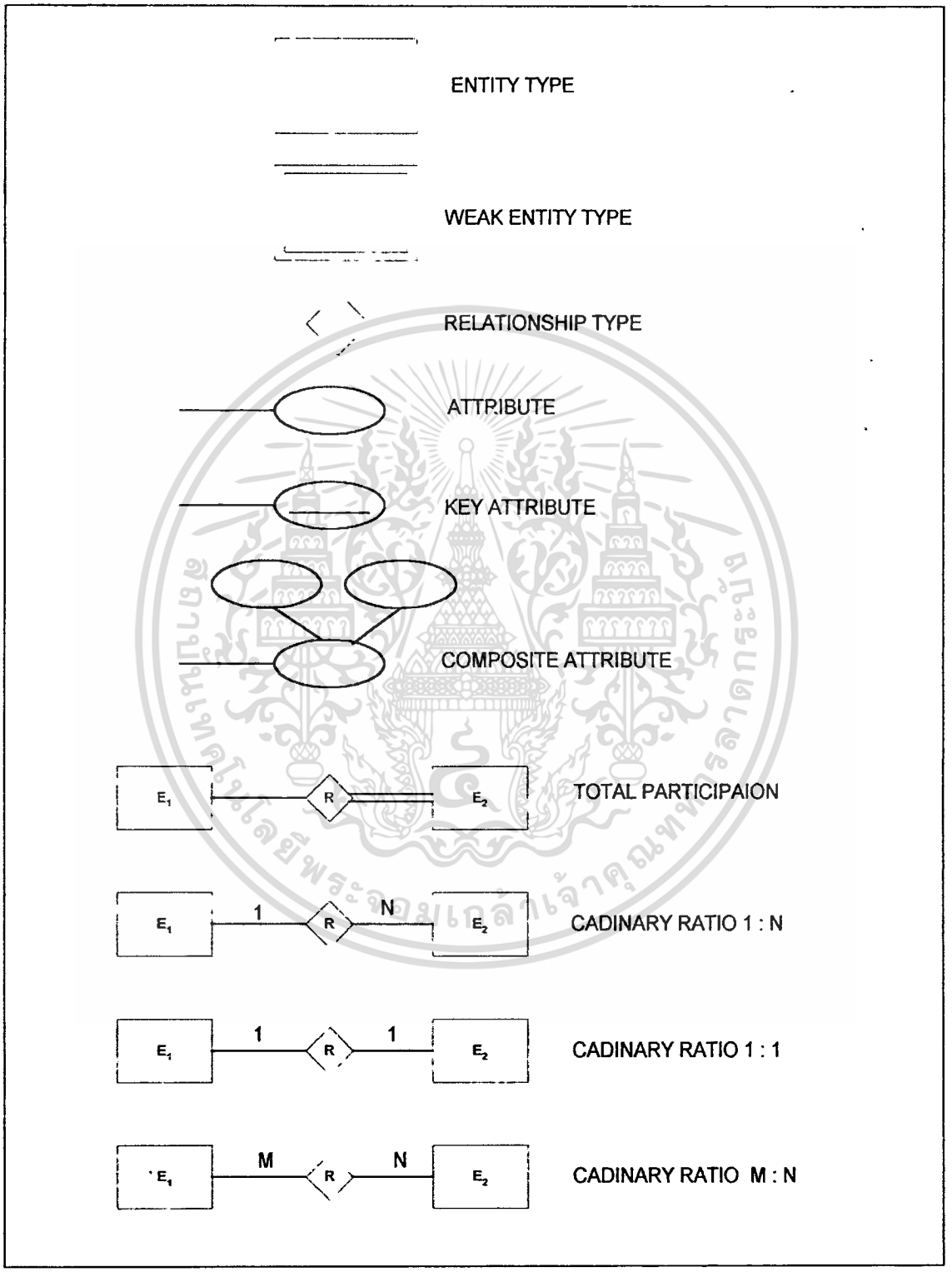
เอนติตี้พนักงาน ความสัมพันธ์ทำงานให้ เอนติตี้บริษัท



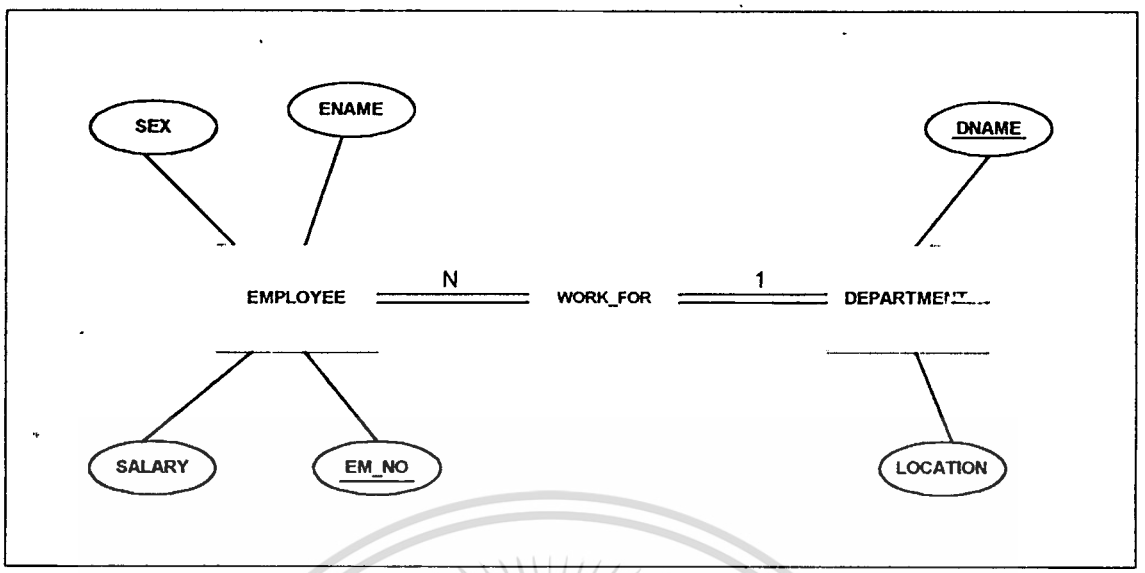
นอกจากนี้ยังมีการกำหนด ลักษณะพิเศษต่างๆ เพิ่มอีก เช่น จากตัวอย่าง ถ้าบริษัทมีการกำหนดว่าพนักงานทุกคนต้องทำงานให้แผนก ดังนั้นเอนติตี้พนักงานจะอยู่ได้ก็ต่อเมื่อมีความสัมพันธ์ทำงานให้เกิดขึ้น เราเรียกความสัมพันธ์ของเอนติตี้พนักงานกับความสัมพันธ์ ทำงานให้ว่าเป็นแบบโทโทล นั่นหมายความว่าทุกเอนติตี้ในเซ็ตของเอนติตี้พนักงานจะต้องมีความสัมพันธ์กับ เอนติตี้แผนก ผ่านทางความสัมพันธ์ทำงานให้

บางเอนติตี้ ก็ไม่มีคีย์ แอททริบิวของตนเอง นั่นแสดงว่า เราไม่สามารถที่จะทำการแยกความแตกต่างระหว่างเอนติตี้ได้เพราะว่าแม้ว่าเราจะใช้เป็น คอมบิเนชั่นคีย์ (combination key) ก็ไม่สามารถแยกแยะได้เพราะต่างก็มีค่าคีย์เหมือนกันเราเรียกกรณีนี้ว่า วีคเอนติตี้ (Weak Entities)

อีอาร์โมเดลไดอะแกรม



รูปที่ 2.4 แสดงสัญลักษณ์ที่ใช้ในการสร้างอีอาร์โมเดลไดอะแกรม



รูปที่ 2.5 ตัวอย่างอีอาร์โมเดลไดอะแกรม

เมื่อทำการออกแบบฐานข้อมูลเรียบร้อยแล้ว ก็ทำการเลือกระบบฐานข้อมูลที่จะนำมาใช้ในการเก็บข้อมูลโดยในโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ ตัดสินใจเลือกที่จะใช้ Interbase ซึ่งเป็นฐานข้อมูลที่อยู่บนเคลไฟในการเก็บข้อมูล เนื่องด้วยเหตุผลต่อไปนี้

1. การติดต่อระหว่างโปรแกรมต่างๆ ภายในโครงงานสามารถทำได้ง่าย เนื่องจากเขียนด้วยภาษาเคลไฟเหมือนกัน

2. การเลือกใช้ Database Server ที่มีประสิทธิภาพสูงจะมีปัญหาในการจัดหาเครื่องมา ติดตั้งโปรแกรม

3. หากเลือกใช้ Database Server ที่มีประสิทธิภาพในการเก็บข้อมูล อาจมีปัญหาในการเชื่อมต่อทางด้าน Database (ODBC)

4. เนื่องจากข้อมูลที่ผู้จัดทำโครงงาน สามารถรวบรวมมาได้ในขณะนี้ยัง มีปริมาณไม่มาก ดังนั้นการเลือกใช้ Interbase ในเคลไฟ จึงเป็นการสะดวกในการศึกษาและจัดทำโครงงาน

2.3.5 อินเทอร์เบส (Interbase)

อินเทอร์เบสคือ Relational Database Management System (RDBMS) ที่สามารถทำงานได้หลายลักษณะไม่ว่าจะเป็นการทำงานในลักษณะที่มี ผู้ใช้งานเพียงคนเดียวหรือผู้ใช้งานหลายคน อินเทอร์เบส สามารถรันได้บนแพลทฟอร์มต่อไปนี้ วินโดว์ 95 , วินโดว์ เอ็นที , โนเวล เน็ตแวร์ หรือสามารถทำงานได้บน ระบบยูนิกซ์ การทำงานของอินเทอร์เบสสามารถทำงานได้ 2 ระบบ

1. Local Interbase, Single-user

ใช้เป็นโลคอล เซิร์พเวอร์สำหรับกรณีที่มีผู้ใช้งานเพียงคนเดียว ซึ่งการทำงานของ โลคอล อินเทอร์เบสนี้ สามารถเชื่อมต่อได้กับ เดลไฟ หรือ วิชวลเบสิค โดยการทำงานของโลคอล อินเทอร์เบสนี้ สามารถทำการติดต่อได้ผ่าน Desktop SQL Application หรือทำการติดต่อมาจาก รีโมท เซิร์พเวอร์

2. Interbase Server

ในกรณีที่มี ผู้ใช้งานมากกว่า 1 คนจำเป็นที่จะต้องใช้การทำงานแบบ อินเทอร์เบส เซิร์พเวอร์ โดยจะมีการทำงานเหมือนเป็นรีโมทเซิร์พเวอร์ ซึ่งสามารถที่จะทำงานได้อย่างมี ประสิทธิภาพ รวมทั้งยังสามารถที่จะติดต่อกับเซิร์พเวอร์ ของคาต้าเบส เช่น ออราเคิล , เอสคิวแอล เซิร์พเวอร์

เมื่อเราทำการออกแบบฐานข้อมูล และเลือกเซิร์พเวอร์ ที่จะใช้ทำงาน แล้วก็จัดทำการสร้าง ตามที่ได้ออกแบบไว้ หลังจากนั้นก็ทำการเตรียมส่วนที่จะใช้ในการดึงข้อมูล ที่ต้องการออกมา ซึ่ง การดึงข้อมูลออกจากฐานข้อมูล เราจะใช้ภาษาที่มีชื่อเรียกว่า ภาษาเอสคิวแอล

2.3.6 ภาษาเอสคิวแอล

Structured Query Language หรือที่เรียกกันว่าภาษา เอสคิวแอล เป็น Standard Relational Languare ที่ใช้ในการจัดการข้อมูลที่เก็บอยู่บนฐานข้อมูลแบบ รีเลชันแนล คาต้าเบส ภาษาเอสคิวแอลเป็นภาษา ที่ไม่ case sensitive คือไม่ว่าจะพิมพ์เป็นตัวเล็ก หรือ ตัวใหญ่ ก็จะได้ตัวใหญ่ออกมาเสมอ เราสามารถแบ่งประเภทการทำงานของ ภาษาเอสคิวแอลได้ 2 mode การทำงาน

1. Interactive Mode คือ Mode ที่ให้ผู้ใช้งานใส่คำสั่ง SQL ทางคีย์บอร์ด เมื่อรับคำสั่งแล้ว ก็จะมีการแสดงผลลัพธ์ออกมาตรง ๆ ซึ่งการทำงานใน mode นี้จะมีประโยชน์มาก ในกรณีที่เราจะใช้ในการทดสอบคำสั่ง การตอบคำถามเฉพาะกิจในงานลักษณะของ MIS

2. Embedded Mode คือ การใช้คำสั่ง SQL ร่วมกับภาษา host เช่น หากอยากได้หน้าจอสวย ก็ต้องใช้ภาษาอื่นเข้าช่วย

รูปแบบของภาษาเอสคิวแอล

คำสั่งของภาษาเอสคิวแอลจะรูปแบบที่แน่นอน ดังจะมีรูปแบบโดยรวมดังนี้

```
Select < colname > | * |
From < table name >
```

Where < row condition >

Group by < colname >

Having < group condition >

Order by < column list >

ต่อไปก็จะทำการอธิบายหลักการโดยรวมของภาษา เอสคิวแอล โดยสองคำสั่งแรกคือ คำสั่ง Select , From จะต้องมีเสมอทุกครั้งที่มีการใช้คำสั่ง SQL ส่วนคำสั่ง Where เป็นคำสั่งที่ช่วยให้เราสามารถที่จะระบุ ได้ว่า แถวใดบ้างที่เราต้องการเป็นคำตอบ และแถวประเภทใดที่ต้องแยกออก ส่วนคำสั่ง Group by , Having จะเป็นคำสั่งที่ใช้ในการจัดกลุ่ม แต่แตกต่างกันที่การใช้งานและเงื่อนไข ที่ใช้ สำหรับในกรณีที่เราต้องการที่จะจัดลำดับการทำงานของผลลัพธ์ เราสามารถทำได้โดยการใช้คำสั่ง Order by

2.4 ส่วนหน้าจอรับข้อมูลและแสดงผล

ส่วนหน้าจอรับข้อมูลและแสดงผลนั้น บนระบบเครือข่ายคอมพิวเตอร์ (อินเทอร์เน็ต) นี้หมายถึง โฮมเพจ ซึ่งเป็น เอกสารบนเว็บ เราจะต้องทำการจัดสร้างโฮมเพจของเซิร์สเอนจิน ขึ้นมาเพื่อที่จะใช้ในการติดต่อรับข้อมูลจากผู้ใช้บริการ โดยข้อมูลที่รับก็คือ คำที่ต้องการค้นหา เงื่อนไขที่จะใช้ในการค้นหาข้อมูล

นอกจากนี้ส่วนนี้ยังใช้ในการแสดงผลลัพธ์ที่ได้จากการค้นหาข้อมูลในระบบฐานข้อมูล โดยจะทำการรับส่งค่าต่าง ๆ ผ่านทางโปรแกรมซีจีไอ

การทำการสร้างโฮมเพจนั้นเราจะใช้ภาษาที่เรียกว่า ภาษาเอ็ชทีเอ็มแอล ดังนั้นผู้จัดทำโครงงานจึงศึกษาวิธีการเขียนโฮมเพจโดยใช้ภาษาเอ็ชทีเอ็มแอล ดังนี้

2.4.1 ภาษาเอ็ชทีเอ็มแอล

HTML (HyperText Markup Language) ถูกออกแบบมาเพื่อใช้กำหนดรูปแบบของเอกสารในระดับ โลจิคอล ซึ่งแตกต่างจากเวิร์ด โพรเซสเซอร์โดยทั่วไป เนื่องจากเอกสารเดียวกันนี้สามารถดูได้จากบราวเซอร์ต่างชนิดกัน

คำสั่ง ในHTML ถูกเรียกว่า `elements' สามารถแบ่งออกได้เป็น 2 ประเภท

-- กำหนดการแสดงผลของเอกสารในส่วน บอดี้ โดยบราวเซอร์

กำหนดข้อมูลเกี่ยวกับเอกสารเช่นความสัมพันธ์กับเอกสารอื่นหรือหัวข้อของเอกสาร

อีลีเมนในเอกสารแบบเอ็ชทีเอ็มแอล

คำสั่งที่สามารถทำงานได้โดย HTML ถูกเรียกว่า HTML อีลีเมนท์ หรือเรียกย่อ ๆ ว่า Tag ใช้ในการกำหนดรูปแบบของข้อมูล ซึ่งมีรูปแบบพื้นฐาน ดังนี้

<element_name>

เป็น Tag ที่อยู่ในตำแหน่งเริ่มต้นของบล็อคของข้อมูลที่ต้องการกำหนดรูปแบบ โดยมีชื่อของอีลีเมนท์อยู่ระหว่าง Angle brackets

</element_name>.

เป็น Tag ที่อยู่ในตำแหน่งสิ้นสุดของบล็อคของข้อมูล

เอ็มตี้อีลีเมนท์ (Empty Elements)

บางอีลีเมนสามารถเป็นอีลีเมนว่างได้ นั่นคือไม่ส่งผลกระทบกับบล็อคของข้อมูล อีลีเมนท์ ชนิดนี้ไม่จำเป็นต้องใช้ Tag ปิดท้ายบล็อคของข้อมูล

อัพเพอร์และโลวเวอร์เคสอีลีเมน (UpperandLowerCase Element)

การเขียน อีลีเมนท์เนม สามารถใช้ได้ทั้งตัวใหญ่และตัวเล็ก ซึ่งให้ผลเหมือนกัน เช่น การขีดเส้นในแนวนอนสามารถเขียนได้ ดังนี้

, <Em> หรือ <EM>.

อีลีเมนท์ หลายชนิดสามารถมีอาร์กูเมนท์ เพื่อใช้ในการผ่านค่าพารามิเตอร์ไปต่าง ๆ ไปยังอีลีเมนท์ได้อาร์กูเมนท์ เหล่านี้ถูกเรียกว่า แอททริบิวของอีลีเมนท์ (Attributes of the element)

ตัวอย่าง

<A HREF="http://www.somewhere.ca/file.html"> marked text .

จากตัวอย่าง a เป็น Element name ซึ่งปิดท้ายด้วย Tag

HREF เป็นแอททริบิวซึ่งใช้กำหนดค่า

โครงสร้างของเอกสารแบบเอ็ชทีเอ็มแอล

เอกสารแบบเอ็ชทีเอ็มแอล ประกอบด้วย 2 ส่วน คือเฮดและบอดี้

-- เฮด ประกอบด้วยข้อมูลเกี่ยวกับเอกสาร

-- บอดี้ ประกอบด้วยส่วนเนื้อข้อมูลที่จะแสดง

โครงสร้างของเอกสาร

<HTML>

<HEAD>

.

<element_name> ..... </element_name>

.

</HEAD>

<BODY>

.

<element_name> ..... </element_name>

.

</BODY>

</HTML>

การกำหนดลักษณะของเอกสารแบบเอ็ชทีเอ็มแอล

เมื่อ HTML บราวเซอร์ (Mosaic, TkWWW, Lynx .....) ได้รับไฟล์ มันต้องรู้ว่าจะทำอย่างไรกับไฟล์เหล่านั้น วิธีการที่ง่ายที่สุดก็คือ การดู filename extension นั่นคือ HTML ไฟล์จะมี extension เป็น html สำหรับยูนิกซ์และเป็น htm สำหรับดอสซึ่งมีสามารถกำหนดได้ไม่เกิน 3 ตัว

Extension มาตรฐานที่ใช้กันทั่ว ๆ ไป

- .html สำหรับยูนิกซ์ หรือ .htm สำหรับดอส
- .txt or .text
- .gif
- .xbm
- .xpm
- .jpeg
- .mpeg
- .au.
- .Z

บราวเซอร์ใช้ เอ็มไอเอ็มอีไทพ์ (Multipurpose Internet Mail Extension) ของ เอกสาร ในการกำหนดว่า ไฟล์ชนิดใด ต้องใช้วิธีการใดในการแสดงผล เช่น .GIF ต้องใช้ Image viewer เป็นต้น HTTP เซิร์พเวอร์จะเพิ่ม เอ็มไอเอ็มอี contents-types เข้าไปในส่วนเฮดเดอร์ ของทุกไฟล์ที่ถูกร้องขอให้กับบราวเซอร์เพื่อให้บราวเซอร์ รู้ชนิดของไฟล์และวิธีที่จะจัดการกับไฟล์เหล่านี้

2.4.2 เฮด

ในส่วนเฮดนี้จะประกอบไปด้วยข้อมูลเกี่ยวกับ เอกสารแต่ไม่มีการแสดงผลเหมือน กับในส่วนบอดี้

Mark-up อีลีเมนท์ที่สามารถใช้ในส่วนเฮดนี้ ได้แก่

- TITLE
- ISINDEX
- NEXTID
- LINK
- BASE

2.4.2.1 TITLE

ส่วน Title ของเอกสาร ถูกกำหนดโดย TITLE อีลีเมนในส่วนเฮด ซึ่งมีได้เพียง Title เดียวในแต่ละเอกสาร มักใช้เป็น Label ของวินโดว์ในการแสดงผล การกำหนด Title ของเอกสาร ควรจะมีความหมาย และไม่ยาวมากนัก (ควรจะน้อยกว่า 64 ตัวอักษร)

2.4.2.2 ISINDEX

อีลีเมนท์ นี้จะบอกตัวแสดงผลว่าเอกสาร นี้เป็นเอกสารดรรชนี (index document) นั่นคือ ในการแสดงหรืออ่านเอกสารสามารถทำการค้นหาข้อมูลโดยตามคีย์เวิร์ดที่กำหนดได้ โดยการเติมเครื่องหมาย “?” ตามหลังยูอาร์แอลนั้นและถ้าหาก คีย์เวิร์ดมีหลายคำ ให้ใช้ เครื่องหมาย “+” ในการแยกคำ เช่น

ยูอาร์แอลที่สามารถค้นหาข้อมูลได้

http://www.here.ca/cgi-bin/srch

การค้นหาคำว่า 'instructional' ทำได้โดยใช้เปลี่ยนแปลงยูอาร์แอลเป็น

http://www.here.ca/cgi-bin/srch?instructional

โดยปกติ ISINDEX อีลีเมนท์ นี้จะถูกสร้างโดย server-side script โดยอัตโนมัติ สำหรับเซิร์พเวอร์ที่สามารถทำการค้นหาได้เท่านั้น ดังนั้นจึงไม่ควรใส่อีลีเมนท์นี้เข้าไปเองในเอกสาร

2.4.2.3 NEXTID

Tag นี้เป็นแอททริบิวเดี่ยวอยู่ในส่วนเฮดของเอกสาร ถูกออกแบบมาเพื่อใช้สำหรับ Automatic HyperText Editors

Old anchor ids จะไม่ถูกนำกลับมาใช้ใหม่เมื่อมีการเปลี่ยนแปลงเอกสารขณะที่ เอกสารอื่นอ้างอิงถึง เอกสาร เหล่านี้

2.4.2.4 LINK

LINK อีลีเมนท์ใช้สำหรับบอกความสัมพันธ์ระหว่างเอกสารกับเอกสาร หรือ ออปเจคอื่น ในปัจจุบันไม่ค่อยใช้กันแล้ว

LINK อีลีเมนท์เป็นอีลีเมนท์ว่างมี แอททริบิวต่าง ๆ ดังนี้

- HREF กำหนดเอกสารหรือส่วนหนึ่งของเอกสารที่ต้องการติดต่อ
- NAME สำหรับตั้งชื่อ LINK เพื่อเป็นปลายทางสำหรับเอกสารอื่น
- REL อธิบายความสัมพันธ์ที่ถูกกำหนดโดยการเชื่อมต่อนี้
- REV คล้าย ๆ กับ REL แต่เป็นความสัมพันธ์ที่ตรงข้ามกัน
- URN กำหนดชื่อของ Uniform รีซอร์สเนม
- TITLE ใช้เป็น title ของ HREF แอททริบิวของ LINK อีลีเมนท์
- METHOD อธิบาย HTTP เมธอททีออปเจค อ้างอิงถึงใน HREF ของ LINK อีลีเมนท์

2.4.2.5 BASE

อีลีเมนนี้ใช้สำหรับเก็บออริจินอล ยูอาร์แอลของ เอกสาร ซึ่งจะทำให้สามารถย้าย เอกสารไปยังไดเรคทอรีหรือไซท์อื่น มีแอททริบิวเดียว คือ HREF ใช้สำหรับเก็บ เบสยูอาร์แอลของเอกสารเพื่อใช้กับ Partial ยูอาร์แอลภายในเอกสารโดยใช้ Base ยูอาร์แอลเป็นจุดเริ่มต้น

2.4.2.6 META

อีลีเมนท์ที่ใช้ในการกำหนดเมต้าอินฟอร์เมชั่น (Information about Information) ภายในเอกสารโดยมีแอททริบิวต่าง ๆ ดังนี้

- HTTP-EQUIV แอททริบิวนี้ใช้ติดต่อระหว่าง META อีลีเมนท์ นี้เพื่อตอบสนองต่อ โปรโตคอลเฉพาะ ซึ่งถูกสร้างโดย HTTP เซิร์พเวอร์ที่เก็บเอกสาร

- NAME กำหนดชื่อของอินฟอร์เมชันในเอกสารไม่เหมือนกับ TITLE เนื่องจาก "meta name" เป็นการแบ่งประเภทของอินฟอร์เมชัน

- CONTENT "meta name" สำหรับ content ที่เกี่ยวข้องกับชื่อที่กำหนดโดย NAME แอททริบิวหรือการตอบสนองที่ถูกกำหนดโดย HTTP-EQUIV

2.4.3. บอดี้

บอดี้อีลีเมนท์ประกอบด้วย อินฟอร์เมชันเกี่ยวกับเอกสาร และ mark-up อีลีเมนท์ ต่าง ๆ ที่ใช้ในการกำหนดรูปแบบของตัวอักษร, รูปภาพ และอื่น ๆ

2.4.3.1 Section Headings (H1, H2, ..., H6)

HTML ยอมให้มีเฮดดิง ถึง 6 ระดับ โดยใช้ Tag H1, H2, ..., H6 ในการใช้ไม่จำเป็นต้องเรียงตามลำดับ แต่เพื่อความเข้าใจที่ตรงกัน จึงควรที่จะใช้ <H1> เพื่อเป็น เ
เฮดดิงที่สำคัญที่สุด และใช้ Tag อื่น กับเฮดดิงที่สำคัญรองลงมา

2.4.3.2 Marking Paragraphs with HTML

อีลีเมนท์ <P> ใช้สำหรับแบ่งพารากราฟ และใช้เมื่อต้องการแบ่ง Text 2 Blocks ออกเป็น 2 พารากราฟ ซึ่งในการใช้ต้องทำเครื่องหมายทุกอีลีเมนท์และ ข้อความทุกอย่างที่ต้องการให้อยู่ในพารากราฟเดียวกัน

2.4.3.3 Line Breaks

อีลีเมนท์
 ใช้สำหรับกำหนดให้ข้อความที่ตามหลัง Tag ขึ้นบรรทัดใหม่

2.4.3.4 IMG (In-line Images) Element

อีลีเมนท์นี้ใช้สำหรับให้ Graphical browser แสดงรูปภาพในตำแหน่งที่ Tag ปรากฏ มีชนิดเป็นอีลีเมนท์ว่างจึงไม่จำเป็นปิดท้ายด้วย </IMG> ประกอบด้วยแอททริบิวต่าง ๆ ดังนี้

- SRC="image_url" ใช้กำหนดยูอาร์แอลของ image
- ALIGN=BOTTOM (MIDDLE or TOP) ใช้บอกตำแหน่งของรูปภาพในการแสดง
- ALT="alternative text" ใช้กำหนดข้อความที่แสดงผลสำหรับ browser ที่ไม่สามารถแสดงผลแบบกราฟฟิคได้
- ISMAP ใช้สำหรับกำหนดให้รูปภาพเป็น active image map นั่นคือ สามารถให้ผู้ใช้คลิกเมาส์ไปในตำแหน่งที่แตกต่างกันสามารถให้ผลตอบสนองที่แตกต่างกัน

2.4.3.5 Hypertext Anchors

Anchor คือ การกำหนดให้ข้อความหรือออปเจคชนิดต่าง ๆ ที่อยู่ระหว่าง <A> และ </A> ใช้สำหรับ Hypertext link

Anchor Attributes เหมือนกับ LINK Attributes

- HREF (link to object) *
- NAME (link from object) *
- REL (relationship between objects)
- REV (relationship between objects)
- URN (URN for the document)
- TITLE (TITLE of document)
- METHOD (how to link)

2.4.3.6 Lists

HTML สนับสนุนการสร้างลิสต์โดยมีอีลีเมนท์ต่าง ๆ ให้เลือกใช้ ซึ่งแบ่งออกเป็น 2 ประเภท คือ Glossary Lists : <DL>, Regular lists : <UL>, <OL>, <MENU> and <DIR>.

2.4.3.7 Horizontal Ruled Line

HR อีลีเมนท์ใช้สำหรับวาดเส้นในแนวนอน ใช้ในการแบ่งบล็อคข้อมูล จัดเป็นประเภท อีลีเมนท์ว่างไม่จำเป็นต้องมี </HR> ปิดท้าย

2.4.3.8 Special Characters in HTML

สัญญลักษณ์พิเศษบางตัวถูกใช้โดย HTML ไปตามวัตถุประสงค์ที่ต่างกัน ดังนั้นเมื่อผู้ใช้ต้องการให้แสดงสัญญลักษณ์พิเศษบนบราวเซอร์จึงมี เอนติตี้ของตัวอักษรพิเศษเพื่อใช้ในการแสดงสัญญลักษณ์พิเศษเหล่านั้น ซึ่งเอนติตี้จะประกอบด้วย 3 ส่วน คือ

- ขึ้นต้นด้วย Ampersand "&"
- ตัวเลขหรือข้อความของตัวอักษรที่ต้องการแสดง
- ลงท้ายด้วย Semicolon ";"

ตาราง : ตัวอย่าง Special Character ที่นิยมใช้

Entity	Special Character	Description
<	<	less than sign
>	>	greater than sign
&	&	The ampersand sign itself
"	"	double quote
		A non-breaking space

2.4.3.9 Character Emphasis Modes

HTML ยอมให้มีการกำหนดรูปแบบของตัวอักษรได้ เช่น ตัวหนา, ตัวเอียง เป็นต้น ซึ่งแบ่งเป็น 2 วิธี คือ

Logical Styles

EM	ตัวเอียง
STRONG	ตัวหนา
CODE	ขนาดตายตัว
SAMP	ลำดับของตัวอักษร
KBD	Text ที่ถูก mark จะเหมือนกับอินพุต จาก คีย์บอร์ด
VAR	ชื่อตัวแปร
DFN	กำหนดค่า
CITE	คำที่ใช้อ้าง มักจะเป็นตัวเอียง

Physical Styles

TT	ใช้รูปแบบตัวอักษรเหมือนพิมพ์ดีด ขนาดคงที่
B	ตัวหนา
I	ตัวเอียง
U	ขีดเส้นใต้

2.4.3.10 Special Text Formatting Modes

HTML มีรูปแบบการแสดงผลข้อความในระดับ logical อยู่ 3 แบบ แบ่งตามลักษณะการแสดงผลของบราวเซอร์ คือ

PRE (Preformatted Text)

แสดงข้อความที่อยู่ระหว่าง Tag <PRE> และ </PRE> ให้มีรูปแบบแน่นอนตามที่กำหนด มี ออปชั่นแอททริบิว คือ WIDTH (default=80) ใช้กำหนดจำนวนตัวอักษรมากที่สุด ที่สามารถแสดงได้ใน 1 บรรทัด

BLOCKQUOTE

ใช้สำหรับกำหนดให้บราวเซอร์จัดการเรียงข้อมูลที่อยู่ระหว่าง Tag <BLOCKQUOTE> และ </BLOCKQUOTE> ตามหน้าจอแสดงผล

ADDRESS

ใช้สำหรับข้อมูลเกี่ยวกับที่อยู่, signatures, authorship เป็นต้น

2.4.3.11 TABLES

การสร้างตารางสามารถทำได้โดยใช้ TABLE อีลีเมนท์ซึ่งมีแอททริบิวต่าง ๆ ดังนี้

- ALIGN ใช้ในการกำหนดตำแหน่งของตาราง ซึ่งมีค่าที่เป็นไปได้ ดังนี้BLEEDLEFT, LEFT, CENTER, RIGHT, BLEEDRIGHT, JUSTIFY

- BORDER ใช้กำหนดขนาดของเส้นขอบของตาราง

- WIDTH ใช้กำหนดความกว้างของตาราง สามารถกำหนดเป็น % ของความกว้างของส่วนที่ใช้ในการแสดงผลได้

- COLSPEC ใช้กำหนดตำแหน่งของไอเทมในคอลัมน์

- CAPTION ใช้สำหรับตั้งชื่อตาราง มีแอททริบิว คือ Align ใช้สำหรับกำหนดตำแหน่งของชื่อตาราง ซึ่งมีค่าต่าง ๆ

- TH (Table Header) และ TD (Table Data) ใช้กำหนดชนิดของข้อมูลที่อยู่ในตารางโดย TH จะเป็นตัวหนา ส่วน TD จะเป็นตัวอักษรธรรมดา ทั้ง 2 อีลีเมนท์นี้มีแอททริบิวที่เหมือนกัน

2.4.3.12 FORMS Element for fill-out Forms

FORMS เป็นรูปแบบที่ทำให้ผู้ใช้สามารถติดต่อกับ HTTP เซิร์ฟเวอร์ ได้มากกว่าการใช้ ISINDEX โดย FORM อีลีเมนท์จะสร้าง fill-out ฟอร์ม ที่ผู้ใช้สามารถกรอกข้อมูลลงไปในช่อง ที่กำหนดมา เพื่อส่งต่อไปให้ ซีจีไอบินสคริปต์บน HTTP เซิร์ฟเวอร์

อีลีเมนท์ที่ใช้ในฟอร์มประกอบด้วย

FORM

ใช้สำหรับกำหนดวิธีการรับข้อมูล และวิธีการจัดการกับอินพุตที่ได้มาเพื่อให้ได้ผลลัพธ์ ที่ต้องการมีแอททริบิวดังนี้

- ACTION ใช้กำหนดยูอาร์แอลของโปรแกรม หรือสคริปต์ที่จะใช้ในการกระทำงาน ถ้าไม่กำหนดส่วนนี้ Base ยูอาร์แอลจะถูกเรียกใช้

- METHOD เป็นการกำหนด HTTP เมธอท ที่ใช้ในการส่งข้อมูลจากฟอร์มไปยังเซิร์พเวอร์ซึ่งมีได้ 2 แบบ คือ

1. GET (default) ข้อมูลจากฟอร์มจะถูกต่อท้ายยูอาร์แอล (หลังเครื่องหมายคำถาม "?") เหมือนกัน ISINDEX queries

2. POST ข้อมูลจากฟอร์มจะถูกส่งไปยังเซิร์พเวอร์ในส่วน บอดี้ ของข่าวสาร การที่จะใช้วิธีนี้ต้องแน่ใจว่า HTTP เซิร์พเวอร์ที่ใช้สนับสนุนวิธีนี้ด้วย

INPUT

ใช้สำหรับรวบรวมข้อมูลจากผู้ใช้ มี แอททริบิวต่าง ๆ ดังนี้

- ALIGN ใช้กับ IMAGE TYPE ในการกำหนดตำแหน่งของ image ที่สัมพันธ์กับ text (TOP, MIDDLE หรือ BOTTOM)

- CHECKED ใช้ร่วมกับ CHECKBOX และ RADIO botton ในการเลือกในตอนเริ่มต้น ถ้าหากไม่กำหนดจะไม่มีตัวใดถูกเลือก

- MAXLENGTH กำหนดจำนวนตัวเลขและตัวอักษรมากที่สุดที่ผู้ใช้สามารถเติมได้ โดยปกติจะไม่ถูกจำกัด

- NAME ชื่อของตัวแปรที่ใช้ในการส่งข้อมูล

- SIZE กำหนดความกว้างของช่องที่ใช้ป้อนข้อมูล

- SRC ใช้กำหนดยูอาร์แอลของ image ไฟล์เมื่อ ไทพ์ถูกกำหนดเป็นแบบ image TYPE กำหนดรูปแบบของฟิลด์ที่ให้ผู้ใช้ป้อนข้อมูล

- CHECKBOX สามารถให้ค่าได้หลายค่า

- HIDDEN สำหรับค่าที่ถูกกำหนดโดยฟอร์มโดยไม่ต้องรับอินพุตจากผู้ใช้

- IMAGE ใช้รูปภาพเป็นตัวรับอินพุตจากผู้ใช้โดยส่งค่า x, y ที่ถูกคลิกโดย เมาส์

- PASSWORD เป็นฟิลด์สำหรับป้อนข้อมูลแบบเท็กซ์แต่ไม่แสดงผล

- RADIO ใช้สำหรับเก็บข้อมูลที่สามารถเลือกได้เพียงค่าเดียวในแต่ละครั้ง

- RESET ใช้เคลียร์ค่าให้กลับเป็นค่ามาตรฐาน

- SUBMIT เป็นปุ่มที่ใช้สำหรับการส่งข้อมูลในฟอร์มใช้ VALUE ในการกำหนดข้อความ

ที่แสดงบนปุ่ม

- TEXT ใช้สำหรับ single-line of text
- TEXT เป็นช่องให้เติมข้อมูลแบบเท็กซ์บรรทัดเดียว
- VALUE ใช้กำหนดค่าเริ่มต้นของฟิลด์หรือ กำหนดค่าของฟิลด์เมื่อถูกเลือก
- SELECT อีลีเมนท์จะแสดงทางเลือกสำหรับให้ผู้ใช้เลือกค่าใดค่าหนึ่งจากหลาย ๆ ค่า โดยใช้ออปชันเป็นตัวกำหนดค่าต่าง ๆ เหล่านั้น มี แอททริบิวต่าง ๆ ดังนี้
- NAME ชื่อที่ใช้ในการส่งข้อมูลที่มาผู้ใช้เลือก
- MULTIPLE โดยปกติผู้ใช้สามารถเลือกได้เพียงอย่างเดียว แต่ MULTIPLE แอททริบิว นี้จะทำให้ผู้ใช้สามารถเลือกได้มากกว่าหนึ่งอย่าง
- SIZE กำหนดจำนวนของไอเทมที่ผู้ใช้สามารถมองเห็นได้ ถ้าไอเทมมีมากกว่าที่กำหนด การแสดงผลจะอยู่ในรูปของลิสต์

OPTION

อีลีเมนท์นี้จะปรากฏภายใน SELECT อีลีเมนท์และใช้ในการแสดงตัวเลือกต่าง ๆ ภายใน SELECT อีลีเมนท์มีแอททริบิวต่าง ๆ ดังนี้

- SELECTED กำหนดให้ออปชันถูกเลือกในตอนเริ่มต้น
- VALUE ใช้กำหนดค่าที่จะถูกส่งกลับไปยังเซิร์พเวอร์ ของแต่ละออปชัน ถ้าไม่มี แอททริบิวนี้ ค่าที่จะถูกส่งกลับไปคือค่าที่ถูกกำหนดโดยออปชันแอททริบิว

TEXTAREA

อีลีเมนท์ นี้จะใช้ในการรวบรวมข้อมูลชนิด เท็กซ์ที่มีหลายบรรทัดจากผู้ใช้ มี Scroll Bar ทำให้สามารถเลื่อนไปยังตำแหน่งต่าง ๆ ได้ TEXTAREA มีแอททริบิวต่าง ๆ ดังนี้

- NAME ชื่อที่จะใช้ในการส่งค่ากลับ
- ROWS จำนวนของแถวที่แสดงผล
- COLS จำนวนของคอลัมน์ที่แสดงผล

บทที่ 3

การคำนวณและการสร้าง

เมื่อทำการศึกษาทฤษฎีต่างๆ ที่เกี่ยวข้องกับการทำงานของ โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์แล้ว เราก็เริ่มที่ทำงานในส่วนของการออกแบบ โดยมีการออกแบบแบ่งเป็นส่วน ๆ ตามฟังก์ชันการทำงาน

3.1 การกำหนดฟังก์ชันการทำงานของเซิร์สเอนจิน

การทำงานของเซิร์สเอนจินเราจะแบ่งฟังก์ชันการทำงานออกได้เป็นหลายฟังก์ชันดังนี้

1. ฟังก์ชันในการดึงข้อมูลและวิเคราะห์ข้อมูล
    - ฟังก์ชันในการตรวจสอบ ยูอาร์แอล
    - ฟังก์ชันในการดึงข้อมูล
    - ฟังก์ชันในการตัดข้อมูลเป็นคำ
    - ฟังก์ชันในการคำนวณค่าความสำคัญและคัดเลือกคำ
    - ฟังก์ชันในการเก็บคำลงในฐานข้อมูล
2. ฟังก์ชันในการติดต่อระหว่างผู้ใช้บริการและเซิร์พเวอร์โดยผ่านโปรแกรมซีจีไอ
    - ฟังก์ชันในการเรียกใช้บริการจากเซิร์พเวอร์
    - ฟังก์ชันในการรับข้อมูลจากผู้ใช้บริการมายังเซิร์พเวอร์
    - ฟังก์ชันในการส่งข้อมูลจากเซิร์พเวอร์กลับไปยังเซิร์พเวอร์
3. ฟังก์ชันในการค้นหาข้อมูลบนฐานข้อมูล

3.2 กำหนดออปชั่นในการค้นหาข้อมูลของเซิร์สเอนจิน

ในการค้นหาข้อมูลต่างๆ บนอินเตอร์เน็ตผ่านทางเซิร์สเอนจิน นั้นสามารถทำได้หลาย รูปแบบดังนั้นจึงต้องมีการกำหนดว่าจะให้เซิร์สเอนจิน ที่สร้างขึ้นสามารถทำการค้นหาข้อมูลได้กี่แบบ

ข้อกำหนดในการค้นหาข้อมูลของเซิร์สเอนจินมีดังนี้

1. สามารถทำการค้นคำได้เฉพาะคำภาษาอังกฤษ
2. สามารถทำการค้นหาคำจากตัวแปรที่ผู้ใช้บริการกำหนดได้
3. สามารถทำการค้นหาข้อมูลได้ทั้งหมด 6 รูปแบบดังนี้
    - คำเดียวพิจารณารูปแบบของคำ (Single Case-Sensitive)

- คำเดียวไม่พิจารณารูปแบบของคำ (Single Noncase-Sensitive)

- หลายคำพิจารณารูปแบบของคำ กรณี ' หรือ ' (Case-Sensitive Or)

- หลายคำไม่พิจารณารูปแบบของคำ กรณี ' หรือ ' (Noncase-Sensitive Or)

- หลายคำพิจารณารูปแบบของคำ กรณี ' และ ' (Case-Sensitive And)

- หลายคำไม่พิจารณารูปแบบของคำ กรณี ' และ ' (Noncase-Sensitive And)

ซึ่งการกำหนดออปชันในการค้นหาข้อมูลนั้น จะสามารถทำการกำหนดได้ โดยผ่านทางโฮมเพจของเซิร์สเอนจินซึ่งมี ช่องสำหรับให้เลือกออปชัน

ต่อไปเราจะทำการอธิบายถึงการออกแบบและการสร้างในแต่ละส่วนของเซิร์สเอนจิน

3.3 การออกแบบในส่วนที่ใช้ดึงข้อมูลและวิเคราะห์ข้อมูล

ในส่วนนี้จะเป็นการออกแบบการทำงาน และกำหนด Algorithm ในการทำงานของฟังก์ชัน ต่างๆ ที่เกี่ยวข้องกับการทำงานของโรบอทในส่วนที่ใช้ดึงข้อมูลและวิเคราะห์ข้อมูล

3.3.1 การเก็บคีย์เวิร์ด (Keywords) จากเว็บเพจลงฐานข้อมูล

ใน 1 เว็บเพจ จะประกอบด้วยคำต่าง ๆ มากมาย ที่สามารถนำมาใช้เป็นคีย์เวิร์ดในการ ค้นหา ข้อมูลของผู้ใช้ได้ การเลือกคีย์เวิร์ด จากคำต่าง ๆ เหล่านั้นจึงเป็นส่วนสำคัญ ที่ส่งผลกระทบต่อการค้นหาข้อมูลของผู้ใช้โดยตรง นั่นคือ การเลือกคีย์เวิร์ดที่ดีจะทำให้ผู้ใช้ สามารถค้นหาข้อมูลได้ถูกต้อง รวดเร็ว และตรงตามความต้องการ

การเลือกเก็บคีย์เวิร์ดเป็นฐานข้อมูล มี 2 วิธี คือ

วิธีที่ 1. เลือกคำทั้งหมดที่ปรากฏในเว็บเพจเป็นคีย์เวิร์ด

วิธีนี้จะทำการเก็บคำทุกคำที่ปรากฏอยู่ภายในเว็บเพจลงฐานข้อมูล เพื่อใช้เป็นคีย์เวิร์ดในการค้นหาไม่ว่าคำ ๆ นั้นจะมีความหมายหรือไม่ก็ตาม

ข้อดี ของวิธีการนี้ คือ

ผู้ใช้สามารถค้นหาข้อมูลที่ตนเองต้องการได้ ถ้าสิ่งที่ผู้ใช้ค้นหามีอยู่ในอินเตอร์เน็ต แม้ว่าจะเป็นส่วนเล็ก ๆ ที่ไม่ค่อยจะมีความสำคัญมากนัก เนื่องจากฐานข้อมูลจะเก็บข้อมูลทุกอย่าง คำทุก คำที่ปรากฏ

ข้อเสีย ของวิธีการนี้ คือ

1. **ฐานข้อมูลมีขนาดใหญ่** เนื่องจากต้องเก็บคำทุกคำ

2. **สิ้นเปลืองเนื้อที่ที่ใช้ในการเก็บ** เนื่องจากคำส่วนใหญ่ มักจะเป็นคำที่ไม่มีผู้ใช้คนใดใช้เป็นคีย์เวิร์ดในการค้นหา เช่น a, an, the, is, am, in, on, and เป็นต้น

3. การค้นหาช้า เนื่องจากฐานข้อมูลมีขนาดใหญ่

4. จำนวนของ เว็บเพจ ที่ได้จากการค้นหามีจำนวนมาก ทำให้เสียเวลาในการตรวจเลือกและหาเว็บเพจที่ต้องการ

วิธีที่ 2. เลือกคำบางคำเก็บเป็นคีย์เวิร์ด

วิธีการนี้จะทำการเลือกเก็บเฉพาะคำที่สามารถใช้เป็นคีย์เวิร์ดได้เก็บเป็นฐานข้อมูล ส่วนคำที่เหลือไม่ต้องสนใจ เนื่องจากคำส่วนใหญ่ที่ปรากฏภายในเว็บเพจนั้นมักจะเป็นคำที่ไม่มีประโยชน์ที่จะใช้ในการค้นหา เช่น

- เป็นคำที่มีปรากฏอยู่ภายในเกือบทุก เว็บเพจ เช่น is, the, and, or เป็นต้น
- เป็นสัญญลักษณ์ เช่น @, ?, เป็นต้น
- เป็นคำที่ไม่สามารถบ่งบอกถึงเนื้อหาของเว็บเพจใด ๆ ได้ เช่น OK, click, here
- อื่น ๆ

ข้อดี ของการเก็บคีย์เวิร์ดโดยใช้วิธีนี้

1. เพื่อช่วยให้ประหยัดเนื้อที่ที่ต้องใช้ในการเก็บข้อมูล เนื่องจากไม่ต้องเก็บคำทุกคำที่ปรากฏ ในเว็บเพจ ขนาดของฐานข้อมูลจึงไม่ใหญ่มากนัก

2. ทำให้การค้นหาเป็นไปได้รวดเร็วขึ้น เนื่องจากฐานข้อมูลมีขนาดเล็กลง

3. ผลลัพธ์ที่ได้จากการค้นหามีความใกล้เคียงกับความต้องการของผู้ค้นหามากกว่าการเก็บข้อมูลทั้งหมด

4. ผลลัพธ์ที่ได้จากการค้นหามีปริมาณน้อยกว่า การเก็บคำทุกคำเป็นคีย์เวิร์ด

ข้อเสีย ของการเก็บคีย์เวิร์ดโดยใช้วิธีคำนวณหาค่าความสำคัญ

1. ผลลัพธ์ที่ได้ไม่สมบูรณ์ เนื่องจากการเลือกคีย์เวิร์ดที่จะเก็บเป็นฐานข้อมูล จะเลือกคำที่มีความสำคัญตามเกณฑ์ที่ตั้งไว้ ดังนั้นคำบางคำสามารถใช้เป็นคีย์เวิร์ดแต่มีความ สำคัญน้อยกว่าที่กำหนดไว้จึงไม่ถูกเก็บไว้ในฐานข้อมูล

2. ไม่สามารถทำการค้นหาข้อมูลที่เป็นประโยคได้ เนื่องจากไม่ได้เก็บคำทุกคำเป็นคีย์เวิร์ด

3.3.2 คุณสมบัติของคำที่สามารถใช้เป็นคีย์เวิร์ดได้

ในการเลือกคำที่จะใช้เป็นคีย์เวิร์ดในขั้นแรกสุด จะใช้การตรวจสอบ โดยกรองเอาคำที่ ไม่มีคุณสมบัติพอที่จะเป็นคีย์เวิร์ดออก และเหลือไว้แต่คำที่มีคุณสมบัติพอที่จะเป็นคีย์เวิร์ดได้ ซึ่งคุณสมบัติต่าง ๆ ของคำที่จะใช้เป็นคีย์เวิร์ดของ เว็บเพจ ใด ๆ ได้จะต้องมีคุณสมบัติ ดังนี้

1. คีย์เวิร์ดจะต้องเป็นคำที่ปรากฏอยู่ในเว็บเพจนั้น

2. เป็นคำที่ประกอบด้วยตัวอักษร “a-z”, “A-Z”, “-”, “_”, “0-9” เท่านั้น และจะต้องไม่ใช่ตัวเลขทั้งคำ เช่น “word6” ใช้ได้

“cgi-bin” ใช้ได้

“95682” ใช้ไม่ได้

“&Quote” ใช้ไม่ได้

3. คำที่ใช้อักษรตัวพิมพ์ใหญ่หมดทั้งคำ (UpperCase) มีโอกาสที่จะเป็นคีย์เวิร์ดได้ ถ้าคำที่ปรากฏในเว็บเพจส่วนใหญ่ใช้อักษรตัวเล็ก (LowerCase) ซึ่งการใช้ตัวพิมพ์ใหญ่ แสดงให้เห็นว่า ผู้เขียนเว็บเพจต้องการเน้นคำ ๆ นั้นให้เห็นได้ชัดเจน จึงเป็นไปได้ว่าคำ ๆ นั้นมีความสำคัญพอที่ จะเป็นคีย์เวิร์ดของเว็บเพจนั้นได้ และนอกจากนั้น คำ ๆ นี้ มีโอกาสสูงที่จะเป็นตัวย่อ เช่น “NECTEC”, “CGI”, “WYSIWYG” เป็นต้น

4. คำที่เป็นคีย์เวิร์ดจะต้องเป็นคำที่มีความยาวของคำตั้งแต่ 3 ตัวอักษรขึ้นไป เนื่องจากคำที่มีความยาวน้อยกว่า 2 ตัวอักษร มักจะเป็นคำที่ไม่มีความหมาย หรือมีความหมาย แต่ไม่สามารถ ที่จะบอกได้ว่า เว็บเพจที่มีคำ ๆ นี้ปรากฏอยู่มีข้อมูลเกี่ยวกับอะไร และโดยปกติ คำที่มีความยาวน้อยกว่า 3 ตัวอักษร มักจะมีปรากฏอยู่ในทุกเว็บเพจ จึงไม่เหมาะอย่างยิ่งที่จะใช้คำ เหล่านี้เป็นคีย์เวิร์ด เช่น “a”, “an”, “he”, “is”, “or”, “on” เป็นต้น

5. คำที่เป็นคีย์เวิร์ดจะต้องไม่อยู่ในลิสต์ของคำที่ไม่ควรจะเป็นคีย์เวิร์ด การตรวจสอบข้อนี้เนื่องจากคำบางคำมีความยาวมากกว่า 2 ตัวอักษร แต่ก็อยู่ในประเภทที่ไม่สามารถบ่งบอก ข้อมูลเกี่ยวกับเว็บเพจใด ๆ ได้ เช่น “the”, “and”, “another”, “they”, “there” เป็นต้น และนอกจากจะเลือกคำที่ไม่สามารถใช้เป็นคีย์เวิร์ดออกไปได้แล้ว วิธีนี้ยังใช้ป้องกันไม่ให้มีการเก็บคีย์เวิร์ดบาง คำที่ไม่ต้องการได้ เช่น คำไม่สุภาพ ต่าง ๆ หรือ คำที่มีความหมายสื่อไปในทางที่เราไม่ต้องการ เช่น “sex”, “nude” เป็นต้น

3.3.3 การคำนวณหาค่าความสำคัญของคำที่ปรากฏในเว็บเพจ

เมื่อตรวจสอบคุณสมบัติของคำ ต่าง ๆ ที่ปรากฏอยู่ในเว็บเพจแล้ว ว่ามีคุณสมบัติพอที่จะใช้เป็นคีย์เวิร์ดของเว็บเพจได้ นำคำเหล่านั้นมาคำนวณหาค่าความสำคัญของแต่ละคำ ที่ปรากฏ อยู่ภายในเว็บเพจ และเลือกคำที่มีความสำคัญมากกว่าเกณฑ์ที่กำหนดไว้ จัดเก็บลงในฐานข้อมูล เพื่อใช้เป็นคีย์เวิร์ดในการค้นหาต่อไป ซึ่งวิธีการคำนวณหาค่าความสำคัญของคำต่าง ๆ สามารถหา ได้โดยใช้สูตร ดังต่อไปนี้

$$Priority_n = S\ (Priority\_Key_i)$$

เมื่อ Priority คือ ค่าความสำคัญของคำใด n ใด ๆ

Priority_Key คือ ค่าความสำคัญของคำที่อยู่ ณ ตำแหน่ง i

จากสูตร การคำนวณหาค่าความสำคัญของคำ ทำได้โดยนำค่า Priority_Key ของตำแหน่งที่ปรากฏคำ ๆ หนึ่งทั้งหมดในเว็บเพจ มารวมกันเพื่อให้ได้ค่า Priority ของคำ ๆ นั้น

ตัวอย่าง

คำว่า "computer" ปรากฏอยู่ในเว็บเพจ 3 ที่

ตำแหน่งที่ 1 มีค่า $Priority\_Key_1 = 100$

ตำแหน่งที่ 2 มีค่า $Priority\_Key_2 = 20$

ตำแหน่งที่ 3 มีค่า $Priority\_Key_3 = 3$

เพราะฉะนั้น $Priority_{computer} = 100+20+3$

$= 123$

การกำหนดค่า Priority_Key ของคำในตำแหน่งต่าง ๆ

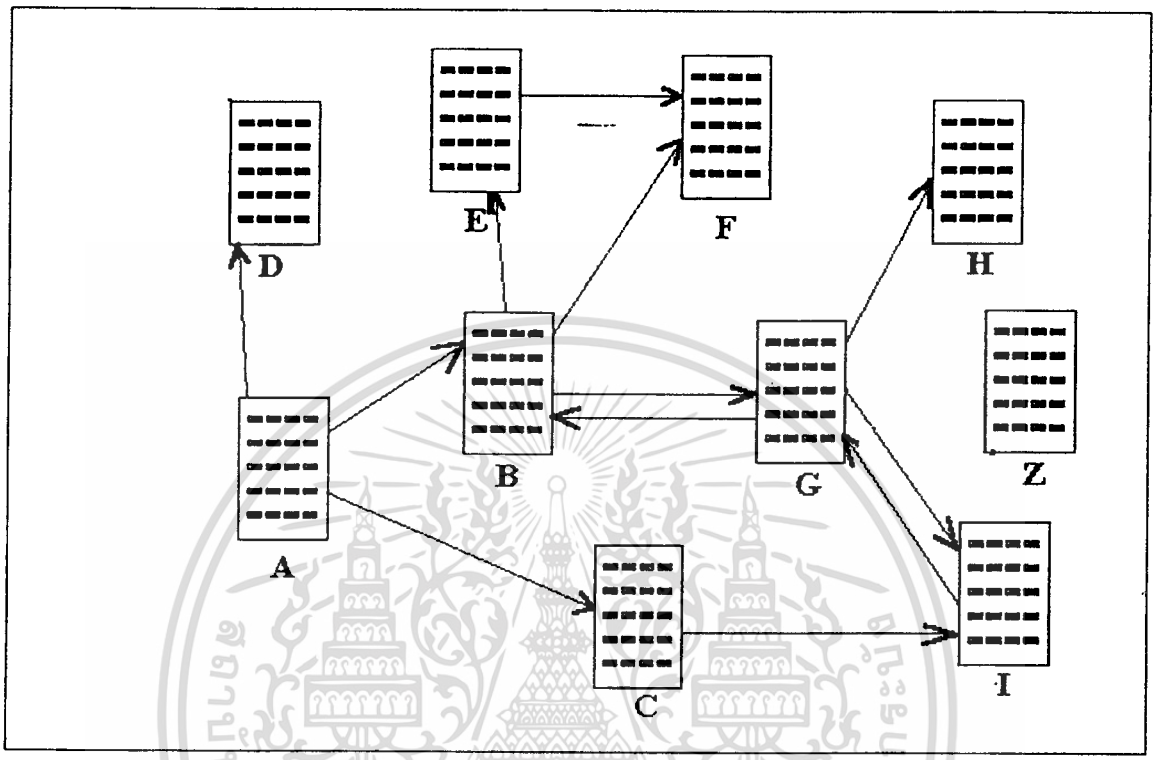
ค่า Priority_Key เป็นค่าที่ใช้กำหนดความสำคัญของคำ ณ. ตำแหน่งใด ๆ ในเว็บเพจ สิ่งที่จะใช้กำหนดค่า Priority_Key นี้คือ แท็ก (Tag) นั่นคือ ค่าของ Priority_Key จะแตกต่างกัน ไปมากน้อยขึ้นอยู่กับว่า ณ. ตำแหน่งนั้นอยู่ระหว่าง แท็กใด เช่น ค่า Priority_Key ณ. ตำแหน่งที่ อยู่ระหว่าง <TITLE> กับ </TITLE> จะมีค่า Priority_Key มากกว่าค่า Priority_Key ณ.ตำแหน่ง ที่อยู่ระหว่าง <H1> กับ </H1> เป็นต้น และแท็กบางแท็ก จะไม่มีผลต่อค่า Priority_Key เลย

3.3.4 วิธีในการท่องเว็บไซท์ของหุ่นยนต์เก็บข้อมูล

ในการเก็บข้อมูล หุ่นยนต์ (Bots) จะต้องติดต่อไปยังเว็บไซต์ (Web Site) ต่าง ๆ ในอินเตอร์เน็ต ที่ถูกกำหนดเอาไว้ในลิสต์ และทำการขอข้อมูลจากเว็บไซท์นั้น ให้ส่งกลับมาที่เซิร์ฟเวอร์ จากนั้นจึงส่งข้อมูลหรือเว็บเพจที่ได้ไปทำการวิเคราะห์หาคีย์เวิร์ดเก็บไว้เป็นฐานข้อมูลเพื่อใช้ในการค้นหา นอกจากนั้นยังต้องเลือก ยูอาร์แอล ที่ปรากฏในเว็บเพจนั้นเอาไปใส่ไว้ในลิสต์เพื่อให้หุ่นยนต์ทำการเก็บข้อมูลจากเว็บเพจนั้นต่อไป

ในการค้นหาข้อมูลจากเว็บไซท์ต่าง ๆ ของหุ่นยนต์เก็บข้อมูลนี้จะใช้วิธีการค้นหาในทางกว้างก่อน (Breadth-First Search) นั่นคือ หุ่นยนต์จะทำการดึงข้อมูลจากเว็บไซท์เริ่มต้น ที่ถูกกำหนดเอาไว้ในลิสต์ของยูอาร์แอลที่จะต้องทำการเก็บข้อมูล เมื่อได้ข้อมูลที่ต้องการแล้วจึงส่งข้อมูลที่ได้ไปให้ส่วนวิเคราะห์ทำหน้าที่ เลือกหาคำที่จะนำมาใช้เป็นคีย์เวิร์ดเก็บไว้ เป็นฐานข้อมูลเพื่อใช้ในการค้นหาต่อไป ส่วนยูอาร์แอลที่ปรากฏอยู่ภายในเว็บเพจที่ติดต่อไปยัง ยูอาร์แอล อื่นจะถูกนำไปตรวจสอบ (อยู่ในประเทศไทย หรืออยู่ในลิสต์ของยูอาร์แอลที่เราสนใจ) และถ้ายูอาร์แอล

นั้นผ่านการตรวจสอบ ก็จะถูกนำไปต่อท้ายเอาไว้ในลิสต์เพื่อให้หุ่นยนต์เก็บ ข้อมูลใช้ในการติดต่อขอข้อมูลต่อไป ดูรูปภาพประกอบ



รูปที่3.1 การเก็บข้อมูลโดยใช้วิธีการค้นหาทางกว้างก่อน

ตัวอย่าง

แสดงลิสต์ของยูอาร์แอลที่หุ่นยนต์ค้นหาจะต้องติดต่อเพื่อขอข้อมูล เมื่อหุ่นยนต์เก็บข้อมูลต้องทำการเก็บข้อมูลจาก ยูอาร์แอลตามรูปข้างบน โดยสมมุติ ให้ ยูอาร์แอล ทั้งหมดที่ปรากฏ ในรูปเป็น ยูอาร์แอล ที่อยู่ภายในประเทศไทย

A	NULL								

ลิสต์ 1. ลิสต์รายการเริ่มต้น เป็นลิสต์ที่กำหนดไว้ตอนเริ่มต้นเพื่อให้หุ่นยนต์เก็บข้อมูลใช้เป็น ยูอาร์แอล เริ่มต้นในการติดต่อร้องขอข้อมูล

B	C	D	NULL				..		

ลิสต์ 2. หลังจากได้ข้อมูลจาก "A" แล้ว นำเอา "A" ออกจาก ลิสต์ และนำ ยูอาร์แอล ที่ "A" ติดต่อไปถึง ไปตรวจสอบ เมื่อผ่านการตรวจสอบแล้ว นำ ยูอาร์แอล ที่ผ่านการตรวจ สอบทั้งหมดไปใส่ไว้ในลิสต์ จากรูป "A" มีไฮเปอร์ลิงก์ (HyperLink) ไปยัง "B", "C" และ "D" เมื่อนำยูอาร์แอลทั้ง 3 ไปตรวจสอบและผ่านการตรวจสอบแล้วจึงนำ ยูอาร์แอล ทั้งหมดไปใส่เข้าไว้ในลิสต์

C	D	E	F	G	NULL				

ลิสต์ 3. หลังจากเก็บข้อมูลจาก "B" แล้วปรากฏว่ามีไฮเปอร์ลิงก์ไปยัง "E", "F" และ "G" เมื่อนำไปตรวจสอบแล้วจึงนำยูอาร์แอลทั้งหมดที่ผ่านการตรวจสอบไปใส่ไว้ในลิสต์

D	E	F	G	I	NULL				

ลิสต์ 4. หลังจากเก็บข้อมูลจาก "C" ปรากฏว่ามีไฮเปอร์ลิงก์ไปยัง "I" จึงนำ "I" ไปตรวจสอบและเพิ่ม "I" เข้าไปใส่ไว้ในลิสต์

E	F	G	I	NULL					

ลิสต์ 5. หลังจากเก็บข้อมูลจาก "D" ปรากฏว่าไม่มีไฮเปอร์ลิงก์ไปยังเว็บเพจใด ๆ ดังนั้น จึงไม่มี ยูอาร์แอลใส่เพิ่มเข้าไปในลิสต์

F	G	I	NULL						

ลิสต์ 6. หลังจากเก็บข้อมูลจาก "E" ปรากฏว่ามีไฮเปอร์ลิงก์ไปยัง "F" ซึ่งเมื่อตรวจสอบแล้ว ปรากฏ ว่า "F" มีอยู่ในลิสต์ของ ยูอาร์แอลที่จะต้องทำการเก็บข้อมูลอยู่แล้ว ดังนั้นจึงไม่ต้องเพิ่มเข้าไปในลิสต์อีก เพื่อไม่ให้เกิดการติดต่อขอข้อมูล 2 ครั้ง ซึ่งจะทำให้เสียเวลา และอาจทำให้เกิดการวนลูป (Loop) ขึ้นได้ ทำให้หุ่นยนต์ ไม่สามารถหยุดการทำงานได้

G	I	NULL							

ลิสต์ 7. หลังจากเก็บข้อมูลจาก "F" แล้ว เนื่องจาก "F" ไม่มีไฮเปอร์ลิงก์ไปยังเว็บเพจใด ๆ จึงไม่ต้องเพิ่มยูอาร์แอลใด ๆ เข้าไปใส่ไว้ในลิสต์

I	H	NULL							

ลิสต์ 8. หลังจากเก็บข้อมูลจาก "G" แล้ว ปรากฏว่า "G" มีไฮเปอร์ลิงก์ต่อไปยัง "B", "I" และ "H" ซึ่งเมื่อตรวจสอบดูแล้วปรากฏว่า "I" อยู่ในลิสต์ของยูอาร์แอลที่หุ่นยนต์เก็บข้อมูลต้อง ทำการติดต่ออยู่แล้ว ดังนั้นจึงไม่ต้องเพิ่ม "I" เข้าไปในลิสต์อีก ส่วน "B" เป็นเว็บเพจที่ หุ่นยนต์เก็บข้อมูลเคยเก็บข้อมูลมาแล้ว ดังนั้นจึงไม่ต้องมีการนำเข้าไปใส่เพิ่ม ไว้ในลิสต์อีก เนื่องจากจะทำให้เสียเวลาในการทำงานมากขึ้นโดยที่ไม่จำเป็น และอาจทำให้เกิดการวนลูป ทำให้หุ่นยนต์ไม่สามารถหยุดการทำงานได้ แต่ "H" เป็นเว็บเพจใหม่ที่ "G" ลิงก์ไปถึง และเป็นเว็บเพจที่ยังไม่เคยถูกเก็บข้อมูล จึงเอาไปใส่เพิ่มไว้ในลิสต์

NULL									

ลิสต์ 9. หลังจากเก็บข้อมูลจาก "I" และ "H" แล้ว ปรากฏว่าลิสต์ว่าง เนื่องจากเว็บเพจทั้งหมดที่อยู่ในลิสต์ถูกเก็บข้อมูลหมดแล้ว ดังนั้น การทำงานของหุ่นยนต์ค้นหาข้อมูลจึงถูกสั่งให้หยุดลง ณ. จุดนี้

ข้อสังเกตุ : จากรูปจะเห็นได้ว่า ยูอาร์แอลของเว็บเพจ "Z" จะไม่มีทางเข้าไปอยู่ในลิสต์ของ ยูอาร์แอล เนื่องจากไม่มีเส้นทางจาก "A" ที่ลิงก์ไปจนถึง "Z" เลย ทำให้ "Z" เป็นเว็บเพจที่ หุ่นยนต์ค้นหา ข้อมูลไม่สามารถเก็บข้อมูลใส่ไว้ในฐานข้อมูลได้ ทำให้ฐานข้อมูลที่ใช้เก็บคีย์เวิร์ด ไม่สมบูรณ์ ครบถ้วน วิธีแก้ไขง่าย ๆ คือ นำยูอาร์แอลของ "Z" ไปใส่ในลิสต์ตั้งแต่ตอนเริ่มต้น เพื่อให้หุ่นยนต์เก็บ ข้อมูลทำการเก็บข้อมูลตั้งแต่ตอนแรก แต่การแก้ไขวิธีนี้ ต้องตรวจสอบแล้วว่า ยูอาร์แอลของ "Z" เป็นยูอาร์แอลที่อยู่ในประเทศหรือเป็นยูอาร์แอลที่สนใจ ก่อนที่จะนำไปใส่ ไว้ในลิสต์ตั้งแต่ตอนเริ่มต้น

การกำหนดจุดสิ้นสุดการทำงานของหุ่นยนต์เก็บข้อมูล

การกำหนดจุดสิ้นสุดของการทำงานของหุ่นยนต์นับว่าเป็นส่วนสำคัญส่วนหนึ่งของการทำงาน เนื่องจากเครือข่ายอินเตอร์เน็ตเป็นเครือข่ายที่มีขนาดใหญ่มาก ถ้าต้องการให้หุ่นยนต์ ค้นหาทำการเก็บ ข้อมูลจากทุกเว็บเพจที่ปรากฏอยู่ในอินเตอร์เน็ตจะต้องใช้เวลาอย่างมาก ดังนั้นจึงมีวิธีการในการ กำหนดจุดสิ้นสุดในการเก็บข้อมูล เพื่อเป็นทางเลือกในการเก็บข้อมูล และ อัพเดท (Update) ข้อมูล ทำให้การเก็บข้อมูลใช้เวลารวดเร็วขึ้นและตรงตามความต้องการของผู้ใช้มากขึ้น ดังนี้ คือ

1. หยุดการทำงานเมื่อลิสต์ว่าง

วิธีนี้เป็นการเก็บข้อมูลที่สมบูรณ์และครบถ้วนที่สุด เนื่องจากหุ่นยนต์จะทำการค้นหา และทำการเก็บข้อมูลจากทุกเว็บเพจที่สามารถติดต่อไปถึงได้ วิธีนี้เป็นวิธีที่มั่นใจได้ว่าเว็บเพจใด ๆ ก็ตามที่มีเส้นทางการลิงก์จากเว็บเพจของยูอาร์แอลเริ่มต้นในลิสต์ไปถึงแล้วจะต้องได้ถูกเก็บข้อมูลไว้เป็นฐานข้อมูล เพื่อใช้ในการค้นหาแน่นอน

2. หยุดการทำงานเมื่อหุ่นยนต์เก็บข้อมูลทำงานจนครบระดับความลึก (Level of Breadth-First Search) ตามที่กำหนดไว้

วิธีนี้เป็นการกำหนดระดับความลึกของการเก็บข้อมูล และให้หุ่นยนต์เก็บข้อมูลสิ้นสุดการทำงานและทำให้ลิสต์ว่างเมื่อระดับความลึกของการเก็บข้อมูลเท่ากับค่าที่กำหนดไว้ โดยระดับความลึกของการเก็บข้อมูล คือ ระดับของการลิงก์จากเว็บเพจหนึ่งไปยังอีกเว็บเพจหนึ่ง เมื่อเทียบกับเว็บเพจเริ่มต้น โดยกำหนดให้ยูอาร์แอลของเว็บเพจเริ่มต้นมีระดับความลึก ของการเก็บข้อมูลเป็น 1 และ ยูอาร์แอลของเว็บเพจที่ถูกลิงก์ไปถึงมีค่าระดับความลึกเป็น 2, 3, ...

ตัวอย่าง

จากรูปข้างบน เมื่อต้องการให้หุ่นยนต์ทำการเก็บข้อมูลจนถึงระดับความลึก 2

เนื่องจาก	A	มีค่าระดับความลึกเท่ากับ	1
	B, C, D	มีค่าระดับความลึกเท่ากับ	2
	E, F, G, I	มีค่าระดับความลึกเท่ากับ	3
	H	มีค่าระดับความลึกเท่ากับ	4
	Z	ไม่มีระดับความลึก เนื่องจาก ไม่สามารถลิงก์ไปถึง	

ดังนั้น A, B, C และ D เท่านั้นที่จะถูกเก็บเป็นฐานข้อมูล

3.3.5 การกำหนดลักษณะการทำงานของหุ่นยนต์เก็บข้อมูล

เนื่องจากอินเตอร์เน็ตเป็นเครือข่ายที่มีขนาดใหญ่ ในการเก็บข้อมูลแต่ละครั้งจึงต้องใช้เวลามาก การที่จะให้หุ่นยนต์ทำการเก็บข้อมูลทุกครั้งจากทุกเว็บเพจที่มีอยู่ในอินเตอร์เน็ตจึงเป็นการกระทำ ที่ต้องใช้เวลามาก และอาจจะทำให้เสียเวลาโดยใช่เหตุ เนื่องจากข้อมูลที่ถูกส่งมาเป็น ข้อมูลที่มีอยู่ใน ฐานข้อมูลอยู่แล้ว และส่วนใหญ่เว็บเพจจะไม่มีการเปลี่ยนแปลงบ่อย ๆ ดังนั้นจึงกำหนดลักษณะการ ทำงานของหุ่นยนต์เก็บข้อมูลออกมา 2 ลักษณะ ดังนี้

1. เก็บข้อมูลจากทุกเว็บเพจ

วิธีนี้จะทำการเก็บข้อมูลจากทุกเว็บเพจที่สามารถลิงก์ไปถึงได้ และเมื่อได้ข้อมูลมาแล้ว จะทำการตรวจสอบในฐานข้อมูล ถ้าปรากฏอยู่ในฐานข้อมูลอยู่แล้ว จะทำการตรวจสอบว่าข้อมูลที่ได้มาใหม่เปลี่ยนแปลงไปจากฐานข้อมูลเดิมที่มีอยู่แล้ว หรือไม่ ถ้าเหมือนกัน หุ่นยนต์จะทำการเก็บข้อมูลจากเว็บเพจอื่นต่อไป แต่ถ้าข้อมูลใหม่ที่ได้ไม่เหมือนกับข้อมูลเก่าที่มีอยู่ หุ่นยนต์เก็บข้อมูลจะส่งข้อมูลใหม่ที่ได้มาให้กับส่วนวิเคราะห์แล้วทำการเปลี่ยนแปลงฐานข้อมูล เพื่อให้เป็นปัจจุบัน

วิธีในการตรวจสอบการอัพเดตของข้อมูล

เนื่องจากการในฐานข้อมูลเก็บเพียงคีย์เวิร์ดของเว็บเพจเท่านั้น ไม่ได้เก็บข้อมูลทั้งหมด ทำให้ไม่สามารถนำข้อมูลที่ได้มาใหม่เปรียบเทียบกับของเก่าได้ตรง ๆ ในฐานข้อมูลจึงมีส่วนหนึ่ง ที่ใช้เก็บข้อมูลเกี่ยวกับเว็บเพจนั้น เช่น ชื่อของยูอาร์แอล (URL Name), หัวข้อของยูอาร์แอล (Title), ขนาดของเว็บเพจ (Size) เป็นต้น และส่วนที่นำมาใช้ในการตรวจสอบว่า ข้อมูลใหม่ที่ได้ มีการเปลี่ยนแปลงไปจากข้อมูลที่มีอยู่ในฐานข้อมูลเดิมหรือไม่ คือ ขนาดของเว็บเพจ นั่นคือถ้า ขนาดของข้อมูลที่ได้มาใหม่ไม่เท่ากับขนาดของเว็บเพจเดิม แสดงว่าข้อมูลที่ได้มีการเปลี่ยนแปลง

ข้อจำกัด วิธีนี้จะใช้ไม่ได้ถ้าเว็บเพจมีการเปลี่ยนแปลงข้อความภายในเว็บเพจแต่ว่ามีขนาดไม่มีการเปลี่ยน แปลง

ข้อดีของการเก็บข้อมูลจากทุกเว็บเพจ คือ

ข้อมูลเป็นปัจจุบันเมื่อมีการเก็บข้อมูลใหม่และทันสมัยทุกครั้งเมื่อมีการเก็บข้อมูล นั่นคือข้อมูลที่อยู่ในฐานข้อมูลจะมีการเปลี่ยนแปลงให้ตรงกับข้อมูลของเว็บเพจเมื่อเว็บเพจมีการเปลี่ยนแปลง

ข้อเสียของการเก็บข้อมูลจากทุกเว็บเพจ คือ

เสียเวลาในการรับข้อมูลจากเว็บเพจไปโดยเปล่าประโยชน์ ถ้าข้อมูลของเว็บเพจส่วนใหญ่ที่ได้มาไม่มีการเปลี่ยนแปลงจากในฐานข้อมูลเดิมที่เคยมีอยู่แล้ว ดังนั้นการเก็บข้อมูลโดย ใช้วิธีนี้จึงไม่ควรใช้บ่อยครั้งมากนัก ควรใช้เมื่อทำการเก็บข้อมูลครั้งแรก หรือใช้เมื่อมั่นใจว่า มีเว็บเพจหลายเว็บเพจเปลี่ยนแปลง อาจจะใช้นาน ๆ ครั้ง เช่น ทุกสัปดาห์ หรือ ทุกเดือน เป็นต้น

2. เก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล

การเก็บข้อมูลวิธีนี้จะมีการตรวจสอบ ยูอาร์แอล ก่อนที่จะนำไปใส่ไว้ในลิสต์ของ ยูอาร์แอลที่จะให้หุ่นยนต์ทำการเก็บข้อมูล เพื่อดูว่ายูอาร์แอลของเว็บเพจที่จะให้หุ่นยนต์ทำ การเก็บข้อมูลนี้มีอยู่ในฐานข้อมูลหรือไม่ ถ้ามียูอาร์แอลที่ตรวจสอบอยู่แล้ว ก็ไม่ต้องนำยูอาร์แอล นี้ไปใส่เพิ่มเข้าไปไว้ในลิสต์อีก แต่ถ้ายังไม่มีในฐานข้อมูล ก็นำยูอาร์แอลนี้ไปใส่เพิ่มไว้ ในลิสต์ ของยูอาร์แอลที่หุ่นยนต์ต้องทำการเก็บข้อมูล

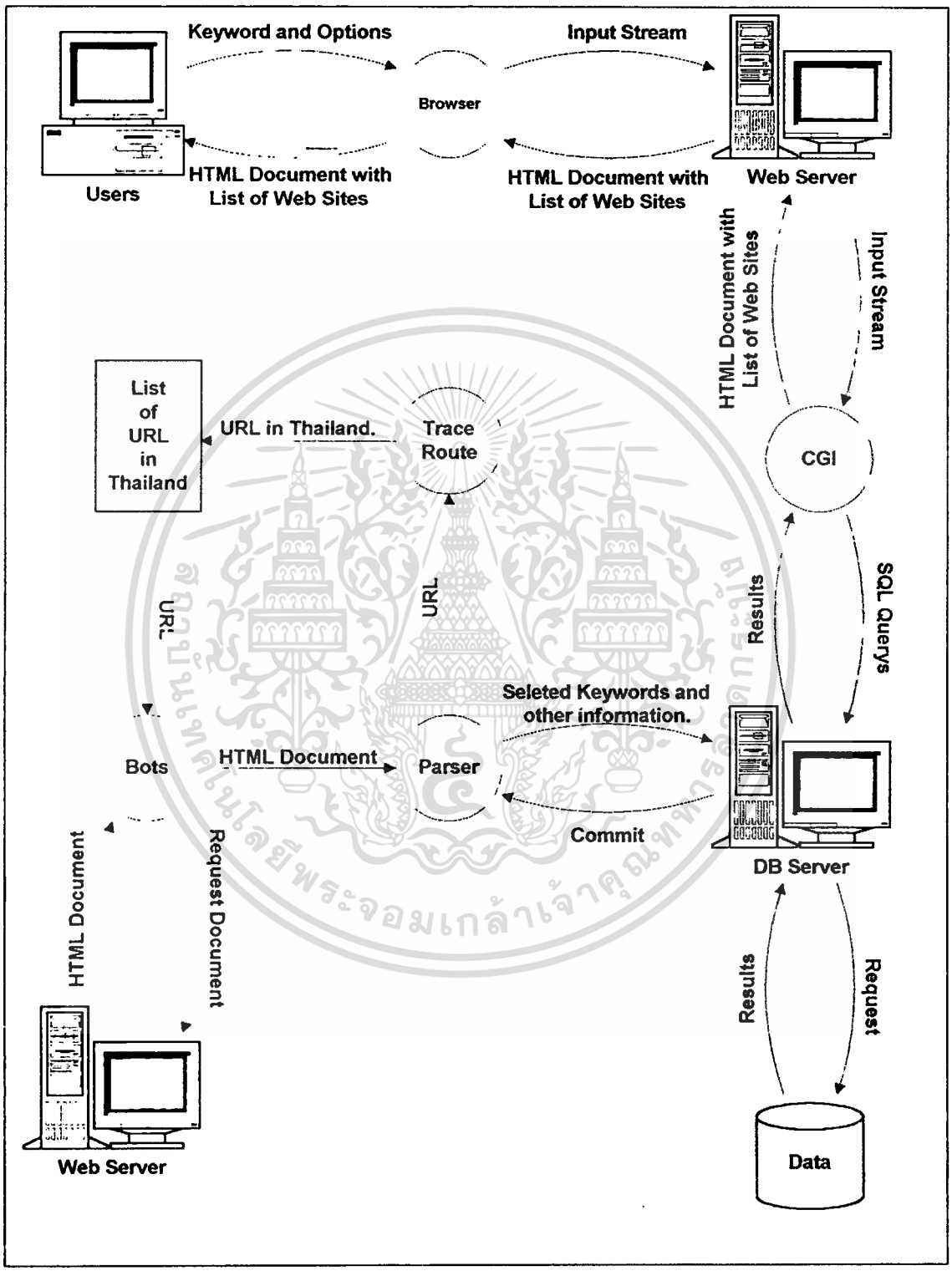
ข้อดีของการเก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล คือ

การเก็บข้อมูลใช้เวลาน้อยลง เนื่องจากไม่ต้องเสียเวลากับการเก็บข้อมูล ที่เคยมีอยู่แล้วในฐานข้อมูล ทำให้สามารถใช้ได้บ่อยครั้งกว่าวิธีแรก เนื่องจากใช้เวลาน้อยกว่า

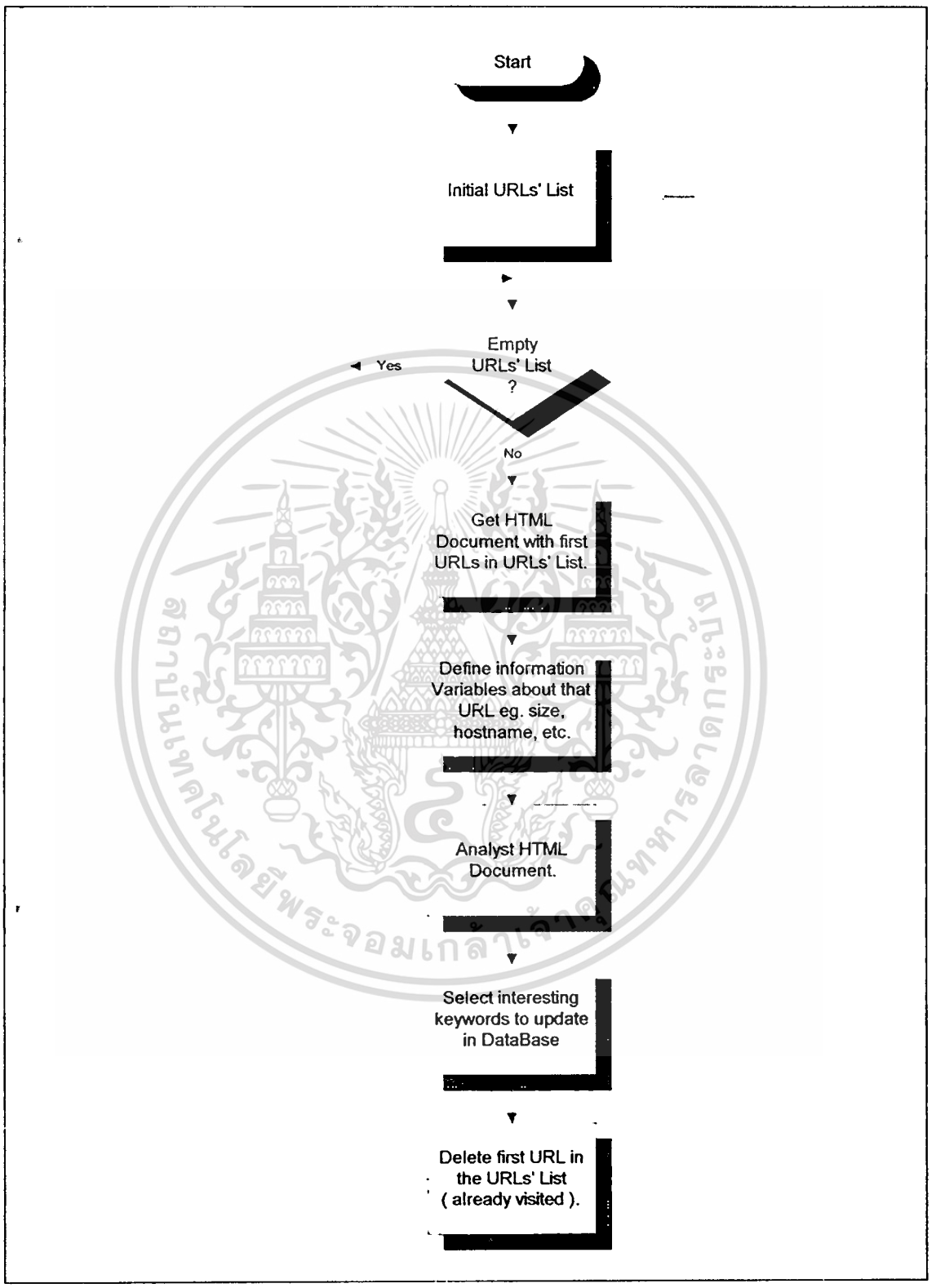
ข้อเสียของการเก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล คือ

ข้อมูลที่มีอยู่ในฐานข้อมูลเป็นข้อมูลเก่า ทำให้การค้นหาข้อมูลได้ข้อมูลไม่ตรงกับความเป็นจริงในปัจจุบัน เนื่องจากเว็บเพจที่ได้จากการค้นหาอาจมีการเปลี่ยนแปลงไปแล้ว หรือ ยูอาร์แอล นั้นถูกยกเลิก การเก็บข้อมูลด้วยวิธีนี้จึงเป็นวิธีที่ควรจะใช้เมื่อต้องการเก็บข้อมูลจาก เว็บเพจที่เกิดขึ้นใหม่ โดยมั่นใจว่าข้อมูลที่มีอยู่ในฐานข้อมูลเดิมไม่มีการเปลี่ยนแปลงมากนัก แต่เมื่อเว็บเพจปัจจุบันมีการเปลี่ยนแปลงไปมาก การเก็บข้อมูลด้วยวิธีนี้จึงไม่เหมาะสมเท่าวิธีแรก

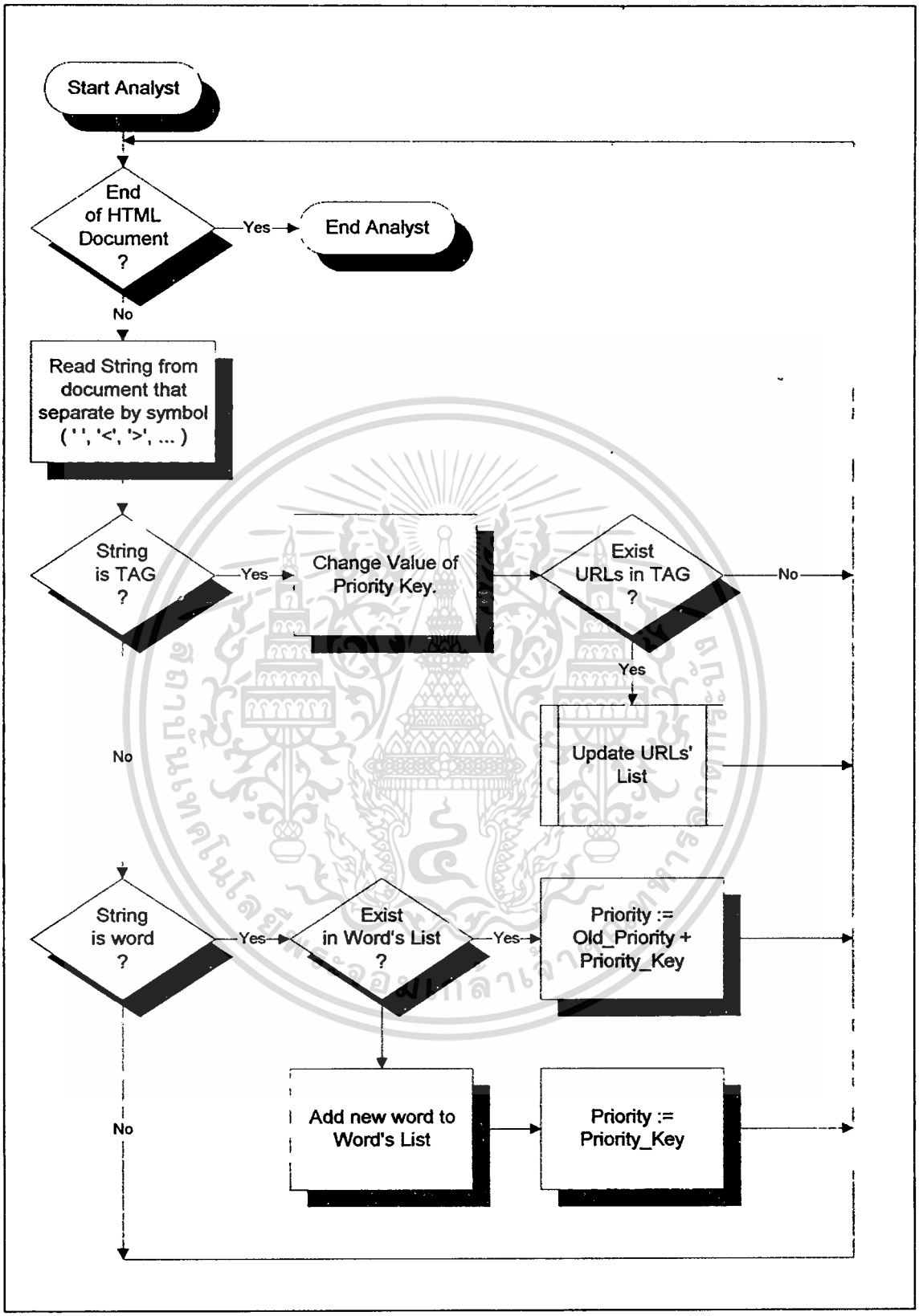
ส่วนแสดงขั้นตอนการทำงานของโรบอททั้งหมด



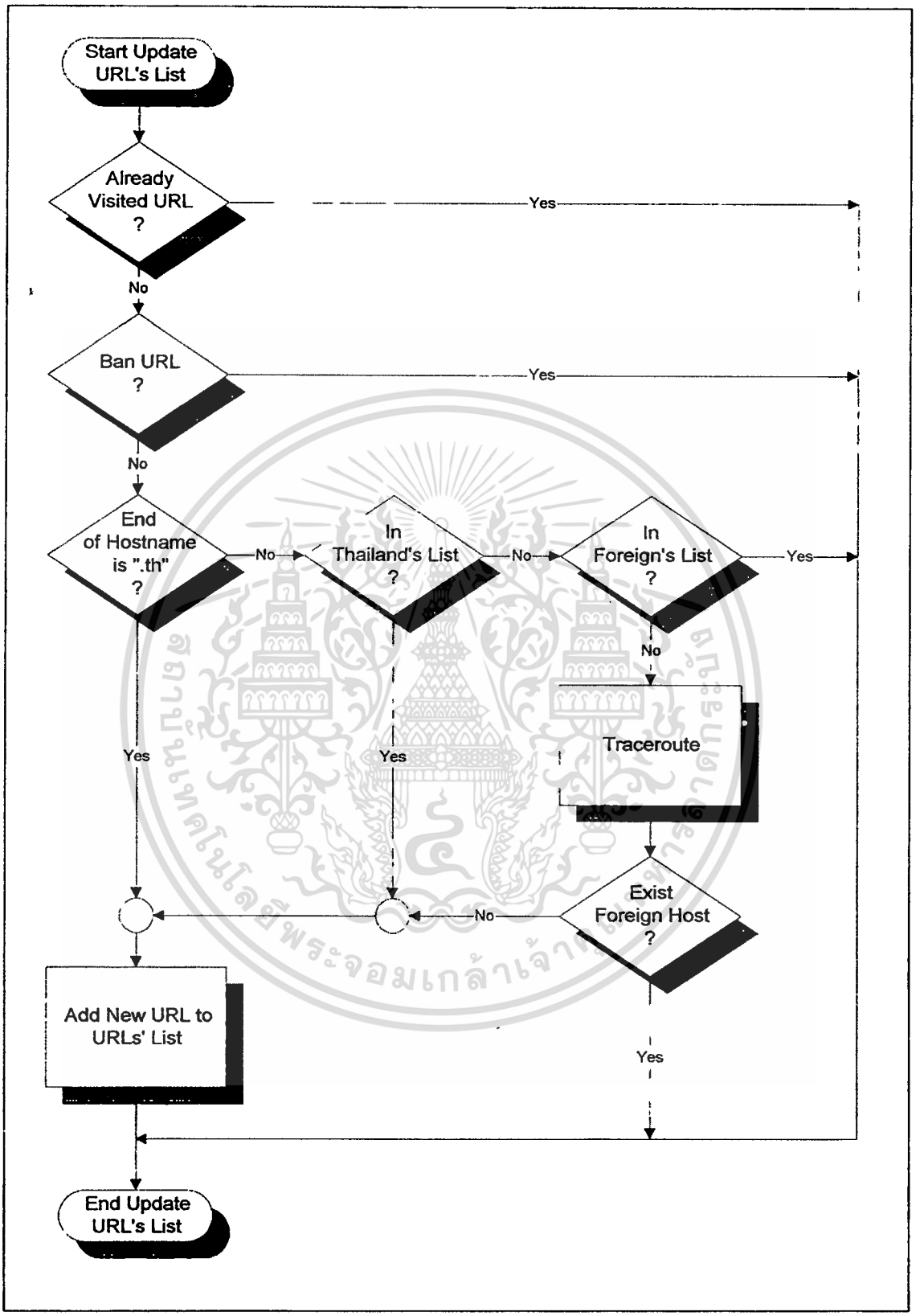
รูปที่ 3.2 ความสัมพันธ์ระหว่างแต่ละส่วนของระบบค้นหาข้อมูล



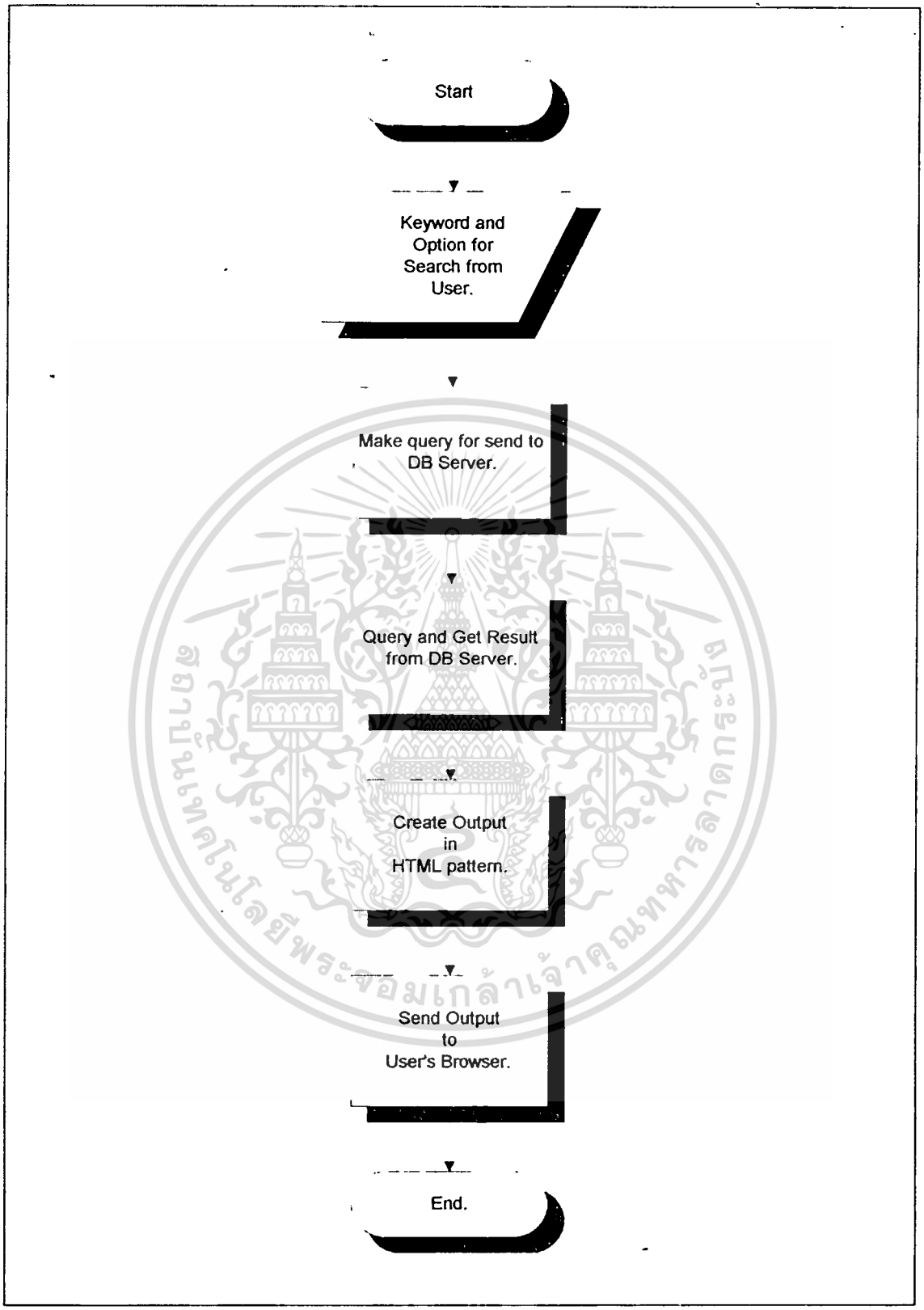
รูปที่ 3.3 แสดงแผนภาพการเก็บข้อมูลของโรบอท



รูปที่ 3.4 การวิเคราะห์เอกสาร HTML



รูปที่ 3.5 การตรวจสอบ URLs

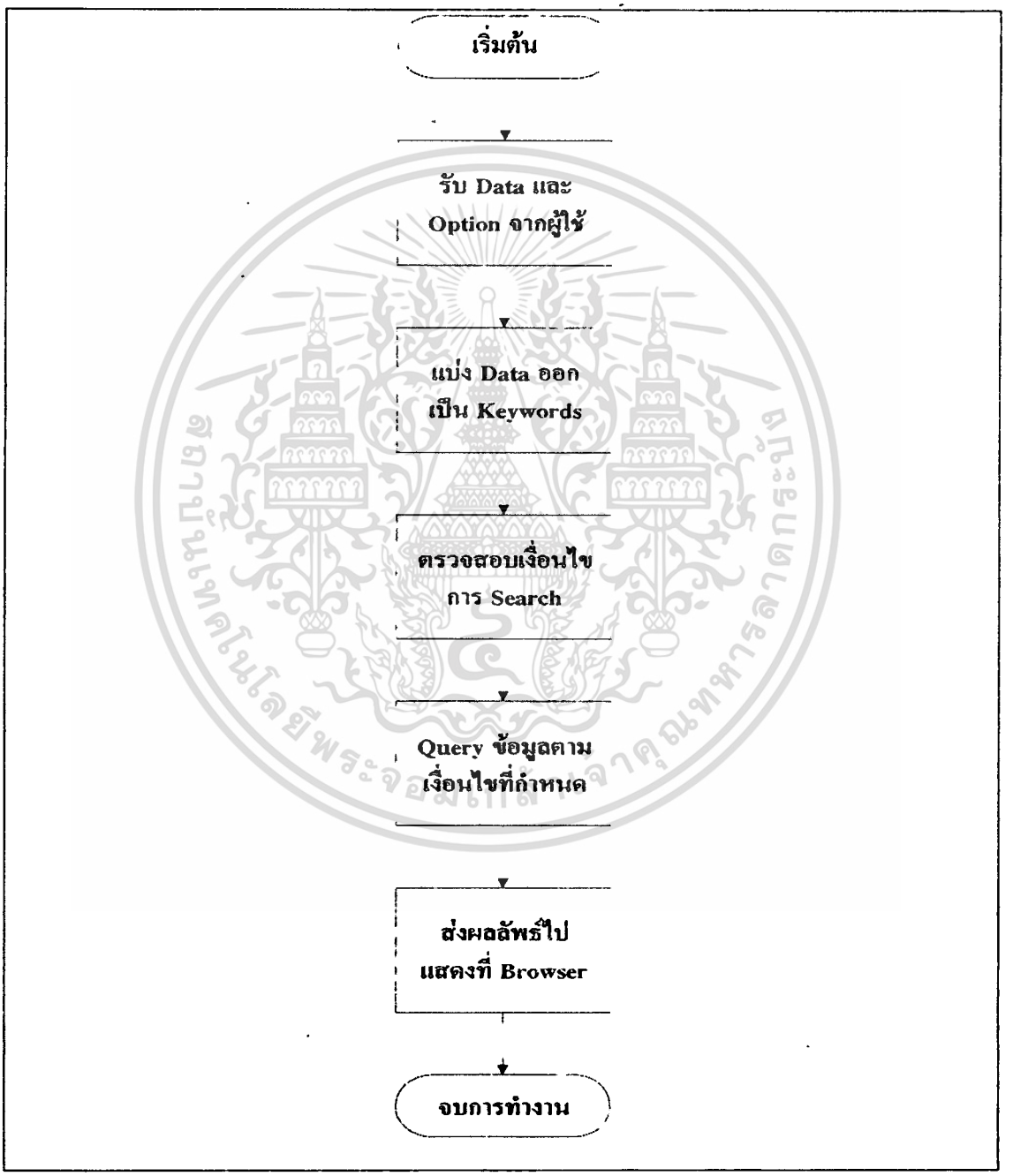


รูปที่ 3.6 ส่วนติดต่อกับผู้ใช้

เมื่อทำการออกแบบหน้าที่และการทำงานของส่วนแรกเรียบร้อยแล้ว เราก็เริ่มที่จะทำการออกแบบในส่วนต่อไปนั่นคือส่วนของซีจีไอ

3.4 การออกแบบในส่วนของการติดต่อระหว่างผู้ใช้งานกับระบบฐานข้อมูล

การทำการติดต่อระหว่างผู้ใช้งานกับระบบฐานข้อมูล สามารถที่จะทำการติดต่อได้ผ่านทางโปรแกรม ซีจีไอ โดยการทำงานหลักของโปรแกรม ซีจีไอ เป็นไปดังนี้

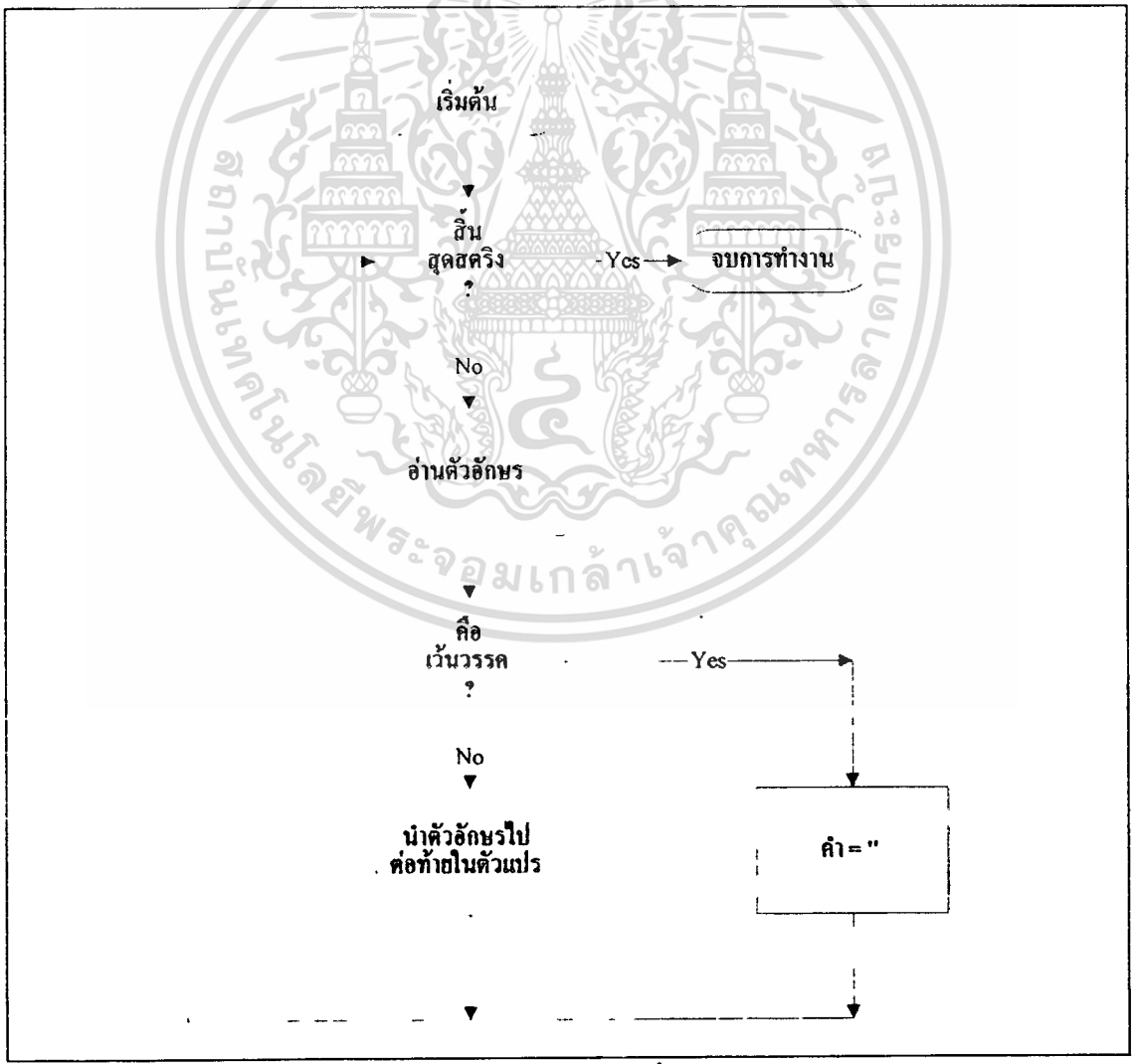


รูปที่ 3.7 แสดงการทำงานหลักของโปรแกรมซีจีไอ

โปรแกรมซีจีไอจะทำหน้าที่ในการส่งค่าข้อมูลที่ได้รับจากผู้ใช้บริการ ให้แก่เซิร์พเวอร์ และ ทำการรับค่าผลลัพธ์ออกมา ส่งกลับให้ผู้ใช้บริการโดยผ่านทางบราวเซอร์ โดยข้อมูลที่ส่งผ่านโปรแกรมซีจีไอในที่นี้มีดังนี้

1. สติงที่เป็นคำที่ผู้ใช้บริการต้องการค้นหา
2. ค่าของออปชันในการเลือกแบบในการค้นหา ซึ่งมีอยู่ 3 ค่า
   - ค่า Case-sensitive
   - ค่าของกรณี OR
   - ค่าของกรณี AND

แต่สำหรับโปรแกรมซีจีไอของโครงงานหุ่นยนต์ค้นหาข้อมูลนี้ จะต้องทำการตัดสตริงคำที่ผู้ใช้ บริการใส่มาเพื่อทำการค้นหา ออกเป็นคำย่อย ๆ ก่อนที่จะส่งให้กับเซิร์พเวอร์ เพื่อใช้ในการค้นหาข้อมูลในระบบฐานข้อมูล โดยส่วนของการตัดคำนี้มี Algorithm ดังนี้



รูปที่ 3.8 แสดงขั้นตอนในการตัดคำที่รับมาจากผู้ใช้บริการ

นอกจากโปรแกรมซีจีไอจะทำหน้าที่ในการส่งข้อมูลให้แก่เซิร์ฟเวอร์แล้ว ก็ยังทำหน้าที่ในการส่งข้อมูลกลับให้ผู้ใช้บริการด้วย โดยค่าที่ส่งกลับมามีดังนี้

1. Http header
2. คำที่ผู้ใช้ต้องการค้นหา
3. จำนวนยูอาร์แอลที่ค้นพบว่ามีคำที่ต้องการ
4. ยูอาร์แอลที่พบคำที่ต้องการ
5. Title ของยูอาร์แอลที่พบคำที่ต้องการ
6. Description ของยูอาร์แอลที่พบคำที่ต้องการ
7. ขนาดของเว็บเพจที่มีคำที่ต้องการ

3.5 การออกแบบในส่วนของการจัดเก็บและค้นหาข้อมูล

ส่วนจัดเก็บข้อมูล

ในการออกแบบส่วนที่จัดเก็บข้อมูลของโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ เราจะทำการออกแบบตามหลักการออกแบบ ที่เรียกว่า อีอาร์โมเดล แต่ก่อนอื่นต้องทำการรวบรวมข้อมูลทั้งหมดที่เกี่ยวข้องกับโครงงานก่อน แล้วหลังจากนั้นจึงค่อยทำการออกแบบ

- ข้อมูลเกี่ยวกับคำที่จะทำการเก็บไว้เพื่อทำการค้นหาต้องมีดังต่อไปนี้
    1. คำที่เก็บไว้สำหรับค้นหา
    2. คำที่ถูกแปลงเป็นตัวเล็กหมดสำหรับการค้นหาแบบไม่พิจารณารูปแบบ
    3. ค่าความสำคัญของคำที่จัดเก็บ
    4. ยูอาร์แอลที่คำดังกล่าวอยู่
- ข้อมูลเกี่ยวกับยูอาร์แอลที่มีคำที่เก็บไว้อยู่
    1. ชื่อยูอาร์แอลที่มีคำอยู่
    2. Title ของยูอาร์แอล
    3. Description ของยูอาร์แอล
    4. ขนาดของไฟล์ที่มีคำอยู่

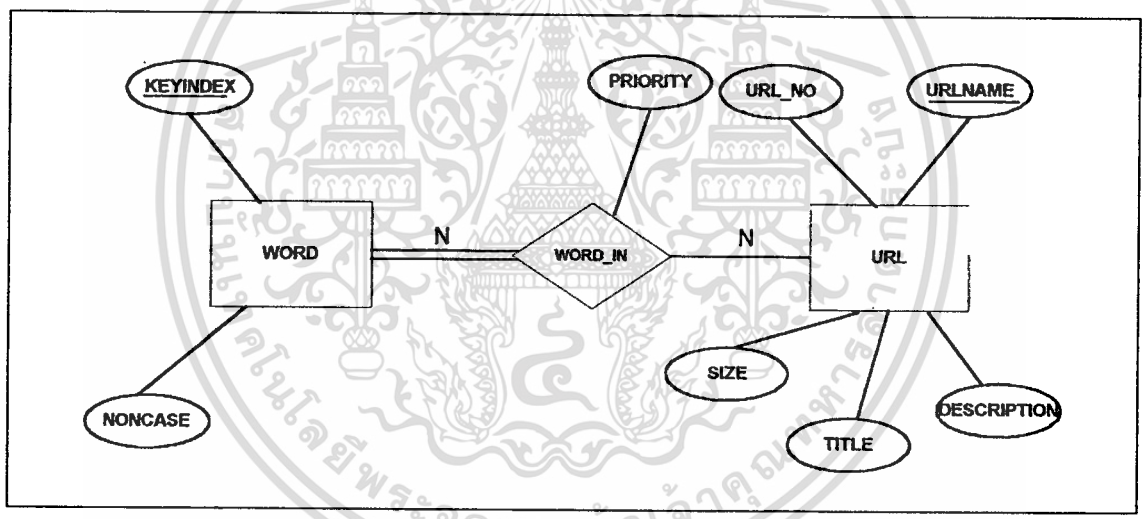
เมื่อทำการรวบรวมข้อมูลได้ครบ ตามที่ต้องการแล้ว ก็ทำการออกแบบตามหลักของ อีอาร์โมเดล เริ่มต้นก็ทำการกำหนด เอนติตี้ โดยในที่นี้มีการกำหนด เอนติตี้ได้ 2 เอนติตี้ คือ

- เอนติตี้ Word
- เอนติตี้ Url

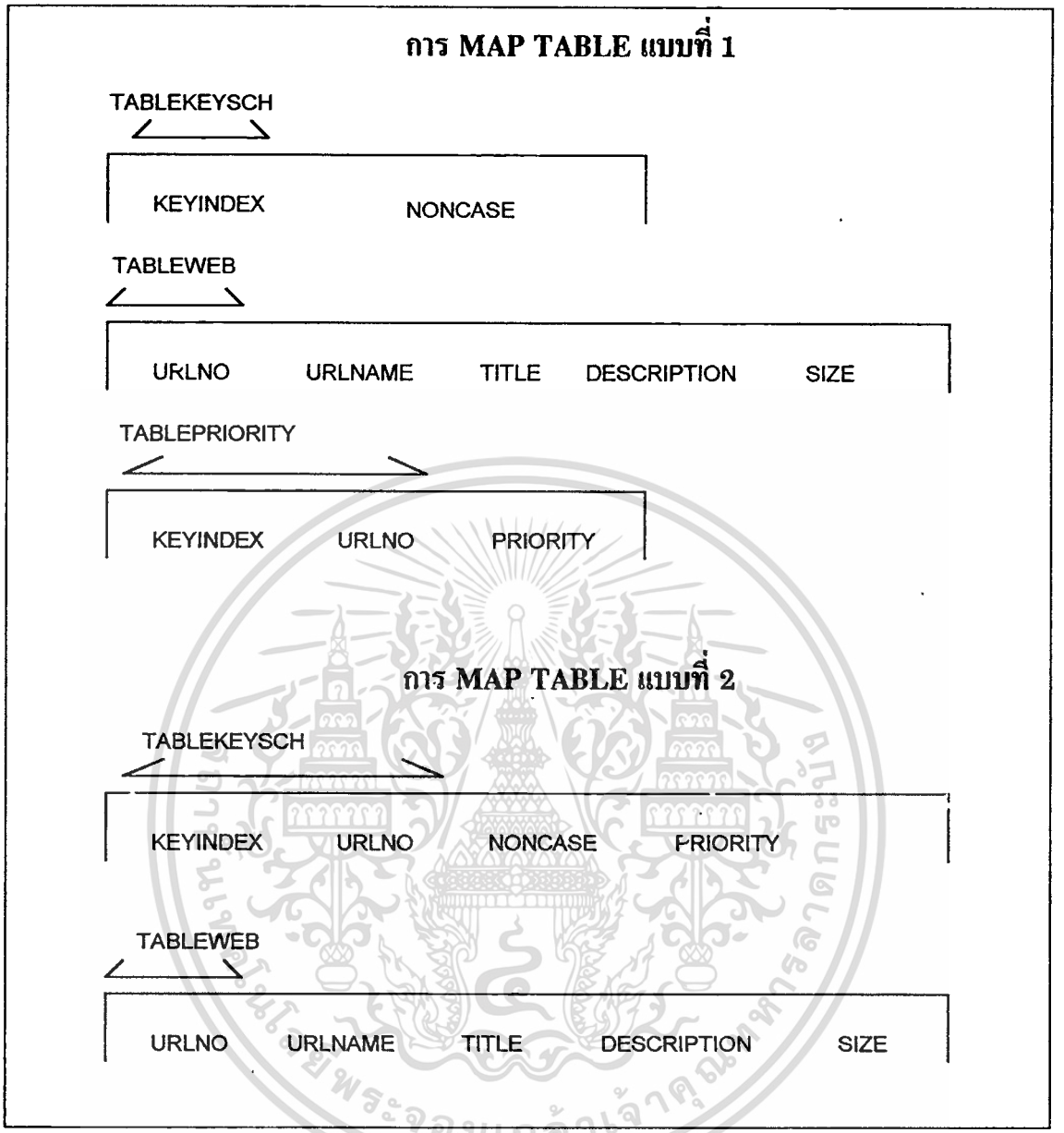
หลังจากนั้นก็ทำการกำหนด ค่า attribute ของแต่ละ เอนติตี้ นอกจากนี้ยังทำการพิจารณา ค่าความสัมพันธ์ระหว่างเอนติตี้ ซึ่งจากค่าความสัมพันธ์นี้ก็ทำให้เกิดค่า priority ขึ้น เมื่อทำการกำหนด ค่าความสัมพันธ์ต่าง ๆ เรียบร้อยแล้วก็ทำการเขียนไดอะแกรมของฐานข้อมูล

เมื่อทำการเขียนไดอะแกรมเสร็จเรียบร้อยแล้ว ก็ทำการแปลงจากอีอาร์ โมเดล ลงมาเป็นตาราง บนฐานข้อมูล เพื่อทำการใช้งานได้จริง การแปลงเป็นตารางสำหรับ ไดอะแกรมนี้ สามารถทำได้ 2 แบบ โดยแบบที่ 1 จะแปลงได้เป็น 3 ตาราง ส่วนแบบที่ 2 จะแปลงได้เป็น 2 ตาราง ทางผู้ทำโครงงาน จึงทำการเปรียบเทียบข้อดีข้อเสียของแบบทั้งสอง ซึ่งผลคือเลือกตารางแบบที่ 2 ดังมีสาเหตุดังนี้

1. ตารางแบบที่สองช่วยให้การเขียนโปรแกรมภาษาเอสคิวแอลทำได้ง่ายขึ้น
2. ตารางแบบที่สองช่วยให้โปรแกรมที่ทำการค้นหาข้อมูลทำงานได้เร็วขึ้น
3. ตารางแบบที่สองช่วยให้การ update ข้อมูลทำได้ง่ายขึ้น



รูปที่ 3.9 ไดอะแกรมของ อีอาร์โมเดล

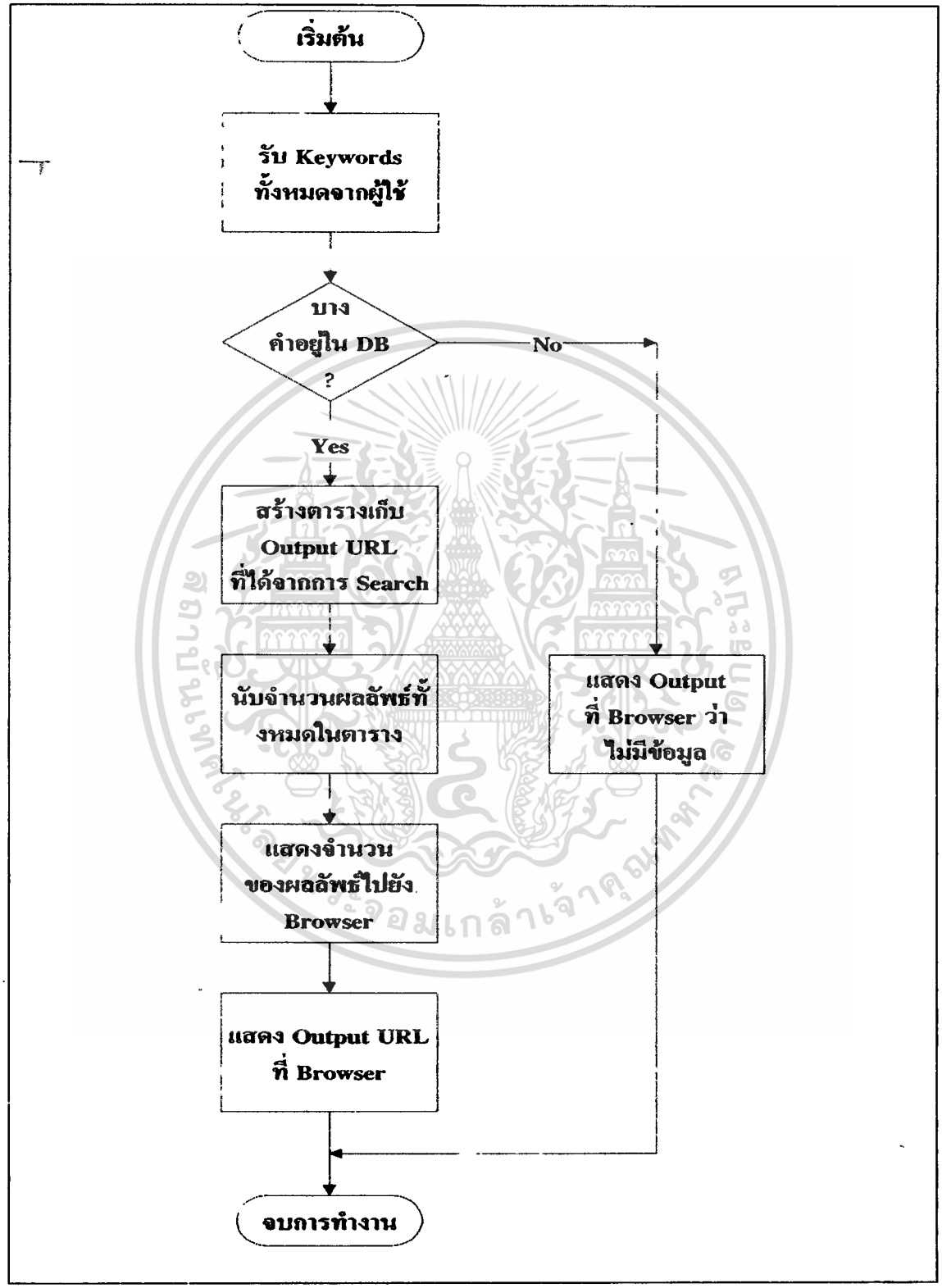


รูปที่ 3.10 ตารางที่แปลงได้จาก อีอาร์โมเดล

ส่วนค้นหาข้อมูล

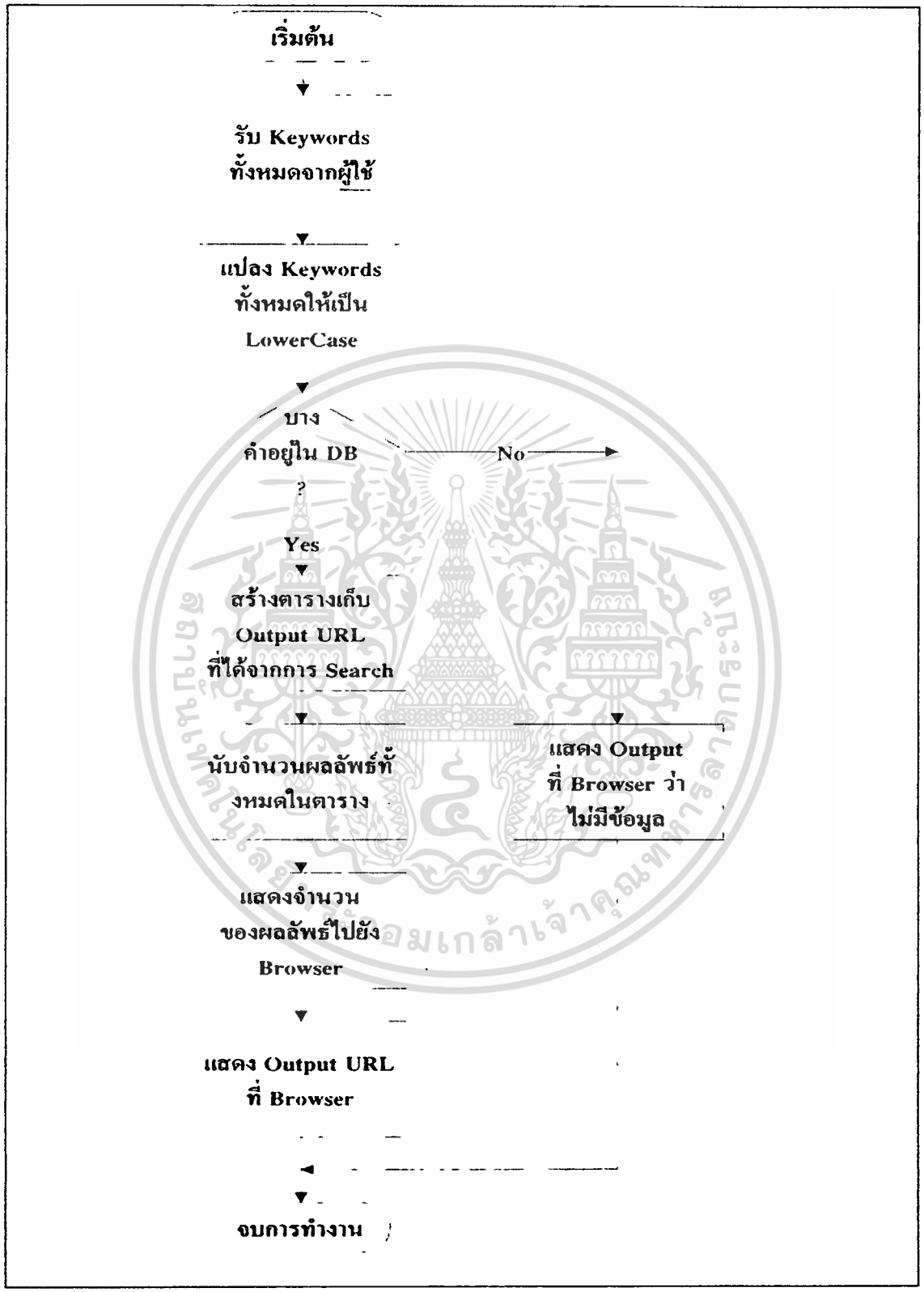
ส่วนค้นหาข้อมูลจะเป็นส่วนที่ทำหน้าที่ในการค้นหาข้อมูลจากระบบฐานข้อมูล โดยจะทำ การค้นหาตามข้อมูลที่ได้รับ อันได้แก่ คำที่ต้องการหา , ออปชั่นในการค้นหา ซึ่งเมื่อเซิร์พเวอร์ได้ รับการรีเควสบริการจากผู้ใช้ ก็จะทำการเรียกโปรแกรมเคลไฟขึ้นมาทำงาน กับระบบฐานข้อมูล เพื่อใช้ ในการค้นหาข้อมูล โดยในการค้นหาข้อมูลแต่ละแบบจะมี algorithm ตามต่อไปนี้

การค้นหาข้อมูลพิจารณารูปแบบ กรณี 'หรือ'



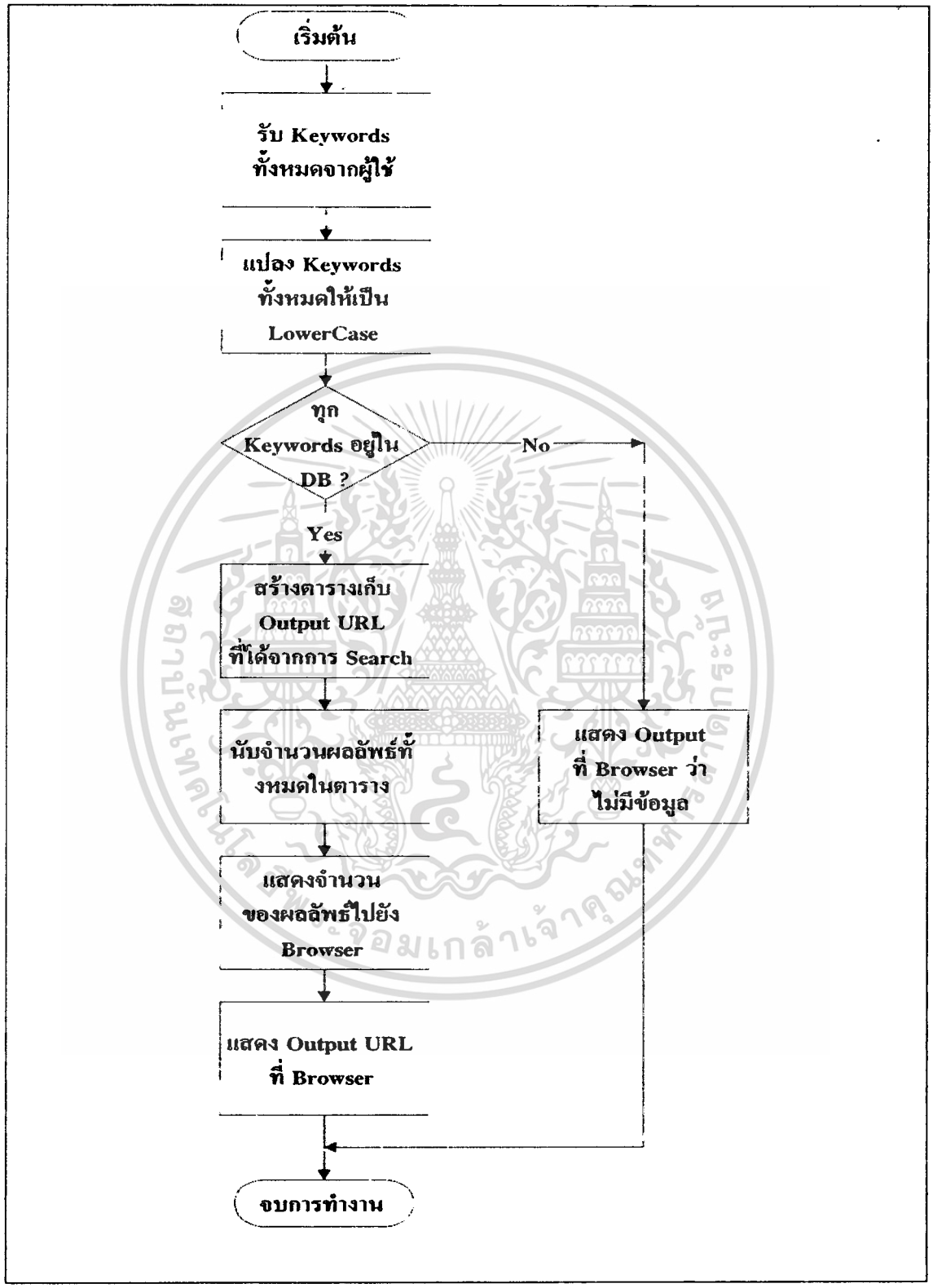
รูปที่ 3.11 ขั้นตอนการค้นหาข้อมูลพิจารณารูปแบบ กรณี ' หรือ '

การค้นหาข้อมูลไม่พิจารณารูปแบบ กรณี ' หรือ '



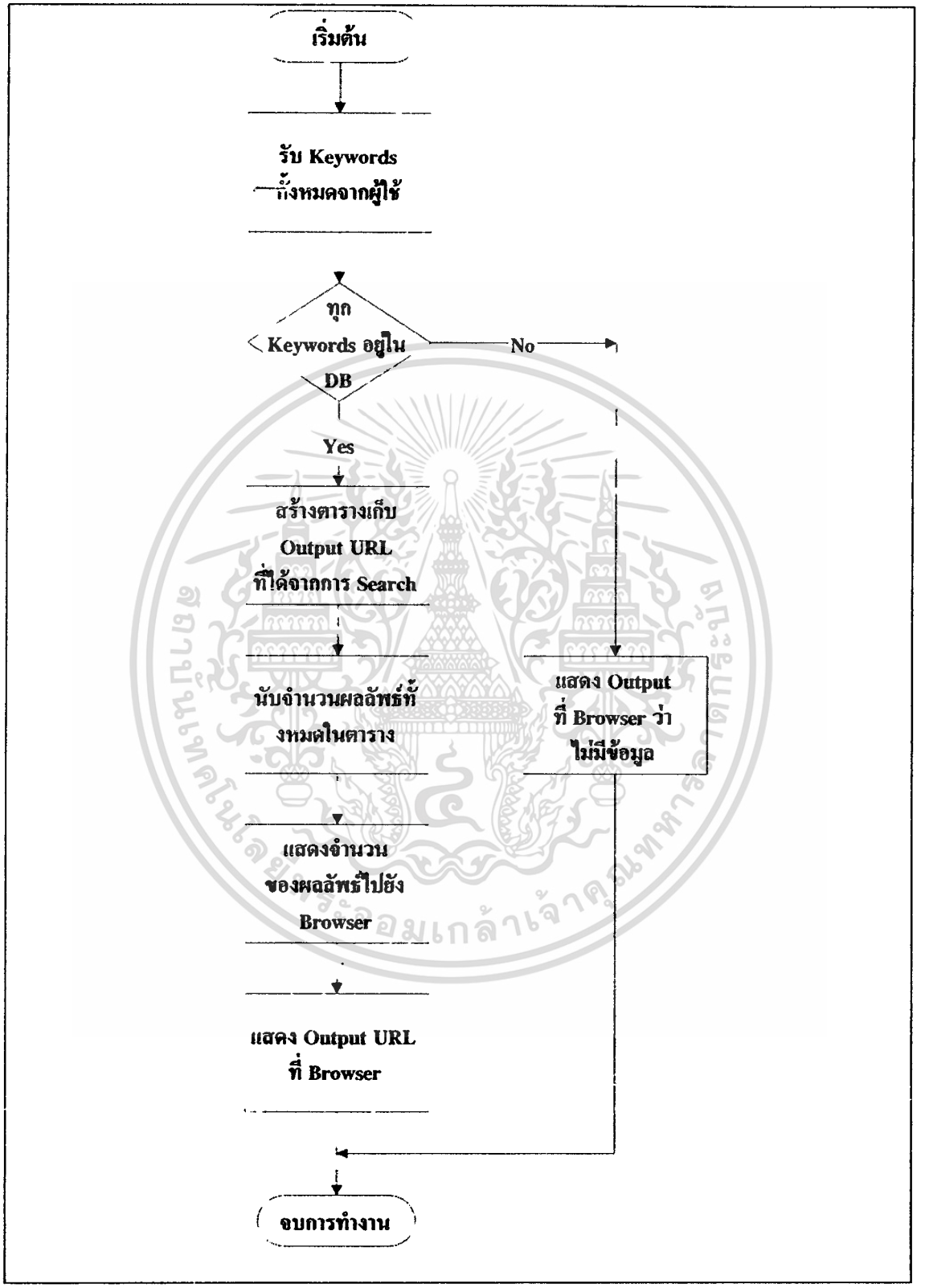
รูปที่ 3.12 ขั้นตอนการค้นหาข้อมูลไม่พิจารณารูปแบบ กรณี ' หรือ '

การค้นหาข้อมูลไม่พิจารณารูปแบบ กรณี ' และ '



รูปที่ 3.13 ขั้นตอนการค้นหาข้อมูลไม่พิจารณารูปแบบกรณี ' และ '

การค้นหาข้อมูลพิจารณารูปแบบ กรณี ' และ '

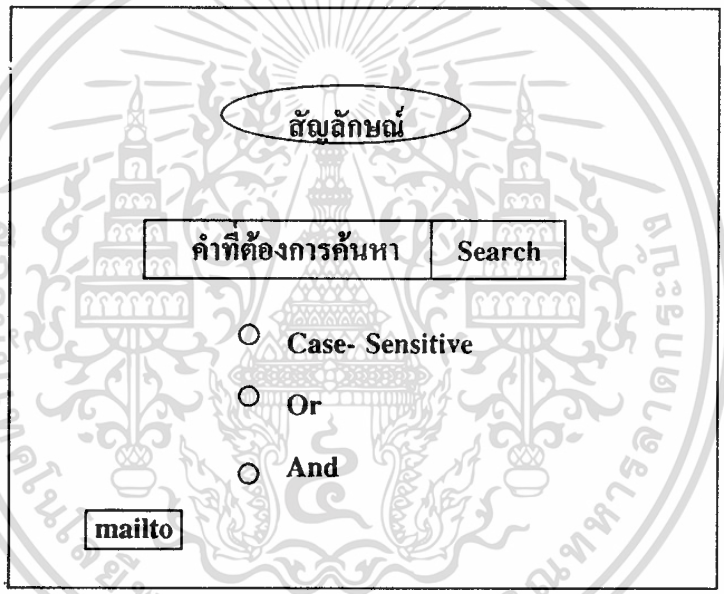


รูปที่ 3.14 ขั้นตอนการค้นหาข้อมูลพิจารณารูปแบบ กรณี ' และ '

3.6 การออกแบบในส่วนของหน้าจอที่ใช้ติดต่อกับผู้ใช้บริการ

ส่วนของหน้าจอที่ใช้ติดต่อกับผู้ใช้บริการ เป็นส่วนที่มีความสำคัญมาก เนื่องจากเป็นส่วนที่ ใช้ในการรับข้อมูลจากผู้ใช้บริการ และ เป็นส่วนที่ใช้แสดงผลลัพธ์ด้วย นอกจากนี้ยังอาจมีส่วนช่วย ในการดึงดูดความสนใจของผู้ใช้บริการ ให้มาใช้บริการ ดังนั้นการออกแบบหน้าจอ ทางผู้จัดทำ โครงงานจึงทำการออกแบบตามหลักการต่อไปนี้

1. หน้าจอต้องมีความสวยงามและเหมาะสมกับลักษณะงานของหน้าจอ
2. การใช้งานผ่านทางหน้าจอต้องสามารถทำได้ง่าย
3. คำอธิบายการใช้งานบนหน้าจอต้องมีความชัดเจนและมีข้อมูลเพียงพอ



รูปที่ 3.15 แสดงรูปแบบของหน้าจอที่ต้องการ

3.7 การเลือกแพลทฟอร์มในการทำโครงงาน

operating system จำนวนมากที่สามารถนำมาจัดทำเป็นแพลทฟอร์มในการทำโครงงาน เช่น ยูนิกซ์ , วินโดว์ เอ็นที , วินโดว์ 95 ซึ่ง os แต่ละตัวก็มีข้อดีข้อเสียที่แตกต่างกัน สำหรับโครงงานหุ่น ยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ เลือกที่จะใช้ วินโดว์ 95 เป็นแพลทฟอร์มในการทำงานเนื่อง จากสาเหตุต่อไปนี้

1. วินโดว์ 95 เป็นแพลทฟอร์มที่สามารถทำงานแบบ Multitasking ได้อย่างมีประสิทธิภาพ

2. โปรแกรมประยุกต์ที่สามารถทำงานบนวินโดว์ 95 ได้มีอยู่เป็นจำนวนมาก จึงเป็นผลดีต่อผู้ทำโครงงานในการที่เลือกใช้โปรแกรมประยุกต์ในการทำโครงงาน

3. ผู้จัดทำโครงงานมีความคุ้นเคยต่อวินโดว์ 95 พอสมควร ทำให้สะดวกในการใช้งานและทำ ให้ประหยัดเวลาในการทำการศึกษาเกี่ยวกับแพลทฟอร์มที่เลือกใช้

4.วินโดว์ 95 เป็นแพลทฟอร์มที่ได้รับความนิยมมาก ดังนั้นจึงเป็นการสะดวกต่อผู้ที่ต้องการ นำโครงงานนี้ไปทำการรันเพื่อการใช้งาน หรือศึกษาใน ภายหลัง

3.8 การเลือกเว็บเซิร์พเวอร์

เช่นเดียวกับแพลทฟอร์ม เว็บเซิร์พเวอร์ในปัจจุบันนี้มีให้เลือกเป็นจำนวนมาก โดยแต่ละตัวก็ มีขีดความสามารถที่แตกต่างกันออกไป ทางผู้จัดทำโครงงานจึงได้ทำการศึกษาถึงขีดความสามารถ แล้วทำการเลือกเว็บเซิร์พเวอร์ ที่เหมาะสมกับโครงงานซึ่งในที่นี้เลือกเว็บเซิร์พเวอร์ที่ชื่อเว็บไซท์

สาเหตุที่เลือกใช้เว็บไซท์เว็บเซิร์พเวอร์ มีดังนี้

1. เป็นเว็บเซิร์พเวอร์ที่มีขนาดเล็กทำให้ไม่เปลืองเนื้อที่ในการจัดเก็บ
2. เป็นเว็บเซิร์พเวอร์ที่มีขนาดเล็กซึ่งสามารถช่วยลดปัญหาในเรื่องของหน่วยความจำ ไม่พอ
3. สามารถทำงานได้ในส่วนที่ผู้จัดทำโครงงานต้องการ
4. เป็นเว็บเซิร์พเวอร์ที่เป็นที่แพร่หลายสามารถหามาใช้งานได้สะดวก
5. ผู้ที่ต้องการนำโครงงานนี้ไปใช้งานหรือทำการศึกษามีความสะดวกในการจัดหา เว็บไซท์มา ติดตั้งเพื่อทำการรันโครงงาน
6. เว็บไซท์สามารถทำงานได้บนวินโดว์ 95

3.8 การเลือกภาษาในการทำโครงงาน

ในการทำงานของโครงงานทั้ง 3 ส่วนอันได้แก่ ส่วนดึงข้อมูลและวิเคราะห์ ข้อมูล , ส่วนติดต่อ ระหว่างผู้ใช้บริการและส่วนเก็บข้อมูลที่เซิร์พเวอร์ , ส่วนจัดเก็บข้อมูล ล้วนแต่ถูกเขียนขึ้นโดยภาษาที่ เรียกว่า ภาษาเดลไฟ

สาเหตุที่เลือกโปรแกรมภาษาเดลไฟ ในการเขียนโปรแกรมทั้ง 3 ส่วนมีดังนี้

1. โปรแกรมภาษาเดลไฟเป็นโปรแกรมที่มีประสิทธิภาพและง่ายต่อการทำการศึกษา

2. โปรแกรมภาษาเดลไฟสามารถที่จะทำ user interface ได้ง่ายและสวยงาม

3. โปรแกรมภาษาเดลไฟสามารถที่จะนำมาเขียนโรบอท , ซีจีไอ รวมทั้งภายใน โปรแกรมเดลไฟก็มีส่วนที่สามารถใช้เก็บฐานข้อมูลได้จึงเป็นการง่าย ต่อการติดต่อการทำงานกันใน แต่ละส่วนการทำงาน

4. ผู้จัดทำโครงงานมีความรู้เกี่ยวกับภาษาเดลไฟอยู่บ้างจึงเป็นการประหยัดเวลาใน การศึกษา

บทที่ 4

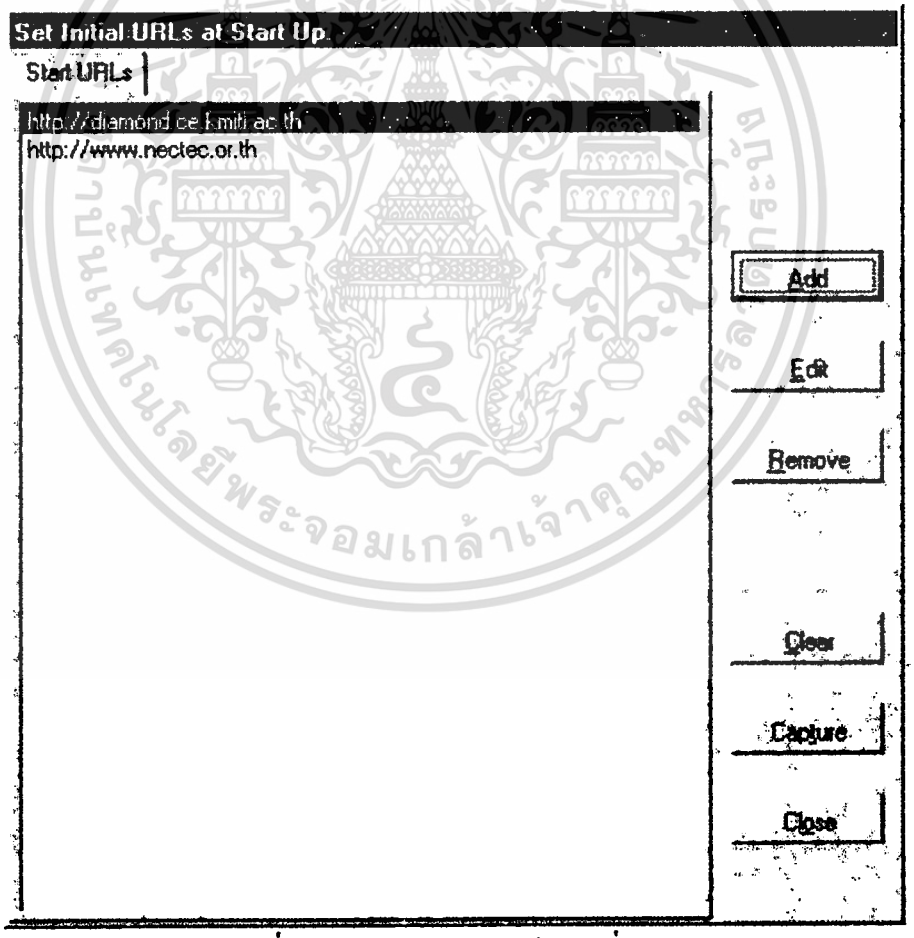
การทดลองและผลการทดลอง

4.1 การทดลองการทำงานโรบอท

เมื่อทำการสร้างโรบอทเสร็จ ก็ทำการทดลองทำงานตามฟังก์ชันที่กำหนดดังต่อไปนี้

4.1.1. การกำหนด ยูอาร์แอล เริ่มต้น

การกำหนด ยูอาร์แอล เริ่มต้นนี้ เพื่อใช้ในการกำหนดจุดเริ่มต้นในการรวบรวมข้อมูลของโรบอท ซึ่งการกำหนด ยูอาร์แอล ที่ต่างกันจะทำให้จำนวนข้อมูลที่รวบรวมได้ต่างกัน ฉะนั้นการกำหนดจุดเริ่มต้นที่เหมาะสมในการค้นหาจึงควรจะกำหนด ยูอาร์แอล ที่สามารถติดต่อไปยัง ยูอาร์แอล อื่น ๆ ได้ในวงกว้าง เพื่อให้ได้ข้อมูลที่ครอบคลุมและครบถ้วน



รูปที่ 4.1 การกำหนด ยูอาร์แอล เริ่มต้น

4.1.2. การตรวจสอบ ยูอาร์แอล ที่อยู่ในประเทศไทย

การตรวจสอบ ยูอาร์แอล มีหลายวิธี ซึ่งวิธีที่ได้ผลดีที่สุด คือการตรวจสอบจากชื่อโฮสต์ โดยชื่อโฮสต์ที่ลงท้ายด้วย .th จะเป็นไซท์ที่อยู่ในประเทศไทยอย่างแน่นอน แต่มีบางไซท์ที่อยู่ในประเทศไทยแต่ชื่อโฮสต์ไม่ได้ลงท้ายด้วย .th จะใช้วิธีการนี้ตรวจสอบไม่ได้ ซึ่งสามารถใช้วิธีอื่น อันได้แก่ การตรวจสอบ ยูอาร์แอล ที่ไม่ได้อยู่ในประเทศไทยเป็นวิธีหนึ่งที่จะช่วยลดการตรวจสอบ ยูอาร์แอล ได้อีกวิธีหนึ่ง และใช้เวลาไม่นานมาก และสามารถกำจัด ยูอาร์แอล ที่ไม่ต้องการได้ดีอีกวิธีหนึ่ง การตรวจสอบเส้นทางของ ยูอาร์แอล นั้น ๆ ซึ่งวิธีนี้เป็นวิธีที่ใช้เวลาในการตรวจสอบนาน

4.1.3. การวิเคราะห์คำและคำนวณค่าความสำคัญ

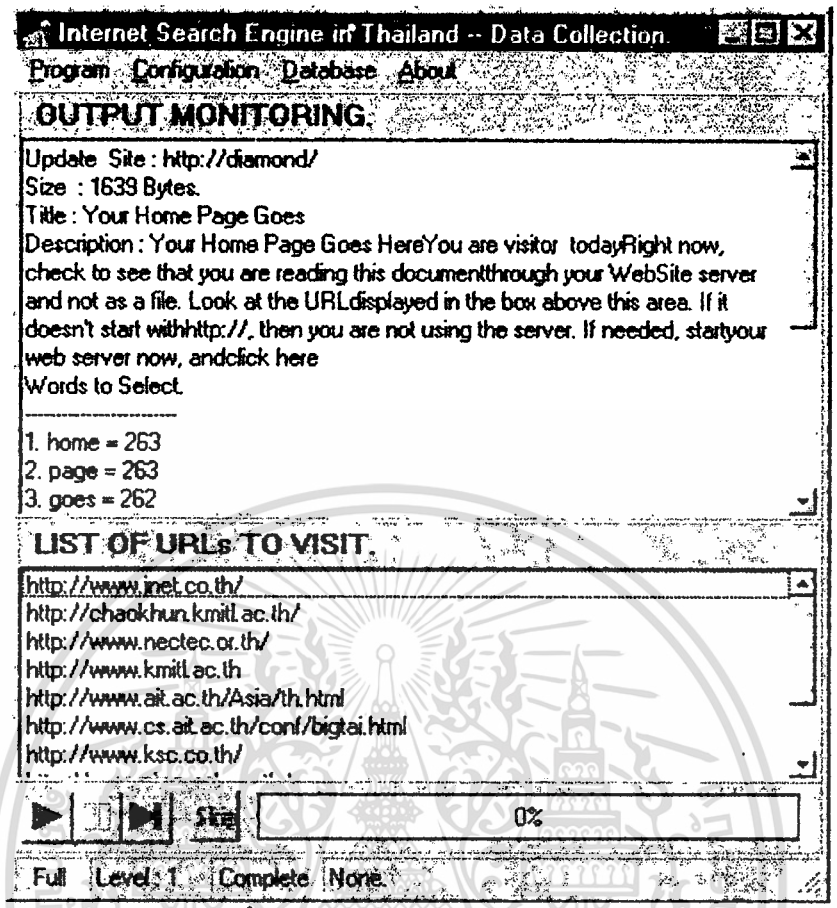
การคำนวณค่าความสำคัญของคำต่าง ๆ ที่ได้จากการวิเคราะห์ข้อมูล ทำให้สามารถเลือกคำที่เป็นคีย์เวิร์ด และเลือกคำที่อาจจะใช้เป็นคีย์เวิร์ดของข้อมูลที่ดึงมาได้ การใช้วิธีนี้จะได้ข้อมูลที่เหมาะสมหรือไม่ขึ้นอยู่กับลักษณะการเขียนโฮมเพจของแต่ละคน นั่นคือถ้าคำที่มีความสำคัญและสามารถใช้เป็นคีย์เวิร์ดได้อยู่ในตำแหน่งที่สำคัญของโฮมเพจ เช่น อยู่ระหว่าง <TITLE> หรือ อยู่ในส่วน <HEAD> โอกาสที่คำนั้นจะถูกเลือกเป็นคีย์เวิร์ดก็มีมากกว่าคำที่อยู่ในส่วนอื่น ๆ

4.1.4. ทดลองทำการดึงข้อมูลจากยูอาร์แอลที่อยู่ในขอบเขต

การดึงข้อมูลจาก ยูอาร์แอล สามารถดึงข้อมูลได้เกือบทุก ยูอาร์แอล ที่กำหนดไว้และตรวจสอบแล้วว่าเป็น ยูอาร์แอล ที่อยู่ในประเทศไทย เนื่องจากเกิดปัญหาระหว่างการติดต่อสื่อสาร ซึ่งเป็นปัญหาที่อยู่นอกขอบเขตที่จะสามารถควบคุมได้ เช่น ไซท์ที่เราติดต่อหยุดการทำงานกลางคัน เกิดความผิดพลาดในการส่ง เป็นต้น และผลที่ได้จากการทดลอง โดยจับเวลาที่ใช้ในขณะที่ทำการเก็บข้อมูล ปรากฏผลดังตารางที่แสดงข้างล่าง นี้

จำนวนเว็บไซท์	จำนวนคำ	เวลาที่ใช้ (นาที)
50	957	4
100	1659	14
150	2241	21
200	3538	38
250	4235	52

ตารางที่ 4.1 ตารางแสดงเวลาที่ใช้ในการเก็บข้อมูล



รูปที่ 4.2 ดึงข้อมูลและวิเคราะห์ข้อมูล

4.2. การค้นหาข้อมูลจากฐานข้อมูลตามเงื่อนไขที่ผู้ใช้บริการกำหนด

ผลลัพธ์ที่ได้จากการค้นหาข้อมูลของผู้ใช้จะขึ้นอยู่กับเงื่อนไขที่ผู้ใช้กำหนด ซึ่งผู้ใช้สามารถกำหนดเงื่อนไขต่าง ๆ ได้ 3 ลักษณะ อันได้แก่

1. จำนวนคำที่ใช้ในการค้นหา
2. ลักษณะของตัวอักษรในคำที่ใช้เป็นคีย์
3. การค้นหาแบบ AND, OR หรือ PHASE

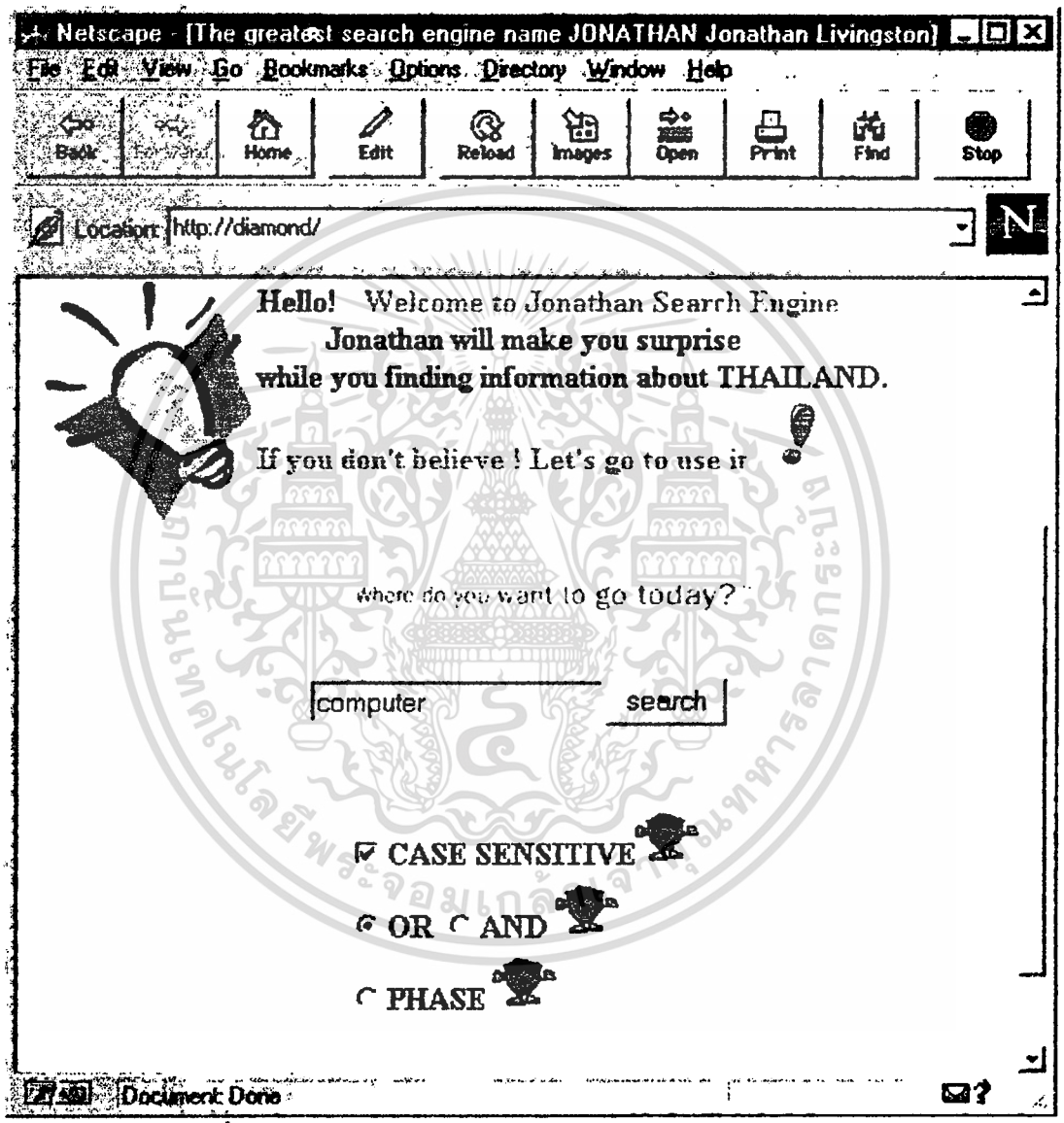
โดยในการกำหนดเงื่อนไข ผู้ใช้สามารถใช้เงื่อนไขต่าง ๆ เหล่านี้มารวมกัน เพื่อให้การค้นหาข้อมูลตรงตามความต้องการของผู้ใช้มากที่สุด ซึ่งสามารถแยกเงื่อนไขการค้นหาด้วยวิธีต่าง ๆ ได้ดังนี้

4.2.1. การค้นหาโดยใช้คีย์เป็นคำเดียว

การค้นหาโดยใช้คำคำเดียวเป็นคีย์จะใช้ลักษณะตัวอักษรเป็นเกณฑ์ในการกำหนดเงื่อนไข ส่วนการกำหนดให้เป็น AND, OR หรือ PHASE จะไม่มีผลต่อการค้นหา นั่นคือไม่ว่าเงื่อนไขจะ

เป็น AND, OR หรือ PHASE ผลลัพธ์ที่ได้ก็จะเป็นเช่นเดียวกัน ดังนั้นการค้นหาแบบคำเดียวจึงแบ่งได้เป็น 2 ลักษณะ คือ

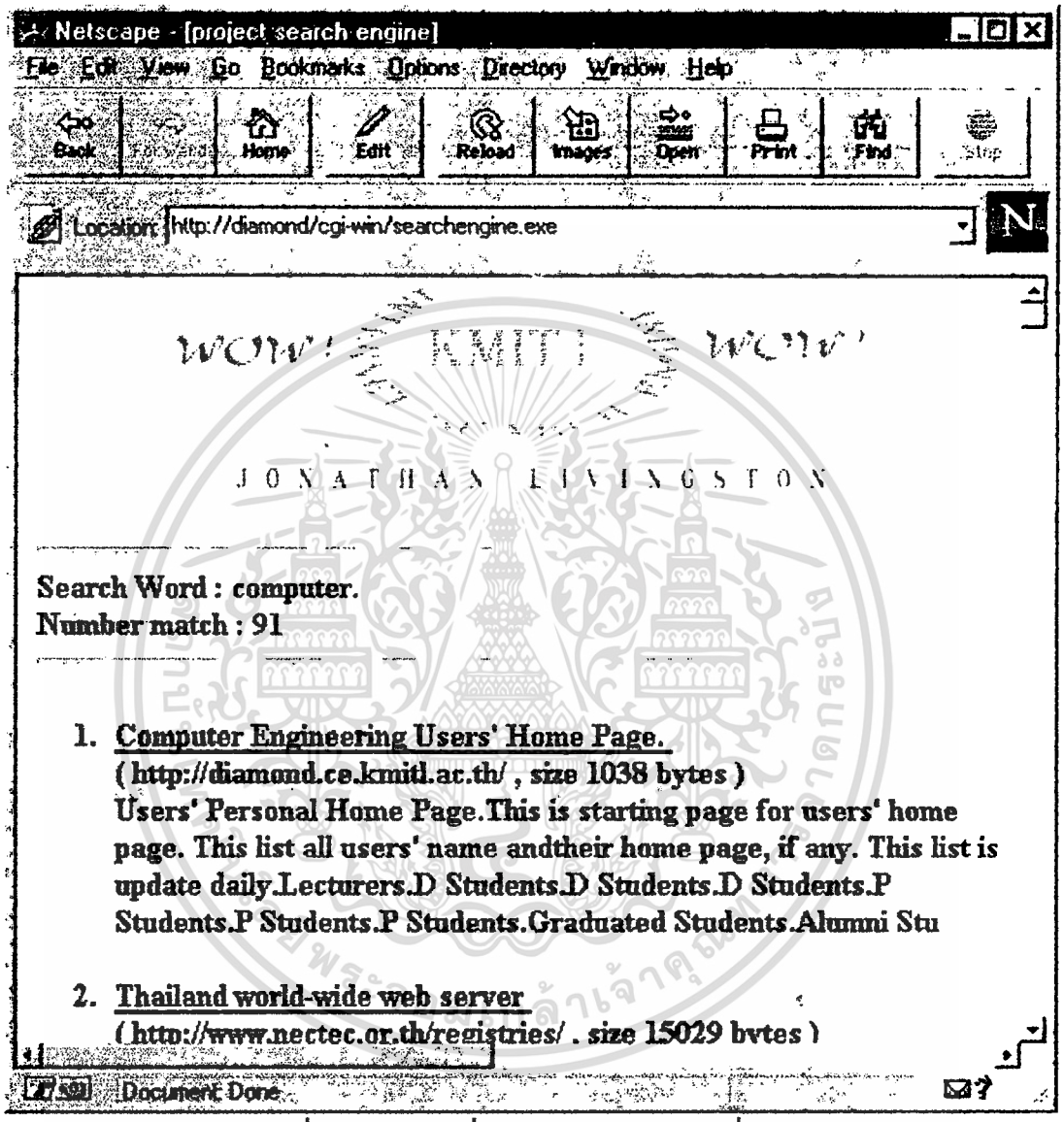
1. การค้นหาแบบพิจารณาลักษณะตัวอักษร
2. การค้นหาแบบไม่พิจารณาลักษณะตัวอักษร



รูปที่ 4.3 ตัวอย่างการค้นหาแบบคำเดียว สนใจลักษณะตัวอักษร

จากตัวอย่างที่แสดงในรูป 4.3 เป็นการค้นหาคำว่า 'computer' โดยสนใจลักษณะของตัวอักษร นั่นคือเป็นตัวเล็กทั้งหมด และในรูป 4.4 เป็นผลลัพธ์ที่ได้จากการค้นหาคำว่า 'computer' ที่

อยู่ในฐานข้อมูล และแสดงรายชื่อ ยูอาร์แอล ทั้งหมดที่มีคำว่า 'computer' เป็นคีย์เวิร์ด โดยไม่สนใจว่าผู้ใช้จะเลือกการค้นหาเป็นแบบ AND, OR หรือ PHASE



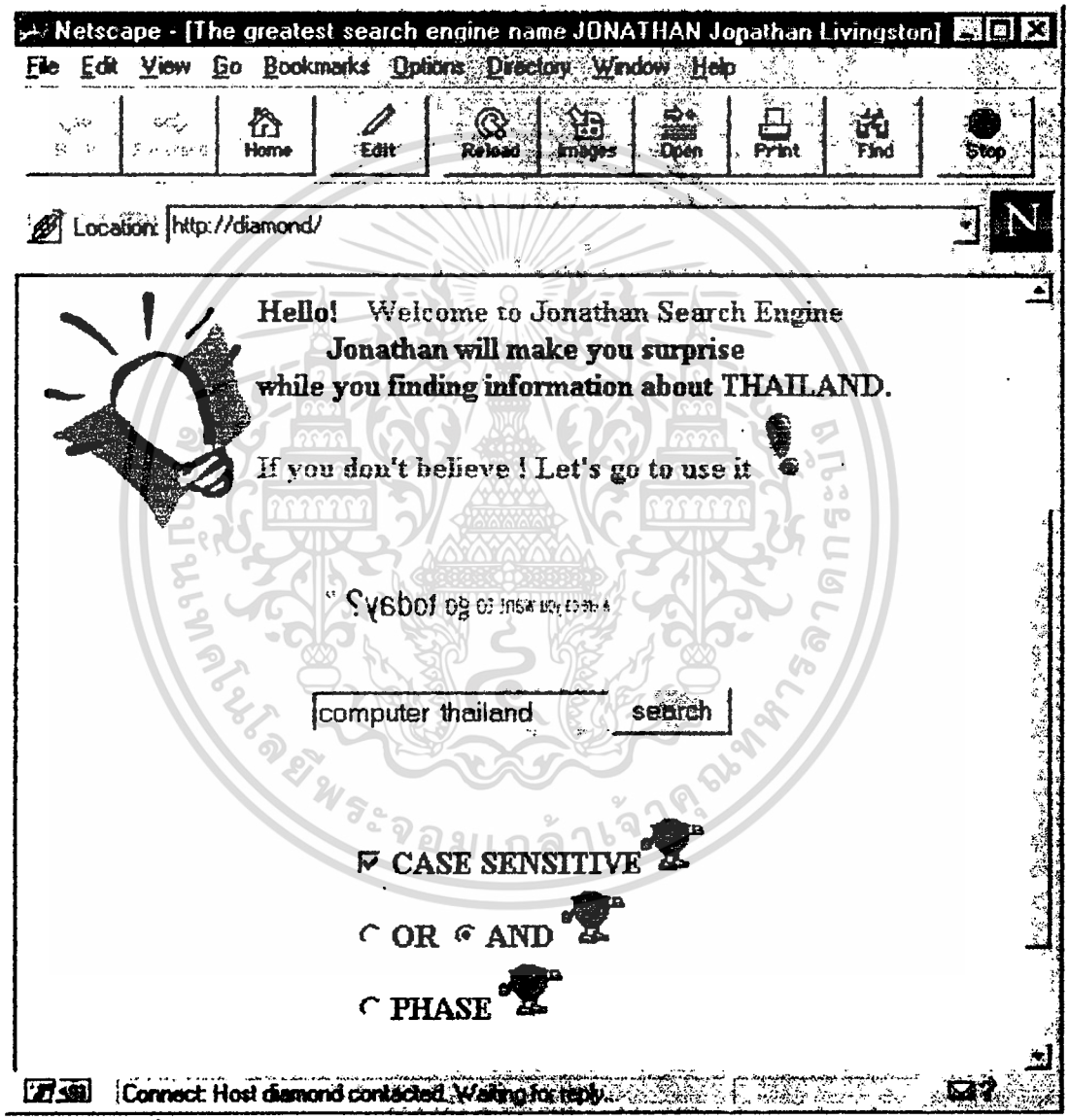
รูปที่ 4.4 ผลลัพธ์ที่ได้จากการค้นหาในรูปที่ 4.3.

4.2.2. การค้นหาโดยใช้คีย์เป็นคำหลายคำ

การหาคำโดยใช้คีย์เวิร์ดหลายคำ ผู้ใช้จะต้องกำหนดเงื่อนไขการค้นหาให้ครบทั้ง 2 ลักษณะ คือ สนใจหรือไม่สนใจลักษณะของตัวอักษร และ การกำหนดให้เป็น AND, OR หรือ PHASE ซึ่งสามารถแบ่งการค้นหาได้เป็น 6 แบบ คือ

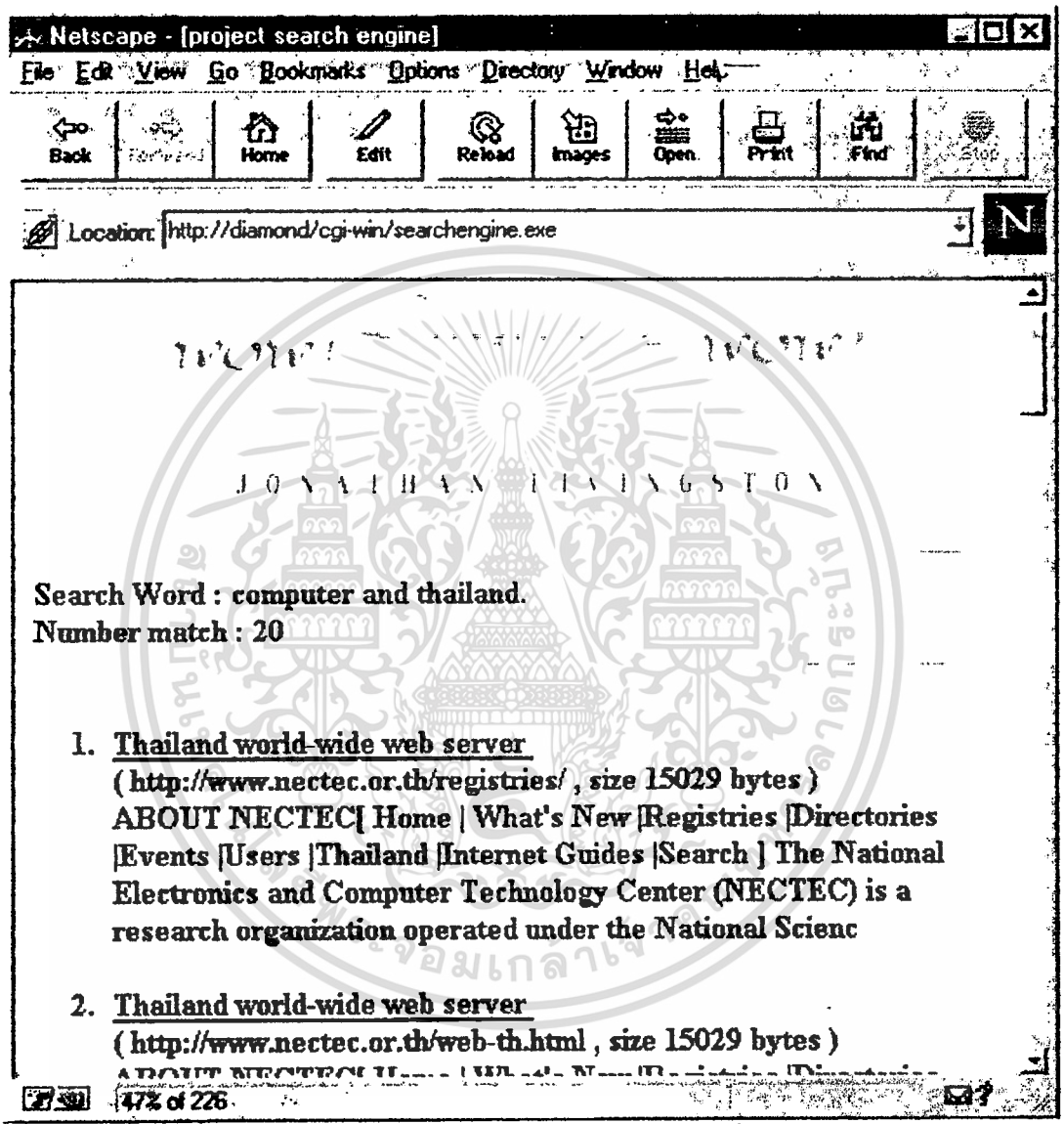
1. การค้นหาเว็บเพจที่มีคำทุกคำเป็นคีย์เวิร์ด โดยสนใจลักษณะตัวอักษร

2. การค้นหาเว็บที่มีคำทุกคำเป็นคีย์เวิร์ด โดยไม่สนใจลักษณะตัวอักษร
3. การค้นหาเว็บเพจที่มีคำบางคำเป็นคีย์เวิร์ด โดยสนใจลักษณะตัวอักษร
4. การค้นหาเว็บเพจที่มีคำบางคำเป็นคีย์เวิร์ด โดยไม่สนใจลักษณะตัวอักษร
5. การค้นหาเว็บเพจที่มีวลีที่กำหนดเป็นคีย์เวิร์ด โดยสนใจลักษณะตัวอักษร
6. การค้นหาเว็บเพจที่มีวลีที่กำหนดเป็นคีย์เวิร์ด โดยไม่สนใจลักษณะตัวอักษร



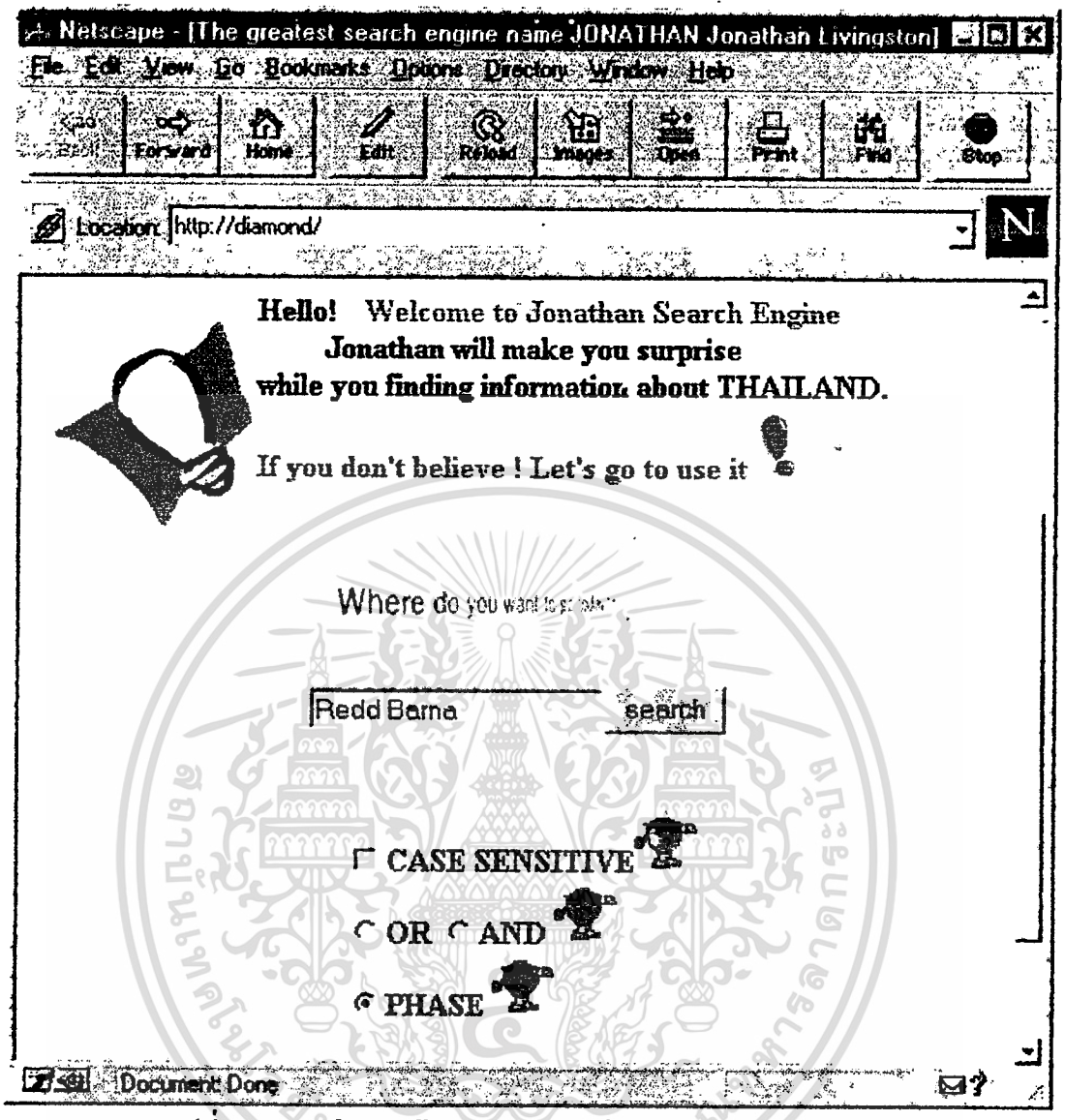
รูปที่ 4.5 ตัวอย่างการค้นหาแบบหลายคำ สนใจลักษณะตัวอักษร แบบ AND

จากรูป 4.5 เป็นตัวอย่างการค้นหาคำว่า 'computer' และคำว่า 'thailand' และเลือกเฉพาะเว็บเพจ ที่มีคำทั้งสองคำนี้ปรากฏ และคำว่า 'computer และ 'thailand' จะต้องเขียนด้วยตัวเล็กทั้งหมด นั่นคือจะไม่สนใจเว็บเพจที่มีคีย์เวิร์ดเป็น 'THAILAND' และ 'COMPUTER'



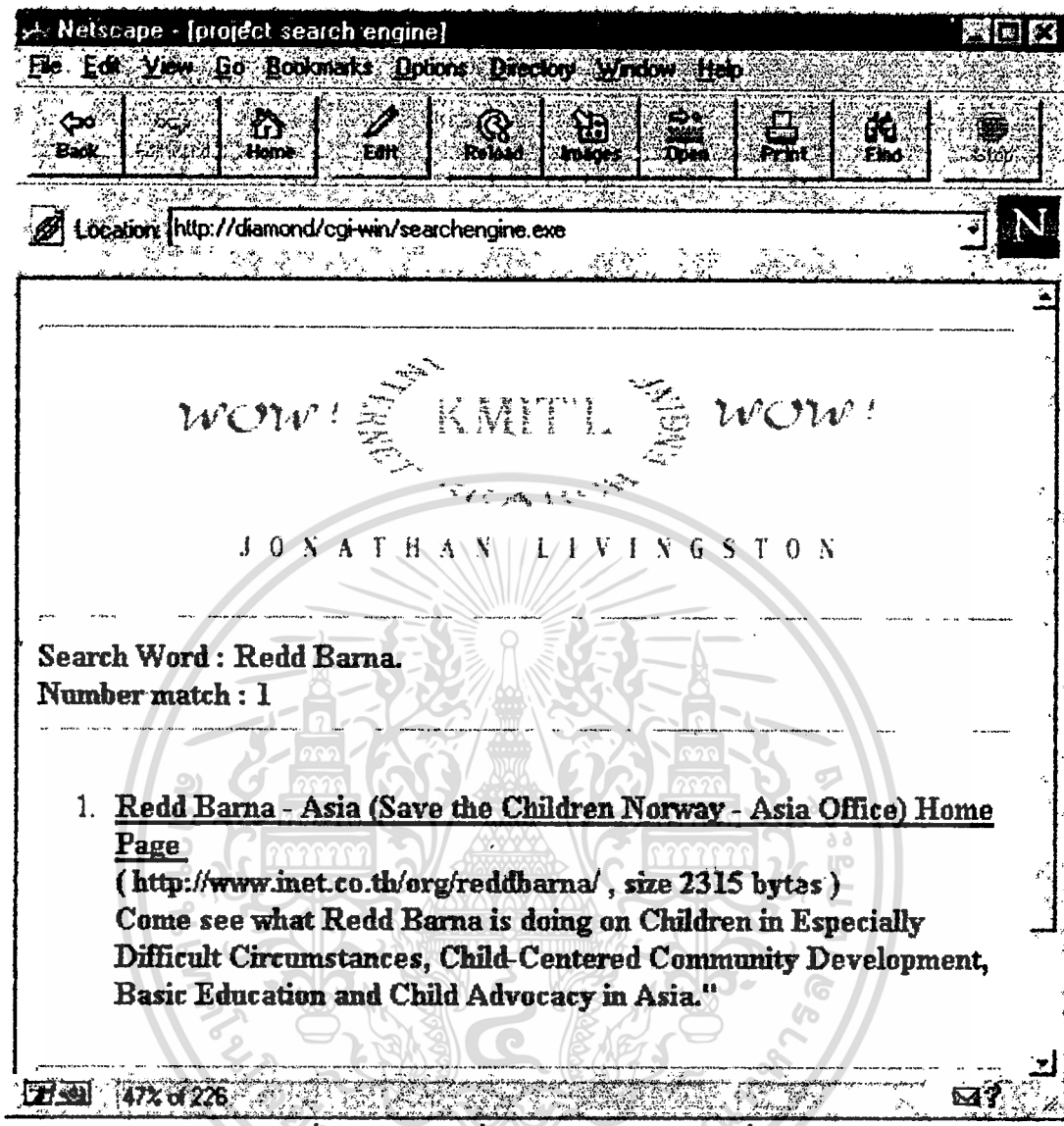
รูปที่ 4.6 ผลลัพธ์ที่ได้จากการค้นหาในรูปที่ 4.5.

รูปที่ 4.6 เป็นผลลัพธ์ที่ได้จากการกำหนดเงื่อนไขในรูปที่ 4.5 ซึ่งการค้นหาด้วยเงื่อนไขแบบนี้จะทำให้ได้ ผลลัพธ์ น้อยกว่าการค้นหาคำว่า 'computer' หรือ 'thailand' เพียงคำเดียว แต่การค้นหาวิธีนี้จะได้ผลลัพธ์ตรงตามความต้องการของผู้ใช้มากขึ้น เมื่อผู้ใช้กำหนดคำที่ใช้ในการค้นหามากขึ้น



รูปที่ 4.7 ตัวอย่างการค้นหาแบบวลี โดยไม่สนใจลักษณะตัวอักษร

จากตัวอย่างการค้นหาในรูปที่ 4.7 เป็นการค้นหาโดยใช้วลี 'Redd Barna' เป็นคีย์ ซึ่งในการค้นหาโปรแกรม จะค้นหาเว็บเพจที่ใช้วลี 'Redd Barna' เป็นคีย์ โดยไม่สนใจลักษณะตัวอักษรของวลีที่ใช้เป็นคีย์ ซึ่งในรูปที่ 4.8 จะเป็นผลลัพธ์ที่ได้จากการค้นหาคำโดยเงื่อนไขนี้ ซึ่งผลลัพธ์ที่ได้จะมีจำนวนน้อย แต่ว่าผลลัพธ์ที่ได้จากการค้นหาแบบวลีนี้ จะตรงกับความต้องการของผู้ใช้มากกว่าเงื่อนไขอื่น ๆ



รูปที่ 4.8 ผลลัพธ์ที่ได้จากการค้นหาในรูปที่ 4.7

บทที่ 5
บทวิจารณ์และสรุป

โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ มีส่วนประกอบและปัญหาโดยสรุปดังต่อไปนี้

5.1 สรุปโครงสร้างของโครงงาน INTERNET SEARCH ENGINE

5.1.1 ส่วนดึงข้อมูลและวิเคราะห์ข้อมูล

เอเจนท์ที่เลือกใช้ในการทำโรบอทของโครงงาน Internet Search Engine เป็นเอเจนท์ประเภท Web Robot ซึ่งจะทำหน้าที่ในการท่องไปตาม WebSite ต่างๆ เพื่อทำการค้นหาข้อมูล โดยโรบอทที่ใช้ทำการดึงข้อมูลนี้จะปฏิบัติตามกฎต่างๆดังต่อไปนี้

ข้อควรปฏิบัติของโรบอท

1.โรบอทของโครงงาน internet search engine จะปฏิบัติตามเงื่อนไขที่ Web Master กำหนดสิทธิให้ในแต่ละ Web Site

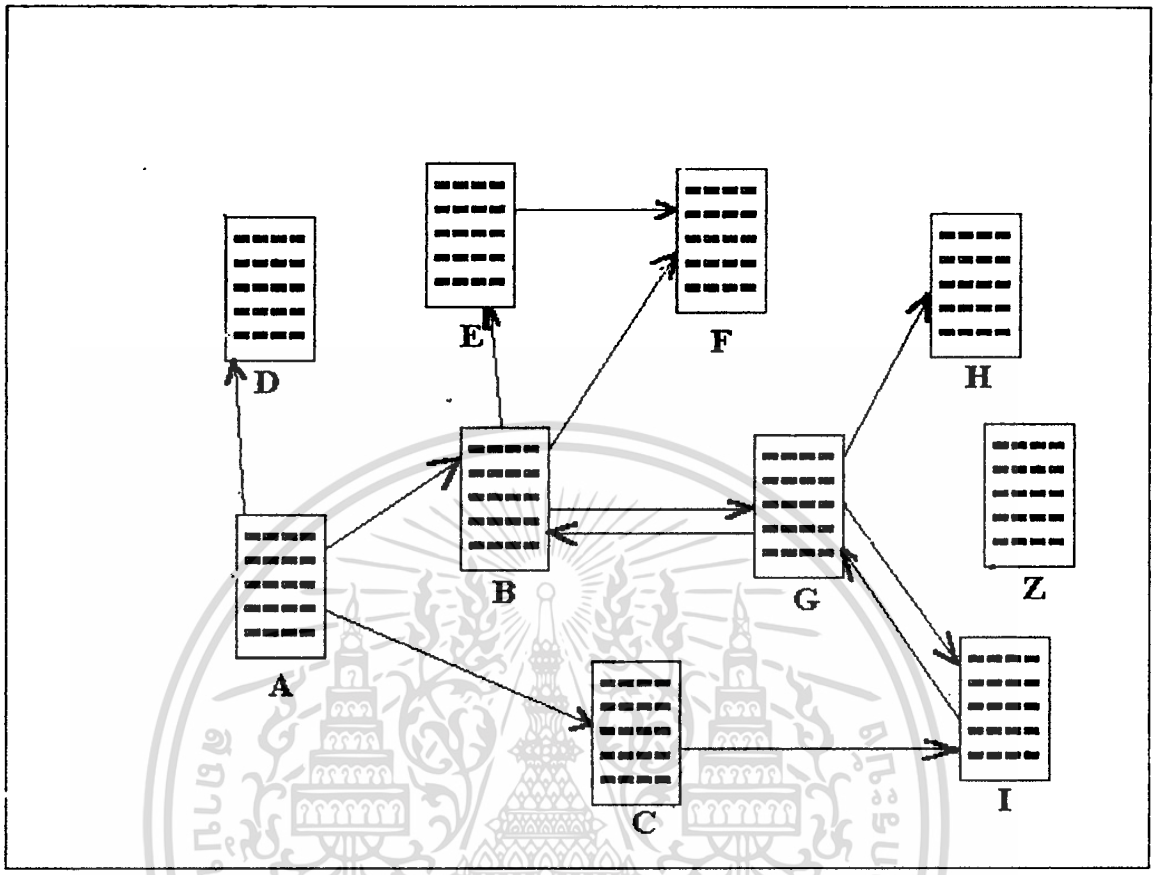
2. โรบอทจะมีการแสดงตนอย่างชัดเจนพร้อมทั้งมีการแสดง address ของ email ของเจ้าของโรบอท เพื่อที่ Web Master จะได้สามารถทำการติดต่อได้ในกรณีที่โรบอทมีข้อ ผิดพลาดขึ้น เช่น s6014115@diamond.ce.kmitl.ac.th

3. ทำการทดสอบการทำงานของโรบอท จากเว็บเซิร์พเวอร์ในระดับโลคอลก่อน ที่จะทำการทดสอบกับเว็บเซิร์พเวอร์ที่อยู่ไกล

4. มีการตรวจสอบข้อผิดพลาดที่อาจจะเกิดขึ้นในการทำงานของโรบอทตลอดเวลา ที่ทำการรันเพื่อที่จะสามารถทำการแก้ไขปัญหาที่เกิดขึ้นได้ทันที

วิธีการท่องเว็บไซท์ของโรบอทในการรวบรวมข้อมูล

การรวบรวมข้อมูลจากเว็บไซท์ต่างๆ ของโรบอทในโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์จะทำการค้นหาในทางกว้างก่อน (Breadth-First Search) นั่นคือ โรบอทจะทำการดึงข้อมูลจากเว็บไซท์เริ่มต้น ที่ถูกกำหนดเอาไว้ในลิสต์ของ ยูอาร์แอล แล้วค่อยขยับไปเก็บข้อมูลยังเว็บไซท์ต่อไปตาม อัลกอริทึมในการค้นหาแบบทางกว้าง ดังรูปต่อไปนี้



รูปที่ 5.1 แสดงการค้นหาแบบทางกว้าง (Breath-First Search)

ลักษณะการเก็บคีย์เวิร์ด (keyword) จากเว็บเพจ

ในการเลือกจัดเก็บคีย์เวิร์ดจากเว็บเพจจะทำการจัดเก็บแบบ เลือกคำบางคำเป็นคีย์เวิร์ด โดยในวิธีการนี้จะทำการเก็บเฉพาะคำที่เหมาะสมในการที่จะนำมาจัดทำเป็น คีย์เวิร์ด เนื่องจากคำส่วนใหญ่ปรากฏภายในเว็บเพจมักจะเป็นคำที่ไม่มีประโยชน์ในการค้นหา เช่น

- สัญลักษณ์ เช่น @
- คำที่เป็นคำเชื่อมหรือไม่มีความสำคัญ เช่น is , the , ok ,or เป็นต้น

ข้อดี ของการเก็บคีย์เวิร์ดโดยใช้วิธีนี้

1. **เพื่อช่วยให้ประหยัดเนื้อที่ที่ต้องใช้ในการเก็บข้อมูล เนื่องจากไม่ต้องเก็บคำทุกคำที่ปรากฏ ในเว็บเพจ ขนาดของฐานข้อมูลจึงไม่ใหญ่มากนัก**
2. **ทำให้การค้นหาเป็นไปได้รวดเร็วขึ้น เนื่องจากฐานข้อมูลมีขนาดเล็กลง**

3. ผลลัพธ์ที่ได้จากการค้นหามีความใกล้เคียงกับความต้องการของผู้ค้นหามากกว่าการเก็บข้อมูลทั้งหมด

4. ผลลัพธ์ที่ได้จากการค้นหามีปริมาณน้อยกว่า การเก็บคำทุกคำเป็นคีย์เวิร์ด

ข้อเสีย ของการเก็บคีย์เวิร์ดโดยใช้วิธีคำนวณหาค่าความสำคัญ

1. ผลลัพธ์ที่ได้ไม่สมบูรณ์ เนื่องจากการเลือกคีย์เวิร์ดที่จะเก็บเป็นฐานข้อมูล จะเลือกคำที่มีความสำคัญตามเกณฑ์ที่ตั้งไว้ ดังนั้นคำบางคำสามารถใช้เป็นคีย์เวิร์ดแต่มีความ สำคัญน้อยกว่าที่กำหนดไว้จึงไม่ถูกเก็บไว้ในฐานข้อมูล

2. ไม่สามารถทำการค้นหาข้อมูลที่เป็นประโยคได้ เนื่องจากไม่ได้เก็บคำทุกคำเป็นคีย์เวิร์ด

คุณสมบัติของคำที่สามารถใช้เป็นคีย์เวิร์ดได้

1. คีย์เวิร์ดจะต้องเป็นคำที่ปรากฏอยู่ในเว็บเพจนั้น

2. เป็นคำที่ประกอบด้วยตัวอักษร " a-z " , " A-Z " , " - " , " 0-9 " และต้องไม่ใช่ตัวเลขทั้งคำ

3. คำที่ใช้ตัวอักษรตัวพิมพ์ใหญ่หมดทั้งคำเนื่องจากมีโอกาสที่จะเป็นคำสำคัญสูง

4. คำที่เป็นคีย์เวิร์ดจะต้องเป็นคำที่มีความยาวของคำตั้งแต่ 3 คำขึ้นไป

5. คำที่เป็นคีย์เวิร์ดจะต้องไม่อยู่ในลิสต์ของคำที่ไม่ควรเป็นคีย์เวิร์ด เช่น คำไม่สุภาพ

การคำนวณหาค่าความสำคัญของคำที่ปรากฏในเว็บเพจ

เมื่อตรวจสอบคุณสมบัติของคำ ต่าง ๆ ที่ปรากฏอยู่ในเว็บเพจแล้ว ว่ามีคุณสมบัติพอที่จะใช้เป็นคีย์เวิร์ดของเว็บเพจได้ นำคำเหล่านั้นมาคำนวณหาค่าความสำคัญของแต่ละคำ ที่ปรากฏ อยู่ภายในเว็บเพจ และเลือกคำที่มีความสำคัญมากกว่าเกณฑ์ที่กำหนดไว้ จัดเก็บลงในฐานข้อมูล เพื่อใช้เป็นคีย์เวิร์ดในการค้นหาต่อไป ซึ่งวิธีการคำนวณหาค่าความสำคัญของคำต่าง ๆ สามารถหา ได้โดยใช้สูตร ดังต่อไปนี้

$$Priority_n = S\ (Priority\_Key_i)$$

เมื่อ Priority คือ ค่าความสำคัญของคำใด n ใด ๆ

Priority_Key คือ ค่าความสำคัญของคำที่อยู่ ณ ตำแหน่ง i

จากสูตร การคำนวณหาค่าความสำคัญของคำ ทำได้โดยนำค่า Priority_Key ของตำแหน่งที่ปรากฏคำ ๆ หนึ่งทั้งหมดในเว็บเพจ มารวมกันเพื่อให้ได้ค่า Priority ของคำ ๆ นั้น

ตัวอย่าง

คำว่า "computer" ปรากฏอยู่ในเว็บเพจ 3 ที่

ตำแหน่งที่ 1 มีค่า $Priority\_Key_1 = 100$

ตำแหน่งที่ 2 มีค่า $Priority\_Key_2 = 20$

ตำแหน่งที่ 3 มีค่า $Priority\_Key_3$ = 3

เพราะฉะนั้น $Priority_{computer}$ = 100+20+3

= 123

การกำหนดค่า Priority_Key ของคำในตำแหน่งต่าง ๆ

ค่า Priority_Key เป็นค่าที่ใช้กำหนดความสำคัญของคำ ณ. ตำแหน่งใด ๆ ในเว็บเพจ สิ่งที่จะใช้กำหนดค่า Priority_Key นี้คือ แท็ก (Tag) นั่นคือ ค่าของ Priority_Key จะแตกต่างกัน ไปมากน้อยขึ้นอยู่กับว่า ณ. ตำแหน่งนั้นอยู่ระหว่าง แท็กใด เช่น ค่า Priority_Key ณ. ตำแหน่งที่ อยู่ระหว่าง <TITLE> กับ </TITLE> จะมีค่า Priority_Key มากกว่าค่า Priority_Key ณ.ตำแหน่ง ที่อยู่ระหว่าง <H1> กับ </H1> เป็นต้น และแท็กบางแท็ก จะไม่มีผลต่อค่า Priority_Key เลย

การกำหนดจุดสิ้นสุดการทำงานของหุ่นยนต์เก็บข้อมูล

การกำหนดจุดสิ้นสุดของการทำงานของหุ่นยนต์นับว่าเป็นส่วนสำคัญส่วนหนึ่งของการทำงาน เนื่องจากเครือข่ายอินเตอร์เน็ตเป็นเครือข่ายที่มีขนาดใหญ่มาก ถ้าต้องการให้หุ่นยนต์ ค้นหาทำการเก็บ ข้อมูลจากทุกเว็บเพจที่ปรากฏอยู่ในอินเตอร์เน็ตจะต้องใช้เวลาอย่างมาก ดังนั้นจึงมีวิธีการในการ กำหนดจุดสิ้นสุดในการเก็บข้อมูล เพื่อเป็นทางเลือกในการเก็บข้อมูล และ อัพเดท (Update) ข้อมูล ทำให้การเก็บข้อมูลใช้เวลารวดเร็วขึ้นและตรงตามความต้องการของผู้ใช้มากขึ้น ดังนี้ คือ

1. หยุดการทำงานเมื่อลิสต์ว่าง

วิธีนี้เป็นการเก็บข้อมูลที่สมบูรณ์และครบถ้วนที่สุด เนื่องจากหุ่นยนต์จะทำการค้นหา และทำการเก็บข้อมูลจากทุกเว็บเพจที่สามารถติดต่อไปถึงได้ วิธีนี้เป็นวิธีที่มั่นใจได้ว่าเว็บเพจใด ๆ ก็ตามที่มีเส้นทางการลิงก์จากเว็บเพจของยูอาร์แอลเริ่มต้นในลิสต์ไปถึงแล้วจะต้องได้ถูกเก็บข้อมูลไว้เป็นฐานข้อมูล เพื่อใช้ในการค้นหาแน่นอน

2. หยุดการทำงานเมื่อหุ่นยนต์เก็บข้อมูลทำงานจนครบระดับความลึก (Level of Breadth-First Search) ตามที่กำหนดไว้

วิธีนี้เป็นการกำหนดระดับความลึกของการเก็บข้อมูล และให้หุ่นยนต์เก็บข้อมูลสิ้นสุดการทำงานและทำให้ลิสต์ว่างเมื่อระดับความลึกของการเก็บข้อมูลเท่ากับค่าที่กำหนดไว้ โดยระดับความลึกของการเก็บข้อมูล คือ ระดับของการลิงก์จากเว็บเพจหนึ่งไปยังอีกเว็บเพจหนึ่ง เมื่อเทียบกับเว็บเพจเริ่มต้น โดยกำหนดให้ยูอาร์แอลของเว็บเพจเริ่มต้นมีระดับความลึก ของการเก็บข้อมูลเป็น 1 และ ยูอาร์แอลของเว็บเพจที่ถูกลิงก์ไปถึงมีค่าระดับความลึกเป็น 2, 3, ...

การกำหนดลักษณะการทำงานของหุ่นยนต์เก็บข้อมูล

เนื่องจากอินเตอร์เน็ตเป็นเครือข่ายที่มีขนาดใหญ่ ในการเก็บข้อมูลแต่ละครั้งจึงต้องใช้เวลามาก การที่จะให้หุ่นยนต์ทำการเก็บข้อมูลทุกครั้งจากทุกเว็บเพจที่มีอยู่ในอินเตอร์เน็ตจึงเป็นการกระทำ ที่ต้องใช้เวลามาก และอาจจะทำให้เสียเวลาโดยใช่เหตุ เนื่องจากข้อมูลที่ถูกส่งมาเป็น ข้อมูลที่มีอยู่ใน ฐานข้อมูลอยู่แล้ว และส่วนใหญ่เว็บเพจจะไม่มีการเปลี่ยนแปลงบ่อย ๆ ดังนั้นจึงกำหนดลักษณะการ ทำงานของหุ่นยนต์เก็บข้อมูลออกมา 2 ลักษณะ ดังนี้

1. เก็บข้อมูลจากทุกเว็บเพจ

วิธีนี้จะทำการเก็บข้อมูลจากทุกเว็บเพจที่สามารถลิงก์ไปถึงได้ และเมื่อได้ข้อมูลมาแล้ว จะทำการตรวจสอบในฐานข้อมูล ถ้าปรากฏอยู่ในฐานข้อมูลอยู่แล้ว จะทำการตรวจสอบว่าข้อมูลที่ได้มาใหม่เปลี่ยนแปลงไปจากฐานข้อมูลเดิมที่มีอยู่แล้ว หรือไม่ ถ้าเหมือนกัน หุ่นยนต์จะทำการเก็บข้อมูลจากเว็บเพจอื่นต่อไป แต่ถ้าข้อมูลใหม่ที่ได้ไม่เหมือนกับข้อมูลเก่าที่มีอยู่ หุ่นยนต์เก็บข้อมูลจะส่งข้อมูลใหม่ที่ได้มาให้กับส่วนวิเคราะห์แล้วทำการเปลี่ยนแปลงฐานข้อมูล เพื่อให้เป็นปัจจุบัน

วิธีในการตรวจสอบการอัพเดตของข้อมูล

เนื่องจากการในฐานข้อมูลเก็บเพียงคีย์เวิร์ดของเว็บเพจเท่านั้น ไม่ได้เก็บข้อมูลทั้งหมด ทำให้ไม่สามารถนำข้อมูลที่ได้มาใหม่เปรียบเทียบกับของเก่าได้ตรง ๆ ในฐานข้อมูลจึงมีส่วนหนึ่ง ที่ใช้เก็บข้อมูลเกี่ยวกับเว็บเพจนั้น เช่น ชื่อของยูอาร์แอล (URL Name), หัวข้อของยูอาร์แอล (Title), ขนาดของเว็บเพจ (Size) เป็นต้น และส่วนที่นำมาใช้ในการตรวจสอบว่า ข้อมูลใหม่ที่ได้ มีการเปลี่ยนแปลงไปจากข้อมูลที่มีอยู่ในฐานข้อมูลเดิมหรือไม่ คือ ขนาดของเว็บเพจ นั่นคือถ้า ขนาดของข้อมูลที่ได้มาใหม่ไม่เท่ากับขนาดของเว็บเพจเดิม แสดงว่าข้อมูลที่ได้มีการเปลี่ยนแปลง

ข้อจำกัด วิธีนี้จะใช้ไม่ได้ถ้าเว็บเพจมีการเปลี่ยนแปลงข้อความภายในเว็บเพจแต่ว่ามีขนาดไม่มีการเปลี่ยนแปลง

ข้อดีของการเก็บข้อมูลจากทุกเว็บเพจ คือ

ข้อมูลเป็นปัจจุบันเมื่อมีการเก็บข้อมูลใหม่และทันสมัยทุกครั้งเมื่อมีการเก็บข้อมูล นั่นคือข้อมูลที่อยู่ในฐานข้อมูลจะมีการเปลี่ยนแปลงให้ตรงกับข้อมูลของเว็บเพจเมื่อเว็บเพจมีการเปลี่ยนแปลง

ข้อเสียของการเก็บข้อมูลจากทุกเว็บเพจ คือ

เสียเวลาในการรับข้อมูลจากเว็บเพจไปโดยเปล่าประโยชน์ ถ้าข้อมูลของเว็บเพจส่วนใหญ่ที่ได้มาไม่มีการเปลี่ยนแปลงจากในฐานข้อมูลเดิมที่เคยมีอยู่แล้ว ดังนั้นการเก็บข้อมูลโดย ใช้วิธีนี้

จึงไม่ควรใช้บ่อยครั้งมากนัก ควรใช้เมื่อทำการเก็บข้อมูลครั้งแรก หรือใช้เมื่อมั่นใจว่า มีเว็บเพจหลายเว็บเพจเปลี่ยนแปลง อาจจะใช้นาน ๆ ครั้ง เช่น ทุกสัปดาห์ หรือ ทุกเดือน เป็นต้น

2. เก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล

การเก็บข้อมูลวิธีนี้จะมีการตรวจสอบ ยูอาร์แอล ก่อนที่จะนำไปใส่ไว้ในลิสต์ของ ยูอาร์แอลที่จะให้หุ่นยนต์ทำการเก็บข้อมูล เพื่อดูว่ายูอาร์แอลของเว็บเพจที่จะให้หุ่นยนต์ทำ การเก็บข้อมูลนี้มีอยู่ในฐานข้อมูลหรือไม่ ถ้ามียูอาร์แอลที่ตรวจสอบอยู่แล้ว ก็ไม่ต้องนำยูอาร์แอล นี้ไปใส่เพิ่มเข้าไปไว้ในลิสต์อีก แต่ถ้ายังไม่มีในฐานข้อมูล ก็นำยูอาร์แอลนี้ไปใส่เพิ่มไว้ ในลิสต์ ของยูอาร์แอลที่หุ่นยนต์ต้องทำการเก็บข้อมูล

ข้อดีของการเก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล คือ

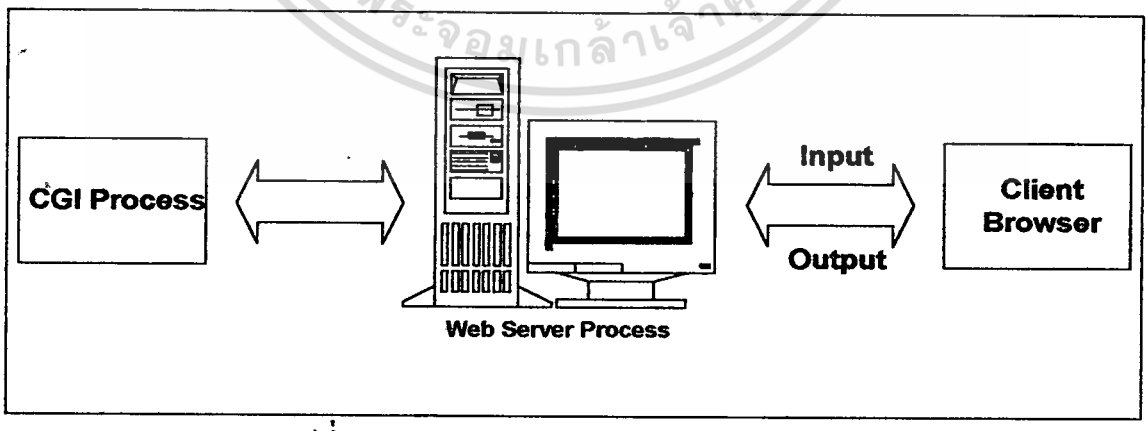
การเก็บข้อมูลใช้เวลาน้อยลง เนื่องจากไม่ต้องเสียเวลากับการเก็บข้อมูล ที่เคยมีอยู่แล้วในฐานข้อมูล ทำให้สามารถใช้ได้บ่อยครั้งกว่าวิธีแรก เนื่องจากใช้เวลาน้อยกว่า

ข้อเสียของการเก็บข้อมูลจากทุกเว็บเพจที่ไม่อยู่ในฐานข้อมูล คือ

ข้อมูลที่มีอยู่ในฐานข้อมูลเป็นข้อมูลเก่า ทำให้การค้นหาข้อมูลได้ข้อมูลไม่ตรงกับความเป็นจริงในปัจจุบัน เนื่องจากเว็บเพจที่ได้จากการค้นหาอาจมีการเปลี่ยนแปลงไปแล้ว หรือ ยูอาร์แอล นั้นถูกยกเลิก การเก็บข้อมูลด้วยวิธีนี้จึงเป็นวิธีที่ควรจะใช้เมื่อต้องการเก็บข้อมูลจาก เว็บเพจที่เกิดขึ้นใหม่ โดยมั่นใจว่าข้อมูลที่มีอยู่ในฐานข้อมูลเดิมไม่มีการเปลี่ยนแปลงมากนัก แต่เมื่อเว็บเพจปัจจุบันมีการเปลี่ยนแปลงไปมาก การเก็บข้อมูลด้วยวิธีนี้จึงไม่เหมาะสมเท่าวิธีแรก

5.1.2 ส่วนเชื่อมต่อระหว่างผู้ใช้งานกับเซิร์พเวอร์ (CGI)

การทำงานของ CGI จะมีการทำงานตามขั้นตอนดังรูปต่อไปนี้



รูปที่ 5.2 แสดงหลักการทำงานของโปรแกรม CGI

5.1.3 ส่วนจัดเก็บข้อมูล (DATABASE)

ในโครงงานหุ่นยนต์ค้นหาข้อมูล เราจะทำการจัดเก็บข้อมูลไว้ภายในระบบฐานข้อมูล เนื่องจากระบบฐานข้อมูลมีข้อดีดังต่อไปนี้

ข้อดีของการจัดเก็บข้อมูลแบบฐานข้อมูล

การจัดเก็บข้อมูลโดยใช้ระบบฐานข้อมูลเข้ามาช่วยนั้นเป็นวิธีการ ที่ได้รับความนิยม อย่างแพร่หลาย เนื่องจากมีประสิทธิภาพและการใช้งานที่เหมาะสมดังมีข้อดีต่อไปนี้

1 สามารถลดความซ้ำซ้อนของข้อมูล

การจัดเก็บข้อมูลโดยวิธีอื่น เช่น การจัดเก็บข้อมูลเป็นไฟล์ อาจทำให้เกิดปัญหาการจัดเก็บข้อมูลซ้ำซ้อนกัน เช่น การที่ผู้ใช้งานทุกคนต่างก็ทำการคัดลอกไฟล์ดังกล่าวเก็บไว้เพื่อใช้งาน ซึ่งการซ้ำซ้อนในการเก็บข้อมูลนี้ จะทำให้เกิดการเสียเนื้อที่ในการจัดเก็บข้อมูล โดยไม่จำเป็นซึ่ง หากใช้การเก็บข้อมูลแบบฐานข้อมูล สามารถที่จะช่วยลดปัญหาดังกล่าวได้ เนื่องจากการจัดเก็บ ข้อมูลโดยใช้ฐานข้อมูล จะมีการออกแบบและควบคุมข้อมูลที่จัดเก็บเพื่อลดความซ้ำซ้อน และ นอกจากนี้จุดประสงค์หลักของการใช้ ฐานข้อมูลก็คือ การลดความซ้ำซ้อนของข้อมูลให้มากที่สุด

2 สามารถควบคุมความถูกต้องของข้อมูลได้

การจัดเก็บข้อมูลโดยวิธีทั่วไป อาจเกิดปัญหาในการทำการควบคุมความถูกต้องของข้อมูล อันอาจจะมีผลเนื่องจากการเก็บข้อมูล ซ้ำซ้อนทำให้ผู้ใช้งาน แต่ละคนสามารถทำการเปลี่ยนแปลงข้อมูลได้ ซึ่งทำให้ข้อมูลในแต่ชุดที่คัดลอกไปอาจจะไม่เหมือนกัน จนไม่อาจจะแยกแยะได้ว่าข้อมูลใดเป็นข้อมูลที่ถูกต้อง แต่หากมีการใช้การเก็บข้อมูลโดยใช้ ฐานข้อมูลจะมีการควบคุมการเปลี่ยนแปลงข้อมูล โดยจะอนุญาตให้มีเพียงผู้ใช้งาน เพียงคนเดียวในการทำการเปลี่ยนแปลงแต่ละครั้ง ทำให้สามารถควบคุมความถูกต้องของข้อมูลได้

3 สามารถใช้ข้อมูลร่วมกันได้

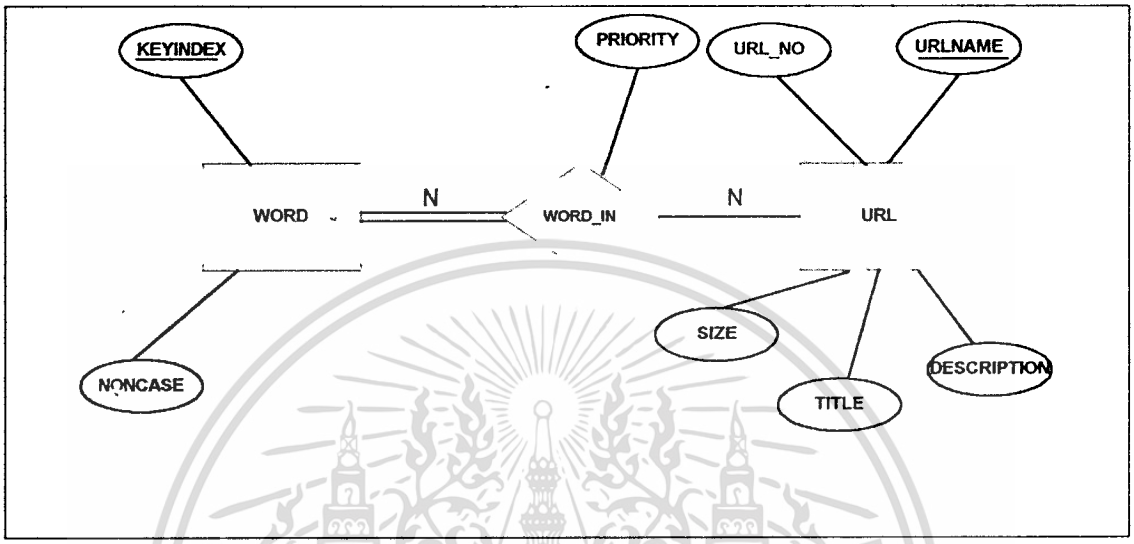
การที่ทำการจัดเก็บข้อมูลโดยใช้ ฐานข้อมูลทำให้ข้อมูลที่จัดเก็บ สามารถที่จะใช้ร่วมกันได้ระหว่างผู้ใช้หลายคน โดยที่สามารถควบคุมความถูกต้องของข้อมูลได้ ในขณะที่หากเราใช้ การเก็บข้อมูลแบบอื่น อาจมีปัญหาในการใช้ข้อมูลร่วมกันในเรื่องของความถูกต้องของข้อมูล

4 สามารถตรวจสอบความขัดแย้งกันของข้อมูลที่จัดเก็บได้

5 การรักษาความปลอดภัยของข้อมูลสามารถทำได้อย่างมีประสิทธิภาพ

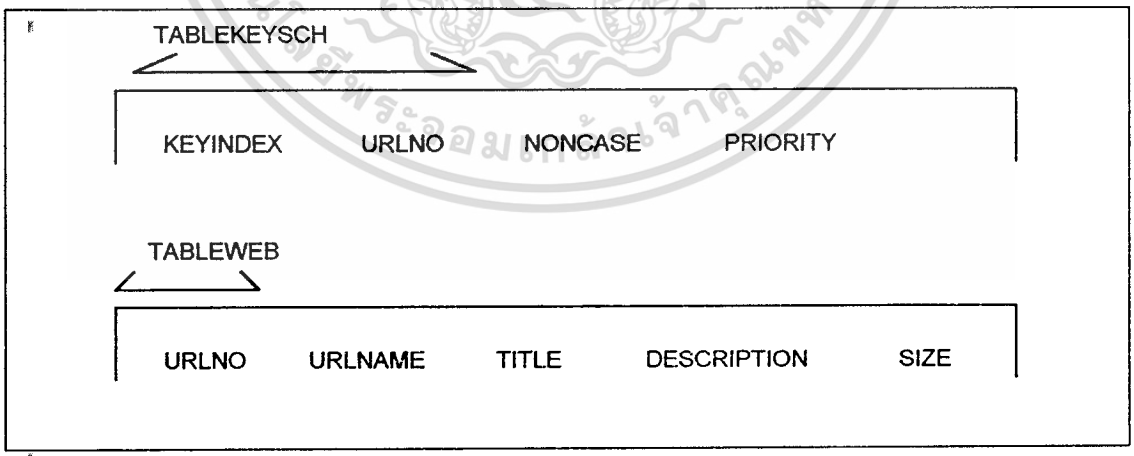
การออกแบบฐานข้อมูลตามหลักการของอีอาร์โมเดล

ในการออกแบบระบบฐานข้อมูลที่จะใช้ภายในโครงงานนี้ เราจะทำการออกแบบตามหลักการของ อีอาร์โมเดล ซึ่งสามารถออกแบบได้ดังนี้



รูปที่ 5.3 แสดงการออกแบบตามหลักอีอาร์โมเดล

เมื่อทำการออกแบบระบบฐานข้อมูลเรียบร้อยแล้ว ก็ทำการ map อีอาร์โมเดลที่ได้เป็นตารางที่จะใช้ทำงานจริงดังรูปต่อไปนี้



รูปที่ 5.4 แสดงตารางที่ได้จากการ MAP อีอาร์โมเดล

5.1.4 ออปชั่นการทำงานของเซิร์สเอนจิน

ในการทำงานของเซิร์สเอนจิน เราทำการกำหนดส่วนของการค้นหาข้อมูลตามเงื่อนไขไว้ให้ผู้ใช้สามารถเลือกได้ดังต่อไปนี้

1. การค้นหาข้อมูลแบบคำเดียวพิจารณารูปแบบ
2. การค้นหาข้อมูลแบบคำเดียวไม่พิจารณารูปแบบ
3. การค้นหาข้อมูลแบบหลายคำกรณี "และ" พิจารณารูปแบบ
4. การค้นหาข้อมูลแบบหลายคำกรณี "และ" ไม่พิจารณารูปแบบ
5. การค้นหาข้อมูลแบบหลายคำกรณี "หรือ" พิจารณารูปแบบ
6. การค้นหาข้อมูลแบบหลายคำกรณี "หรือ" ไม่พิจารณารูปแบบ
7. การค้นหาข้อมูลแบบบวลี พิจารณารูปแบบ
8. การค้นหาข้อมูลแบบบวลี ไม่พิจารณารูปแบบ

5.2 เปรียบเทียบและวิเคราะห์ปัญหารวมทั้งแนวทางในการพัฒนาโครงสร้างของโครงงาน Internet Search Engine กับ Search Engine ที่มีอยู่แล้วในระบบอินเตอร์เน็ต

5.2.1 เวลาในการดึงข้อมูลของโครงงาน Internet Search Engine

จากการทดลองดึงข้อมูลและจับเวลาแสดงให้เห็นว่า เมื่อข้อมูลที่เก็บได้มีขนาดใหญ่ขึ้นจะทำให้ต้องใช้เวลาในการเก็บข้อมูลเฉลี่ยต่อ 1 เว็บเพจมากขึ้น โดยเวลาที่เสียไปในการเก็บข้อมูลแบ่งออกได้เป็น 3 ส่วน คือ

1. เวลาที่ใช้ในการดึงข้อมูล
2. เวลาที่ใช้ในการตัดคำ วิเคราะห์ และเลือกคำที่จะใช้เป็นคีย์เวิร์ด
3. เวลาที่ใช้ในการติดต่อกับฐานข้อมูล อันได้แก่
   - ตรวจสอบเพื่อทดสอบว่ามีข้อมูลจากเว็บไซท์ที่จะทำการดึงข้อมูลนั้น มีอยู่ในฐานข้อมูลแล้วหรือยัง (เฉพาะ Fast Mode)
   - ตรวจสอบว่าข้อมูลที่ได้มาใหม่ตรงกับที่มีอยู่ในฐานข้อมูลเดิมแล้วหรือไม่
   - ลบข้อมูลเก่าที่อยู่ในฐานข้อมูล เพื่อทำการเปลี่ยนแปลงข้อมูลใหม่
   - เพิ่มข้อมูลใหม่เข้าในตาราง

เนื่องจากเวลาที่ใช้ในการดึงข้อมูลเป็นเวลาที่ขึ้นกับความเร็วในการส่งข้อมูลของระบบ ซึ่งต้องแก้ไขทางฮาร์ดแวร์ของระบบทั้งหมด ซึ่งเป็นไปได้ยากมาก ดังนั้นเพื่อให้โปรแกรมสามารถทำการเก็บข้อมูลได้รวดเร็วขึ้น เราจึงต้องทำการปรับปรุงเวลาในส่วนอื่น ๆ แทน ดังนี้

1. แก้ไขเวลาในส่วนของการตัดคำ วิเคราะห์ และเลือกคำที่จะใช้เป็นคีย์เวิร์ด โดยทำการแก้ไขวิธีการในการตัดคำ วิเคราะห์และเลือกคีย์เวิร์ดใหม่

แก้ไขเวลาที่ใช้ในการติดต่อกับฐานข้อมูล โดยแก้ไขวิธีการในการตรวจสอบข้อมูลในฐานข้อมูล และการแก้ไขฐานข้อมูล หรือการเปลี่ยนฐานข้อมูลที่มีความสามารถในการจัดการฐานข้อมูลที่ดีกว่า

การจัดการกับความผิดพลาดที่เกิดขึ้นขณะกำลังดึงข้อมูล

เมื่อเกิดความผิดพลาดบางอย่างขึ้น ในขณะที่บ็อทกำลังทำการเก็บข้อมูลอยู่ ถ้าโปรแกรมสามารถตรวจสอบได้ว่าเกิดความผิดพลาดขึ้น โปรแกรมจะจัดการกับความผิดพลาดที่เกิดขึ้น ซึ่งสาเหตุของความผิดพลาดที่โปรแกรมสามารถจัดการได้มีดังต่อไปนี้ คือ

1. ความผิดพลาดที่เกิดขึ้นจากชื่อยูอาร์แอลที่จะทำการดึงข้อมูล
   1.1. กำหนดชื่อยูอาร์แอลผิดพลาด เช่น เขียนผิด เป็นต้น
   1.2. ไม่มีสิทธิ์ในการเข้าถึงไฟล์หรือไดเร็กทอรี่ที่กำหนดมาในยูอาร์แอล
   1.3. ไฟล์หรือไดเร็กทอรี่ที่กำหนดไว้ในยูอาร์แอลถูกย้ายที่หรือลบไปแล้ว
2. ความผิดพลาดที่เกิดขึ้นในการติดต่อเพื่อขอข้อมูล
   2.1. เว็บไซท์ที่จะติดต่อขอข้อมูลหยุดการทำงาน
   2.2. ไม่มีเส้นทางในการติดต่อกับเว็บไซท์ที่ต้องการ เช่น สายขาด เป็นต้น
3. ความผิดพลาดที่เกิดขึ้นในการส่งข้อมูล

เมื่อเกิดความผิดพลาดจากสาเหตุที่กล่าวมาข้างต้นแล้ว โปรแกรมจะจัดการกับความผิดพลาดนั้น โดยไม่สนใจและกระโดดข้ามไป และจัดการดึงข้อมูลจากเว็บไซท์อื่นต่อไป ตามยูอาร์แอลที่อยู่ในลิสต์

แต่ยังมีสาเหตุบางอย่างซึ่งอยู่นอกเหนือความคาดหมาย ที่ทำให้โปรแกรมในส่วนของการเก็บข้อมูลหยุดการทำงานเพื่อรอการตอบสนองจากเว็บไซท์ปลายทาง เนื่องจากโปรแกรมไม่สามารถตรวจสอบได้ว่าเป็นความผิดพลาด ซึ่งเป็นผลให้โปรแกรมไม่สามารถทำงานได้อัตโนมัติ นั่นคือต้องมีคนคอยนั่งคุมการทำงานของโปรแกรม ซึ่งความผิดพลาดเหล่านี้ ได้แก่ เว็บไซท์ที่ทำการดึงข้อมูลเป็นเว็บไซท์ที่มีความเร็วต่ำ หรือมีการติดต่อมากทำให้การติดต่อเป็นไปได้อย่างล่าช้า ทำให้ส่วนเก็บข้อมูล เป็นต้น

สิ่งที่จะต้องทำการแก้ไขโปรแกรมในส่วนนี้ ได้แก่ การทำให้โปรแกรมสามารถตรวจสอบได้ว่าเกิดความผิดพลาดขึ้น และจัดการกับความผิดพลาดเหล่านี้อย่างเหมาะสม เช่น กำหนดเวลาที่โปรแกรมจะรอการตอบสนองจากเว็บไซท์ปลายทาง เมื่อหมดเวลาก็ให้โปรแกรมถือว่าเกิดความผิดพลาดขึ้น และข้ามไปดึงข้อมูลจากเว็บไซท์อื่นต่อไป

5.2.2 อัลกอริทึมในการตัดคำเพื่อจัดทำอินเด็กซ์

การตัดคำทั่วไปที่ปรากฏในเว็บเพจ

ในการตัดคำทั่วไปนี้จะใช้หลักเกณฑ์ในการตัดคำ 3 ข้อ ต่อไปนี้

1. การตัดคำที่ได้จากเว็บเพจต่าง ๆ นั้น สามารถตัดคำออกมาในรูปแบบของคำที่ประกอบด้วย ตัวอักษร (A-Z และ a-z) ตัวเลข (0-9) และ อักขระพิเศษบางตัว อันได้แก่ '-', '_' ตัวอย่างเช่น Thursday, IC8051, ER-MODEL, JR_VOY เป็นต้น
2. เครื่องหมายที่ใช้ในการแบ่งคำ คืออักขระพิเศษอื่น ๆ ที่อยู่นอกเหนือจากที่กำหนดไว้ในข้อ 1 เช่น เครื่องหมายเว้นวรรค เครื่องหมายขึ้นบรรทัดใหม่ เครื่องหมายคำถาม "?" เป็นต้น ซึ่งเครื่องหมายต่าง ๆ ที่ใช้ในการแบ่งคำนี้จะยกเว้นเครื่องหมาย "<" ซึ่งเป็นเครื่องหมายที่ใช้เป็นแท็กในภาษาเอชทีเอ็มแอล เพราะแท็กนี้สามารถแทรกอยู่ระหว่างคำได้ เช่น

   TEXT Color.

   จากตัวอย่างที่กำหนดจะสามารถตัดคำได้ออกมาเป็น 2 คำ คือ "TEXT" และ "Color" โดยใช้เครื่องหมายเว้นวรรคเป็นตัวแบ่งคำ
3. คำต่าง ๆ ที่ได้มากจากการตัดคำจะต้องไม่ใช่ตัวเลข

จากหลักเกณฑ์ที่ใช้ในการแบ่งคำข้างบนสามารถแบ่งคำได้ดีในระดับหนึ่ง แต่ก็ไม่สามารถครอบคลุมทุกกรณี เช่นคำบางคำที่มีอักขระพิเศษเป็นส่วนประกอบในคำที่อยู่นอกเหนือจากที่กำหนดไว้ในกฎเกณฑ์ข้างต้น เช่น

- OS/2 จะถูกแบ่งออกเป็น 2 คำ คือ "OS" และ "2" แต่คำว่า "2" เป็นตัวเลข ดังนั้นจึงไม่ถือว่าเป็นคำคำหนึ่ง ดังนั้นผู้ใช้ที่ต้องการหาข้อมูลเกี่ยวกับ OS/2 จึงไม่สามารถค้นหาคำนี้ได้

- C++ เนื่องจากเครื่องหมาย '+'เป็นอักขระพิเศษที่อยู่นอกเหนือจากที่กำหนดไว้ดังนั้นคำที่ตัดได้จึงเหลือเพียงแค่ "C" เท่านั้น

จากตัวอย่างข้างบนเป็นกรณีพิเศษ ซึ่งมีอยู่ส่วนน้อย แต่ถ้าอนุญาติให้มีการใช้เครื่องหมาย '/' หรือเครื่องหมาย "+" ประกอบอยู่ในคำได้ จะทำให้มีปัญหาต่าง ๆ เกี่ยวกับการตัดคำอื่นได้เช่น

- "Thai/English" จะถูกมองเป็นคำคำเดียวกัน นั่นคือ "ThaiEnglish" ซึ่งไม่มีความหมาย แต่ควรจะแบ่งออกเป็น 2 คำ คือ "Thai" และ "English" ซึ่งมีความหมายทั้งสองคำ

- "Data+CRC=Packet" ถ้าหากยอมให้สามารถใช้เครื่องหมาย "+" ประกอบในคำได้จะทำให้แบ่งได้ 2 คำ คือ "Data+CRC" และ "Packet" ซึ่งคำแรกที่ได้จะเป็นคำที่ไม่มีความหมาย แต่ถ้าแบ่งประโยคนี้ออกเป็น 3 คำ คือ "Data", "CRC" และ "Packet" จะทำให้ได้คำที่มีความหมายมากกว่า

การตัดคำที่เป็นคีย์เวิร์ดของเว็บเพจที่อยู่ในแท็ก " <META> "

ในภาษาเอชทีเอ็มแอลมีแท็กที่ให้ผู้เขียนเว็บเพจใส่รายละเอียดต่าง ๆ เกี่ยวกับเว็บเพจนั้น ๆ ได้ เช่น รายละเอียดเกี่ยวกับผู้เขียน คำอธิบายเนื้อหาของเว็บเพจนั้น คำที่จะใช้เป็นคีย์เวิร์ดของเว็บเพจนั้น ๆ ได้ ดังมีรูปแบบ ดังนี้ คือ

<meta name="DETAIL" content="STRING">

โดย DETAIL คือ คำเฉพาะที่ใช้บอกว่าเป็นรายละเอียดเกี่ยวกับอะไร

STRING คือ ข้อมูลที่เป็นสตริงค์

สำหรับ DETAIL นี้สามารถมีได้หลายค่า แต่ในที่นี้เราจะสนใจเพียงแค่ 2 อย่าง คือ

1. "description" ข้อมูลที่อยู่ในส่วนของ STRING จะเป็นข้อมูลที่ใช้อธิบายเกี่ยวกับเนื้อหาของเว็บเพจนั้น ๆ ซึ่งใช้เก็บเป็นฐานข้อมูลเพื่อใช้งานต่อไป ในส่วนนี้ยังไม่มีปัญหาใด ๆ เกิดขึ้น
2. "keywords" ซึ่งจะกำหนดคำที่จะใช้เป็นคีย์เวิร์ดของเว็บเพจนั้นได้ โดยคำที่เป็นคีย์เวิร์ดทั้งหมดจะอยู่ใน STRING โดยใช้เครื่องหมาย ',' เป็นตัวแบ่งคำ

เนื่องจาก ในการตัดคำที่เป็นคีย์เวิร์ดจากแท็ก <META> นี้จะใช้เครื่องหมาย ',' เป็นตัวแบ่งคำเพียงตัวเดียว ดังนั้นจะทำให้คำที่ได้มีลักษณะที่แตกต่างจากการตัดคำทั่วไป คือ

1. สามารถมีอักขระพิเศษแทรกอยู่ระหว่างคำได้ เช่น C++, OS/2, DOS6.22 เป็นต้น
2. มีคำที่เป็นคำผสม เช่น 'IP Address', 'Microsoft Office 4.30' เป็นต้น ทำให้สามารถค้นหาข้อมูลแบบ Phase ได้

ข้อเสียที่เป็นปัญหาและต้องทำการแก้ไขต่อไปของการตัดคำด้วยวิธีการนี้ คือ

1. **จำนวนเว้นวรรคของผู้เขียนแต่ละคนไม่เท่ากัน ซึ่งจะทำให้ถูกมองเป็นคำคนละคำกัน เช่น 'IP Address' กับ 'IP Address' ซึ่งอาจแก้ไขได้โดยการทำให้จำนวนเว้นวรรคที่อยู่ระหว่างคำเป็นเครื่องหมายเว้นวรรคเพียงตัวเดียว ในที่นี้คือ 'IP Address'**
2. **คำที่ได้มีอักขระพิเศษที่ไม่ต้องการแทรกอยู่ระหว่างคำได้ เช่น มีเครื่องหมายขึ้นบรรทัดใหม่แทรกอยู่ระหว่างคำตามตัวอย่าง**

<meta name="keywords" content="aaa, bbb, ccc, ddd eee,fff">

เมื่อทำการตัดคำจะได้คีย์เวิร์ด 5 ตัว คือ aaa, bbb, ccc, ***ddd~eee***, fff ทำให้ได้ คีย์เวิร์ดที่มีเครื่องหมายขึ้นบรรทัดใหม่แทรกอยู่ระหว่างคำ

ปัญหาและสิ่งที่จะต้องพัฒนาต่อไปในส่วนของการวิเคราะห์และเลือกคำที่ใช้เป็นคีย์เวิร์ด

1. คำที่เลือกมาเป็นคำทั่วไปไม่มีความหมายพิเศษพอที่จะเป็นคีย์เวิร์ดได้ เช่น can, long เป็นต้น ซึ่งแก้ไขได้โดยการเพิ่มคำที่เราไม่ต้องการใช้เป็นคีย์เวิร์ดเหล่านี้เข้าไปในตาราง Tableunusewd ซึ่งเป็นตารางที่โปรแกรมจะใช้ในการเปรียบเทียบกับคำที่ได้ และจะตัดคำนั้นออกจากการใช้เป็นคีย์เวิร์ด ถ้าคำนั้นปรากฏอยู่ในตาราง
2. คำที่เลือกมาบางคำไม่ใช่คำที่จะใช้เป็นคีย์เวิร์ดของเว็บเพจนั้นได้ เนื่องจากคำเหล่านี้ปรากฏอยู่ในตำแหน่งที่มีค่าความสำคัญมาก ทำให้ค่าความสำคัญของคำเหล่านี้สูงกว่าคำที่ควรจะเป็นคีย์เวิร์ดจริง ๆ ดังนั้นจึงต้องทำการแก้ไขการกำหนดค่าความสำคัญให้เหมาะสมกว่านี้ หรืออาจจะไม่สนใจเนื่องจากเป็นคำส่วนน้อย หรืออาจจะเก็บคำให้มากขึ้นเพื่อให้ครอบคลุมเนื้อหามากขึ้น แต่วิธีนี้จะทำให้เสียเนื้อที่ในการเก็บและเสียเวลาในการติดต่อกับฐานข้อมูลมาก
3. การกำหนดค่าความสำคัญของคำในตำแหน่งต่าง ๆ ในเวอร์ชันนี้จะใช้แท็กที่สำคัญเพียงบางส่วนเท่านั้น ทำให้ค่าความสำคัญของคำไม่ละเอียดมากเท่าที่ควร ดังนั้นในการปรับปรุงโปรแกรมครั้งต่อไปจึงควรจะนำแท็กอื่น ๆ มาร่วมใช้ในการคำนวณค่าความสำคัญของคำด้วยเพื่อให้คำที่ได้ มีความเหมาะสมที่จะใช้เป็นคีย์เวิร์ดมากขึ้น

5.2.3 ระบบฐานข้อมูลที่ใช้ในการเก็บข้อมูล

โครงงานหุ่นยนต์ค้นหาข้อมูลยนเครือข่ายคอมพิวเตอร์จะใช้ Interbase ใน Delphi เป็นตัวที่ใช้ในการจัดเก็บข้อมูล

ข้อดีของการใช้ระบบฐานข้อมูล Interbase

1. สามารถจัดสร้างฐานข้อมูลในการเก็บข้อมูลได้ง่าย
2. ใช้เนื้อที่น้อยในการติดตั้งโปรแกรมรวมทั้งการจัดเก็บข้อมูล
3. สามารถทำการติดต่อการทำงานกับบ็อทที่เขียนด้วยภาษาเคลไฟได้ง่าย
4. ผู้ที่จะเข้ามาดูแลข้อมูลที่จัดเก็บสามารถทำการศึกษาการทำงานของระบบฐานข้อมูล Interbase ได้ง่าย

ข้อเสียของการใช้ระบบฐานข้อมูล Interbase

1. มีขีดจำกัดในการจัดเก็บข้อมูลโดยข้อมูลที่สามารถจัดเก็บได้และไม่มีผลต่อเวลาในการทำการค้นหามากเกินไป คือประมาณ 15,000 แถว

2. ความสามารถในการค้นหาข้อมูลขึ้นอยู่กับจำนวนข้อมูลที่จัดเก็บ

เมื่อเปรียบเทียบกับการทำงานของ Search Engine ที่มีอยู่แล้วในปัจจุบันจะพบว่า ระบบฐานข้อมูลของโครงงานยังสามารถเก็บข้อมูลได้ไม่มากพอรวมทั้งยังมีผลทำให้การค้นหาข้อมูล ใช้เวลามาก

แนวทางในการแก้ปัญหา

1. ทำการเปลี่ยนระบบฐานข้อมูลเป็นระบบที่มีขีดความสามารถมากขึ้น เช่น SQL Server

2. เปลี่ยนรูปแบบของตารางที่ใช้จัดเก็บข้อมูล เพื่อลดความซ้ำซ้อน เนื่องจากสามารถเพิ่มขีด ความสามารถในการทำงานจากการพัฒนาความสามารถของ ฮาร์ดแวร์ได้

5.3 ประเมินผลโครงงาน

ประเมินผลความสามารถของโครงงานได้ดังนี้

1. โรบอทที่สร้างขึ้นสามารถตรวจสอบเว็บไซท์ที่ผ่านไปถึงได้ว่าอยู่ในขอบเขตที่กำหนดไว้หรือไม่

2. โรบอทสามารถดึงข้อมูลจากเว็บไซท์ที่ต้องการได้

3. โรบอทสามารถวิเคราะห์ข้อมูลที่ดึงมาได้ คือสามารถคัดเลือกคำที่เหมาะสม และคำนวณค่าความสำคัญได้

4. โรบอทสามารถเก็บคำที่คัดเลือกและค่าความสำคัญที่คำนวณลงในฐานข้อมูลได้

5. ฐานข้อมูลที่ออกแบบสามารถเก็บข้อมูลได้ครบตามที่ต้องการ

6. สามารถทำการค้นหาข้อมูลในฐานข้อมูลได้ตามเงื่อนไขที่กำหนดจากผู้ใช้บริการ

7. สามารถแสดงผลลัพธ์กลับมายังหน้าจอของผู้ใช้บริการได้ และครบตามที่กำหนดไว้

5.4 วิเคราะห์ประสิทธิภาพของโครงงาน

1. การตรวจสอบเว็บไซท์ของโรบอทยังมีปัญหาในการตรวจสอบ เนื่องจากยังไม่ครอบคลุมข้อมูลทั้งหมด เช่นกรณีที่เว็บไซท์ของไทยแต่ไปตั้งอยู่ในต่างประเทศ

2. การดึงข้อมูลของโรบอทในบางเว็บไซท์ ไม่สามารถทำได้โดยไม่ทราบสาเหตุ แล้วทำให้โรบอทหยุดการทำงาน

3. การดึงข้อมูลในกรณีที่ทำการดึงหลายระดับในบางครั้งจะใช้เวลานานมาก

4. การวิเคราะห์ข้อมูลในส่วนของการเลือกคำยังไม่สมบูรณ์แบบ

5. การค้นหาข้อมูลในฐานข้อมูลสามารถทำได้ถูกต้อง

6. การแสดงผลลัพธ์ของข้อมูลที่ออกมาทางหน้าจอของผู้ใช้บริการยังไม่สมบูรณ์

7. ระบบฐานข้อมูลที่เลือกใช้มีขีดจำกัดในเรื่องของจำนวนเมื่อ มากเกินไปจะมีผลต่อการความเร็วในการค้นหาข้อมูล

การพัฒนาโครงงานในอนาคต

โครงงานหุ่นยนต์ค้นหาข้อมูลสามารถที่จะพัฒนาขีดความสามารถได้อีกในเรื่องต่อไปนี้

1. พัฒนาหลักการในการคัดเลือกคำและคำนวณค่าความสำคัญให้เหมาะสมยิ่งขึ้น
2. เปลี่ยนระบบฐานข้อมูลที่ใช้เก็บข้อมูลให้มีขีดความสามารถมากขึ้น
3. พัฒนาส่วนที่แสดงผลให้มีความสวยงามและง่ายต่อการใช้งานมากขึ้น

5.5 สรุปโครงงาน

โครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ สามารถทำงานอันเป็นประโยชน์ในด้านการช่วยค้นหาข้อมูลที่เกี่ยวกับประเทศไทยบนเครือข่ายคอมพิวเตอร์ได้ ตามขอบเขตและจุดประสงค์ที่ตั้งไว้ แต่ก็พบปัญหาที่เกิดขึ้นในการทำงานในด้านของประสิทธิภาพของการทำงาน เช่น ความเร็ว , จำนวนข้อมูลที่สามารถจัดเก็บ อันควรที่จะได้รับการพัฒนาต่อไปในอนาคต ตามแนวทางที่ได้ศึกษาไว้

ภาคผนวก

ภาคผนวก ก

อินเตอร์เน็ต

อินเตอร์เน็ต คือ เครือข่ายคอมพิวเตอร์ของเครือข่ายคอมพิวเตอร์หลาย ๆ เครือข่าย ซึ่งเครือข่าย คือ ระบบที่มีคอมพิวเตอร์ตั้งแต่ 2 เครื่องขึ้นไปมาเชื่อมต่อกัน และใช้ข้อมูลร่วมกัน โดยติดต่อสื่อสารกันผ่านโปรโตคอล ทีซีพี/ไอพี นั่นคือการสื่อสารข้อมูลจะเป็นการส่งแพคเก็ต (Packet) กล่าวคือข้อมูลที่ได้ส่งจะได้รับการตัดะแบ่งเป็นแพคเก็ต นำมากำหนดแอดเดรสปลายทางที่ต้องการส่งไปและกำหนดแอดเดรสต้นทางของผู้ส่ง โดยระบบแอดเดรสนี้ จะใช้หมายเลขไอพีที่ได้ รับการกำหนดและออกแบบมาเพื่อใช้ในเครือข่ายคอมพิวเตอร์ การกำหนดหมายเลขไอพีเป็นมาตรฐานกลางสำหรับเครือข่าย โดยกำหนดให้เครื่องคอมพิวเตอร์ทุกเครื่อง ที่ต่ออยู่บนเครือข่ายมีรหัสหมายเลขไอพีไม่ซ้ำกัน

ดังนั้น อินเตอร์เน็ตจึงเป็นการเชื่อมต่อคอมพิวเตอร์จากทั่วทุกมุมโลกเข้าด้วยกัน คอมพิวเตอร์ที่อยู่ในอินเตอร์เน็ตอาจเป็นได้ทั้งคอมพิวเตอร์ที่ใช้งานร่วมกัน หรือคอมพิวเตอร์ ส่วนบุคคลก็ได้ที่ต้องการเชื่อมต่อเข้าด้วยกัน เพื่อใช้บริการบางอย่างร่วมกัน อินเตอร์เน็ตจะ อนุญาติให้ผู้ใช้บนเครือข่ายสามารถเข้าถึงข้อมูลของผู้ใช้คนอื่น หรือข้อมูลบนเครือข่ายที่ได้รับ อนุญาติแล้วได้ ด้วยความสามารถนี้ จึงทำให้มีผู้คนนับล้านใช้อินเตอร์เน็ตในทุกวันนี้

บริการต่าง ๆ ที่มีอยู่ในอินเตอร์เน็ต

- จดหมายอิเล็กทรอนิกส์ (Electronic Mail)
- ข่าวสาร (Usenet News)
- การท่องไปบนเวิร์ลไวด์เว็บ
- โปรโตคอลการส่งผ่านไฟล์ หรือ เอฟทีพี
- เทลเน็ต (Telnet)
- ฟิงเกอร์ (Finger)
- ฮูอิส (Whois)
- พิง (Ping)

ไซท์และเว็บไซท์

ไซท์ คือ คอมพิวเตอร์ที่อยู่บนเครือข่ายอินเตอร์เน็ต ในเครือข่ายอินเตอร์เน็ต คอมพิวเตอร์ที่กำลังใช้งานอยู่จะเรียกว่า โลคอลไซท์ (Local Site) ส่วนคอมพิวเตอร์ที่กำลังติดต่อสื่อสารอยู่ด้วยขณะนั้นเรียกว่า รีโมตไซท์ (Remote Site) ซึ่ง คอมพิวเตอร์ทั้ง 2 ฝ่ายจะต้องถูกกำหนดชื่อที่ทั้ง 2 ฝ่ายรู้จักโดยใช้ หมายเลขไอพี

เว็บไซท์ คือ คอมพิวเตอร์ที่อยู่บนเครือข่ายอินเตอร์เน็ต ที่ให้เป็นแหล่งรวมและ บริการเอกสารบนเว็บ (Web Document) ซึ่งอยู่ในรูปแบบที่เรียกว่า ไฮเปอร์เท็กซ์ (HyperText) ซึ่งเขียนโดยใช้ภาษา เอชทีเอ็มแอล (HTML or HyperText Mark-up Language) ซึ่งเป็นภาษามาตรฐาน ที่ใช้บนเว็บ เอชทีเอ็มแอล คือข้อความที่เขียนโดยใช้สัญญลักษณ์ในการกำหนด โครงสร้างและ รูปแบบการแสดงผลของเอกสาร

เวิร์ล ไวด์ เว็บ (World Wide Web)

เวิร์ล ไวด์ เว็บ หรือเรียกสั้น ๆ ว่า เว็บ (Web) คือ แหล่งรวมเอกสารข้อมูลต่าง ๆ ที่อยู่ในที่อยู่ตามไซท์ต่าง ๆ บนอินเตอร์เน็ต การนำเสนอข้อมูล บนอินเตอร์เน็ต ซึ่ง เวิร์ล ไวด์ เว็บ นี้จะประกอบด้วยส่วนต่าง ๆ ดังต่อไปนี้

1. เว็บไคลเอ็นท์ (Web Clients)

โปรแกรมที่ทำหน้าที่เป็นเว็บไคลเอ็นท์ จะต้องสามารถติดต่อและเข้าถึงข้อมูลบนเว็บเซิร์ฟเวอร์อื่น ๆ ที่อยู่บนอินเตอร์เน็ต

2. เว็บเซิร์ฟเวอร์ (Web server)

โปรแกรมเว็บเซิร์ฟเวอร์ ส่วนใหญ่จะทำงานบนเครื่องที่ทำงานได้พร้อมกันหลาย ๆ งาน (Multitasking Workstation) ที่มีความสามารถมากเพียงพอ

3. เว็บพร็อกซี่ (Web Proxies)

พร็อกซี่ มักเป็นเว็บเซิร์ฟเวอร์ที่จะทำงานบนเครื่องที่มีระบบรักษาความปลอดภัย (Firewall Machine) เพื่อคอยรักษาความปลอดภัยของข้อมูล และป้องกันอันตรายที่อาจเกิดขึ้น ในการติดต่อสื่อสารกันระหว่าคอมพิวเตอร์ในเครือข่ายแบบท้องถิ่น (Local Area NetWork หรือ LAN) ขนาดเล็ก กับ เครือข่ายอินเตอร์เน็ต

นอกจากนี้ พร็อกซี่ยังใช้เป็นแคช (Cache) ของเอกสารบนเว็บ ซึ่งจะเป็นประโยชน์มาก เมื่อมีผู้ใช้หลายคนในองค์กรต้องการใช้เว็บเพจ (Web Page) เดียวกัน

4. ชื่อของเว็บ (Web รีซอร์ส Naming)

เวิร์ล ไวด์ เว็บ จะมีการทำกำหนดชื่อของไซต์ต่าง ๆ ที่อยู่ในระบบทั้งหมดโดยใช้ ยู อาร์ แอล (URL) หรือ ยู อาร์ ไอ (ยูอาร์ไอ)

5. โปรโตคอล (Protocol)

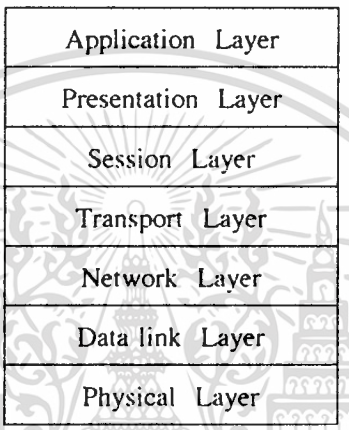
โปรโตคอลที่ใช้กันมากก็คือ เอชทีทีพี (HTTP) เพราะเป็นโปรโตคอลบน อินเตอร์เน็ต (Internet Protocol) สำหรับติดต่อกับเว็บเซิร์ฟเวอร์ มีความสามารถในการอ่านข้อมูลในรูปแบบของตัวอักษร โปรโตคอลเอชทีทีพี ไม่ใช่โปรโตคอล สำหรับการส่ง ไฮเปอร์เท็กซ์ เท่านั้น แต่เป็นโปรโตคอลสำหรับส่งข้อมูลใด ๆ ก็ได้ที่ ไฮเปอร์เท็กซ์ ติดต่อไปถึง

6. ซี จี ไอ (CGI)

เป็นโปรแกรมที่ถูกเรียกให้ใช้งานโดย เว็บเซิร์ฟเวอร์ ทำงานบนเว็บเซิร์ฟเวอร์ และส่งผลลัพธ์ที่ได้ผ่านทางเว็บเซิร์ฟเวอร์ไปยังไคลเอ็นท์

โอเอสไอโมเดล (OSI Model)

อินเตอร์เนชันแนล สแตนดาร์ด ออร์กาไนซ์เซชั่น International Standards Organization (ISO) ได้กำหนดแบบจำลองการทำงานของระบบเครือข่ายคอมพิวเตอร์ (Computer networking model) เรียกว่า open system interconnection model (OSI Model) ได้เป็น 7 ชั้น ดังนี้



OSI 7 layer model

1. ฟิสิคอล เลเยอร์ (Physical layer) ทำหน้าที่ในการสื่อสารทางกายภาพ ระดับสัญญาณ ทางไฟฟ้า วัสดุตัวนำที่ใช้ในการสื่อสาร การต่อเชื่อมทางกายภาพ มาตรฐานในระดับนี้ ได้แก่ RS-232, DIX-Ethernet(IEEE 802.3), Token-ring (IEEE 802.5) เป็นต้น อุปกรณ์ใน เครือข่ายที่ทำงานในชั้นนี้ เรียกว่า รีพีทเตอร์ (Repeater) หรือ ตัวทวนสัญญาณ

2. ดาต้าลิงค์ เลเยอร์ (Data link layer) ทำหน้าที่ควบคุมการสื่อสาร แบบจุด ต่อจุดที่ติดกัน (Point to Point) ให้สามารถได้รับข้อมูลได้อย่างถูกต้อง ปราศจากข้อมูลที่ผิดพลาด ในโหนดที่ติดกัน มาตรฐานในระดับนี้ อุปกรณ์ในเครือข่ายที่ทำงานในชั้นนี้ เรียกว่า Bridge

3. เน็ตเวิร์ค เลเยอร์ (Network layer) ทำหน้าที่ควบคุมการสื่อสารระหว่าง ต้นทาง กับปลายทาง (End to End) ซึ่งในระหว่างทางจะมีเครือข่ายอยู่หรือไม่ก็ได้ รวมทั้งการหาเส้นทางที่เหมาะสมในการเดินทางของข้อมูลจากต้นทางไปยังปลายทาง (Routing) มาตรฐานในระดับนี้ ได้แก่ X.25, IP, SPX เป็นต้น อุปกรณ์ในเครือข่ายที่ทำงานในชั้นนี้ เรียกว่า Router

4.ทรานสปอร์ต เลเยอร์ (Transport layer) ทำหน้าที่ควบคุมการสื่อสารระหว่างต้นทางและปลายทางให้สามารถได้รับข้อมูลที่ถูกต้องปราศจากข้อผิดพลาด (Error free) มาตรฐานในระดับนี้ ได้แก่ TCP, IPX เป็นต้น

5. เซสชั่น เลเยอร์ (Session layer) ทำหน้าที่ควบคุมการจัดการจราจรในการสื่อสาร เช่น การหยุดส่งข้อมูลชั่วคราวเมื่อฝ่ายรับรับข้อมูลไม่ทัน หรือประมวลผลข้อมูลนั้นยังไม่เสร็จ เป็นต้น ซึ่งโดยปรกติจะจัดการโดยโปรแกรมประยุกต์เอง

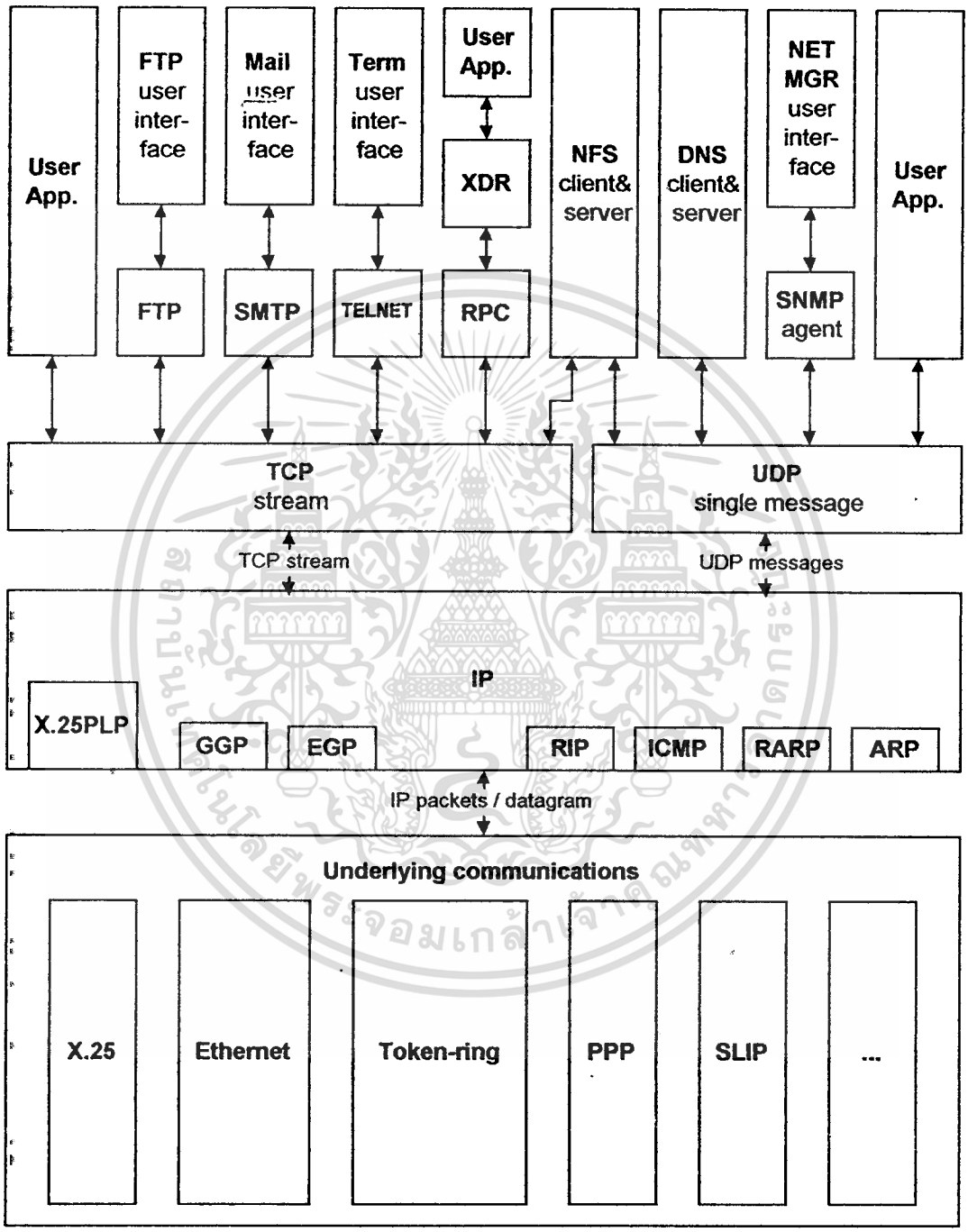
6. พรีเซนต์เตชั่น เลเยอร์ (Presentation layer) ทำหน้าที่ควบคุมการแสดงผลข้อมูล การแทนค่าข้อมูล การให้ความหมายของข้อมูล เช่น จะให้ข้อมูลชุดนี้แทนรหัส ASCII หรือ EBCDIC, การตีความกลุ่มของข้อมูลว่าเป็น เลขจำนวนเต็ม หรือ ตัวอักษร หรือ จำนวนจริง เป็นต้น ซึ่งโดยปรกติจะจัดการโดยโปรแกรมประยุกต์เอง

7. แอพพลิเคชั่น เลเยอร์ (Application layer) เป็นระดับโปรแกรมประยุกต์ เช่น การส่งข้อความร้องขอข้อมูล การส่งข้อมูลตอบกลับ เป็นต้น

จะเห็นว่าสามารถแบ่งแบบจำลองออกได้เป็น 2 ส่วนใหญ่ๆ ตามลักษณะการใช้งานจริงๆ คือ 1. ส่วนที่รับผิดชอบโดยระบบปฏิบัติการ (O.S.) คือ ชั้นที่ 1-4 จะเรียกว่าผู้ให้บริการขนส่ง
2. ส่วนที่อยู่ในโปรแกรมประยุกต์ คือชั้นที่ 5-7 จะเรียกว่า โปรแกรมประยุกต์

ดังนั้นในการสร้างโปรแกรมประยุกต์ที่จะสามารถติดต่อสื่อสารผ่านเครือข่ายได้ จะต้องเรียกใช้บริการของผู้ให้บริการขนส่ง เหมือนการเขียนโปรแกรมบนดอส จะเรียกใช้บริการของดอสในการติดต่อกับดิสก์ โดยไม่ต้องเขียนโปรแกรมติดต่อกับดิสก์โดย ตรงให้ยุ่งยากและผูกติดกับดิสก์แบบนั้น หรือการเขียนโปรแกรมบนวินโดว์ ก็จะเรียกใช้บริการของวินโดว์ ซึ่งลักษณะการเรียกใช้บริการแบบนี้จะเรียกว่าเป็นการติดต่อผ่าน API (Application Program Interface) ในระบบเครือข่ายก็เช่นกันที่จะต้องมี API เช่นในระบบยูนิกส์ที่เป็นที่นิยมมีอยู่ 2 วิธี คือ BSD Sockets และ System V TLI

ทีซีพี/ไอพี



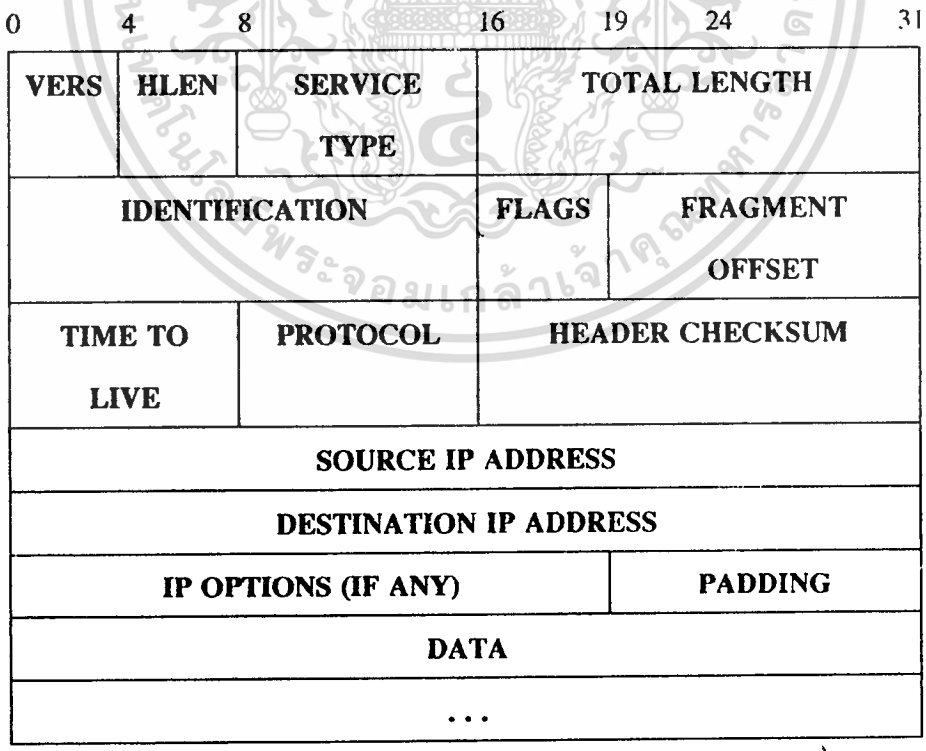
TCP/IP protocol suites

ทีซีพีไอพีโปรโตคอลซุท (TCP/IP Protocol suite) คือ ชุดโปรโตคอลที่มีทีซีพี ไอพี เป็นหลักและโปรโตคอลอื่นๆที่ทำงานร่วมกับทีซีพีไอพี ในชั้นอื่นๆของโอเอสไอ โมเดล (ยกเว้น ชั้นที่ 1 และ 2) ซึ่งจะมีทั้งที่เป็นโปรโตคอลช่วยเหลือ (facility) และโปรโตคอลหลัก ซึ่งโปรโตคอลเหล่านี้จะมีเป็นมาตรฐานในทีซีพีไอพี ที่มีขายทั่วไป ในส่วนของชั้นที่ 1 และ 2 นั้นจะไม่มีการกำหนดไว้ แต่โดยทั่วไปจะเป็น อีเธอร์เน็ต ใน แลนและ X.25 ใน แวน เนื่องจาก ว่าโปรโตคอลเหล่านี้ได้มีมาก่อนที่ ทีซีพีไอพี จะเกิดขึ้นในรายงานนี้จะกล่าวถึง โปรโตคอลที่ สำคัญๆ เท่านั้นดังนี้

อินเตอร์เน็ต โปรโตคอล (IP : Internet Protocol)

ไอพีเป็นโปรโตคอลที่ทำหน้าที่หาเส้นทาง(routing)คือ ส่งข้อมูลระหว่างต้นทาง และปลายทาง แต่ไม่มีการประกันว่าข้อมูลที่ผ่านจากไอพี ขึ้นไปนั้นจะถูกต้อง อาจจะมีการสูญหาย การซ้ำซ้อนของข้อมูล การไม่เรียงลำดับของข้อมูล ซึ่งจะเป็นหน้าที่ของโปรโตคอลชั้น บนขึ้นไปจะเป็นส่วนจัดการ

ในการหาเส้นทาง นั้น ไอพี จำเป็นจะต้องมีความสามารถในการแลกเปลี่ยนข่าวสารที่ใช้ในการทำด้วย เช่น RIP : Routing Information Protocol, GGP : Gateway-Gateway Protocol, EGP : External Gateway Protocol เป็นต้น แต่โปรโตคอลเหล่านี้อาจจะไม่มีใน โหนดทั่วไปจึง ไม่ได้กล่าวไว้ด้วย และโปรโตคอลที่จำเป็นจะต้องมี ซึ่งจะได้กล่าวถึงต่อไป

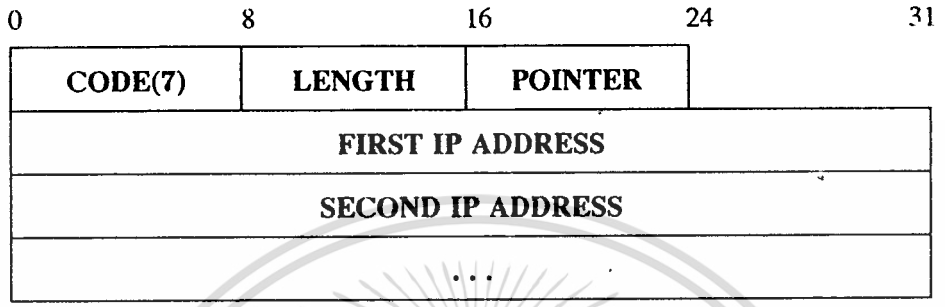


ไอพี ไดอะแกรม

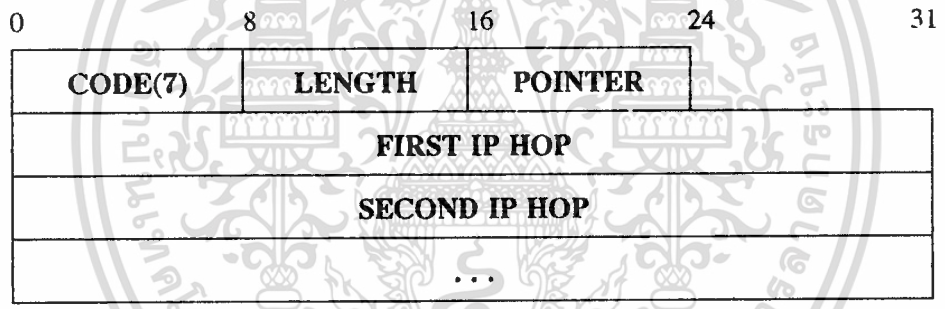
ในการพัฒนาซอร์ฟแวร์ทีซีพีไอพี หรือการทดสอบเครือข่ายไอพี ก็ได้จัดให้มี แพคเก็ต packet ที่ใช้ในงานเหล่านี้ได้ โดยจะต้องเขียนโปรแกรมเพื่อสร้าง ไอพีแพคเก็ต เอง เพื่อส่งไปในเครือข่าย ความสามารถนี้จะอยู่ในส่วนไอพีออปชั่น ซึ่งปกติจะไม่ใช้ (ขนาดเป็น 0)

ไอพี ออปชั่น

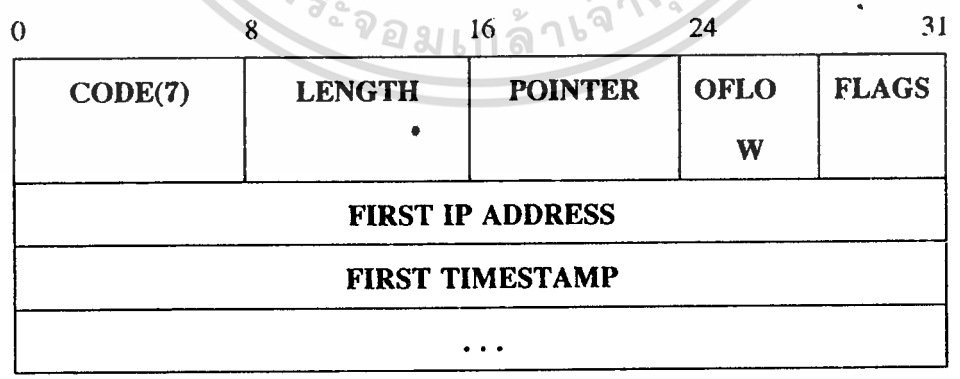
- จะบันทึก ไอพี แอดเดรส ของเครื่องตามเส้นทางที่ข้อมูลผ่านไป



- จะกำหนดไอพี แอดเดรส ของเครื่องตามเส้นทางที่จะให้ข้อมูลผ่านไป



- เวลาที่บันทึกไว้ จะเพิ่มเวลาที่ข้อมูลไปถึงด้วย



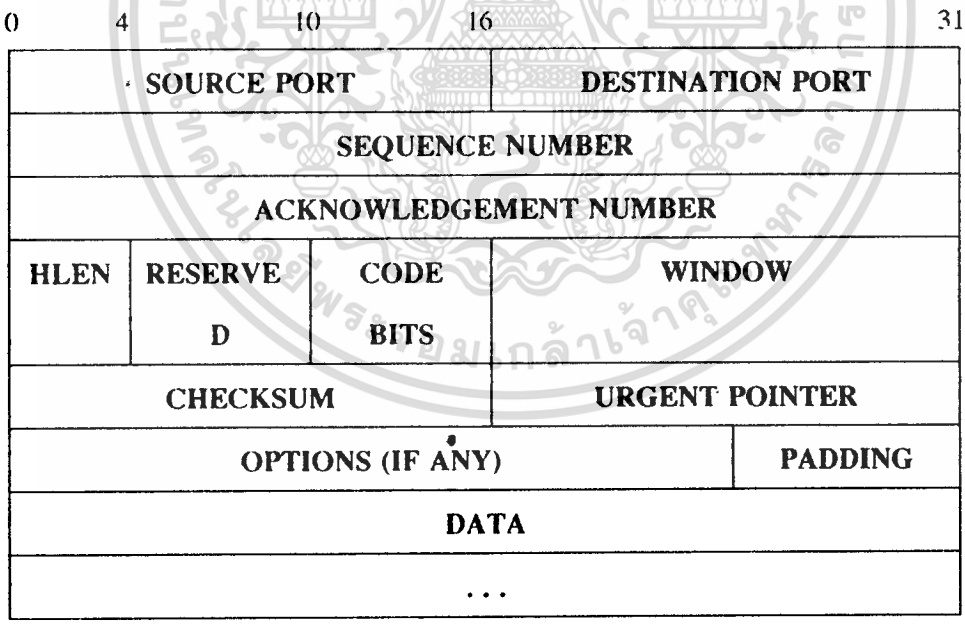
ทีซีพี (TCP : Transmission Control Protocol)

ทีซีพี จะทำงานในชั้นที่ 4 ของโอเอสไอ โมเดล ซึ่งใช้เพื่อตรวจสอบความถูกต้องของข้อมูลที่รับส่งโดย ไอพีในชั้นที่ 3 เช่น การเรียงข้อมูล (Re-assembling) ในกรณีที่ข้อมูลเข้ามาไม่มีการเรียงกัน การขอให้ส่งข้อมูลใหม่ (Retransmit) ในกรณีที่ไม่ได้รับข้อมูล การตัดข้อมูลทิ้ง (Discard) ในกรณีที่ได้รับข้อมูลซ้ำเป็นต้น ข้อมูลที่ผ่านชั้นนี้ไปจะถือเป็นข้อมูลที่ปราศจาก ข้อผิดพลาด (Error free) และข้อมูลที่ส่งขึ้นไปบนชั้นบนจะเป็นลักษณะไบท์เรียงกันไป (Byte stream)

การติดต่อสื่อสารกันจะเป็นลักษณะ(Connection oriented)คือ ต้องมีการสร้าง การเชื่อมต่อขึ้นมาก่อน เปรียบได้กับโทรศัพท์ต้องมีการหมุนโทรศัพท์และต้องมีคนรอรับด้วย เหมาะกับการสื่อสารที่ใช้เวลานานและมีข้อมูลมาก

สลิดดิง วินโดว์ (Sliding windows)

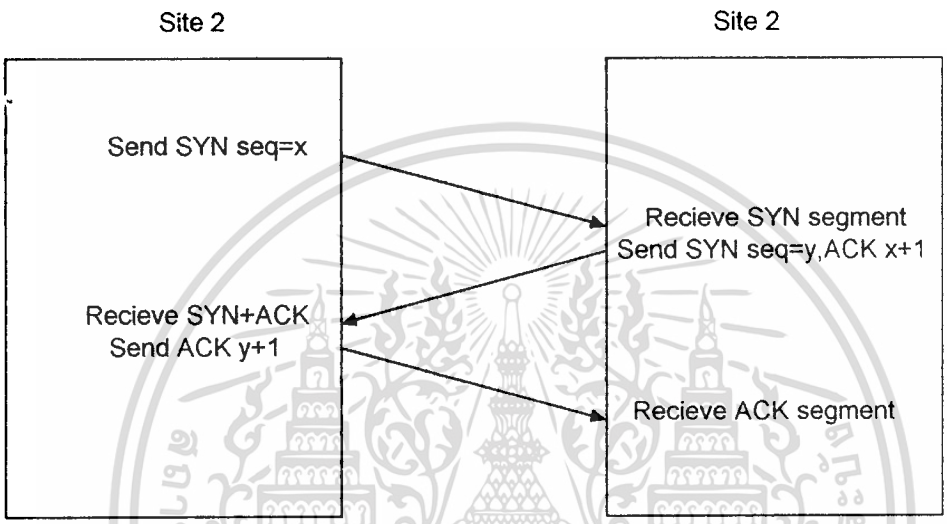
ฝ่ายที่ส่งข้อมูลเมื่อส่งข้อมูลไปแล้วจะต้องรอรับ ACK ก่อนจึงจะส่งข้อมูลต่อได้ เรียกว่า Positive acknowledgement จะเสียเวลาว่างมากจึงมี sliding windows ซึ่งจะสามารถส่งข้อมูล ไปเรื่อยๆ จนกระทั่งข้อมูลที่ส่งมีจำนวนแพคเก็ต เท่ากับขนาดของวินโดว์ จึงจะหยุดส่งหรือส่งใหม่ถ้ารอการตอบรับนานเกินไป (time out) แต่โดยปกติขณะที่ส่งข้อมูลออกไปก็จะมีการ ACK ตอบกลับมาทำให้สามารถส่งไปเรื่อยๆได้



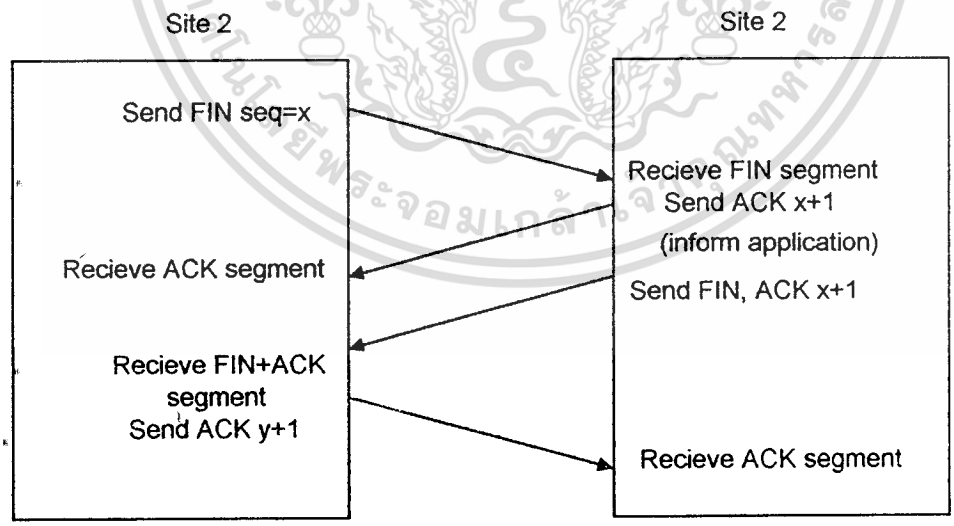
TCP segment format.

Code bits (left to right)

URG	Urgent field is valid
ACK	Acknowledgement field is valid
PSH	Push
RST	Reset
SYN	Synchronize sequence number
FIN	Sender has reached end of its byte stream



การสร้าง connection



การยกเลิก connection

ภาคผนวก ง

ยูนิฟอร์มรีซอร์สโลเคเตอร์ (Uniform Resource Locators)

ยูอาร์แอล เป็นวิธีการกำหนดที่อยู่ของเอกสารหรือข้อมูลในเวิร์ลไวด์เว็บ ซึ่งใน ยูอาร์แอล จะประกอบด้วยข้อมูลต่าง ๆ ดังนี้

- ชื่อของสถานที่ที่มีแหล่งข้อมูลอยู่ (เอกสาร หรือ ข้อมูล)
- ชนิดของบริการที่บริการแหล่งข้อมูลนั้น
- เบอร์พอร์ทของบริการ ถ้าไม่มีการกำหนด บราวเซอร์จะกำหนดให้เป็นค่ามาตรฐาน
- ตำแหน่งของแหล่งข้อมูลนั้นในโครงสร้างไดเร็กทอรี่ ที่ทำหน้าที่เป็นเซิร์ฟเวอร์

ชนิดของบริการ

ในส่วนแรกนี้คือตัวกำหนดชนิดของบริการ (ในที่นี้ คือ การบริการแบบเอ็ชทีทีพี) ที่จะกำหนดวิธีในการเข้าถึงข้อมูลส่วนนี้จะอยู่หน้าเครื่องหมาย ":" ตัวอย่างของบริการได้แก่

- เอ็ชทีทีพี
- โกฟเฟอร์
- เอฟทีพี
- ดับบลิวเอไอเอส
- เทลเน็ต
- ยูสเน็ต
- เมลทู

ที่อยู่และหมายเลขพอร์ท (Address and Port number)

ในส่วนที่ 2 จะเป็น อินเตอร์เน็ตแอดเดรส ของเครื่องที่ให้บริการ กำหนดไว้หลัง เครื่องหมาย "//" ซึ่งที่อยู่นี้อาจจะประกอบด้วย หมายเลขพอร์ท ที่ใช้ในการให้บริการ หรือไม่ก็ได้ จากตัวอย่างข้างบน ส่วนที่เป็นที่อยู่ คือ "//www.address.edu:1234" ซึ่ง ":1234" คือ หมายเลข พอร์ท ถ้าไม่กำหนดจะหมายถึงหมายเลขพอร์ททั่วไป เช่น บริการแบบเฮ็ชทีทีพี จะมีหมายเลข พอร์ท คือ

ที่อยู่ของแหล่งข้อมูล (Resource Location)

เครื่องหมาย "/" ที่อยู่หลังจากการกำหนดชื่อเครื่องและหมายเลขพอร์ท จะเป็นตัวเริ่มต้นในการกำหนดตำแหน่งของข้อมูลที่จะถูกติดต่อใช้งาน

กรณีพิเศษ

ในบางกรณี ไม่จำเป็นต้องกำหนด ที่อยู่อินเตอร์เน็ต และ ตำแหน่งของแหล่งข้อมูล ตัวอย่างเช่น news (สำหรับติดต่อขอใช้กับข้อมูลข่าวสาร) และ mailto (สำหรับส่งจดหมาย)

ยูอาร์แอลสำหรับเฮ็ชทีทีพีเซิร์พเวอร์

เนื่องจากภาษาเฮ็ชทีเอ็มแอล ส่วนใหญ่อยู่ภายใต้การบริการของ เฮ็ชทีทีพี เซิร์พเวอร์ รูปแบบของ ยูอาร์แอล แบบนี้จึงเป็นที่คุ้นตามากกว่าแบบอื่น

ส่วนแรก 'http:' หมายถึง เอกสารชิ้นนี้อยู่ภายใต้การบริการของเฮ็ชทีทีพี เซิร์พเวอร์ ส่วนต่อมาคือ เครื่องหมาย '//' หมายถึง ส่วนที่ตามหลังคือ ชื่อของเครื่องเซิร์ฟเวอร์ซึ่งสามารถ มีได้ 2 ส่วน คือ หมายเลขอินเตอร์เน็ตของเครื่อง (จำเป็นต้องมี) และหมายเลขพอร์ท (ไม่จำเป็น ต้องมี) จากตัวอย่างแรก ไม่มีการกำหนดหมายเลขพอร์ท ดังนั้นบราวเซอร์จะกำหนดให้เป็น หมายเลขพอร์ทปกติ (หมายเลข 80) ส่วนตัวอย่างที่สอง กำหนดหมายเลขพอร์ทเป็น 3232 ซึ่ง หมายเลขพอร์ทนี้ถูกกำหนดให้อยู่หลังชื่อเซิร์ฟเวอร์และใช้เครื่องหมาย ':' เป็นตัวแยก

ส่วนสุดท้ายได้แก่ตำแหน่งของข้อมูลที่ต้องการ ซึ่งอยู่หลังเครื่องหมาย '/' ที่ตามหลังที่อยู่และหมายเลขพอร์ทของเซิร์ฟเวอร์ ตำแหน่งของข้อมูลนี้จะสัมพันธ์กับ รูทไดเร็กทอรี ของเซิร์ฟเวอร์ที่ถูกกำหนดไว้

แฟ้มข้อมูลหรือแหล่งข้อมูลที่ถูกหนดให้เริ่มต้นด้วย '/cgi-bin/' จะเป็นกรณีพิเศษ เนื่องจาก 'cgi-bin' จะเป็นไดเร็กทอรี่ที่ใช้เก็บโปรแกรมหรือสคริปต์ ที่สามารถทำงานได้โดยเซิร์ฟเวอร์ ซึ่งจะกล่าวต่อไป

ในกรณีที่ชื่อแฟ้มข้อมูลไม่ได้ถูกกำหนดไว้ใน ยูอาร์แอลเซิร์ฟเวอร์จะกำหนดให้เป็นชื่อไฟล์ที่ถูกกำหนดไว้ (ชื่อนี้สามารถเปลี่ยนแปลงได้ขึ้นอยู่กับการกำหนดของผู้ดูแลเซิร์ฟเวอร์)

การส่งผ่านค่าไปยังเซิร์ฟเวอร์

โปรโตคอล เฮ็ชทีทีพี สนับสนุนการส่งผ่านค่าไปยังเซิร์ฟเวอร์ ซึ่งมีรูปแบบง่าย ๆ คือ ส่งตัวแปรต่อท้ายไปกับ ยูอาร์แอลโดยใช้เครื่องหมาย '?' เป็นตัวคั่น เหตุผลของการทำเช่นนี้ ส่วนใหญ่ใช้เมื่อต้องการค้นหาข้อมูลจากฐานข้อมูลและค่าที่ส่งผ่านไปให้เซิร์ฟเวอร์ คือตัวแปร ที่จะใช้ในการค้นหา.ซึ่งมีรูปแบบดังนี้

http://some.site.edu/cgi-bin/foo?arg1+arg2+arg3

เราสามารถแบ่งได้เป็น 2 ส่วน คือ

ซีจีไอ-บิน

คือชื่อของไดเร็กทอรี่พิเศษที่เซิร์ฟเวอร์กำหนดไว้ เพื่อเก็บโปรแกรม หรือ สคริปต์ ที่สามารถทำงานได้ สำหรับเหตุผลที่ต้องมีการผ่านค่าตัวแปรบางอย่างให้กับ โปรแกรม หรือ สคริปต์ เพื่อทำงานบางอย่างซึ่งโปรแกรม หรือ สคริปต์เหล่านี้จะใช้ในการติดต่อกับ เวิร์ลไวด์ เว็บเพื่อให้ตอบสนองกับผู้ใช้ผ่านบราวเซอร์ได้

การส่งผ่านอาร์กูเมนท์ (passed arguments)

ส่วนนี้จะอยู่ต่อท้ายจาก ยูอาร์แอลโดยใช้เครื่องหมาย '?' เป็นตัวคั่น ในการส่งผ่านค่าสามารถทำได้มากกว่า 1 ค่า โดยใช้เครื่องหมาย '+' เป็นตัวแยกค่าแต่ละค่า เช่นจากตัวอย่างข้างบนโปรแกรม (หรือ สคริปต์) จะได้รับค่า 3 ค่า คือ arg1, arg และ arg3

ไดเร็กทอรี่ส่วนตัวของผู้ใช้ (Personal HTML directories)

ผู้ใช้ทั่วไปสามารถมีเอกสารแบบเฮ็ชทีเอ็มแอล ในไดเร็กทอรี่ของตนเองได้ โดยขึ้นอยู่กับระดับความสามารถของเซิร์ฟเวอร์ที่ใช้ วิธีการที่ใช้กันทั่วไปคือ ผู้ใช้ต้องสร้างไฟล์พิเศษ เก็บไว้ในไดเร็กทอรี่ของตนเอง ที่ใช้ในการบอกถึงตำแหน่งของไดเร็กทอรี่ส่วนตัวที่ใช้เป็นจุดเริ่มต้น ทำให้สามารถติดต่อกับไฟล์นั้นได้โดยผ่านเส้นทาง

โกฟเฟอร์ ยูอาร์แอล (Gopher URLs)

โกฟเฟอร์ เซิร์พเวอร์ สามารถติดต่อกับ ยูอาร์แอล คล้าย ๆ กับเฮ็ชทีทีพี เซิร์พเวอร์ข้อแตกต่างที่สำคัญอยู่ที่การกำหนดเกี่ยวกับไฟล์ ซึ่งจะต้องมีตัวบอกชนิดของไฟล์ด้วยในโปรโตคอลโกเฟอร์ โดยตัวกำหนดนี้จะใช้รหัสตัวเลขกำหนดก่อนชื่อไฟล์ ซึ่งมีดังต่อไปนี้ คือ

0 - ไฟล์เอกสารทั่วไป (Text file)

1 - ไดเร็กทอรี่ (Directory)

2 - ซีเอสโอ (CSO Name Server)

4 - แม็ค (Mac *.hqx file)

5 - ไฟล์เกี่ยวกับเสียง (Sound file)

7 - ดัชนีเอกสาร (Full text index)

8 - เทลเน็ต (Telnet session)

9 - ไฟล์ที่ใช้รหัสเลขฐาน 2 (Binary file)

การดึงรายการหลักจากโกเฟอร์เซิร์ฟเวอร์ ทำงานผ่านพอร์ททั่วไป (หมายเลข 70)

การดึงรายการ 'adaptive.technology' จากโกเฟอร์เซิร์ฟเวอร์ ทำงานบนพอร์ทปกติ (70)

การติดต่อเพื่อทำการค้นหาดัชนี 'fonebook.txt' จากโกเฟอร์เซิร์ฟเวอร์ ทำงานผ่านพอร์ท หมายเลข 151

การใช้บริการเอฟทีพีผ่านทางยูอาร์แอล (Anonymous FTP via URLs)

ไฟล์สามารถทำการติดต่อบริการเอฟทีพีโดยใช้ยูอาร์แอลมีรูปแบบทั่วไป คือ

ftp://internet.address.edu/file/path/file.txt

ถ้ากำหนดชื่อไดเร็กทอรี่แทนที่จะเป็นไฟล์ บราวเซอร์ส่วนใหญ่จะแสดงรายการ ของไฟล์ทั้งหมดที่อยู่ภายใต้ไดเร็กทอรี่นั้นเพื่อให้ผู้ใช้เลือกไฟล์หรือไดเร็กทอรีต่อไป

ในการติดต่อขอใช้บริการนี้เราสามารถทำการติดต่อได้โดยตรง โดยไม่ต้องใช้ชื่อ ก็ได้โดยการใส่ชื่อผู้ใช้และรหัสลับของผู้เป็นเจ้าของไฟล์ที่เราต้องการติดต่อขอใช้ตัวอย่างเช่น

ftp://joe_bozo:bl123@internet.address.edu/path/file.gz

จากตัวอย่างจะทำให้สามารถติดต่อไฟล์ที่อยู่บนเครื่อง 'internet.address.edu' ซึ่งมี 'joe_bozo' เป็นเจ้าของ และมีรหัสลับ คือ 'bl123'

คำเตือน

จะเกิดอะไรขึ้นถ้าในเอกสารแบบเฮ็ชทีเอ็มแอล ปรากฏข้อความบรรทัดนี้อยู่ภายในนั่นคือทุกคนที่เข้ามาใช้เอกสารนี้จะรู้รหัสลับของ 'joe_bozo' นี่จึงเป็นวิธีการที่ไม่ปลอดภัยและควรหลีกเลี่ยงเป็นอย่างยิ่ง

การติดต่อดับบลิวเอไอเอส โดยใช้ยูอาร์แอล (WAIS access via URL)

ดับบลิวเอไอเอส เซิร์พเวอร์ สามารถทำการติดต่อโดยใช้ ยูอาร์แอล ได้เช่นเดียวกับ เอ็ชทีทีพี เซิร์พเวอร์ แต่ข้อแตกต่างที่เห็นได้ชัด อยู่ที่การกำหนดเกี่ยวกับไฟล์ โดยจะต้องส่ง คำสั่งที่ใช้ในการค้นหาไปยังดับบลิวเอไอเอส เซิร์พเวอร์ ให้ถูกต้อง ในส่วนของข้อมูลที่เพิ่ม เข้ามานี้จะใช้ในการค้นหา มีรูปแบบมาตรฐานดังนี้ คือ

wais://host_and_port/database[? search]

โดยที่ '?search' คือ รายการของคำสั่งที่ใช้ในการค้นหาที่ส่งผ่านไปยังดับบลิวเอไอเอส เซิร์พเวอร์

ยูอาร์แอลของการใช้บริการเทลเน็ต

เราสามารถใช้บริการของเทลเน็ต เพื่อติดต่อไปยังเครื่องที่อยู่ไกลออกไป โดยผ่านทาง ยูอาร์แอลได้ดังตัวอย่าง คือ

telnet://flober.rodent.edu { สำหรับเครื่องทั่วไป }

tn3270://flober.rodent.edu { สำหรับเครื่อง 'ไอบีเอ็ม3270' }

การติดต่อโดยใช้ 'rlogin' ผ่านทางยูอาร์แอลกำหนดโดย

rlogin://username@flober.rodent.edu

โดยที่ยูสเซอร์เนม คือ ชื่อของผู้ที่มีสิทธิ์ในการเข้าใช้บริการ

นิวส์ ยูอาร์แอล News URL

เพื่อให้สามารถอ่านข่าวสารต่าง ๆ จาก ยูสเน็ต นิวกรุ๊ป ที่ให้บริการโดย เอ็นเอ็นทีพี นิวเซิร์พเวอร์ สามารถทำได้โดยใช้ นิวยูอาร์แอล ซึ่งมีรูปแบบ คือ

news:news.group

โดยที่ 'news.group' คือ ชื่อของกลุ่มข้อมูลข่าวสาร

เมลทู ยูอาร์แอล Mailto URL

เมลทูยูอาร์แอล ทำให้สามารถส่งจดหมายไปยังจุดหมายที่ได้กำหนดไว้ มีรูปแบบ ดังนี้

mailto:user@host

โดยที่ 'ยูสเซอร์' คือ ชื่อของผู้รับจดหมาย

'โฮส' คือ ชื่อของเครื่อง

'เมลทูยูอาร์แอล' นี้จะไม่สามารถใช้งานได้กับบราวเซอร์ทุกชนิดที่มีอยู่

ภาคผนวก จ

เอ็ชทีทีพี โปรโตคอล

เอ็ชทีทีพี โปรโตคอล เป็นการทำงานในรูปของ รีเควส/เรสปอน ส่วนของ รีเควส โปรแกรม หรือที่เรียกว่า ไคลเอนท์ จะทำการติดต่อกับส่วนให้บริการโปรแกรมหรือที่เรียกว่า เซิร์พเวอร์ แล้วทำการส่ง รีเควส ไปยังเซิร์พเวอร์ ส่วนรับบริการโปรแกรมอาจจะเป็น ในส่วนของไคลเอนท์ หรือ เซิร์พเวอร์ ก็ได้ ซึ่งสิ่งที่กล่าวมานี้จะมีการจัดแบ่งตามบทบาทในการทำงานของโปรแกรมต่างๆ ในขณะที่ทำการติดต่อกันอยู่

รีเควส จะถูกส่งไปยังเซิร์พเวอร์ ในรูปแบบต่างๆ ขึ้นอยู่กับ วิธีในการ รีเควส, ยูอาร์ไอ, เวอร์ชั่นของโปรโตคอล แล้วตามด้วย เอ็มไอเอ็มอี - ไลค์ ข่าวสารที่จะมีข้อมูลที่เกี่ยวกับการ รีเควส, ข้อมูลของ ไคลเอนท์ และอื่นๆ หลังจากนั้นเซิร์พเวอร์ก็จะทำการตอบสนองมาด้วย status line ซึ่งประกอบด้วย เวอร์ชั่นของโปรโตคอลพร้อมทั้งความสำเร็จ หรือ ข้อผิดพลาด แล้วตามด้วย เอ็มไอเอ็มอี - like ข่าวสารที่ประกอบด้วย เซิร์พเวอร์ในรูปแบบแอคชั่น ,เมต้า (meta) เอนติตี้ ในรูปแบบแคชั่นและอื่นๆ

การทำการติดต่อบนอินเตอร์เน็ต มักจะมีการทำการติดต่อโดยใช้การเชื่อมต่อที่อาศัย ทีซีพี ไอพี โดยมีพอร์ต มาตราฐาน คือ TCP 80 (RP 94) แต่พอร์ต ก็สามารถใช้งานได้เอ็ชทีทีพี โปรโตคอล 1.0 จะมีการทำงานอยู่ในชั้นที่สูงกว่าโปรโตคอลตัวอื่นบนอินเตอร์เน็ต หรือระบบ เครือข่ายอื่นๆ ในการทำงานโดยทั่วไปแล้วไคลเอนท์จะทำการสร้างการติดต่อก่อนที่จะมีการ รีเควส และเซิร์พเวอร์ ก็จะทำการสิ้นสุดการติดต่อหลังจากที่ได้ส่งเรสปอนแล้ว ซึ่งการทำงาน ไม่จำเป็นต้องมีในการใช้งานเอ็ชทีทีพี เวอร์ชัน 1.0 เพราะ ทั้งไคลเอนท์และเซิร์พเวอร์สามารถที่ จะทำการสิ้นสุดการติดต่อได้จากสาเหตุต่อไปนี้ ซึ่งเมื่อมีการสิ้นสุดการติดต่อแล้ว ก็จะมีการทำ การระงับการ รีเควส รวมทั้งแสดงสถานะปัจจุบัน

การส่งข่าวสารของเอ็ชทีทีพี

เอ็ชทีทีพี ข่าวสารประกอบด้วย รีเควส จากไคลเอนท์และเรสปอนส์จากเซิร์พเวอร์ ซึ่งข่าวสาร นี้อาจจะเป็นรีเควสเต็มรูปแบบ, เรสปอนส์เต็มรูปแบบ หรือ อาจจะเป็นรีเควส, เรสปอนส์ก็ได้ ซึ่งถ้าหากว่ามีการใช้แบบเต็มรูปแบบก็จะมีการจัด ข่าวสาร ตามมาตรฐานของ RCF 822 สำหรับทำการส่งเอนติตี้ โดยข่าวสารทั้งสองแบบนี้สามารถที่จะรวมออปชั่น เฮดเดอร์ ฟิลด์ และเอนติตี้ บอดี้ เข้าไปด้วย โดยจะมีการอาศัย null line ในการที่จะจัดแบ่งเอนติตี้ บอดี้ ออกจากส่วนของเฮดเดอร์

ข่าวสารเฮดเดอร์

เอ็ชทีทีพี เฮดเดอร์ฟิลด์ จะประกอบด้วยเฮดเดอร์, รีเควส เฮดเดอร์ , เอนติตี้ เฮดเดอร์ โดยในแต่ ฟิลด์ ของ เฮดเดอร์จะต้องมีชื่อแล้วตามด้วย (:) แล้วตามด้วย ค่าของฟิลด์ ดังกล่าว เฮดเดอร์ ฟิลด์ สามารถที่จะขยายไปได้หลายบรรทัด นอกจากนี้ยังสามารถที่จะใส่คอมมอนด์ เข้าไปใน เอ็ชทีทีพี เฮดเดอร์ ได้ด้วยโดยใช้วงเล็บในการแยก

ฟิลด์ต่างๆ

เป็นส่วนของ เฮดเดอร์ ฟิลด์ ที่รวมอยู่ในรีเควสและเรสปอนส์ ข่าวสารแต่ไม่ใช้ในการ ติดต่อระหว่าง party เพราะจะใช้ก็ต่อเมื่อมีการอ้างถึงเท่านั้น

- วันที่ (Date) เป็นส่วนที่บอกว่าข่าวสารได้ถูกสร้างขึ้นเมื่อใด ดังตัวอย่าง

Date : Tue , 15 Apr 1995 07:45:20 GMT

- ฟอร์เวิร์ด (Forwarded)เป็นส่วนที่บอกเกี่ยวกับการติดต่อระหว่าง ยูสเซอร์ เอเจนท์ และเซิร์พเวอร์ รวมทั้งระหว่างเซิร์พเวอร์และไคลเอนท์

- ข่าวสาร - ไอดี เป็นตัวที่ใช้ในการกำหนดข่าวสาร เช่น

Message - ID : < 95050318836.AA00266@เอเจนท์.com >

- เอ็มไอเอ็มอีเวอร์ชัน (MIME - version) เนื่องจาก เอ็ชทีทีพี ไม่ใช่เอ็มไอเอ็มอี ฟอร์มโปรโตคอล เอ็ชทีทีพี 1.0 ข่าวสารอาจจะต้องรวมเอ็มไอเอ็มอี เวอร์ชันเฮดเดอร์ ฟิลด์ (single) เข้าไปด้วยเพื่อเป็นการแสดงถึงเวอร์ชันของ เอ็มไอเอ็มอี โปรโตคอล ที่ใช้ในการสร้างข่าวสารโดยมาตราฐานของ เอ็ชทีทีพี 1.0 ก็คือ เอ็มไอเอ็มอี - เวอร์ชัน 1.0

รีเควสข่าวสาร

เวิร์ลไวท์เว็บไคลเอนท์ สามารถที่จะทำการรีเควสไปยัง เวิร์ลไวท์เว็บ เซิร์พเวอร์ เพื่อทำ operation ที่ต้องการโดย รีเควส ข่าวสารจาก ไคลเอนท์ ไปยัง เซิร์พเวอร์ จะรวมข้อมูลต่อไปนี้ไว้ในบรรทัดแรกของ รีเควสไลน์ คือเมธอท ที่จะทำกับ รีซอร์ส รีเควส , โปรโตคอล เวอร์ชัน ซึ่ง เมธอท ก็คือตัวที่จะกำหนดวิธีการที่จะใช้ในการกระทำกับรีซอร์สโดยอาศัย ยูอาร์ไอรีเควส เมธอท เหล่านี้จะเป็นพวกตรงตามรูปแบบ ซึ่งมี เมธอท ที่น่าสนใจดังต่อไปนี้

- **Get** ทำการดึงข้อมูลกลับมาจากตำแหน่งที่ระบุโดย ยูอาร์ไอที่ต้องการในรูปของ เอนติตี้

- **Head** ใช้ในการเก็บข้อมูลเกี่ยวกับ เมต้า อินฟอร์เมชัน โดยอาศัยยูอาร์ไอซึ่งจะเป็นส่วนของ เฮดเดอรอินฟอร์เมชั่น

- **Post** ใช้ในการ รีเควส โดยที่ทางปลายทางเซิร์พเวอร์จะรับเอนติตี้ ที่อยู่ใน รีเควส มาพิจารณา

- **Put** เป็นรีเควสที่มี เอนติตี้ ติดไปด้วยเป็นส่วนที่ใช้ในการสร้างหรือเปลี่ยนแปลง แหล่งข้อมูล

- **Delete** เป็นการที่เซิร์พเวอร์ทำการลบ รีซอร์ส ในส่วนที่ถูกรีเควส ยูอาร์ไอ มา

- **Link** เป็นตัวที่ทำหน้าที่ในการสร้างลิงค์จากรีซอร์สหนึ่งไปยังอีกรีซอร์สหนึ่ง ซึ่งการลิงค์นี้จะทำกี่รีซอร์สก็ได้โดยอาศัยการรีเควส ยูอาร์ไอ

- **Unlink** ทำการลบลิงค์ที่มีอยู่ของ รีซอร์ส ออกโดยอาศัยการอ้างอิง ยูอาร์ไอ

เรสปอนส์ข่าวสาร

เมื่อทำการรับ รีเควสข่าวสารเข้ามาก็จะทำการแปลรีเควสข่าวสารดังกล่าวโดย เรสปอนส์จะถูกส่งเฉพาะ ในส่วนที่มีการรีเควสแบบ เอ็ชทีทีพี 0.9 หรือในกรณีที่เซิร์พเวอร์ สนับสนุนเฉพาะ เอ็ชทีทีพี โปรโตคอล 0.9 ในส่วนของไคลเอนท์ที่มีการส่ง รีเควส เอ็ชทีทีพี 1.0 แบบเต็มรูปแบบ แล้วได้รับเรสปอนสที่ไม่มีในส่วนของสถานะที่บรรทัดแรก มันจะทำการสมมติว่า เรสปอนส์ที่มันได้รับเป็นแบบง่ายแล้วทำการส่งต่อไป เพราะเรสปอนส์จะมีเฉพาะในส่วนของ เอนติตี้ และจะถูกปิดไปเมื่อเซิร์พเวอร์ทำการสิ้นสุดการติดต่อ

ภายในสถานะไลน์หรือบรรทัดแรกของ เรสปอนส์ ข่าวสารจะประกอบด้วยสิ่งเหล่านี้

- โปรโตคอล เวอร์ชัน
- ตัวบอกสถานะ (A numeric status code)
- การส่งผ่านค่า (the associated textual phrase)

การแบ่งประเภทเอ็ชทีทีพี เรสปอนส์

Digit	Type	Description
1xx	ข่าวสาร	ไม่ใช้งานแต่จองไว้สำหรับอนาคต
2xx	สำเร็จ	แอคชั่นได้รับเรียบร้อยแล้ว สามารถทำความเข้าใจและทำได้
3xx	ย้อนกลับ	ต้องมีการทำให้ รีเควส สมบูรณ์โดยมี action เพิ่ม
4xx	ไคลเอนท์ ข้อผิดพลาด	รีเควส ที่ส่งไปมีรูปแบบ ข้อผิดพลาด เกิดขึ้น
5xx	เซิร์พเวอร์ ข้อผิดพลาด	เซิร์พเวอร์หยุดทำงานในการที่จะตอบสนองสิ่งที่ รีเควส

เอนติตี้

รีเควสเต็มรูปแบบและเรสปอนสเต็มรูปแบบ สามารถที่จะรวมส่วนของเอนติตี้ เข้าไปในข่าวสารได้ด้วย โดยในส่วนของเอนติตี้ จะประกอบด้วย เอนติตี้ เฮดเดอร์ และส่วนของเอนติตี้บอดี้ในที่นี้จะไม่มีการกำหนดว่า ไคลเอนท์และเซิร์พเวอร์ ตัวใดเป็น ผู้ส่ง หรือ ผู้รับ ที่แน่นอน แต่จะอาศัยการพิจารณาในการส่งและรับเอนติตี้ แทน

เอนติตี้ เฮดเดอร์ ฟิลด์

เอนติตี้ เฮดเดอร์ ฟิลด์จะมีการกำหนดเมต้า อินฟอร์เมชันเกี่ยวกับ เอนติตี้ บอดี้ หรือเกี่ยวกับรีซอร์สที่ถูกกำหนดโดย รีเควสเอนติตี้ เฮดเดอร์

เอนติตี้ บอดี้

เอนติตี้ บอดี้ จะถูกส่งไปพร้อมกับเอ็ชทีทีพี 1.0 รีเควส หรือ เรสปอนส์ ในรูปแบบ ที่ได้กำหนดไว้ที่ เอนติตี้ เฮดเดอร์ฟิลด์ เอนติตีบอดี้ ที่ถูกรวมไปใน รีเควส แมสเสสจะมีการนำมา ใช้เพียงหนึ่งเดียว ถ้ามีการร้องขอที่จะใช้เพียงหนึ่งเดียว แต่ในที่นี้ POST และ PUT รีเควส เมธอท จะอนุญาตให้มีส่วนของ เอนติตี้บอดี้ ร่วมอยู่ด้วย

บรรณานุกรม

ก. <u>เอกสารอ้างอิงที่เป็นวารสารภาษาไทย</u> จัดเรียงลำดับดังนี้

1. ยืน ภู่สุวรรณ, "ชื่อนั้นสำคัญไฉน", วารสารอินเทอร์เน็ต แมกกาซีน, ฉบับที่ 1, พฤษภาคม 2539, หน้า 25-28
2. ธเนศ สุวรรณผ่อง, "Server Power", วารสารพีซี แมกกาซีน, ปีที่ 4, ฉบับที่ 43, ตุลาคม 2539, หน้า139-147
3. กฤติ อัจจิมากร, "ผู้ท่องไปบน Web", วารสารพีซี แมกกาซีน, ปีที่ 4, ฉบับที่ 44, พฤศจิกายน 2539, หน้า 151-162
4. บุญเลิศ เอี่ยมทัศนา, "เตาะแตะถึง CGI", วารสารอินเทอร์เน็ต แมกกาซีน, ฉบับที่ 5, กันยายน 2539, หน้า 61-68

ข. <u>เอกสารอ้างอิงที่เป็นหนังสือภาษาไทยและภาษาอังกฤษ</u> จัดเรียงลำดับดังนี้

1. สุชาดา รัตนคงเนตร, "เรียนรู้เทอร์โบปาสคาล สู่หลักการเขียนโปรแกรม", ไอบิซ พับลิชิ่ง, 518 หน้า, 2536
2. Jonathan Matcho, Brian Salmanowitz, Scott Strool, Brent Biely, Scott T. Jurkouich, Susan Berry, Lawrence Sleeper, Dan Dumbrill, Eric Uber, "Using Delphi 2", Que, 700 p. , 1996
3. Mark Watsom, "Programming Intelligent Agents for the Internet", McGraw-Hill, 204 p. , 1996
4. Graham Ian S, "HTML Sourcebook", John Wiley & Sons, 348 p. , 1995
5. Lan H. Witten, Alistair Moffat, Timothy C. Bell, "Managing Gigabytes", Van Nostrand Reinhold, 409 p. , 1994
6. C. J. Date, "An Introduction to Database System", Addison Wesley, 709 p. , 1995
7. Davis Chapman, "Building Internet Applications with Delphi 2", Que, 471 p. 1996
8. Douglase E. Comer, David L. Stevens, "Internetworking wih TCP/IP Volume II", Prentice Hall International, 551 p. , 1994

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้จะไม่มีทางสำเร็จลงได้ ถ้าขาดบุคคลเหล่านี้

- อ.อภิเนตร อูนากูล อาจารย์ที่ปรึกษากลุ่มโปรเจ็กต์ ผู้ช่วยให้คำแนะนำ เอาใจใส่ดูแล และควบคุมการทำงานอย่างใกล้ชิด ทำให้การทำงานดำเนินไปได้ด้วยดี
- ดร.สุรพันธ์ -- NECTEC ผู้ให้คำปรึกษาในเรื่องการจัดการกับข้อมูลขนาดมหาศาล
- คุณพุธิศานต์ ปราชญ์พงศ์ ผู้ให้ความอนุเคราะห์ Printer และกระดาษที่ใช้ในการพิมพ์
- เพื่อน ๆ ที่ช่วยจุดประกายความคิดในการแก้ไขปัญหาต่าง ๆ

ทางคณะผู้จัดทำโครงงานหุ่นยนต์ค้นหาข้อมูลบนเครือข่ายคอมพิวเตอร์ในประเทศไทย จึงขอแสดงความขอบคุณมา ณ. โอกาสนี้ด้วย ขอบคุณมาก (ครับ, ค่ะ)