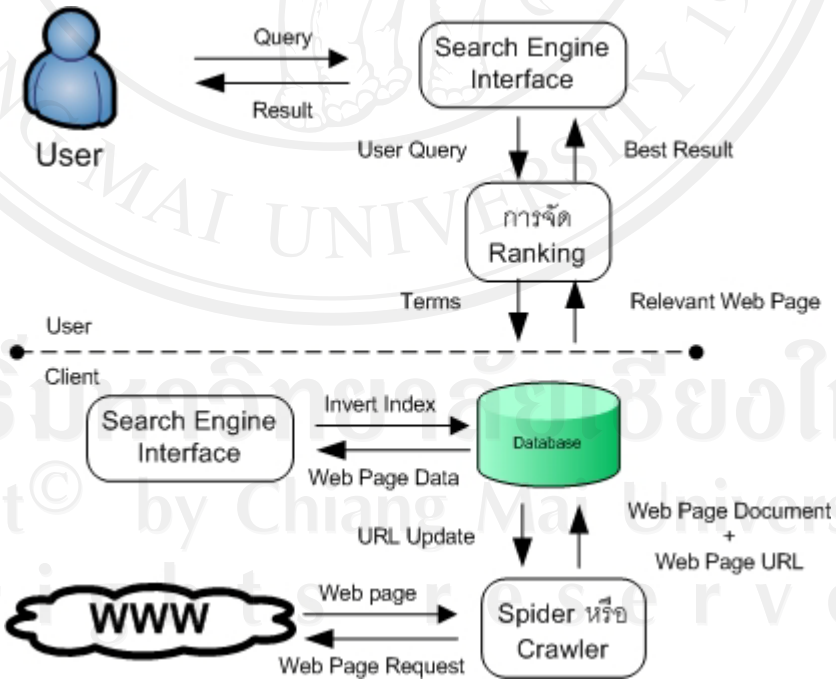
บทที่ 3

แนวคิดในการออกแบบของเว็บสไปเดอร์

ในบทนี้กล่าวถึงแนวคิดในการออกแบบ การทดลองการทำงานของเว็บสไปเดอร์ และการเตรียมข้อมูลก่อนการจัดกลุ่มเอกสาร โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

3.1 แนวคิดในการออกแบบของเว็บสไปเดอร์

เสิร์ชเอนจินส่วนมากจะใช้ในการค้นหาข้อมูลบนอินเตอร์เน็ตโดยทั่วไปผู้ใช้จะทำการใส่คีเวิร์ดผ่านหน้าจอค้นหา (Search Engine Interface) ระบบเสิร์ชเอนจินจะทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เสิร์ชเอนจินต้องการและนำไปค้นหา ซึ่งผลลัพธ์ที่ได้จะเป็นข้อมูลเกี่ยวกับเว็บไซต์ต่าง ๆ โดย เว็บไซต์เหล่านี้จะถูกจัดลำดับ และกลับมาแสดงผลลัพธ์ต่อไปทางด้านฝั่งผู้ใช้บริการ (Client) ประกอบ ด้วยส่วนของ สไปเดอร์มีหน้าที่จัดการเก็บข้อมูลเกี่ยวกับเว็บไซต์ต่าง ๆ โดยนำมาเก็บในฐานข้อมูล ส่วนตัวทำดัชนี (Indexing) มีหน้าที่จัดทำการสร้างดัชนี (Index) สำหรับการค้นหาเพื่อให้การค้นหาข้อมูลเป็นไปด้วยความรวดเร็ว



รูป 3.1 สถาปัตยกรรมของระบบ

ประเภทของเสิร์ชเอนจินสามารถแบ่งประเภทของเสิร์ชเอนจิน (Sullivan, 2005) ได้ดังนี้

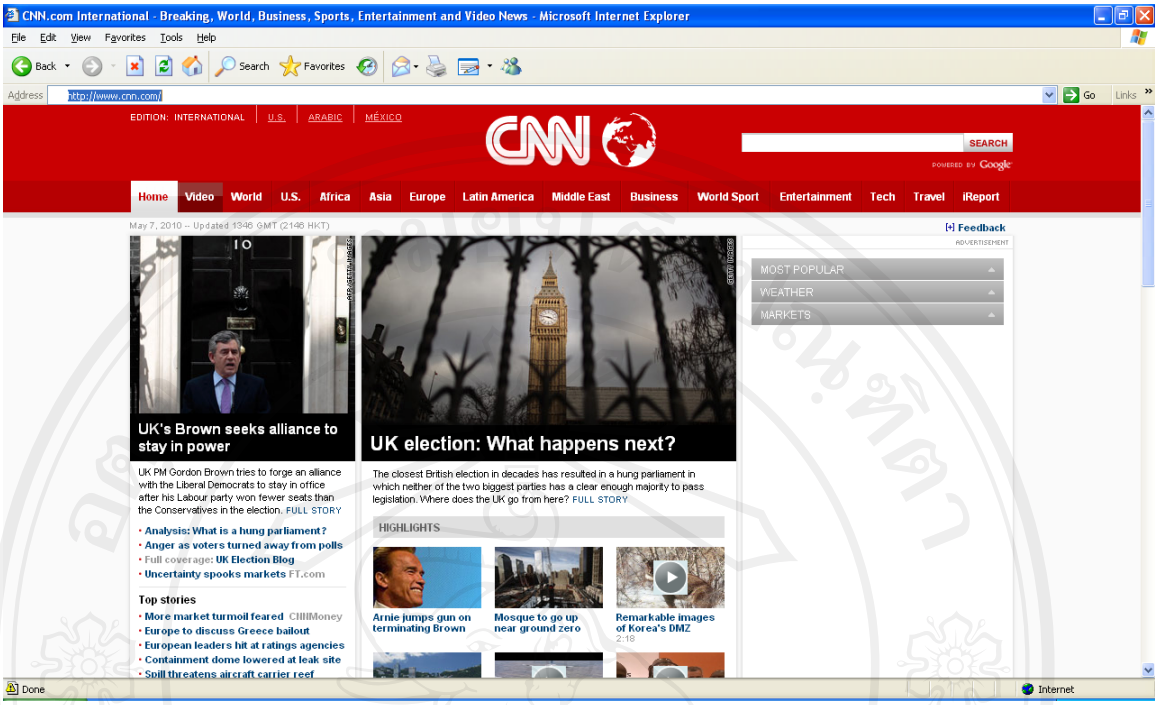
1. Human Powered Directories จะจัดประเภทของข้อมูลโดยใช้บุคคลเว็บไซต์ใดที่ต้องการมีรายชื่อในไดเรคทอรี่ (Directory) เหล่านี้จะต้องทำการติดต่อผู้ดูแลไดเรคทอรี่นั้น ๆ ซึ่งผู้ดูแล Directory จะทำการจัดประเภทของเว็บไซต์ให้อยู่ในประเภทที่เหมาะสม ซึ่งการค้นหาข้อมูลจะตรงต่อความต้องการมาก แต่มีปัญหาคือปริมาณของข้อมูลมีจำนวนน้อย ทำการปรับปรุงได้ช้า และ Web บางตัวสามารถจัดได้หลายประเภท ตัวอย่าง เสิร์ชเอนจินที่ให้บริการในลักษณะนี้ได้แก่ Yahoo Directory , Uncover the Net ,Web Atlas , SevenSeek , Wow Directory เป็นต้น

2. Crawler Based Search Engines จะใช้โรบอต (Robot) ในการเก็บข้อมูลซึ่งบางครั้งเรียกว่าสไปเดอร์โดยจะท่องไป เว็บไซต์ต่าง ๆ และนำข้อมูลจากเว็บไซต์ต่าง ๆ มาสร้างดัชนี ข้อดีคือสามารถเก็บข้อมูลได้เป็นจำนวนมาก การปรับปรุงข้อมูลทำได้รวดเร็ว ซึ่งการปรับปรุงข้อมูลมีผลให้ Ranking ของข้อมูลที่ต้องการเปลี่ยนลำดับได้ตามกาลเวลา ตัวอย่าง เสิร์ชเอนจินที่ให้บริการในลักษณะนี้ได้แก่ Google, Altavita เป็นต้น

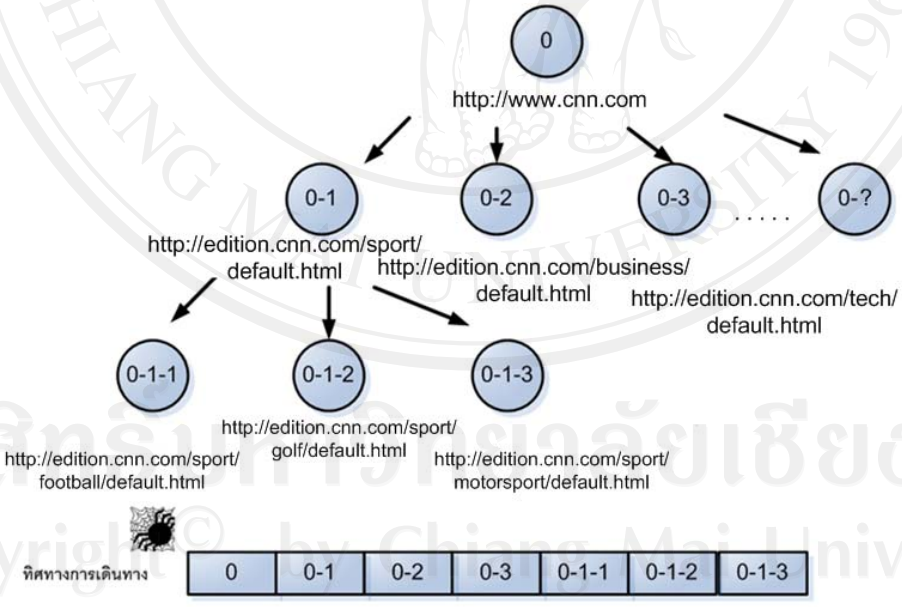
3. Hybrid Search Engines เป็นการผสานการทำงานทั้งแบบ Human Powered Directories และ Crawler Based Search Engines ตัวอย่าง เสิร์ชเอนจินที่ให้บริการในลักษณะนี้ ได้แก่ Msn เป็นต้น

3.1.1 การวิเคราะห์โดเมน (Domain) ของ www.cnn.com

เวิลด์ไวด์เว็บ (World Wide Web, WWW, W3) หรือที่เรียกกันสั้นๆ ว่า "เว็บ" คือพื้นที่เก็บข้อมูลข่าวสารที่เชื่อมโยงกันทางอินเทอร์เน็ตเป็นแบบเครือข่ายคล้ายใยแมงมุมโดยใช้การกำหนดยูอาร์แอลซึ่งผู้ใช้สามารถเชื่อมต่อเข้าถึงแหล่งข้อมูลที่เก็บไว้ภายในของคอมพิวเตอร์แต่ละเครื่องได้ผ่านทางบราวเซอร์ (Browser) โดยทางผู้วิจัยจะทำการวิเคราะห์โดเมนของ “www.cnn.com” ซึ่งเป็นเว็บไซต์ที่เกี่ยวกับข่าวสารดังรูป 3.2



รูป 3.2 ตัวอย่างโดเมน “www.cnn.com”



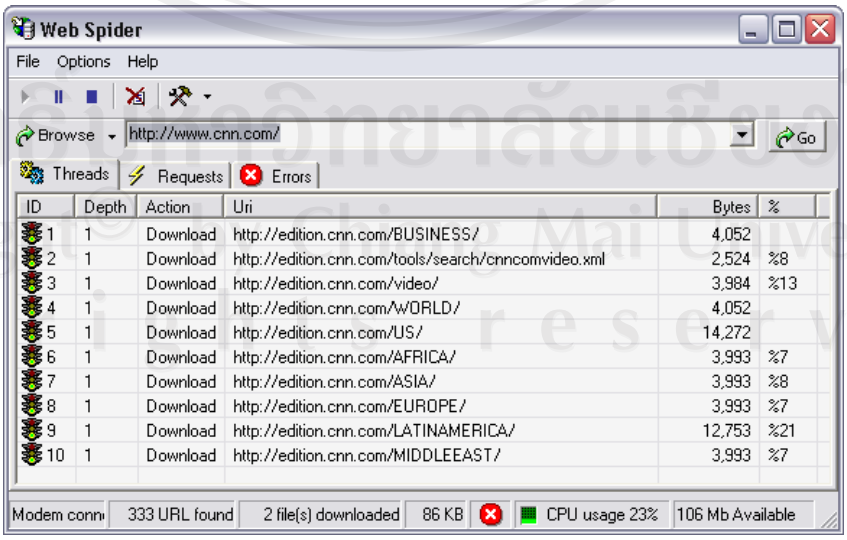
รูป 3.3 วิธีการเดินทางเพื่อเก็บรวบรวมข้อมูลของสไปเดอร์หรือครอว์เลอร์

การทำงานเพื่อเก็บรวบรวมข้อมูลของ ครอว์เลอร์ดังรูป 3.3 จะเริ่มจากเว็บไซต์ หรือ ยูอาร์แอล ที่ถูกตั้งเอาไว้ โดยจะทำงานอ่านข้อมูล ดังเช่นตัวอย่างนี้ เว็บไซต์ http://www.cnn.com จะมีลิงค์ของเอกสารประกอบไปด้วย http://edition.cnn.com/sport/default.html,

http://edition.cnn.com/business/default.html และ http://edition.cnn.com/tech/default.html ซึ่งจะถือว่าเป็นระดับชั้นความลึกชั้นที่ 1 โดยจะเดินทางจาก 0, 0-1, 0-2, 0-3 ตามลำดับ ในขณะเดียว กันเมื่อเข้าไปถึงระดับความลึกชั้นที่ 1 แล้วก็จะมีลิงค์ไปยังหน้าเว็บเพจอื่นๆ อีก เช่น http://edition.cnn.com/sport/football/default.html ซึ่งจะเรียกว่า ระดับความลึกชั้นที่ 2

3.1.2 เว็บสไปเดอร์ หรือ ครอว์เลอร์

ในงานค้นคว้าอิสระนี้จะใช้เว็บสไปเดอร์ซึ่งพัฒนาเป็น Open Sourceโดย Hatem Mostafa ในปี 2006 ใช้วิชวลซีชาร์ป (Visual C#) ซึ่งมีการตั้งค่าในการทดลอง โดยกำหนดยูอาร์แอลเริ่มต้นที่ต้องการ ซึ่งกำหนดยูอาร์แอลเป็น "http://www.cnn.com" ส่วนประเภทของ MIME (Multipurpose Internet Mail Extensions) เป็น text/richtext, text/html, text/htm, text/asp, text/aspx, text/php, text/mspx ซึ่งหมายถึงไฟล์ที่มีนามสกุลเป็น html, htm, php, php2, asp, aspx, mspx เช่น default.html, index.php ฯลฯ เพื่อให้ตรงกับเงื่อนไขในการเตรียมเอกสารในหัวข้อ 3.2 โดยโดเมนที่ตัดออกคือ .org; .gov จำนวนของการทำงานของภาพของเทรดในเว็บครอว์เลอร์เท่ากับ 10 เทรด การส่งและรับ Timeout ทุกซ็อกเก็ตของครอว์เลอร์เท่ากับ 20 ส่วนความลึกของนำทางในการรวบรวมข้อมูลเท่ากับ 3 โดยจำกัดการรวบรวมข้อมูลยูอาร์แอลเดียวกันของต้นฉบับยูอาร์แอลและลักษณะข้อมูลในการจัดเก็บของโปรแกรมขึ้นอยู่กับประเภทของ MIME ที่กำหนดไว้ตั้งแต่ต้น เวลาในการทดสอบเว็บสไปเดอร์ 1 ชั่วโมง ซึ่งได้ผลการทดสอบในตาราง 3.1



รูป 3.4 เว็บสไปเดอร์

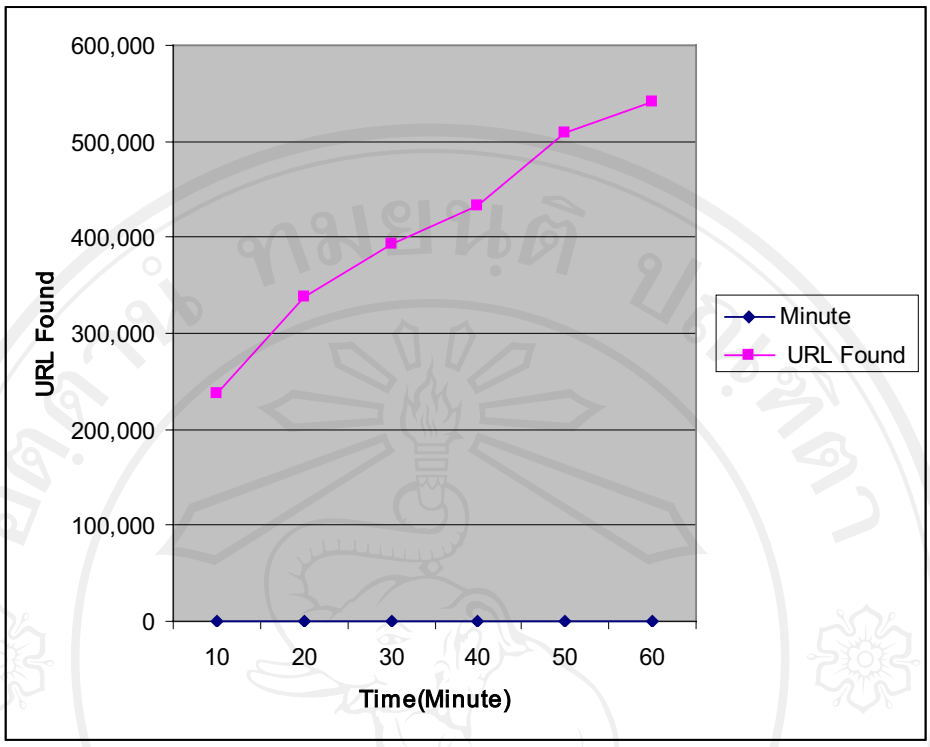
การจัดเก็บของโปรแกรมเว็บสไปเดอร์จะจัดเก็บในลักษณะ HTML ซึ่งสามารถขยายหรือดูได้ตามรูป 3.5

```
<html>
<head>
<title>Marketing your business school - CNN.com</title>
<meta name="TITLE" content="Marketing your business school -
CNN.com">
<meta name="KEYWORDS" content="California , Cable News Network
(CNN) , Duke University , David Miller , Asia , Schools , Military"></head>
<body>
<p><b>LONDON, England</b> (CNN) -- A great deal of attention is paid to
the tricky business of how aspirant MBAs can identify and then get into the
best courses at top business schools.</p>
<p>But there is another side to this, one that is less often remarked on but is
equally crucial for the whole process -- how business schools can attract
students.</p>
<p>At the sharp end of this equation, business schools have to make sure they
have sufficient enrollments each year to pay the wages of their professors, not
to mention the host of other bills involved in running such an institution.</p>
</body>
</html>
```

รูป 3.5 ตัวอย่างการเก็บข้อมูลในลักษณะ HTML

ตาราง 3.1 การทดสอบการใช้งานเว็บสไปเดอร์

Minute	URL Found	Files Downloaded	Errors
10	236,923	1,044	21
20	338,390	1,929	30
30	392,675	2,993	31
40	433,226	4,161	61
50	508,604	5,603	87
60	541,417	6,538	90



จากการทดสอบการทำงานของเว็บสไปเดอร์เป็นที่น่าพอใจ ซึ่งสามารถค้นหายูอาร์แอลได้ถึงห้าแสนยูอาร์แอลและสามารถดาวน์โหลดไฟล์ได้หกพันกว่าไฟล์ภายในเวลาหนึ่งชั่วโมง โดยมีไฟล์ที่ไม่สามารถดาวน์โหลด (Errors) เนื่องจากการส่งและรับ Timeout ทุกซ็อกเก็ตของครอว์เลอร์มากกว่าที่ได้กำหนดไว้คือ 20 Timeout การจะเพิ่มประสิทธิภาพของโปรแกรมเว็บสไปเดอร์ให้มากยิ่งขึ้นสามารถเปลี่ยนแปลงการตั้งค่าของโปรแกรมได้ ซึ่งในการวิจัยการจัดกลุ่มเอกสารโดยใช้ซัพพอร์ตเวก เตอร์แมชชีนในบทที่ 4 จะนำไฟล์ที่ได้ดาวน์โหลดจากการทำงานของเว็บสไปเดอร์ไปสร้างโมเดล 1,000 เอกสาร และทดสอบ 500 เอกสาร

3.2 การเตรียมเอกสาร

เมื่อได้ข้อมูลจากการครอว์เลอร์ของเว็บสไปเดอร์เรียบร้อยแล้ว โดยกำหนด ยูอาร์แอลเริ่มต้นเป็น “http://www.cnn.com” จะนำข้อมูลที่ได้ในลักษณะข้อมูลที่มีนามสกุลที่กำหนดประเภทของ MIME ทำการสกัดสัญลักษณ์ สคริปต์ แท็ค (Tag) และแอททริบิวท์ (Attribute) ที่ไม่ต้องการเพื่อต้องการออกมาในรูปของ Text file ซึ่งแต่ละเอกสารที่ได้จะการวิเคราะห์คำในประโยค (Parsing) ประกอบด้วยส่วนของ Title, Meta, Keyword และ Body ใช้ในการจัดกลุ่มเอกสารโดยซัพพอร์ตเวกเตอร์แมชชีนในบทที่ 4 ต่อไป ขั้นตอนนี้ใช้ภาษา Perl ในการจัดการโดยมีการกำหนดค่าและรายละเอียดของโค้ดดังนี้

กำหนดข้อมูลที่ต้องการ

@FILE_EXTENSIONS = (""html", "htm", "php", "php2", "asp", "aspx","mspx",);

กำหนดคำที่ต้องการตัดออก

$IGNORE_TERMS_FILE = $CONFIGURATION_DIR.'stop_terms.txt';

เช่น a,about,above,according,across,adj,after,afterwards,again,against,all,almost,an

เป็นต้น

สกัด Tittle

$contents =~ s/<\s*TITLE\s*>\s*(.*?)\s*<\s*\/TITLE\s*>;

สกัดสคริปต์ และ สไตล์

$contents=~s/(<\s*script[^>]*>.*?<\s*\/script\s*>)|(<\s*style[^>]*>.*?<\s*\/style\s*>);

แสดงรายละเอียดของแต่ละไฟล์

Starting indexer at c:/inetpub/wwwroot/search/files/

Indexed c:/inetpub/wwwroot/search/files/news.bbc.co.uk/default.html

Last Updated: Mar 26, 2010

File Size: 94 KB

Title: BBC NEWS | News Front Page

Description: Accessibility links Low graphics Skip to content Skip to local navigation

Saved file contents to Flat File database

```
<html>
<head>
<title>Marketing your business school - CNN.com</title>
<meta name="TITLE" content="Marketing your business school - CNN.com">
<meta name="KEYWORDS" content="California , Cable News Network (CNN)
, Duke University , David Miller , Asia , Schools , Military"></head>
<body>
<p><b>LONDON, England</b> (CNN) -- A great deal of attention is paid to the
tricky business of how aspirant MBAs can identify and then get into the best
courses at top business schools.</p>
<p>But there is another side to this, one that is less often remarked on but is
equally crucial for the whole process -- how business schools can attract
students.</p>
<p>At the sharp end of this equation, business schools have to make sure they
have sufficient enrollments each year to pay the wages of their professors, not to
mention the host of other bills involved in running such an institution.</p>
</body>
</html>
```

รูป 3.6 ตัวอย่างก่อนการวิเคราะห์คำในประโยคของเอกสาร

Marketing your business school - CNN.com LONDON, England (CNN) -- A great deal of attention is paid to the tricky business of how aspirant MBAs can identify and then get into the best courses at top business schools. But there is another side to this, one that is less often remarked on but is equally crucial for the whole process -- how business schools can attract students. At the sharp end of this equation, business schools have to make sure they have sufficient enrollments each year to pay the wages of their professors, not to mention the host of other bills involved in running such an institution.

รูป 3.7 ตัวอย่างหลังการวิเคราะห์คำในประโยคของเอกสาร

เมื่อเราได้เอกสารในลักษณะ Text File ที่ต้องการก็จะนำไปเข้าขั้นตอนการการจัดกลุ่มเอกสารโดยซัพพอร์ตเวกเตอร์แมชชีนในบทที่ 4 ต่อไป